UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Obtención, procesamiento y uso de datos en una Herramienta de BI

INFORME DE ACTIVIDADES PROFESIONALES

Que para obtener el título de

Ingeniero Mecatrónico

P R E S E N T A

Reynoso Ávila Abraham Alejandro

ASESOR DE INFORME

M.I. Serafín Castañeda Cedeño

Ciudad Universitaria, Cd. Mx., 2024

ÍNDICE

Introducción

A lo largo de este informe se verán de forma breve las actividades profesionales que desempeñé durante mi empleo en Huawei Technologies de México durante el primer semestre del año 2021, bajo el cargo de Ingeniero Analista de Datos.

En esta empresa, me vi involucrado el proyecto: LATAM-App, cuyo objetivo era la creación de una herramienta que proveyera a gobiernos y entes privados de Latinoamérica el análisis e introspectiva de negocios necesaria para la evaluación de propuestas que involucrasen el despliegue de infraestructura de telecomunicaciones en zonas de difícil acceso, tomando como punto de partida la conexión para puestos educativos y de salud.

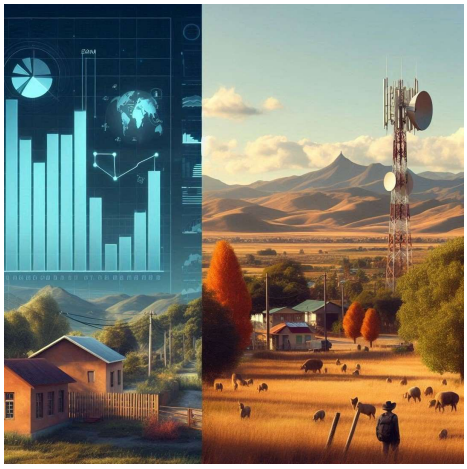


Figura 1: Sinergia de la tecnología en la comunidad (Imagen generada con Inteligencia Artificial)

Durante mi labor en dicho proyecto, apliqué distintas áreas del conocimiento propias de la ingeniería, incluyendo principal, pero no exclusivamente, programación, geometría, estadística y manejo de bases de datos. Estas habilidades fueron aplicadas en distintas ocasiones en las que la particularidad de cada caso ameritaba la búsqueda de una solución adecuada.

El reporte aquí presentado se encuentra dividido en tres secciones principales:

1. Búsqueda de Información
2. Procesamiento y transformación
3. Despliegue

Dichas secciones, que debido a la naturaleza del proyecto fueron secuenciales durante la creación de un dashboard (interfaz gráfica personalizable para la visualización de datos; Few, 2006), comprenden a su vez distintos subprocesos que podían ser realizados haciendo uso de diversas técnicas. Estas eran utilizadas conforme a las características específicas de cada caso, las instrucciones recibidas o bien, los conocimientos que poseía durante su implementación.

A continuación, estos procedimientos serán desarrollados y expuestos con el fin de demostrar el uso de los conocimientos lógico-matemáticos adquiridos durante mi formación académica, al mismo tiempo que se exponen las capacidades de adaptación y mejora continua, inherentes al estudiante de ingeniería.

Empresa: "Huawei Technologies"

Huawei Technologies es una empresa multinacional de origen chino fundada en la cuidad de Shenzhen en 1987 por Rhen Zhengfei, un exoficial del Ejército Nacional de Liberación, quien actualmente preside la compañía (Huawei Technologies Co., Ltd., 2024).

Enfocada inicialmente en la fabricación de conmutadores telefónicos, Huawei ha llegado a posicionarse en años recientes como una de las empresas de telecomunicaciones con mayor importancia a nivel mundial, siendo incluso, la empresa con mayor número de patentes tecnológicas registradas en la Administración Nacional de Propiedad Intelectual de China y en la Oficina Europea de Patentes durante el 2021 (Huawei Blog, 2024).

Aunque la compañía es mayormente conocida por la fabricación de teléfonos móviles y dispositivos inteligentes, esta no se limita únicamente a estos campos, teniendo una extensa variedad de negocios que van desde la infraestructura de telecomunicaciones hasta almacenamiento de energía y cómputo en la nube. A lo largo de este reporte, me enfocaré principalmente al sector de telecomunicaciones, ya que mis actividades dentro de la empresa estuvieron relacionadas únicamente con este.

Hoy en día, las tecnologías que Huawei fabrica y da soporte, tienen presencia en más de 170 países alrededor del mundo, pudiendo incluso declarar que al menos una de cada tres llamadas que se

realizan, hacen uso de su infraestructura. Es por esto por lo que la compañía, también es reconocida como la tercera empresa más grande en cuanto a fabricación de Routers, Switches y equipos de telecomunicaciones, estando únicamente por detrás de Alcatel-Lucent y Cisco.

La misión de la empresa es llevar la digitalización a cada persona, hogar y organización, para así formar un mundo totalmente conectado e inteligente. Para llevar a cabo este objetivo, Huawei destina más del 20% de sus ganancias a la investigación y desarrollo, además de que el 54.8% de su plantilla laboral está enfocada en estos mismos campos.

El proyecto explicado a lo largo de este informe es un claro reflejo de la misión y visión de Huawei, teniendo como objetivo último, facilitar el uso de tecnología en lugares cuyas características geográficas imposibilitan el uso de soluciones tradicionales.

Figura 2: Logotipo Oficial de Huawei Technologies

Organigrama

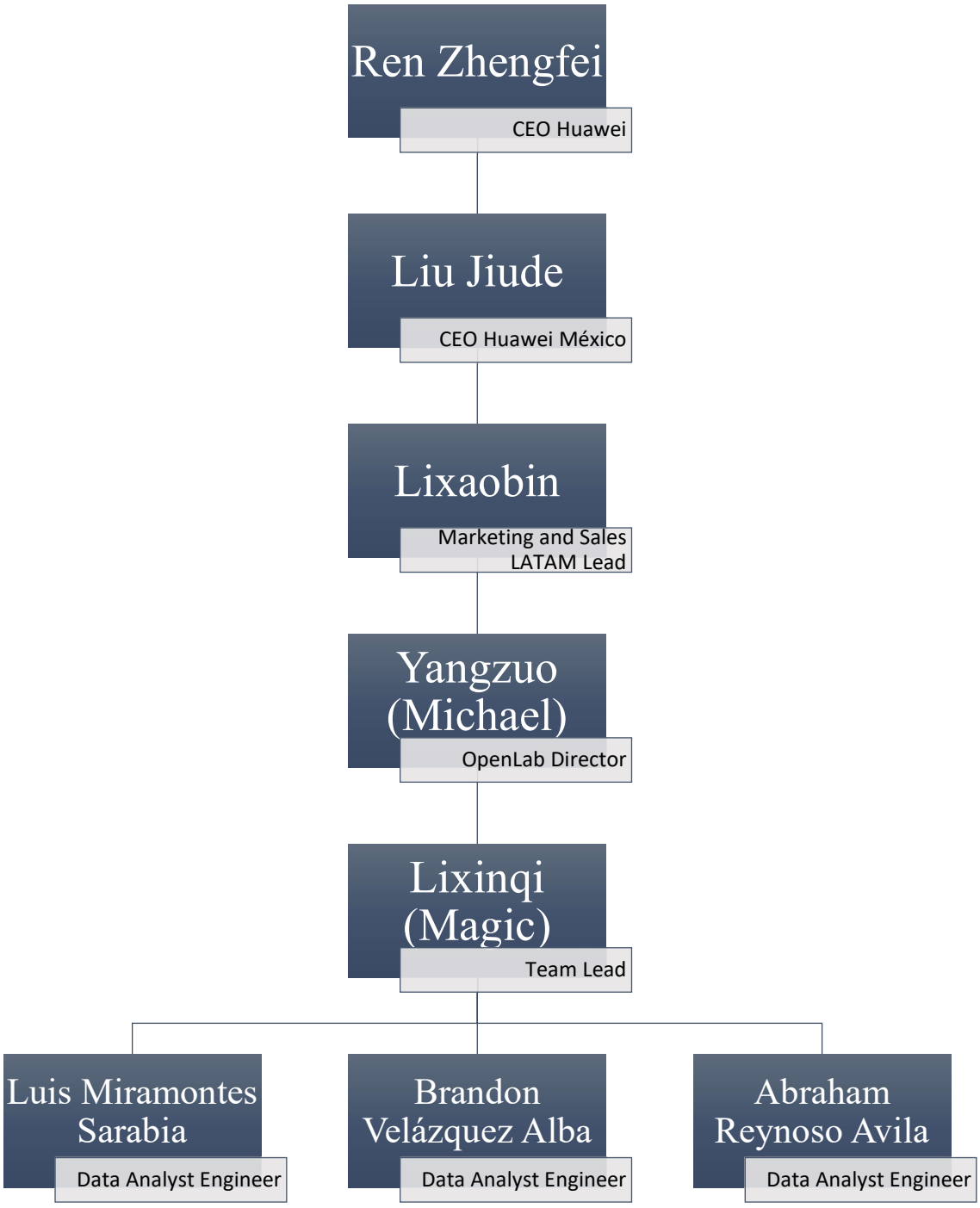


Figura 3: Estructura corporativa de Huawei y mi posición en la empresa durante el desarrollo del proyecto LATAM-App (Comunicación personal, 30 de octubre del 2023)

Descripción del puesto de trabajo: Ingeniero Analista de Datos

Durante mi labor en Huawei Technologies de México, desempeñé tareas principalmente relacionadas al análisis de datos, como lo son: recopilación, limpieza, transformación, visualización, storytelling (narrativa y conclusiones basadas en datos), etc.

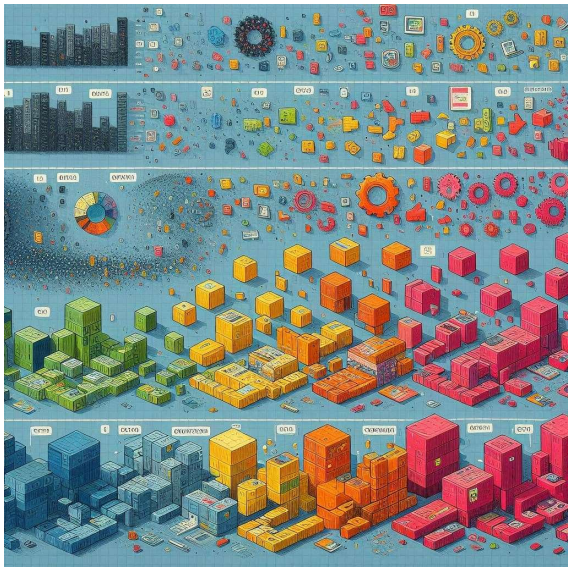


Figura 4: Analogía del proceso de transformación de datos (Imagen generada con Inteligencia Artificial)

Al momento de mi entrada a la empresa, fui asignado a un equipo de 3 integrantes, junto con un líder de proyecto. Mis dos compañeros y yo desempeñábamos tareas similares, mientras que nuestro líder se encargaba de la supervisión de nuestros avances, brindarnos ayuda para problemas específicos a nuestra labor individual y, en un principio, enseñarnos los usos y alcances de las diversas plataformas que utilizábamos en las distintas etapas del proyecto.

En un principio, nuestra tarea principal era obtener datos geográficos y estadísticos acerca de escuelas y hospitales en diversos países de Latinoamérica, para lo cual se realizaba una intensa búsqueda por internet. En esta etapa, siempre nos era encomendado comprobar las fuentes de información, su veracidad y congruencia entre las mismas.

Una vez terminada la fase de obtención de datos, era preciso realizar una limpieza de estos, pues en una gran cantidad de ocasiones podíamos encontrar errores que dificultaban la visualización de la información o su correcto procesamiento en etapas futuras.

Posteriormente, procedía a obtener registros faltantes en los conjuntos de datos obtenidos, específicamente, coordenadas nulas o incorrectas para establecimientos en los que únicamente se tenía una dirección o referencia geográfica. Para esto, hacía uso principalmente de Python (un lenguaje de programación) y la librería especializada en geocodificación, Geopy. Aunque también se llegaron a utilizar otras herramientas como Google Earth o API's (Interfaces de Programación de Aplicaciones) específicas para la información que se buscaba.

Cuando la información se encontraba completa y en el formato correcto, procedía a cargarla en una plataforma propia de la empresa, en la cual se realizaban operaciones que facilitarían su procesamiento y posterior sinergia con otros conjuntos de datos relativos a la calidad de conexión en distintas zonas del país. Al combinarlos, proveerían el análisis necesario para determinar de una forma medianamente precisa, la velocidad de conexión para una escuela, hospital o zona del territorio en cuestión. En las siguientes secciones se profundizará en cada una de las actividades realizadas.

Antecedentes

La brecha tecnológica en el mundo representa un problema para alcanzar la igualdad en distintos ámbitos: educativo, social e incluso económico. En distintas regiones del mundo, así como dentro de un mismo país, el acceso a servicios de telecomunicaciones en zonas rurales o comunidades poco accesibles significa un fuerte impedimento para el desarrollo integral de la población (World Bank, 2016).

En los últimos años se ha realizado un esfuerzo conjunto por parte de gobiernos e iniciativas privadas en distintos países con el fin de reducir la desigualdad en el ámbito tecnológico, tal como se refleja en programas como "Internet para todos" en México (Secretaría de Comunicaciones y Transportes, s.f.). Particularmente para la región de América Latina, el presupuesto que los gobiernos destinan a los proyectos relativos a infraestructura de telecomunicaciones, específicamente la respectiva a internet, ha aumentado considerablemente derivado del aumento exponencial de dispositivos que requieren de este para su correcto funcionamiento (IDC Corporate, 2022). Sin embargo, el despliegue de infraestructura no es una tarea sencilla; desde el momento

de la planeación del proyecto, hasta su implementación y mantenimiento, se requiere de una gran cantidad de mano de obra especializada que pueda llevar a cabo la gestión necesaria para cada etapa del proyecto.

Como se ha mencionado anteriormente, Huawei Technologies, al ser una de las compañías con más fuerza a nivel mundial respecto a la infraestructura de telecomunicaciones, ha tenido una presencia e influencia inconmensurable para el avance de estas tecnologías en la región de LATAM. Según un comunicado reciente de Huawei (2024), la empresa incluso hoy sigue trabajando en el desarrollo de infraestructuras TIC en la región para promover una conectividad más amplia.

El proyecto en el que laboré se enfoca en la etapa de planeación para este tipo de iniciativas. En este, mediante el uso de datos, tanto geográficos como aquellos correspondientes a miles de pruebas de conexión, se podía hacer un análisis medianamente profundo acerca de la necesidad de implementación de tecnologías de transmisión en distintas comunidades, o bien, si es que estas ya contaban con conexión, determinar la posible necesidad de implementar nuevas tecnologías que permitieran alcanzar mayores velocidades de carga o descarga, dependiendo de las intenciones del organismo que estuviese realizando el análisis.

Figura 5: Importancia de facilitar el uso de tecnología en zonas rurales (Imagen generada con Inteligencia Artificial)

El proyecto "Latam-APP"

Este proyecto, perteneciente al departamento de Ventas y Marketing para América Latina, surge a partir de la intención del gigante tecnológico "Huawei" de ser el proveedor de la tecnología necesaria a los gobiernos y/o empresas latinoamericanas para la mejora de su infraestructura de telecomunicaciones o bien, para la transformación de centros educativos y de salud mediante tecnologías de Smart Campus. Bajo esta premisa, y la intención de la empresa de siempre brindarle al cliente una solución adecuada a sus necesidades, se optó por desarrollar una herramienta que fuese auxiliar para determinar aquellas zonas que fueran prioritarias en el planteamiento de un proyecto de infraestructura, ya fuese por su localización estratégica, número de habitantes o rezago tecnológico.

De esta forma, y también tomando en cuenta la creciente tendencia del uso de datos para una toma de decisiones informada, el departamento optó por desarrollar un dashboard interactivo por cada país de Latinoamérica, basado en información de fuentes oficiales para exponer las necesidades de conectividad en ciertas zonas geográficas y una posible solución a las mismas.

Mi equipo, conformado por 3 ingenieros recién egresados bajo el liderazgo de un ingeniero especialista de China, fue el encargado de realizar el dashboard mencionado para este proyecto, desde la recopilación de los datos, su entendimiento, transformación y posteriormente su traducción en gráficos entendibles y útiles al cliente. Siendo, igualmente, nosotros los que presentaban y exponían su uso y conclusiones al interesado final.

El proyecto tuvo una duración aproximada de un año, durante el cual llegamos a realizar aproximadamente 15 dashboards distintos con información variada y eventualmente con objetivos un tanto distintos a los que planteaba el objetivo inicial del proyecto. Aunado a esto, llegamos a exponer estos, a una cantidad considerable de clientes siendo muy elogiados debido a su información completa y fácil de entender.

Este proyecto vio su fin debido a que agotamos los países de interés y a que no existe gran variabilidad en el panorama de conectividad en el corto plazo para las zonas rurales de la región.

Figura 6: Ejemplo a grandes rasgos de LATAM APP (Imagen generada con Inteligencia Artificial)

Definición del problema

El conflicto principal por resolver en el proyecto se trata de la falta de conectividad en determinadas zonas de los distintos países en Latinoamérica. La aproximación que se tomó como propuesta de solución a este problema, involucra gran cantidad de otros obstáculos que se fueron sorteando gradualmente conforme a las distintas situaciones que enfrentaba.

En primera instancia, me encontraba con la interrogante: ¿Qué zonas son de interés para el proyecto?; para cuya respuesta había distintos aspectos a tomar en cuenta, tales como:

- Número de establecimientos en la localidad
- Población
- Tipo de establecimiento y su nivel
- Conectividad existente
- Para algunos casos, la cercanía con infraestructura ya existente

La solución encontrada para sortear cada uno de estos aspectos y poder tener una visión general sobre la situación de una zona analizada, se encontró en el Análisis de Datos como disciplina. El uso de esta, a su vez involucra algunos otros problemas a resolver propios de la misma, como lo son la recopilación de los datos, entendimiento, limpieza, procesamiento, despliegue y finalmente estructuración de una historia.

El segundo conflicto enfrentado durante mi trabajo se puede describir como: ¿A quiénes les interesa tener esta información?; Durante mi labor en la empresa, constantemente nos entrevistábamos con clientes potenciales a los que se tenía que mostrar el producto terminado, el proceso para su obtención y sus posibles usos. Durante estas negociaciones, a causa de la dificultad de mis supervisores inmediatos para hablar nuestro idioma, tuve la oportunidad de hablar con representantes de gobiernos y empresas Latinoamericanos, dispuestos a hacerse con la analítica generada para determinar la viabilidad de sus proyectos, no únicamente hablando en términos de infraestructura, sino también para llevar sus negocios a zonas de potencial interés.

Metodología utilizada

Obtención de datos

Figura 7: Analogía del procesamiento de datos: Obtención (Imagen generada con Inteligencia Artificial)

En la obtención de los datos requeridos para nuestro análisis utilizábamos diversas técnicas, que van desde lo más sencillo e intuitivo como lo es la búsqueda en portales abiertos, hasta algunas que requerían de un nivel medianamente complejo de programación para hacer uso de métodos como web scraping. A continuación, explicaré detalladamente su uso y las diversas consideraciones para tener en cuenta al elegir su uso:

- Portal de datos:

  Esta fue la técnica que utilicé con mayor frecuencia, debido a su facilidad, relativa rapidez y a que generalmente los datos obtenidos mediante este método solían contener una menor cantidad de errores.

  El objetivo de este método era buscar en portales gubernamentales los datos sobre la geolocalización de escuelas y hospitales, verificando por lo menos en dos fuentes alternativas la congruencia de los datos con cifras oficiales en censos nacionales. Generalmente, el procedimiento que se seguía era:

  1. Realizar una búsqueda sencilla en Google, especificando los datos a obtener y el país de interés
  2. Seleccionar algún portal gubernamental de entre los resultados obtenidos (identificados en su mayoría por contar con un dominio ". gob")
  3. Verificar la utilidad de los datos. En caso de que estos no cumplieran con las características requeridas, se podía buscar nuevamente en el catálogo de datos del mismo sitio. Dependiendo de la calidad de los datos encontrados, podía llegar a realizar varias iteraciones de este paso y los anteriores.
  4. Una vez obtenidos los datos, procedía a buscar cifras oficiales en comunicados gubernamentales o bien en artículos periodísticos que pudiesen ratificar de manera aproximada, la veracidad de estos.

  Para algunos países, la información también podía ser obtenida de los sitios correspondientes a la autoridad educativa o de salud del país en cuestión. En estos casos, la información podía de igual forma ser confiable, puesto que era la misma institución que llevaba un registro de los establecimientos existentes y de los servicios que ofrecía.

País	Sitio web utilizado
Ecuador	https://educacion.gob.ec/datos-abiertos/
Perú	https://www.datosabiertos.gob.pe/group/ministerio-de-educaci%C3%B3n
Brasil	https://dados.gov.br/home
Guatemala	https://datosabiertos.mineduc.gob.gt/
Panamá	https://www.meduca.gob.pa/direccion-plane/estadisticas
Argentina	https://www.argentina.gob.ar/educacion/evaluacion-e-informacion-educativa/padron-oficial-de-establecimientos-educativos
Uruguay	https://www.dgeip.edu.uy/listado-escuelas-publicas/
Colombia	https://www.datos.gov.co/Educaci-n/Listado-colegios-oficiales/48xt-tjyj/data
Chile	https://centroestudios.mineduc.cl/datos-abiertos/
Costa Rica	https://dgth.mep.go.cr/directorio-de-centros-educativos/

Figura 8: Ejemplos de portales oficiales consultados para obtención de información

En algunos otros casos, los datos obtenidos provenían de los portales de datos abiertos nacionales, los cuales en múltiples ocasiones proveían información que distaba muy levemente de los censos nacionales. Sin embargo, al no haber una diferencia significativa entre ambas, su uso de igual manera se consideraba adecuado.

También, hubo ocasiones donde los datos eran obtenidos desde el portal del organismo encargado de los estudios estadísticos del país en cuestión, por lo que la verificación de estos datos se hacía mediante la búsqueda de artículos periodísticos o notas provenientes de fuentes confiables, que proveyeran un número aproximado de los establecimientos que se buscaban.

Ya que, por la naturaleza del contenido, muchas veces los datos de una fuente no proporcionaban toda la información deseada, se tenía que buscar y complementar la misma desde diversos sitios.

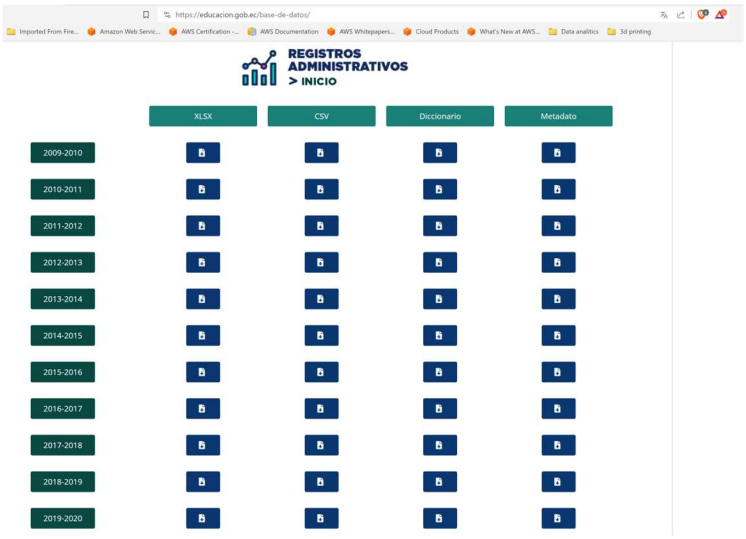


Figura 9: Sitio de datos abiertos del Ministerio de Educación de Ecuador (Ministerio de Educación de Ecuador, 2023.).

- Web Scraping: Requests y Beautiful Soup

Otro de los métodos utilizados para obtener datos fue la técnica popularmente conocida como web scraping (raspado web), mediante la cual es posible simular la navegación humana en un sitio de internet para extraer información de este. En este caso, la técnica fue implementada haciendo uso de los módulos Requests y BeautifulSoup, ambos del lenguaje de programación Python, los cuales fueron recomendados por parte del líder del proyecto. La elección de ellos, finalmente se debió a que ambos cuentan con una amplia documentación y gran comunidad, además de que su sintaxis se podría considerar sencilla y de relativamente fácil comprensión para personas que no están familiarizadas con peticiones HTTP.

Está técnica se utilizó en algunos países que no contaban con un portal de datos abiertos, o bien en los que otros métodos se volvían demasiado complejos al momento de obtener los datos deseados.

Haciendo uso de esta metodología, a su vez, utilicé dos variantes. Una en la cual realizaba peticiones a una API en un formato similar al utilizado en SQL (Lenguaje de Consulta Estructurada) y como respuesta obtenía tablas de datos; y otra en la cual, al utilizar el

módulo Requests en conjunto con la URL de la página en cuestión, obtenía el código fuente de esta para después analizar su estructura y obtener los datos correspondientes mediante sus etiquetas.

Como un ejemplo de la primera aproximación, escribiré acerca del caso particular de Panamá, país para el cual, a pesar de existir un portal de datos abiertos, no se encontró un documento que proveyera las coordenadas de los sitios deseados.

A pesar de que los datos existían, estos se encontraban dentro la página perteneciente al Instituto Nacional de Estadística donde se desplegaba una aplicación web que mostraba un mapa con todas las escuelas en el territorio nacional.

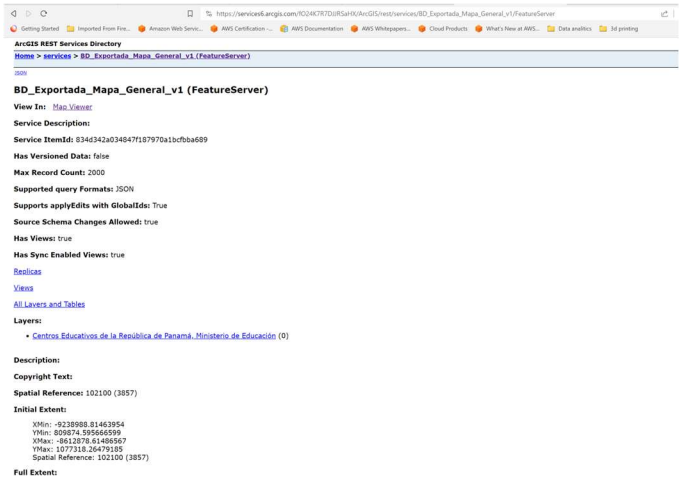


Figura 10: Interfaz de la API desde la cual se obtuvieron los datos de Panamá

Al hacer un breve análisis de la página haciendo uso de las herramientas del desarrollador del navegador en uso y realizando algunas búsquedas específicas, nos pudimos dar cuenta que las peticiones se hacían a una web API accesible desde el navegador.

Una vez habiendo localizado la URL a la que podíamos realizar consultas, se procedió a hacer un código en Python que tomara una URL base y pudiese realizar múltiples peticiones cambiando el sufijo de estas conforme a los requisitos específicos de la consulta realizada. El resultado final, era algo muy similar a una consulta de SQL.

Los resultados de las peticiones hechas se recibían en un documento JSON (JavaScript Object Notation), un formato de texto sencillo de la forma: "{'Llave': 'Valor'}", el cual era posteriormente procesado de forma similar a un diccionario de Python para extraer la información obtenida.

En cuanto la segunda aproximación, se tomará el ejemplo de Brasil, país para el cual se tuvieron que localizar establecimientos, además de educativos y de salud, también algunos pertenecientes a una firma privada.

Para la localización de estos datos se intentó en primera instancia tomar un enfoque similar al de Panamá. Sin embargo, la web-API desde donde se podían extraer estos datos únicamente aceptaba búsqueda mediante coordenadas, lo cual significaba que, para obtener el total de sucursales se tenía que realizar un barrido completo del país.

Debido a que para obtener el número íntegro de establecimientos buscados hubiese tenido que realizar consultas que variaran la distancia de Este a Oeste y de Norte a Sur en aproximadamente 2.5 kilómetros por todo el territorio de Brasil y la gran cantidad de tiempo que esto conllevaba, la opción fue determinada como inviable.

Al buscar vías alternativas de solución a este problema, encontramos una página web que desplegaba todos los sitios de interés en una interfaz interactiva. En esta, se comenzó obteniendo el código HTML para localizar los elementos que nos permitirían redirigir la navegación escalonadamente hacia a los datos deseados.

Ya que la información deseada se encontraba dividida por estados y estos a su vez, se encontraban divididos por municipalidades, era necesario acceder a cada una de las segmentaciones de forma iterativa para obtener el total de datos. Esto se puede comparar a la navegación en profundidad de un árbol.

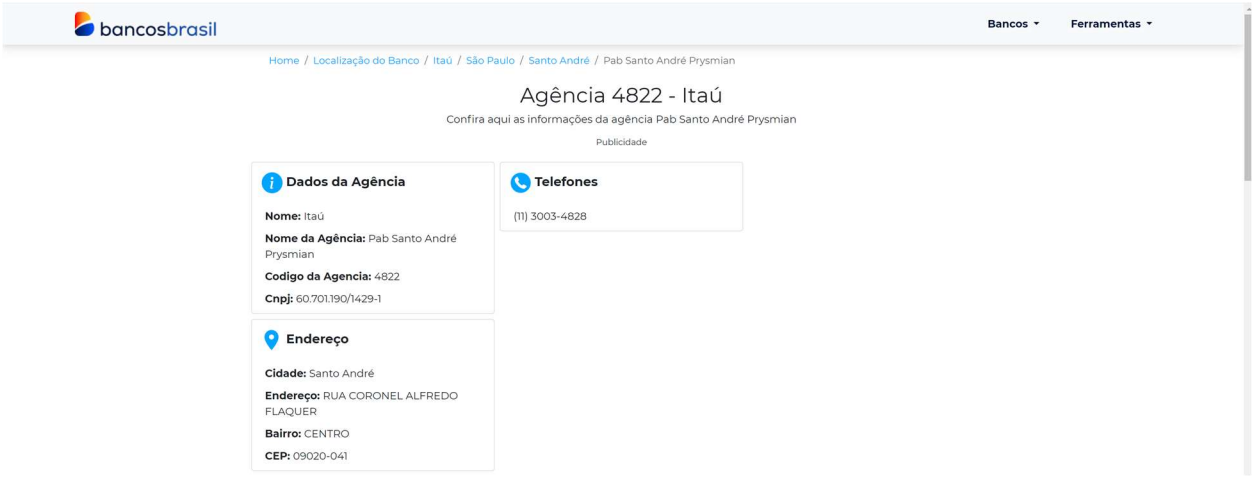


Figura 11: Ejemplo de uno de los sitios web donde los datos podían ser obtenidos

Finalmente, los datos eran localizados mediante las etiquetas en las que se encontraban envueltos, para ser colocados en listas de Python que serían posteriormente procesadas. Cabe aclarar que, en esta solución, no fue posible adquirir las coordenadas de los sitios deseados, sino su dirección. Esta sería posteriormente utilizada para localización de las coordenadas mediante geo codificación haciendo uso de la biblioteca Geopy.

Mediante el uso de ambas aproximaciones hice un uso extensivo de las estructuras de control básicas en Python al obtener todos los datos posibles existentes en la fuente. Gracias a esto, y a la constante búsqueda de formas para optimizar mis métodos, tuve un excelente progreso en la construcción de mis códigos que, a su vez, se traducían en una mejora para la forma en que estos realizaban sus funciones.

Por último, es necesario mencionar que en ambas variantes hice uso de Pandas, otro módulo en Python, utilizado ampliamente para el análisis y ciencia de datos, con el cual, me era posible entender mejor los datos obtenidos, así como visualizar fácilmente los errores que estos podían contener. En otras fases de este trabajo también hice uso de este módulo, las cuáles se abordarán más adelante.

	Name	State	Neighborhood	Zip-code	City	Address	Latitude	Longitude
0	SALVADOR CAJAZEIRAS	Bahia	CAJAZEIRAS	41342-315	Salvador	ESTRADA DO COQUEIRO GRANDE	-12.902433	-38.403966
1	SALVADOR/PARIPE	Bahia	PARIPE	40800-570	Salvador	AV. AFRANIO PEIXOTO,	-12.853022	-38.474910
2	ITURAMA - MG	Minas Gerais	CENTRO	38280-000	Iturama	RUA ITUIUTABA, 558	-19.729062	-50.199098
3	BALSA NOVA/PR	Paraná	CENTRO	83650-000	Balsa Nova	AV. BRASIL,611	-25.581073	-49.629527
4	CURITIBA/FRANCISCO DEROSSO	Paraná	XAXIM	81830-285	Curitiba	RUA FRANCISCO DEROSSO	-25.522887	-49.255387
...	...	...	...	...	...	...	...	...
2968	RIO LUCAS	Rio de Janeiro	VIGÁRIO GERAL	21250-381	Rio de Janeiro	RUA BULHÕES MARCIAL, 317-A SALAS 201, 202, 203...	-22.875483	-43.415116
2969	RIO/BANDEIRA	Rio de Janeiro	PRAÇA DA BANDEIRA	20270-001	Rio de Janeiro	RUA MARIZ E BARROS, 32/40	-22.582283	-43.168136
2970	RIO/BANDEIRA	Rio de Janeiro	PRAÇA DA BANDEIRA	20270-001	Rio de Janeiro	RUA MARIZ E BARROS, 318 B - PARTE	-22.582283	-43.168136
2971	RIO/RUA BUENOS AIRES	Rio de Janeiro	CENTRO	20061-001	Rio de Janeiro	RUA BUENOS AIRES, 318/320	-22.582283	-43.168136
2972	TRAJANO DE MORAIS (RJ)	Rio de Janeiro	CENTRO	28750-000	TRAJANO DE MORAIS	PRAÇA NILO PEÇANHA, 4	-22.066522	-42.067048

2973 rows × 8 columns

Figura 12: Ejemplo del resultado final

- Web Scraping: Selenium

Otra de las herramientas utilizadas fue Selenium, una biblioteca de Python ampliamente utilizada para la automatización de procesos, principalmente en cuanto a pruebas de desempeño de sitios web. Esta herramienta fue elegida debido a que, durante mi búsqueda de soluciones para interactuar con el contenido de una página de internet, encontré una gran cantidad de recursos sobre el uso de esta, además de diversos tutoriales y ejemplos en los que me pude basar para alcanzar mi objetivo.

El uso de esta herramienta en el web scraping incluye un controlador de distintos navegadores usado para poder extraer el código HTML de la página web en cuestión, interactuar con ella y de esta forma obtener los datos que sean de interés.
En esta aproximación, se trató de simular en la forma más fiel posible, el comportamiento humano dentro de una página. Esto resultaba muy útil cuando en las webs visitadas se tenía que realizar scrolling (desplazamiento con el ratón) para la aparición de nuevos elementos. La gran virtud del uso de Selenium es que se puede programar la navegación web hacia un sitio en específico, contando con la posibilidad de interactuar con los distintos elementos

del navegador y teniendo la capacidad de realizar acciones como un clic, clic derecho, scrolling e incluso ingresar texto.

Un ejemplo de uso de esta técnica fue igualmente para Panamá, con el fin de obtener las coordenadas de localidades en un sitio distinto al descrito en el ejemplo anterior.
Para poder ejecutar esta aproximación primero tuve que documentar paso a paso las acciones que llevaba a cabo al momento de extraer los datos de forma manual, para poder crear un algoritmo que pudiese interactuar de la misma forma y obtener la información deseada de forma repetitiva.
Una vez teniendo el algoritmo de comportamiento para acceder a los elementos que eran de mi interés, procedí a programar la segunda parte de esta técnica, la cual es correspondiente a análisis de los elementos mediante sus etiquetas, almacenando los datos deseados en listas de Python que posteriormente eran usadas para alimentar un DataFrame en Pandas. Sin embargo, al usar este método me pude dar cuenta que no era útil para la extracción de grandes volúmenes de información, ya que, al imitar un comportamiento humano en un navegador, se ve limitado por los tiempos de carga y descarga de la página en cuestión, dependiendo también de la cantidad de gráficos en la web, videos y demás elementos.

Esta forma de extracción de datos no fue muy utilizada a lo largo del proyecto, ya que al tener que extraer volúmenes de datos generalmente superiores a los 10 000 elementos, la obtención podía tardar varios días. Aun así, es importante mencionarla para señalar las distintas aproximaciones que se trataron de dar al problema y cómo se intentaron distintas soluciones con el fin de encontrar la más adecuada al caso en cuestión.

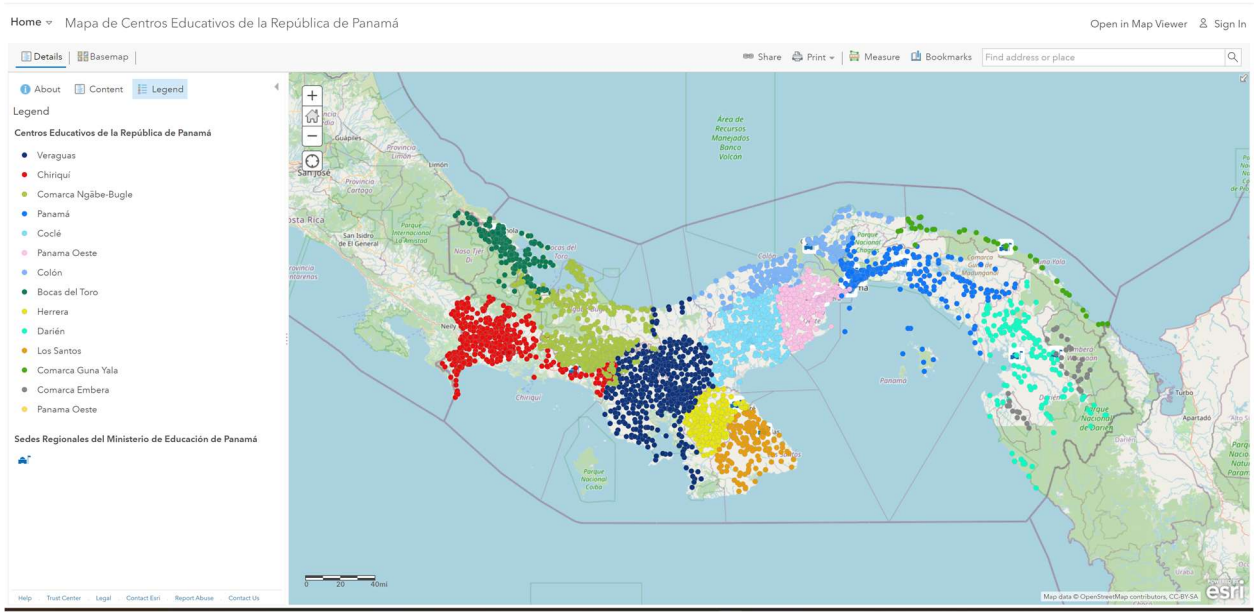


Figura 13: Ejemplo del sitio desde el cuál se obtuvieron los datos para Panamá

- <u>Mediante análisis de texto en documentos</u>

  Esta técnica es un poco más sencilla que el resto. Sin embargo, tuvo que ser utilizada debido a que, para un país en particular se encontraron únicamente documentos en formato PDF que contenían datos acerca de la localización de entidades de salud.

  Para esto, tuve que emplear una herramienta web denominada: "Convertidor PDF a Excel" que precisamente ejecutaba la tarea explícita en su nombre. Aunque los resultados proveídos por la misma distaban mucho de ser perfectos, sí facilitaron la tarea en gran medida ya que únicamente tuve que copiar el texto que había sido detectado por el convertidor y colocarlo en una tabla adicional que contaba con el formato deseado.

Los registros obtenidos mediante el uso de documentos en PDF no alcanzaban los 1000 establecimientos, por lo que se optó por el uso del método manual arriba descrito en lugar de desarrollar un código que se encargara del reconocimiento de texto en los documentos.

Gobierno de Reconciliación y Unidad Nacional
El Pueblo, Presidente!

RED DE SALUD PÚBLICA – MUNICIPIO DE BOACO

SILAIS BOACO	MUNICIPIO	NOMBRE DE LA UNIDAD	DIRECCION
BOACO	BOACO	Centro de Salud Ramón Guillen Navarro (Sede Municipal)	Frente al Parque José Nieborowsky
BOACO	BOACO	Centro de Salud Ramón Toledo	Antiguas instalaciones Hospital Jose Nieborowski
BOACO	BOACO	Puesto de Salud Tierra Azul	Comarca Tierra Azul
BOACO	BOACO	Puesto de Salud San Buenaventura	Comarca San Buenaventura
BOACO	BOACO	Puesto de Salud Santa Elisa	Asentamiento Santa Elisa
BOACO	BOACO	Puesto de Salud Yula Sacal	Comarca Yula Sacal
BOACO	BOACO	Puesto de Salud San José de la Vega	Comarca San José de la Vega
BOACO	BOACO	Puesto de Salud El Paraíso	Comarca El Paraíso
BOACO	BOACO	Puesto de Salud Boaco Viejo	Comarca Boaco Viejo
BOACO	BOACO	Puesto de Salud Las Lagunas	Comarca Las Lagunas
BOACO	BOACO	Puesto de Salud San Nicolás	Contiguo a la Iglesia Juan Pablo II.
BOACO	BOACO	Puesto de Salud Santa Inés	Comarca Santa Inés
BOACO	BOACO	Puesto de Salud Erikson Campos Sotelo. (Mercado)	Mercado Municipal de Boaco

FE, FAMILIA Y COMUNIDAD!
CRISTIANA, SOCIALISTA, SOLIDARIA!
MINISTERIO DE SALUD

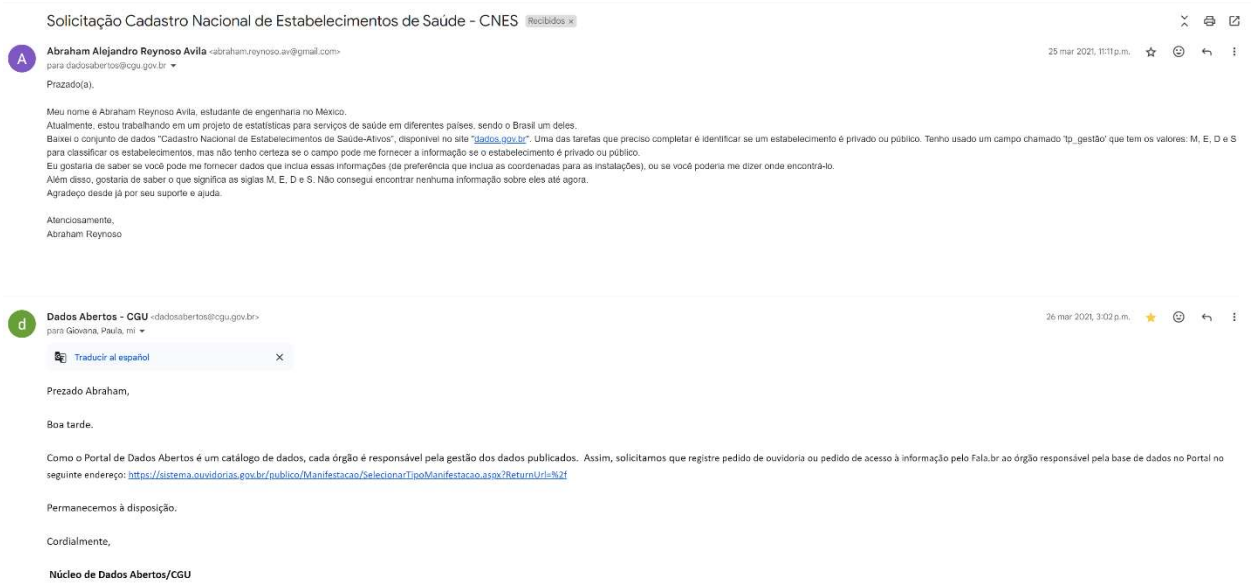
Figura 14: Documento electrónico con datos de establecimientos en Nicaragua

- Contacto directo

Para terminar con los distintos métodos de obtención de datos, mencionaré de forma muy breve esta aproximación que fue utilizada en dos ocasiones para países de Centroamérica. Su uso se debió a la falta de un portal de datos abiertos y a la inexistencia de los datos requeridos en sitios oficiales.

La solución propuesta por nuestro líder de equipo fue contactar mediante correo electrónico a las dependencias correspondientes con el fin de exponer brevemente el objetivo del proyecto, presentarnos como desarrolladores de este, mostrar a manera de ejemplo algunos de los resultados obtenidos para otros países y finalmente hacer una requisición formal de

los datos. Esta idea fue exitosa en ambas ocasiones que fue implementada, con el único detalle que en la segunda ocasión nos fue requerido realizar una llamada con la persona responsable de la protección de estos datos con el fin de asegurar que estos fueran usados únicamente con el fin que se había señalado.



Solicitação Cadastro Nacional de Estabelecimentos de Saúde - CNES Recibidos ×

Abraham Alejandro Reynoso Avila <abraham.reynoso.av@gmail.com> 25 mar 2021, 11:11 p.m.
para dadosabertos@cgu.gov.br

Prazado(a),

Meu nome é Abraham Reynoso Avila, estudante de engenharia no México.
Atualmente, estou trabalhando em um projeto de estatísticas para serviços de saúde em diferentes países, sendo o Brasil um deles.
Baixei o conjunto de dados "Cadastro Nacional de Estabelecimentos de Saúde-Ativos", disponível no site "dados.gov.br". Uma das tarefas que preciso completar é identificar se um estabelecimento é privado ou público. Tenho usado um campo chamado 'tp_gestão' que tem os valores: M, E, D e S para classificar os estabelecimentos, mas não tenho certeza se o campo pode me fornecer a informação se o estabelecimento é privado ou público.
Eu gostaria de saber se você pode me fornecer dados que inclua essas informações (de preferência que inclua as coordenadas para as instalações), ou se você poderia me dizer onde encontrá-lo.
Além disso, gostaria de saber o que significa as siglas M, E, D e S. Não consegui encontrar nenhuma informação sobre eles até agora.
Agradeço desde já por seu suporte e ajuda.

Atenciosamente,
Abraham Reynoso

Dados Abertos - CGU <dadosabertos@cgu.gov.br> 26 mar 2021, 3:02 p.m.
para Giovana, Paula, mi

Traducir al español ×

Prezado Abraham,

Boa tarde.

Como o Portal de Dados Abertos é um catálogo de dados, cada órgão é responsável pela gestão dos dados publicados. Assim, solicitamos que registre pedido de ouvidoria ou pedido de acesso à informação pelo Fala.br ao órgão responsável pela base de dados no Portal no seguinte endereço: https://sistema.ouvidorias.gov.br/publico/Manifestacao/SelecionarTipoManifestacao.aspx?ReturnUrl=%2f

Permanecemos à disposição.

Cordialmente,

Núcleo de Dados Abertos/CGU

Figura 15: Correo electrónico con solicitud de datos al gobierno de Brasil

Los métodos mencionados fueron utilizados en su mayoría para obtener datos de geolocalización con coordenadas para los sitios de interés. Sin embargo, hubo ocasiones en que los sitios encontrados no contenían una coordenada geoespacial, por lo que se tuvo que realizar un paso adicional con el fin de encontrar estas. A continuación, describiré brevemente este paso adicional, así como algunas de las distintas implementaciones de código que se hicieron conforme el proyecto avanzaba.

Obtención de coordenadas con GeoPy

GeoPy (GeoPy Contributors, 2024) es una biblioteca de Python especializada en la geolocalización, es decir, obtención de coordenadas mediante los datos proporcionados a una web API.

Esta biblioteca cuenta con una gran flexibilidad en cuanto a motores de búsqueda, ya que cuenta con más de 30 opciones a elegir de ellos, siendo algunos libres de pago y otros sujetos a una suscripción. Cada uno de ellos cuenta con distintas fortalezas y debilidades, y su elección depende completamente del objetivo del usuario y necesidades del proyecto. Geopy fue elegida frente a otras opciones como el módulo de geocodificación debido a la gran cantidad de documentación por parte de su comunidad y variedad de ejemplos que se pueden encontrar en la web.

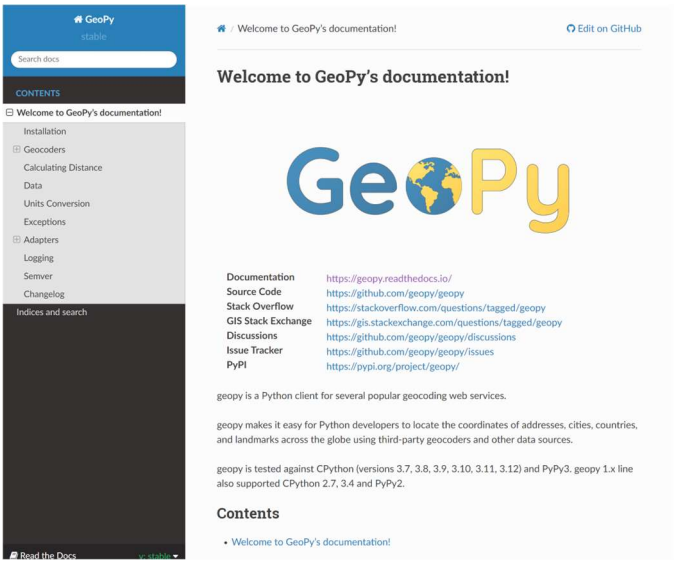


Figura 16: Documentación de Geopy

Para hacer uso de esta herramienta, nos apoyamos principalmente con el motor 'Nominatim', ya que este proveía una alta confianza en los resultados que brindaba, además de ser uso libre tanto personal, como comercialmente. Sin embargo, en un par de ocasiones, también realizamos algunas búsquedas haciendo uso de la API de Google Maps (Google Developers, s.f.), siendo sus resultados aún más precisos y confiables que la anterior, pero con la desventaja que, debido a los términos y condiciones de uso señalados en su documentación, no era posible el uso de sus resultados con un fin comercial.

Como fue mencionado en la sección pasada, este método, “Geolocalización”, era utilizado constantemente al momento de la recolección de datos para los distintos países que nos eran asignados, ya que en todos estos, aun cuando los datos hubiesen sido recolectados directamente en páginas web de dependencias oficiales, se presentaban irregularidades tales como la falta de coordenadas para algunos sitios, coordenadas fuera del país en cuestión, coordenadas con formato irregular e incluso para algunos, ausencia total de las mismas.

En un principio, cuando empecé a laborar en esta empresa y fui asignado en este proyecto, tenía desconocimiento total acerca de esta herramienta, además de que mis conocimientos en el lenguaje de programación Python eran muy limitados. Fue de esta forma, que un compañero más experimentado en cuestiones de programación brindó al equipo un primer código de geolocalización que posteriormente se iría adaptando y evolucionando de acuerdo con las necesidades especiales para la aplicación del país de interés, los datos de entrada e inclusive los conocimientos que íbamos adquiriendo conforme nuestra labor avanzaba.

Este primer código era una aproximación sencilla que necesitaba de un archivo CSV (Valores Separados por Comas) con un formato específico como entrada y posteriormente escribía un segundo archivo CSV, el cuál incluía ya las coordenadas. Este mismo contaba con manejo de excepciones o errores, evitando que la ejecución terminara inesperadamente cada que apareciera un sitio para el cual no se encontraran coordenadas.

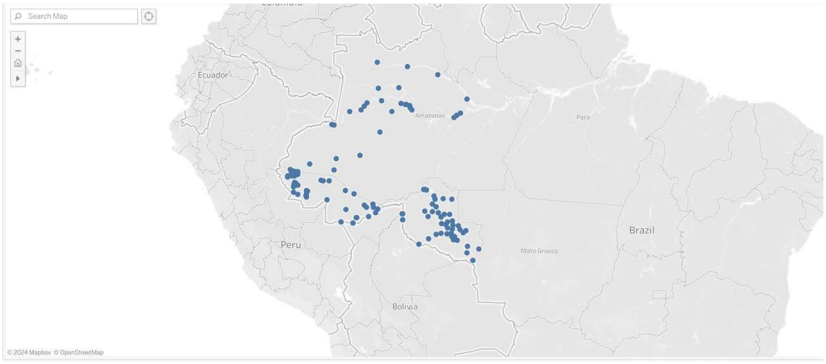


Figura 17: Visualización de coordenadas obtenidas con Geopy usando Tableau

La primera modificación importante que se hizo a este código se dio cuando nos dimos cuenta de que, para escuelas y hospitales localizados en zonas rurales, en rara ocasión se encontraban

coordenadas cuando se hacía uso de la dirección completa, habiendo ocasiones en las que incluso a más de 10,000 registros se les asignaba una coordenada con valor nulo.

Para solucionar este problema, se aprovechó la flexibilidad del proyecto en cuanto a la localización de un establecimiento en específico, ya que, por la naturaleza de este, no requeríamos de una localización exacta sino únicamente de una ubicación aproximada. Tomando esto en cuenta y una vez que contamos con la autorización de nuestro líder de proyecto, optamos por hacer un uso muy burdo de las excepciones en Python para que, en caso de que no se localizaran las coordenadas de un establecimiento, se pudiese efectuar una segunda acción en la cual se geolocalizara usando parámetros distintos.

Cuando llevamos esta acción a cabo, nos dimos cuenta de que aún algunos registros eran procesados con una coordenada nula, por lo que se optó por tomar como referencia la primera excepción anteriormente mencionada, para ejecutarla repetidamente de forma anidada hasta que, en última instancia, se localizaran únicamente las coordenadas de la municipalidad del establecimiento en cuestión.

Posteriormente, derivado de lo descrito anteriormente, se tuvo que llevar a cabo otra modificación importante. Esto se realizó cuando, debido a la gran cantidad de peticiones de geolocalización que se llevaban a cabo, surgió la necesidad de acortar los tiempos de obtención de los datos.

Una de las mayores desventajas que el motor Nominatim posee, es que las peticiones realizadas están limitadas a una por segundo, por lo que, en países con un gran número de registros como lo eran Brasil, Colombia, Chile o Argentina, en donde el número de establecimientos era mayor a 30,000 y que además contaban con un alto número de establecimientos en zonas rurales en las cuales se tenían que efectuar hasta 3 peticiones antes de encontrar los valores deseados, los tiempos podían extenderse desde unas cuantas horas hasta un incluso, un par de días.

Tomando esto en cuenta y también que para una gran cantidad de establecimientos rurales la localización se basaba en la localidad y no en la dirección a nivel de calle, se hizo una modificación en la que se tomaban establecimientos cercanos entre sí, específicamente aquellos para los que no se había podido obtener una coordenada con la dirección completa, y se asignaba la misma coordenada para ambos.

En un principio, para que este algoritmo funcionara correctamente, se tenía que someter a un preprocesamiento a los datos recolectados, ordenándolos alfabéticamente de acuerdo con la localidad donde se encontraban ubicados, para de esta forma, almacenar los datos recabados de un registro anterior y en dado caso que la localidad fuese la misma, asignar las mismas coordenadas sin necesidad de realizar una nueva búsqueda.

Haciendo uso de esta solución, los tiempos de obtención de datos se redujo más de un 33%, sobre todo para aquellos países en los cuales la mayor parte de los establecimientos se encontraban en zonas rurales.

Por último, la modificación de mayor relevancia a este algoritmo se llevó a cabo cuando se empezó a implementar en conjunto con otra librería que ofrece una gran flexibilidad en cuanto al procesamiento de datos: Pandas. En este caso, se creó una función personalizada para la obtención de las coordenadas con GeoPy, que realizaba un mapeo de las distintas filas del DataFrame (estructura tabular de datos orientada a columnas), al mismo tiempo que procesaba, limpiaba y rellenaba los datos sin la necesidad de ejecutar varias operaciones distintas.

Al implementar esta mejora en la geolocalización y después de tomar algunos cursos respecto al análisis de datos, también empecé a utilizar Jupyter Notebook o bien, Notebooks en línea como una herramienta adicional para el desarrollo de nuevos scripts. Esto debido a la gran facilidad que estos proporcionan al momento de realizar pruebas y la practicidad que estos tienen cuando es preciso ejecutar código separado por bloques.

Name	State	Municipality	Type1	Type2	Type3	Type4	Type5	Type6	...	Type9	Type10	Scale1	Scale2	Scale3	Scale4	Scale5	Scale6	Latitude	Longitude
Escuela Secundaria Puerto Armuelles	Chiriqui	Barú	School	-	NaN	Compañía Distribuidora	Si	1	...	Fibra Óptica	Si	1296	10 Mb	NaN	NaN	NaN	NaN	8.272584	-82.859693
C.E.B.G. Cristobal Alabarca	Cocle	Penonomé	School	-	NaN	Panel Solar	Si	2	...	Sin Internet	No	143	SN/SV	NaN	NaN	NaN	NaN	8.845915	-80.241638
C.E.B.G. Los Elegidos	Cocle	Penonomé	School	-	NaN	Panel Solar	Si	3	...	Sin Internet	No	19	SN/SV	NaN	NaN	NaN	NaN	8.904804	-80.310249
C.E.B.G. San Cristobal	Cocle	Penonomé	School	-	NaN	Panel Solar	Si	4	...	Sin Internet	No	144	SN/SV	NaN	NaN	NaN	NaN	8.912581	-80.159971
Escuela La Mina de Rio Indio	Cocle	Penonomé	School	-	NaN	Panel Solar	Si	5	...	Sin Internet	No	28	SN/SV	NaN	NaN	NaN	NaN	8.939526	-80.145804

Figura 18: Ejemplo de DataFrame en pandas

Limpieza de Datos

Figura 19: Analogía del procesamiento de datos: Limpieza (Imagen generada con Inteligencia Artificial)

Una vez obtenidos los datos haciendo uso de los métodos anteriormente mencionados, se procedió a la limpieza y refinamiento de estos. Proceso para el cual también se contaba con diversos métodos que se utilizaban conforme a la operación a realizar o la cantidad de datos a modificar. A continuación, haré una breve descripción de ellos.

Microsoft Excel

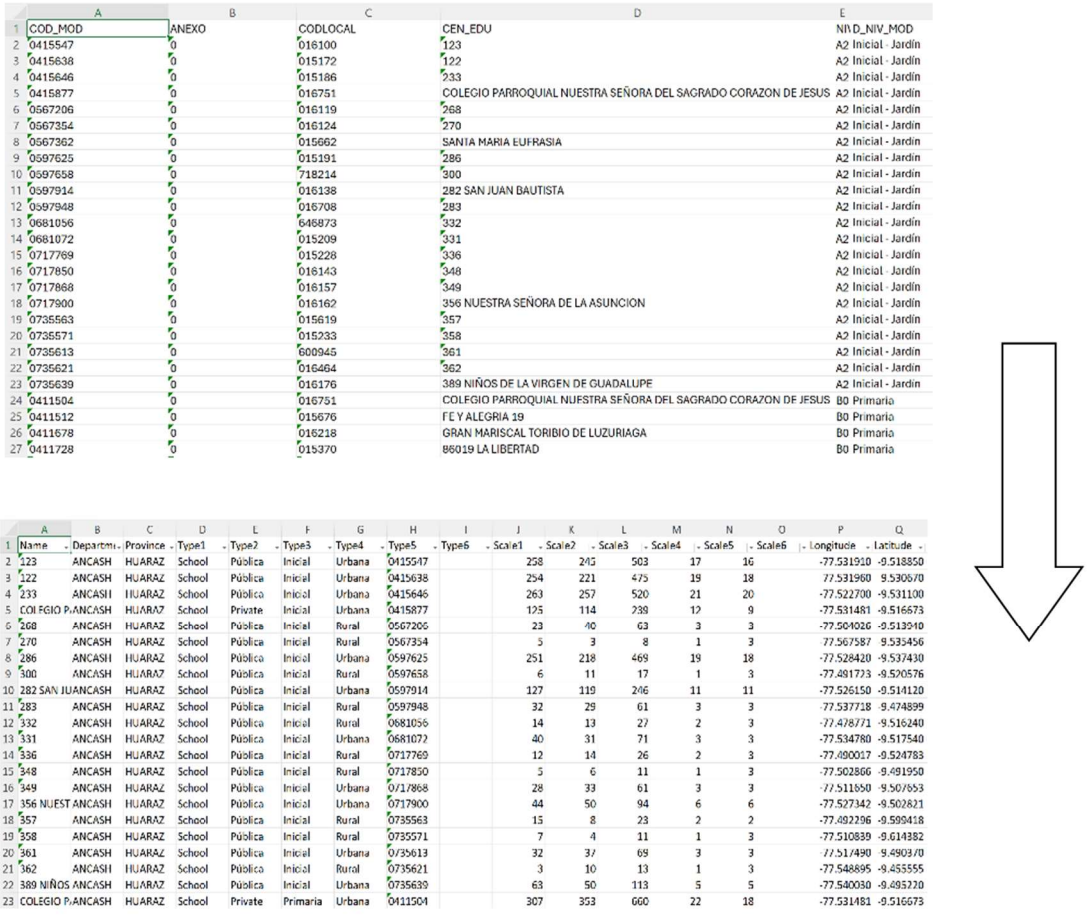
En un principio, cuando me integré a este proyecto, mi conocimiento acerca de procesamiento de datos y las herramientas usadas para este eran algo completamente nuevo para mí. Por lo que, para ejecutar las tareas más básicas de limpieza, utilizábamos hojas de cálculo en Microsoft Excel desde el cual podíamos ejecutar diversas tareas tales como reemplazamientos, eliminación de datos nulos, obtención de valores atípicos, etc. Aunque Excel es un software muy extenso, con mucha flexibilidad en su uso y muy fácil de manipular a alto nivel, suele tener problemas al momento de ejecutar tareas sobre grandes volúmenes de información o con gran variabilidad dentro de esta.

Otra principal razón por la que hacíamos uso de Excel era por que la plataforma donde posteriormente se subían los datos para una segunda etapa de procesamiento, admitía archivos

únicamente en formato XLSX, por lo que se optaba por tomar los archivos CSV generados en etapas anteriores, procesarlos dentro de Excel y posteriormente guardarlos en el formato adecuado.

No obstante, el uso de este software era generalmente necesario una vez que ya se tenían las coordenadas de los establecimientos de cada país, ya que se procedía a corroborar los locales para los cuales no se habían encontrado coordenadas, aquellos en cuyo nombre tenía algún carácter especial o bien, aquellos que contuvieran valores atípicos. Para dar solución a estos problemas, se hacía uso principalmente de las herramientas de filtrado, de eliminación de valores repetidos, de auto llenado, entre otras.

Se optó por dejar gradualmente de lado el uso de este software una vez que los volúmenes de estos empezaron a aumentar. Esto debido a que los tiempos de procesamiento para una operación sencilla eran de varios minutos, además de que, en ocasiones, las computadoras quedaban congeladas sin dejar más opción que reiniciarlas, perdiendo así toda la labor no guardada hasta el momento.



	A	B	C	D	E
1	COD_MOD	ANEXO	CODLOCAL	CEN_EDU	NIVD_NIV_MOD
2	0415547	0	016100	123	A2 Inicial - Jardín
3	0415638	0	015172	122	A2 Inicial - Jardín
4	0415646	0	015186	233	A2 Inicial - Jardín
5	0415877	0	016751	COLEGIO PARROQUIAL NUESTRA SEÑORA DEL SAGRADO CORAZON DE JESUS	A2 Inicial - Jardín
6	0567206	0	016119	268	A2 Inicial - Jardín
7	0567354	0	016124	270	A2 Inicial - Jardín
8	0567362	0	015662	SANTA MARIA EUFRASIA	A2 Inicial - Jardín
9	0597625	0	015191	286	A2 Inicial - Jardín
10	0597658	0	718214	300	A2 Inicial - Jardín
11	0597914	0	016138	282 SAN JUAN BAUTISTA	A2 Inicial - Jardín
12	0597948	0	016708	283	A2 Inicial - Jardín
13	0681056	0	646873	332	A2 Inicial - Jardín
14	0681072	0	015209	331	A2 Inicial - Jardín
15	0717769	0	015228	336	A2 Inicial - Jardín
16	0717850	0	016143	348	A2 Inicial - Jardín
17	0717868	0	016157	349	A2 Inicial - Jardín
18	0717900	0	016162	356 NUESTRA SEÑORA DE LA ASUNCION	A2 Inicial - Jardín
19	0735563	0	015619	357	A2 Inicial - Jardín
20	0735571	0	015233	358	A2 Inicial - Jardín
21	0735613	0	600945	361	A2 Inicial - Jardín
22	0735621	0	016464	362	A2 Inicial - Jardín
23	0735639	0	016176	389 NIÑOS DE LA VIRGEN DE GUADALUPE	A2 Inicial - Jardín
24	0411504	0	016751	COLEGIO PARROQUIAL NUESTRA SEÑORA DEL SAGRADO CORAZON DE JESUS	B0 Primaria
25	0411512	0	015676	FE Y ALEGRIA 19	B0 Primaria
26	0411678	0	016218	GRAN MARISCAL TORIBIO DE LUZURIAGA	B0 Primaria
27	0411728	0	015370	86019 LA LIBERTAD	B0 Primaria

	A	B	C	D	L	F	G	H	I	J	K	L	M	N	O	P	Q
1	Name	Departm	Province	Type1	Type2	Type3	Type4	Type5	Type6	Scale1	Scale2	Scale3	Scale4	Scale5	Scale6	Longitude	Latitude
2	123	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0415547		258	245	503	17	16		-77.531910	-9.518850
3	122	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0415638		254	221	475	19	18		-77.531960	-9.530670
4	233	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0415646		263	257	520	21	20		-77.522700	-9.531100
5	COLEGIO P	ANCASH	HUARAZ	School	Private	Inicial	Urbana	0415877		125	114	239	12	9		-77.531481	-9.516673
6	268	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0567206		23	40	63	3	3		-77.504026	-9.513940
7	270	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0567354		5	3	8	1	3		-77.567587	-9.535456
8	286	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0597625		251	218	469	19	18		-77.528420	-9.537430
9	300	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0597658		6	11	17	1	3		-77.491733	-9.520576
10	282 SAN JU	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0597914		127	119	246	11	11		-77.526150	-9.514120
11	283	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0597948		32	29	61	3	3		-77.537718	-9.474899
12	332	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0681056		14	13	27	2	3		-77.478771	-9.516240
13	331	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0681072		40	31	71	3	3		-77.534780	-9.517540
14	336	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0717769		12	14	26	2	3		-77.490017	-9.524783
15	348	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0717850		5	6	11	1	3		-77.502866	-9.491950
16	349	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0717868		28	33	61	3	3		-77.511650	-9.507653
17	356 NUEST	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0717900		44	50	94	6	6		-77.527342	-9.502821
18	357	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0735563		15	8	23	2	2		-77.492296	-9.599418
19	358	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0735571		7	4	11	1	3		-77.510839	-9.614382
20	361	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0735613		32	37	69	3	3		-77.517490	-9.490370
21	362	ANCASH	HUARAZ	School	Pública	Inicial	Rural	0735621		3	10	13	1	3		-77.548895	-9.455555
22	389 NIÑOS	ANCASH	HUARAZ	School	Pública	Inicial	Urbana	0735639		63	50	113	5	5		-77.540030	-9.495220
23	COLEGIO P	ANCASH	HUARAZ	School	Private	Primaria	Urbana	0411504		307	353	660	22	18		-77.531481	-9.516673

Figura 20: Resultado del Procesamiento de Datos

Jupyter Notebooks y Pandas

La razón por la que menciono estas dos herramientas en un solo apartado es porque a lo largo de todo el proyecto fueron utilizadas de forma conjunta. Esto fue debido a que, por la naturaleza del desarrollo de soluciones en Pandas, los notebooks de Jupyter son ideales para hacer visualizaciones continuas y una depuración fluida.

Jupyter es un ambiente de desarrollo interactivo basado en web para poder realizar anotaciones, escribir código y hacer operaciones centradas en datos. Su interfaz permite al usuario configurar y administrar flujos de trabajo en Ciencia de Datos, Computación científica, periodismo

Computacional y Machine Learning. Esta herramienta ha estado aumentando en gran medida su popularidad en años recientes, derivado también de la creciente demanda por el Análisis de datos en diversas industrias.

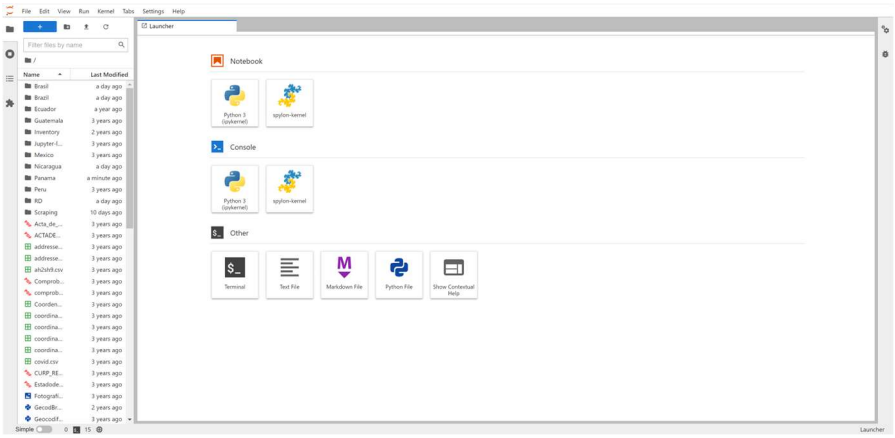


Figura 21: Interfaz de Usuario de Jupyter Notebook

Pandas (NumFOCUS, Inc., 2024), por otro lado, es una biblioteca de Python especializada en la manipulación y análisis de datos. Aunque actualmente existen otras bibliotecas o plataformas con mejores capacidades de rendimiento para el manejo masivo de información y que proveen un mayor abanico de herramientas para la extracción de características, Pandas sigue siendo una de las más populares debido a su flexibilidad para trabajar en conjunto con otras bibliotecas de Python, su despliegue sencillo para la visualización y su velocidad de procesamiento. La elección de esta herramienta se debió a mi familiaridad con ella después de haber trabajado un par de ocasiones en proyectos académicos con esta.

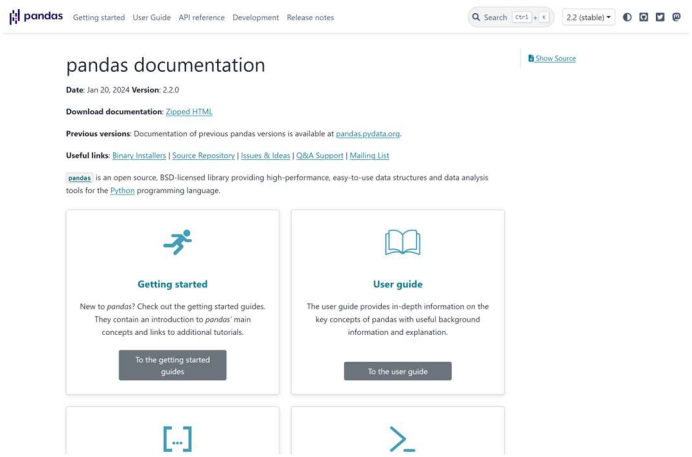


Figura 22: Documentación de Pandas

La forma en que empecé a utilizar Pandas y Jupyter como herramientas del día a día fue debido a que tuve un módulo de aprendizaje acerca de Python en un diplomado enfocado al de Análisis de Datos que empecé en noviembre de 2020. Aquí, se mencionó en numerosas ocasiones de la utilidad del uso de Notebooks en conjunto con Pandas al momento de realizar procesamiento sobre datos, visualización de estos, entre otras operaciones. Sin embargo, no fue sino hasta el cuarto mes de este diplomado que empezamos con el uso de estas herramientas, momento en que empecé a aplicar este conocimiento en mi labor diaria.

Jupyter empezó siendo para mí un simple editor de texto, cuyo uso, incluso era bastante incómodo debido a la carencia de varias utilidades presentes en otras herramientas similares. Sin embargo, conforme me fui adentrando más en su manejo, pude darme cuenta de las facilidades que este ofrecía al momento de realizar desarrollo de código, inclusive para tareas que no estuvieran relacionadas con los datos.

Debido a la capacidad de esta plataforma para realizar depuración por bloques, los despliegues que permite para el análisis exploratorio y la facilidad para elaborar documentación mientras se desarrolla código, fui sustituyendo gradualmente mi uso de editores de texto tradicionales por esta.

En cuanto al uso de Pandas, en un principio migré únicamente la lectura y escritura de archivos CSV haciendo uso de la facilidad que provee Pandas para este menester. Sin embargo, de la misma forma que pasó con Jupyter Notebooks, una vez que fui profundizando en las facilidades que esta

biblioteca brindaba para las operaciones por lotes de datos, así como la alta velocidad en que se terminaban las tareas, fui reemplazando gradualmente el uso de Excel hasta el punto en que en ningún momento tenía que hacer uso de la plataforma de Microsoft.

A continuación, describiré brevemente algunas de las tareas efectuadas haciendo uso de Pandas con un breve ejemplo de cada una de ellas:

- Reemplazamiento de caracteres especiales

  Esta tarea fue realizada haciendo uso simplemente de la función *replace* para las cadenas de texto en Python, específicamente dentro de una estructura de Pandas. Mediante una función definida por mí como usuario, era capaz de iterar elemento a elemento del DataFrame para que, en caso de que se encontrara un carácter que incluyera un acento o bien, un símbolo que se tenía por sentado que no sería admitido por la siguiente plataforma del proceso, era reemplazado ya fuese por una consonante sin acento o bien, por un espacio en blanco. Esta operación fue uno de los puntos más fuertes que observé una vez que decidí cambiar mi procesamiento de datos a Python, ya que usando el método anteriormente utilizado se tenían que ejecutar por lo menos 5 pasos adicionales antes de obtener el mismo resultado.

```
df.head()
```

	REGION	LEVEL ID	SCHOOL NAME	SCHOOL GRADE	ADDRESS
0	01	NUR	PETER & PAUL /C	O	PETER & PAUL PRIMARY SCHOOL
1	01	NUR	FALL TOP/C	O	ECLIPSE FALLS, PORT KAITUMA
2	01	NUR	WARAMURI	D	WARAMURI, MORUCA
3	01	NUR	BUMBURY	D	BUMBURY HILL, BARIMA/WAINI
4	01	NUR	SANTA ROSA	B	SANTA ROSA MISSION, KUMAKA, MORUCA

```
special_characters = "@#$%^&*/"
df.applymap(lambda x: ''.join([char for char in x if char not in special_characters])).head()
```

	REGION	LEVEL ID	SCHOOL NAME	SCHOOL GRADE	ADDRESS
0	01	NUR	PETER PAUL C	O	PETER PAUL PRIMARY SCHOOL
1	01	NUR	FALL TOPC	O	ECLIPSE FALLS, PORT KAITUMA
2	01	NUR	WARAMURI	D	WARAMURI, MORUCA
3	01	NUR	BUMBURY	D	BUMBURY HILL, BARIMAWAINI
4	01	NUR	SANTA ROSA	B	SANTA ROSA MISSION, KUMAKA, MORUCA

Figura 23: Ejemplo de un procedimiento estándar de reemplazamiento de caracteres especiales

- Filtrado de datos atípicos

Otra de las tareas que podía realizar fácilmente haciendo uso de Pandas era el filtrado de establecimientos que contaban con valores erróneos en sus coordenadas. Esto era realizado haciendo uso de las herramientas que esta biblioteca posee apara el análisis de datos. Una vez teniendo una referencia acerca de las latitudes y longitudes máximas y mínimas de cada país (las cuales podían ser fácilmente encontradas mediante una búsqueda en internet), se procedía a usar la función *filter* incluida en Pandas.

Otro ejemplo de filtrado de datos que también se llegó a utilizar fue cuando se tenían datos irregulares en cuanto a la población en una localidad. Para esto, se utilizaba la media y desviación estándar para detectar los datos que se encontraban a una y media desviación estándar del rango intercuartílico de los datos.

Generalmente, cuando se encontraban datos atípicos, ya fuese dentro de coordenadas o en algún otro campo, se procedía a hacer un rápido reprocesamiento de la información con el objetivo de recuperar el registro ya con información corregida.

```
[35]: df2
```

[35]:

	Name	Department	Province	Latitude	Longitude
0	Universidad Nacional Mayor de San Marcos	Lima	Lima	-12.056158	-77.084520
1	Universidad Nacional de San Cristóbal de Huamanga	Ayacucho	Huamanga	-13.161248	-74.225772
2	Universidad Nacional de San Antonio Abad del C...	Cusco	Cusco	-13.521930	-71.958321
3	Universidad Nacional de Trujillo	La Libertad	Trujillo	-8.115007	-79.038305
4	Universidad Nacional de San Agustín de Arequipa	Arequipa	Arequipa	-16.397139	-71.537144
...	...	...	...	...	...
138	Universidad Santo Tomás de Aquino de Ciencia e...	Junín	Huancayo	-12.084040	-75.208442
139	Universidad Privada SISE	Lima	Lima	-12.017367	-77.004572
140	Universidad Seminario Evangélico de Lima (*12)	Lima	Lima	-12.063959	-76.960074
141	Universidad Seminario Bíblico Andino (*12)	Lima	Lima	-12.069621	-77.053398
142	Universidad Católica San José	Lima	Lima	30.248333	-81.625278

143 rows × 5 columns

```
[27]: Q1_Latitude = df2['Latitude'].quantile(0.25)
Q3_Latitude = df2['Latitude'].quantile(0.75) #Calculo de cuantiles 0.25 y 0.75 para Longitud y Latitud
Q1_Longitude = df2['Longitude'].quantile(0.25)
Q3_Longitude = df2['Longitude'].quantile(0.75)
IQR_Latitude = Q3_Latitude - Q1_Latitude #Calculo de Rangos intercuartílicos
IQR_Longitude = Q3_Longitude - Q1_Longitude
umbral_inferior_Latitude = Q1_Latitude - 1.5 * IQR_Latitude
umbral_superior_Latitude = Q3_Latitude + 1.5 * IQR_Latitude
umbral_inferior_Longitude = Q1_Longitude - 1.5 * IQR_Longitude  #Umbrales límite para considerar datos atípicos
umbral_superior_Longitude = Q3_Longitude + 1.5 * IQR_Longitude
```

Figura 24.1: Calculo de rango intercuartílico para el filtrado de datos

```
df2[((df2['Latitude']>=umbral_inferior_Latitude) & (df2['Latitude']<=umbral_superior_Latitude)) &
                    ((df2['Longitude']>=umbral_inferior_Longitude) & (df2['Longitude']<=umbral_superior_Longitude))]
```

	Name	Department	Province	Latitude	Longitude
0	Universidad Nacional Mayor de San Marcos	Lima	Lima	-12.056158	-77.084520
1	Universidad Nacional de San Cristóbal de Huamanga	Ayacucho	Huamanga	-13.161248	-74.225772
2	Universidad Nacional de San Antonio Abad del C...	Cusco	Cusco	-13.521930	-71.958321
3	Universidad Nacional de Trujillo	La Libertad	Trujillo	-8.115007	-79.038305
4	Universidad Nacional de San Agustín de Arequipa	Arequipa	Arequipa	-16.397139	-71.537144
...	...	...	...	...	...
137	Universidad Global del Cusco	Cusco	Cusco	-13.524219	-71.943694
138	Universidad Santo Tomás de Aquino de Ciencia e...	Junín	Huancayo	-12.084040	-75.208442
139	Universidad Privada SISE	Lima	Lima	-12.017367	-77.004572
140	Universidad Seminario Evangélico de Lima (*12)	Lima	Lima	-12.063959	-76.960074
141	Universidad Seminario Bíblico Andino (*12)	Lima	Lima	-12.069621	-77.053398

126 rows × 5 columns

Figura 24.2: Procedimiento estándar utilizado para el filtrado de datos atípicos

- Separación de valores

  Esta labor era especialmente útil al momento que nos encontrábamos con tablas de datos para establecimientos escolares o de salud, en cuya locación había más de un nivel académico o tipo de atención. A lo largo del proyecto existieron necesidades distintas para los diferentes clientes a los cuales se les ofrecía la información, por lo que hubo un momento en que se necesitaron estos valores por separado, tomando los distintos niveles académicos dentro de una locación como si se tratara de establecimientos distintos.

  Se utilizaron expresiones regulares sencillas en una función de filtrado personalizada para uso en Pandas, con el objetivo de determinar si era que en el campo correspondiente al nivel académico o tipo de atención se encontraba un valor en específico, para cuyo caso se separaba el registro en dos o más registros distintos. Más adelante, en una sección posterior donde se hará mención acerca del uso que se hizo de Access también como herramienta de procesamiento, se expondrá un procedimiento inverso a este, en el cual la intención era juntar registros cuando los campos de estos se encontraban como entidades separadas.

```
df
```

	Name	State	Municipality	Codigo	Type1	Type2	Type3	Type4	Type5	Type6	...	Scale4	Scale5	Scale6	Latitude	Longitude	Namemod	Santo Domingo	Municipiomod	Latitude.1	Longitude.1
0	13769 - DE ATENCION NINOS FELICES	11 - PUERTO PLATA	1105 - ALTAMIRA	13769	School	PRIVADO	Inicial 2014	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	De Atencion Ninos Felices	Puerto Plata	Altamira	19.650998	-70.792987
1	10224 - MIS PRIMEROS PASOS	03 - AZUA	0301 - AZUA	10224	School	PRIVADO	Inicial 2014	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Mis Primeros Pasos	Azua	Azua	18.453259	-70.732401
2	11576 - MI HOGAR DIVINO	03 - AZUA	0301 - AZUA	11576	School	PRIVADO	Inicial 2014, Primario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Mi Hogar Divino	Azua	Azua	18.453259	-70.732401
3	14136 - COLEGIO WALDORF	03 - AZUA	0301 - AZUA	14136	School	PRIVADO	Inicial 2014, Primario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Colegio Waldorf	Azua	Azua	18.453259	-70.732401
4	16877 - COLEGIO CRISTIANO IAS	03 - AZUA	0301 - AZUA	16877	School	PRIVADO	Inicial 2014, Primario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Colegio Cristiano las	Azua	Azua	18.453259	-70.732401
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1901	09168 - GENERAL EUSEBIO MANZUETA	17 - MONTE PLATA	1701 - YAMASA	9168	School	PUBLICO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	General Eusebio Manzueta	Monte Plata	Yamasa	18.773498	-70.026632
1902	09175 - GENERAL EUSEBIO MANZUETA	17 - MONTE PLATA	1701 - YAMASA	9175	School	PUBLICO	No aplica	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	General Eusebio Manzueta	Monte Plata	Yamasa	18.773498	-70.026632
1903	12867 - COLEGIO RENUEVO	17 - MONTE PLATA	1701 - YAMASA	12867	School	PRIVADO	Inicial 2014, Primario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Colegio Renuevo	Monte Plata	Yamasa	18.773498	-70.026632
1904	14367 - JOSE DE LA LUZ GUILLEN	17 - MONTE PLATA	1701 - YAMASA	14367	School	PUBLICO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Jose De La Luz Guillen	Monte Plata	Yamasa	18.773498	-70.026632
1905	15099 - HOMERO TAVERAS MARTINEZ	17 - MONTE PLATA	1701 - YAMASA	15099	School	PUBLICO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Homero Taveras Martinez	Monte Plata	Yamasa	18.773498	-70.026632

Figura 25: DataFrame antes de una operación de separación de filas por nivel educativo

```
df_Inicial = df[df['Type3'].str.contains('inicial', case=False)].assign(Type3='Inicial')
df_Primario = df[df['Type3'].str.contains('primario', case=False)].assign(Type3='Primario')
df_Secundario = df[df['Type3'].str.contains('secundario', case=False)].assign(Type3='Secundario')
```

```
pd.concat([df_Inicial, df_Primario, df_Secundario], ignore_index = True)
```

	Name	State	Municipality	Codigo	Type1	Type2	Type3	Type4	Type5	Type6	...	Scale4	Scale5	Scale6	Latitude	Longitude	Namemod	Santo Domingo	Municipiomod	Latitude.1	Longitude.1
0	13769 - DE ATENCION NINOS FELICES	11 - PUERTO PLATA	1105 - ALTAMIRA	13769	School	PRIVADO	Inicial	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	De Atencion Ninos Felices	Puerto Plata	Altamira	19.650998	-70.792987
1	10224 - MIS PRIMEROS PASOS	03 - AZUA	0301 - AZUA	10224	School	PRIVADO	Inicial	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Mis Primeros Pasos	Azua	Azua	18.453259	-70.732401
2	11576 - MI HOGAR DIVINO	03 - AZUA	0301 - AZUA	11576	School	PRIVADO	Inicial	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Mi Hogar Divino	Azua	Azua	18.453259	-70.732401
3	14136 - COLEGIO WALDORF	03 - AZUA	0301 - AZUA	14136	School	PRIVADO	Inicial	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Colegio Waldorf	Azua	Azua	18.453259	-70.732401
4	16877 - COLEGIO CRISTIANO IAS	03 - AZUA	0301 - AZUA	16877	School	PRIVADO	Inicial	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Colegio Cristiano las	Azua	Azua	18.453259	-70.732401
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3539	15217 - CENTRO EDUCATIVO LUCERITOS DEL SEÑOR	10 - SANTO DOMINGO	1001 - VILLA MELLA	15217	School	PRIVADO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Centro Educativo Luceritos Del Señor	Santo Domingo	Villa Mella	18.531096	-69.906438
3540	07425 - LAS AGUITAS	13 - MONTE CRISTI	1303 - VILLA VASQUEZ	7425	School	PUBLICO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Las Aguitas	Monte Cristi	Villa Vasquez	19.808855	-71.442957
3541	09168 - GENERAL EUSEBIO MANZUETA	17 - MONTE PLATA	1701 - YAMASA	9168	School	PUBLICO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	General Eusebio Manzueta	Monte Plata	Yamasa	18.773498	-70.026632
3542	14367 - JOSE DE LA	17 - MONTE	1701 -	14367	School	PUBLICO	Secundario	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	Jose De La	Monte	Yamasa	18.773498	-70.026632

Figura 26: DataFrame después de una operación de separación de filas por nivel educativo

- Funciones personalizadas

  Una de las grandes ventajas que provee Pandas es la capacidad para crear funciones personalizadas para el procesamiento de los datos, pudiendo de esta manera aplicar colectivamente esta función a los datos con el fin de modificarlos, o bien, crear nuevos datos.

  Para ejemplificar esto último, se retomará brevemente el uso que se hizo de pandas en conjunto de Geopy. Primero se intentó usar la función que Pandas tiene integrada para su uso con Geopy. Sin embargo, al aplicar el algoritmo de búsqueda iterativa con distintos parámetros, se tuvo que crear una función adecuada que utilizaba la primera función mencionada, con la modificación adecuada para que esta pudiese ser repetida en múltiples ocasiones.

Figura 27: Función personalizada para reemplazar una cadena de texto

- Left, Right e Inner Join

  Al ser una herramienta enfocada en el análisis de datos, Pandas puede ejecutar operaciones Join o de unión, tal como se hace en distintos motores SQL. Aunque Pandas no tiene los niveles de optimización para la ejecución de esta tarea como los tienen los servidores SQL, también es una función muy útil al momento de tratar volúmenes de datos de tamaño mediano.

  Estas funciones únicamente fueron utilizadas con el fin de añadir datos a los establecimientos cuando se llegaban a encontrar data sets que contuvieran los mismos establecimientos con información complementaria acerca de ellos.

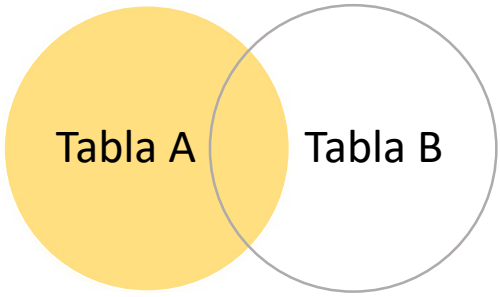


Figura 28: Diagrama de Venn ejemplificando una operación Left Join

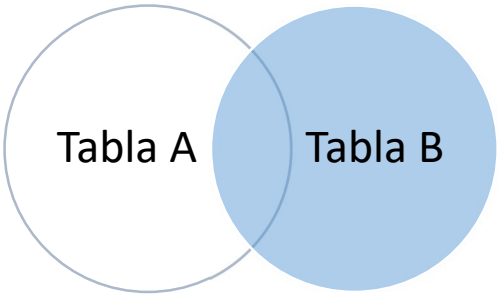


Figura 29: Diagrama de Venn ejemplificando una operación Right Join

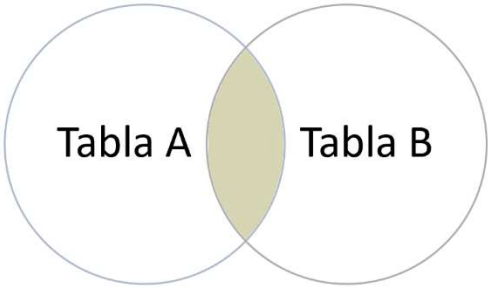
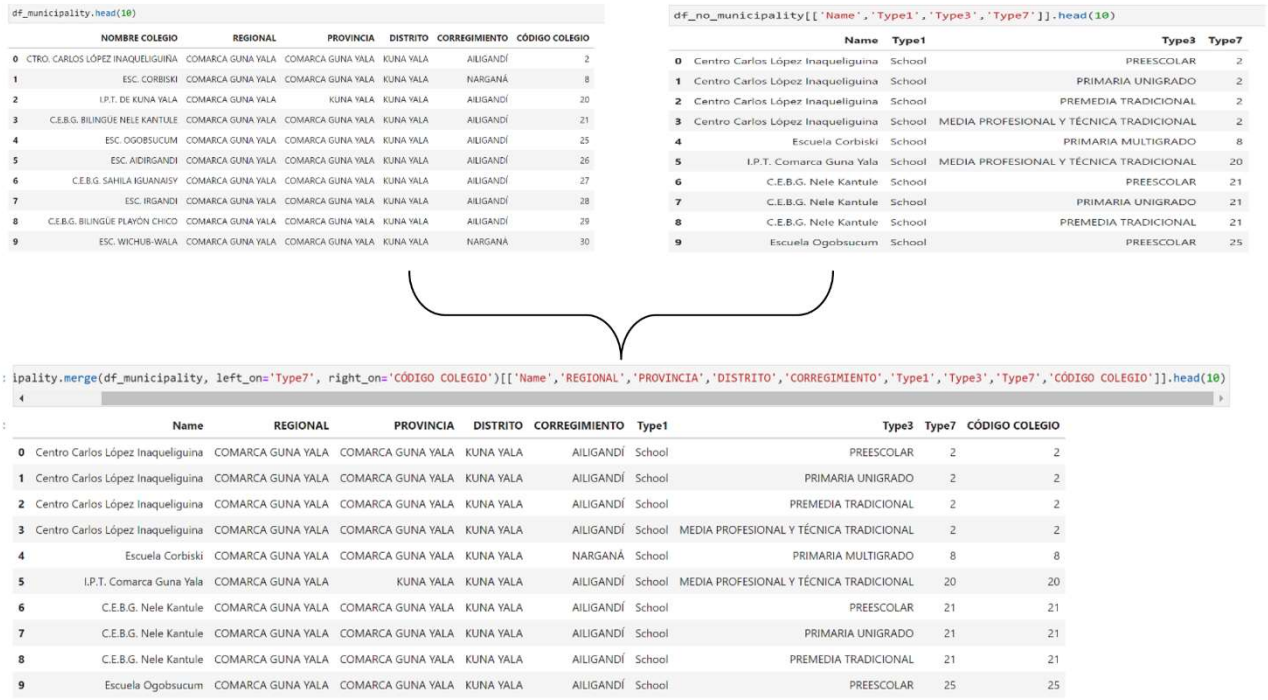


Figura 30: Diagrama de Venn ejemplificando una operación Inner Join



```
df_municipality.head(10)
```

	NOMBRE COLEGIO	REGIONAL	PROVINCIA	DISTRITO	CORREGIMIENTO	CÓDIGO COLEGIO
0	CTRO. CARLOS LÓPEZ INAQUELIGUIÑA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	2
1	ESC. CORBISKI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	8
2	I.P.T. DE KUNA YALA	COMARCA GUNA YALA	KUNA YALA	KUNA YALA	AILIGANDÍ	20
3	C.E.B.G. BILINGÜE NELE KANTULE	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	21
4	ESC. OGOBSUCUM	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	25
5	ESC. AIDIRGANDI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	26
6	C.E.B.G. SAHILA IGUANAISY	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	27
7	ESC. IRGANDI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	28
8	C.E.B.G. BILINGÜE PLAYÓN CHICO	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	29
9	ESC. WICHUB-WALA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	30

```
df_no_municipality[['Name','Type1','Type3','Type7']].head(10)
```

	Name	Type1	Type3	Type7
0	Centro Carlos López Inaqueliguina	School	PREESCOLAR	2
1	Centro Carlos López Inaqueliguina	School	PRIMARIA UNIGRADO	2
2	Centro Carlos López Inaqueliguina	School	PREMEDIA TRADICIONAL	2
3	Centro Carlos López Inaqueliguina	School	MEDIA PROFESIONAL Y TÉCNICA TRADICIONAL	2
4	Escuela Corbiski	School	PRIMARIA MULTIGRADO	8
5	I.P.T. Comarca Guna Yala	School	MEDIA PROFESIONAL Y TÉCNICA TRADICIONAL	20
6	C.E.B.G. Nele Kantule	School	PREESCOLAR	21
7	C.E.B.G. Nele Kantule	School	PRIMARIA UNIGRADO	21
8	C.E.B.G. Nele Kantule	School	PREMEDIA TRADICIONAL	21
9	Escuela Ogobsucum	School	PREESCOLAR	25

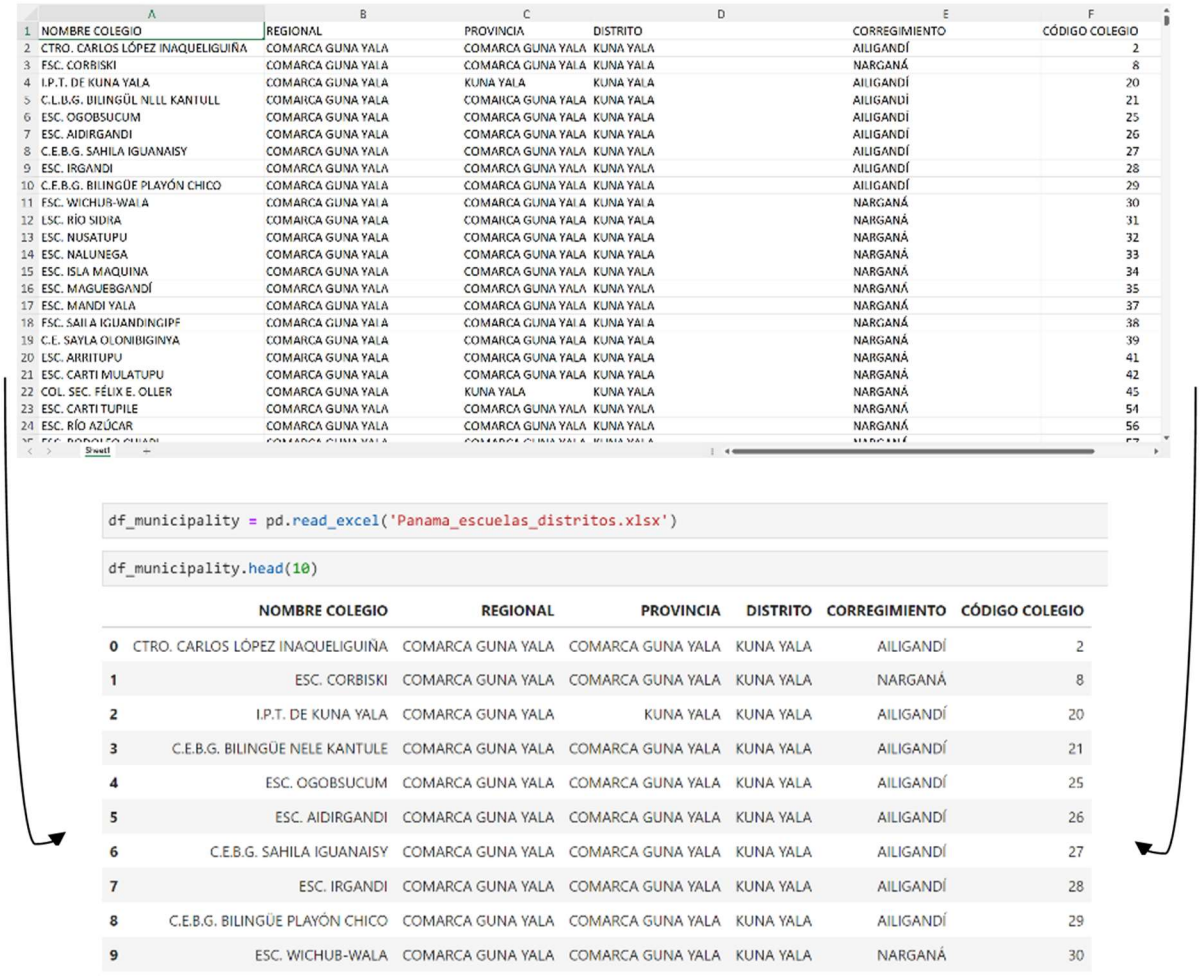
```
ipality.merge(df_municipality, left_on='Type7', right_on='CÓDIGO COLEGIO')[['Name','REGIONAL','PROVINCIA','DISTRITO','CORREGIMIENTO','Type1','Type3','Type7','CÓDIGO COLEGIO']].head(10)
```

	Name	REGIONAL	PROVINCIA	DISTRITO	CORREGIMIENTO	Type1	Type3	Type7	CÓDIGO COLEGIO
0	Centro Carlos López Inaqueliguina	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PREESCOLAR	2	2
1	Centro Carlos López Inaqueliguina	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PRIMARIA UNIGRADO	2	2
2	Centro Carlos López Inaqueliguina	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PREMEDIA TRADICIONAL	2	2
3	Centro Carlos López Inaqueliguina	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	MEDIA PROFESIONAL Y TÉCNICA TRADICIONAL	2	2
4	Escuela Corbiski	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	School	PRIMARIA MULTIGRADO	8	8
5	I.P.T. Comarca Guna Yala	COMARCA GUNA YALA	KUNA YALA	KUNA YALA	AILIGANDÍ	School	MEDIA PROFESIONAL Y TÉCNICA TRADICIONAL	20	20
6	C.E.B.G. Nele Kantule	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PREESCOLAR	21	21
7	C.E.B.G. Nele Kantule	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PRIMARIA UNIGRADO	21	21
8	C.E.B.G. Nele Kantule	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PREMEDIA TRADICIONAL	21	21
9	Escuela Ogobsucum	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	School	PREESCOLAR	25	25

Figura 31: Ejemplo de una operación Inner Join en Pandas

- Lectura, escritura de archivos y conexión SQL

Una de las mejores utilidades de Pandas es la facilidad para la obtención y escritura de DataFrames, ya que es posible importar y exportar estos, desde archivos con distintos formatos tales como xls, xlsx, json, yaml, csv, pkl entre otros. Esto facilita bastante las tareas, ya que únicamente es necesario saber la ubicación del archivo para importarlo y posteriormente empezar a trabajar con él. De la misma forma la escritura se puede hacer con un solo paso, exportando a cualquiera que sea el formato deseado. Esta utilidad fue empleada para prácticamente todos los conjuntos de datos que fueron procesados haciendo uso de Pandas.



	A	B	C	D	E	F
1	NOMBRE COLEGIO	REGIONAL	PROVINCIA	DISTRITO	CORREGIMIENTO	CÓDIGO COLEGIO
2	CTRO. CARLOS LÓPEZ INAQUELIGUIÑA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	2
3	ESC. CORBISKI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	8
4	I.P.T. DE KUNA YALA	COMARCA GUNA YALA	KUNA YALA	KUNA YALA	AILIGANDÍ	20
5	C.E.B.G. BILINGÜE NELE KANTULE	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	21
6	ESC. OGOBSUCUM	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	25
7	ESC. AIDIRGANDI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	26
8	C.E.B.G. SAHILA IGUANAISY	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	27
9	ESC. IRGANDI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	28
10	C.E.B.G. BILINGÜE PLAYÓN CHICO	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	29
11	ESC. WICHUB-WALA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	30
12	ESC. RÍO SIDRA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	31
13	ESC. NUSATUPU	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	32
14	ESC. NALUNEGA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	33
15	ESC. ISLA MAQUINA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	34
16	ESC. MAGUEBGANDÍ	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	35
17	ESC. MANDI YALA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	37
18	ESC. SAILA IGUANDINGIPE	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	38
19	C.E. SAYLA OLONIBIGINYA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	39
20	ESC. ARRITUPU	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	41
21	ESC. CARTI MULATUPU	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	42
22	COL. SEC. FÉLIX E. OLLER	COMARCA GUNA YALA	KUNA YALA	KUNA YALA	NARGANÁ	45
23	ESC. CARTI TUPILE	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	54
24	ESC. RÍO AZÚCAR	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	56

```
df_municipality = pd.read_excel('Panama_escuelas_distritos.xlsx')
```

```
df_municipality.head(10)
```

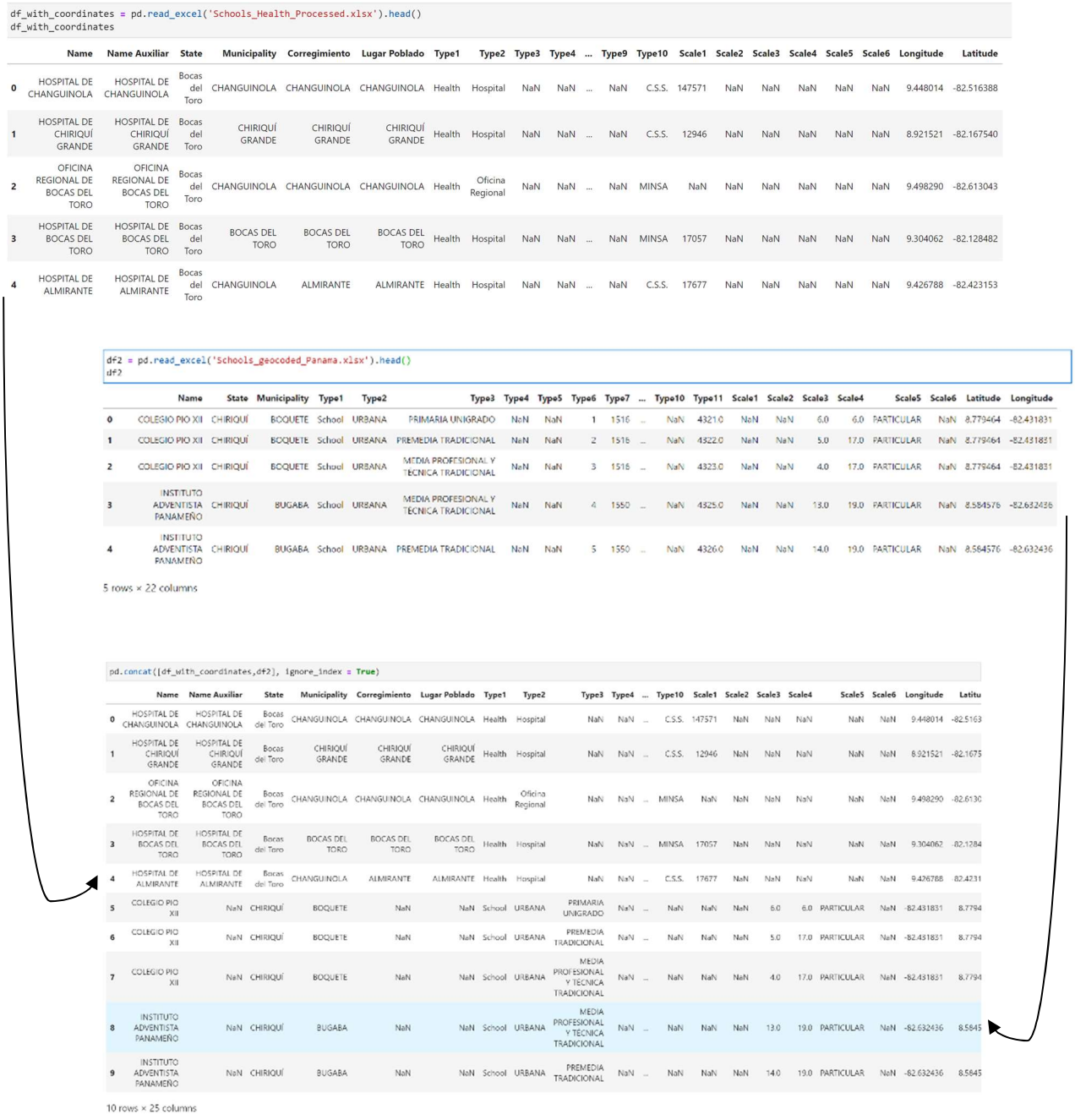
	NOMBRE COLEGIO	REGIONAL	PROVINCIA	DISTRITO	CORREGIMIENTO	CÓDIGO COLEGIO
0	CTRO. CARLOS LÓPEZ INAQUELIGUIÑA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	2
1	ESC. CORBISKI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	8
2	I.P.T. DE KUNA YALA	COMARCA GUNA YALA	KUNA YALA	KUNA YALA	AILIGANDÍ	20
3	C.E.B.G. BILINGÜE NELE KANTULE	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	21
4	ESC. OGOBSUCUM	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	25
5	ESC. AIDIRGANDI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	26
6	C.E.B.G. SAHILA IGUANAISY	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	27
7	ESC. IRGANDI	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	28
8	C.E.B.G. BILINGÜE PLAYÓN CHICO	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	AILIGANDÍ	29
9	ESC. WICHUB-WALA	COMARCA GUNA YALA	COMARCA GUNA YALA	KUNA YALA	NARGANÁ	30

Figura 32: Lectura de Excel a Pandas

Otra característica digna de mencionar es la capacidad de Pandas para la conexión con servidores de Bases de Datos haciendo uso de otra librería llamada SQLalchemy. Haciendo uso de esta es posible cargar, descargar datos e incluso realizar operaciones y consultas a la base deseada.

- Concatenación de datos

  En ocasiones en que los datos relacionados a establecimientos de salud y centros educativos eran obtenidos y procesados en forma separada, estos podían ser fácilmente unidos mediante la función append o concat de Pandas.



```
df_with_coordinates = pd.read_excel('Schools_Health_Processed.xlsx').head()
df_with_coordinates
```

	Name	Name Auxiliar	State	Municipality	Corregimiento	Lugar Poblado	Type1	Type2	Type3	Type4	...	Type9	Type10	Scale1	Scale2	Scale3	Scale4	Scale5	Scale6	Longitude	Latitude
0	HOSPITAL DE CHANGUINOLA	HOSPITAL DE CHANGUINOLA	Bocas del Toro	CHANGUINOLA	CHANGUINOLA	CHANGUINOLA	Health	Hospital	NaN	NaN	...	NaN	C.S.S.	147571	NaN	NaN	NaN	NaN	NaN	9.448014	-82.516388
1	HOSPITAL DE CHIRIQUÍ GRANDE	HOSPITAL DE CHIRIQUÍ GRANDE	Bocas del Toro	CHIRIQUÍ GRANDE	CHIRIQUÍ GRANDE	CHIRIQUÍ GRANDE	Health	Hospital	NaN	NaN	...	NaN	C.S.S.	12946	NaN	NaN	NaN	NaN	NaN	8.921521	-82.167540
2	OFICINA REGIONAL DE BOCAS DEL TORO	OFICINA REGIONAL DE BOCAS DEL TORO	Bocas del Toro	CHANGUINOLA	CHANGUINOLA	CHANGUINOLA	Health	Oficina Regional	NaN	NaN	...	NaN	MINSA	NaN	NaN	NaN	NaN	NaN	NaN	9.498290	-82.613043
3	HOSPITAL DE BOCAS DEL TORO	HOSPITAL DE BOCAS DEL TORO	Bocas del Toro	BOCAS DEL TORO	BOCAS DEL TORO	BOCAS DEL TORO	Health	Hospital	NaN	NaN	...	NaN	MINSA	17057	NaN	NaN	NaN	NaN	NaN	9.304062	-82.128482
4	HOSPITAL DE ALMIRANTE	HOSPITAL DE ALMIRANTE	Bocas del Toro	CHANGUINOLA	ALMIRANTE	ALMIRANTE	Health	Hospital	NaN	NaN	...	NaN	C.S.S.	17677	NaN	NaN	NaN	NaN	NaN	9.426788	-82.423153

```
df2 = pd.read_excel('Schools_geocoded_Panama.xlsx').head()
df2
```

	Name	State	Municipality	Type1	Type2	Type3	Type4	Type5	Type6	Type7	...	Type10	Type11	Scale1	Scale2	Scale3	Scale4	Scale5	Scale6	Latitude	Longitude
0	COLEGIO PIO XII	CHIRIQUÍ	BOQUETE	School	URBANA	PRIMARIA UNIGRADO	NaN	NaN	1	1516	...	NaN	4321.0	NaN	NaN	6.0	6.0	PARTICULAR	NaN	8.779464	-82.431831
1	COLEGIO PIO XII	CHIRIQUÍ	BOQUETE	School	URBANA	PREMEDIA TRADICIONAL	NaN	NaN	2	1516	...	NaN	4322.0	NaN	NaN	5.0	17.0	PARTICULAR	NaN	8.779464	-82.431831
2	COLEGIO PIO XII	CHIRIQUÍ	BOQUETE	School	URBANA	MEDIA PROFESIONAL Y TECNICA TRADICIONAL	NaN	NaN	3	1516	...	NaN	4323.0	NaN	NaN	4.0	17.0	PARTICULAR	NaN	8.779464	-82.431831
3	INSTITUTO ADVENTISTA PANAMEÑO	CHIRIQUÍ	BUGABA	School	URBANA	MEDIA PROFESIONAL Y TECNICA TRADICIONAL	NaN	NaN	4	1550	...	NaN	4325.0	NaN	NaN	13.0	19.0	PARTICULAR	NaN	8.584576	-82.632436
4	INSTITUTO ADVENTISTA PANAMEÑO	CHIRIQUÍ	BUGABA	School	URBANA	PREMEDIA TRADICIONAL	NaN	NaN	5	1550	...	NaN	4326.0	NaN	NaN	14.0	19.0	PARTICULAR	NaN	8.584576	-82.632436

5 rows × 22 columns

```
pd.concat([df_with_coordinates,df2], ignore_index = True)
```

	Name	Name Auxiliar	State	Municipality	Corregimiento	Lugar Poblado	Type1	Type2	Type3	Type4	...	Type10	Scale1	Scale2	Scale3	Scale4	Scale5	Scale6	Longitude	Latitu
0	HOSPITAL DE CHANGUINOLA	HOSPITAL DE CHANGUINOLA	Bocas del Toro	CHANGUINOLA	CHANGUINOLA	CHANGUINOLA	Health	Hospital	NaN	NaN	...	C.S.S.	147571	NaN	NaN	NaN	NaN	NaN	9.448014	-82.5163
1	HOSPITAL DE CHIRIQUÍ GRANDE	HOSPITAL DE CHIRIQUÍ GRANDE	Bocas del Toro	CHIRIQUÍ GRANDE	CHIRIQUÍ GRANDE	CHIRIQUÍ GRANDE	Health	Hospital	NaN	NaN	...	C.S.S.	12946	NaN	NaN	NaN	NaN	NaN	8.921521	-82.1675
2	OFICINA REGIONAL DE BOCAS DEL TORO	OFICINA REGIONAL DE BOCAS DEL TORO	Bocas del Toro	CHANGUINOLA	CHANGUINOLA	CHANGUINOLA	Health	Oficina Regional	NaN	NaN	...	MINSA	NaN	NaN	NaN	NaN	NaN	NaN	9.498290	-82.6130
3	HOSPITAL DE BOCAS DEL TORO	HOSPITAL DE BOCAS DEL TORO	Bocas del Toro	BOCAS DEL TORO	BOCAS DEL TORO	BOCAS DEL TORO	Health	Hospital	NaN	NaN	...	MINSA	17057	NaN	NaN	NaN	NaN	NaN	9.304062	-82.1284
4	HOSPITAL DE ALMIRANTE	HOSPITAL DE ALMIRANTE	Bocas del Toro	CHANGUINOLA	ALMIRANTE	ALMIRANTE	Health	Hospital	NaN	NaN	...	C.S.S.	17677	NaN	NaN	NaN	NaN	NaN	9.426788	-82.4231
5	COLEGIO PIO XII	NaN	CHIRIQUÍ	BOQUETE	NaN	NaN	School	URBANA	PRIMARIA UNIGRADO	NaN	...	NaN	NaN	NaN	NaN	6.0	6.0	PARTICULAR	NaN	-82.431831	8.7794
6	COLEGIO PIO XII	NaN	CHIRIQUÍ	BOQUETE	NaN	NaN	School	URBANA	PREMEDIA TRADICIONAL	NaN	...	NaN	NaN	NaN	NaN	5.0	17.0	PARTICULAR	NaN	-82.431831	8.7794
7	COLEGIO PIO XII	NaN	CHIRIQUÍ	BOQUETE	NaN	NaN	School	URBANA	MEDIA PROFESIONAL Y TECNICA TRADICIONAL	NaN	...	NaN	NaN	NaN	NaN	4.0	17.0	PARTICULAR	NaN	-82.431831	8.7794
8	INSTITUTO ADVENTISTA PANAMEÑO	NaN	CHIRIQUÍ	BUGABA	NaN	NaN	School	URBANA	MEDIA PROFESIONAL Y TECNICA TRADICIONAL	NaN	...	NaN	NaN	NaN	NaN	13.0	19.0	PARTICULAR	NaN	-82.632436	8.5845
9	INSTITUTO ADVENTISTA PANAMEÑO	NaN	CHIRIQUÍ	BUGABA	NaN	NaN	School	URBANA	PREMEDIA TRADICIONAL	NaN	...	NaN	NaN	NaN	NaN	14.0	19.0	PARTICULAR	NaN	-82.632436	8.5845

10 rows × 25 columns

Figura 33: Concatenación de DataFrames

Posterior a la limpieza y preparación de los datos haciendo uso de las herramientas previamente mencionadas, se procedía a subir estos a una plataforma en línea propia de la empresa, de la cual no se mencionará el nombre por políticas de privacidad. A continuación, describiré de forma breve el uso de esta y algunos de sus alcances.

Huawei's ETL

Desde esta, se podían ejecutar operaciones de ETL (Extracción, Transformación y Carga) haciendo uso de herramientas sencillas que iban desde Impala SQL (una variante del leguaje SQL) hasta algunas utilidades más complejas que permitían transformar y complementar los datos enriqueciendo su significado.

A continuación, explicaré a grandes rasgos el funcionamiento de esta plataforma, omitiendo detalles que pudiesen infringir políticas de la empresa respecto a la divulgación de información sensible.

Al subir los datos a esta plataforma, la primera operación a ejecutarse era hacer una conversión de tipos de datos por aquellos que fueran adecuados al sistema. Para esto se tomaba una pequeña parte de los datos contenidos en el dataset y se previsualizaban con la conversión correspondiente a cada uno de ellos. Al final de esta operación, únicamente era necesario solicitar la aprobación para la conservación de estos datos, aceptando un acuerdo en el que se estipulaba que los datos agregados provenían de fuentes públicas.

Posteriormente, era necesario procesarlos haciendo uso de un algoritmo que trazaba el territorio del país en cuestión en una cuadrícula de dimensiones programables. En este proceso, se optó por usar un tamaño de cuadrícula de 200m x 200m para tener resultados confiables en el cálculo que se explicará más adelante.

Una vez teniendo el territorio dividido en sectores, se procedía a calcular el centroide de cada uno de estos, asignando a cada uno de ellos una coordenada. De la misma forma, haciendo uso de información propia del sistema, se determinaba el Estado/Departamento, Municipalidad e incluso localidad a la que pertenecía cada uno de estos sectores, específicamente, el centroide de este.

Teniendo estos nuevos datos, y en conjunto con los anteriormente recabados, se hacía una operación para ver qué y cuantos establecimientos se localizaban dentro de los sectores generados. Operación que sería de utilidad en pasos posteriores.

A continuación, se procedía a importar datos estadísticos adquiridos por la empresa, los cuales contenían la información objetivo para hacer los cálculos que serían de mayor importancia en el

proyecto. Es relevante mencionar que estas tablas generalmente contenían varios millones de registros, por lo que regularmente el trabajo con estas era complicado, sobre todo en cuestiones de limpieza y estandarización.

Una vez importadas las tablas estadísticas, procedía a realizar una limpieza de estas, ya que en los nombres de los organismos con los cuales se hacían cálculos, en muchas ocasiones se encontraban diferencias para cada uno de los registros. En estos procesos, se hizo uso de Impala SQL, expresiones regulares y Python.

Ya que el objetivo final de la tabla era conservar únicamente alrededor de 15 organismos distintos, se tomaban aquellos que siguieran un cierto patrón en su nombre para asignarles un nombre estandarizado. Sin embargo, en múltiples ocasiones se tenían cientos o incluso miles de diferencias en los nombres, por lo que únicamente un patrón de expresiones regulares no era suficiente. De esta forma, se optó por realizar un script de Python que escribiera en un archivo de texto el query o consulta de SQL para todos los casos distintos, recibiendo como parámetro dos listas: una con los distintos nombres que se tenían que cambiar y otra con los nombres por los que se tenía que reemplazar.

Inicialmente el script de SQL fabricado, hacía uso de estructuras de control condicionales IF, junto con técnicas de programación recursiva para de esta forma cubrir todos los casos posibles, anidando las decisiones tomadas. Sin embargo, posteriormente migramos esta técnica para que hiciera uso de la estructura CASE WHEN (en caso de), cuya sintaxis resulta mucho más sencilla y entendible para el usuario. Este cambio entre métodos y la razón principal por la que inicialmente se hizo uso de la estructura IF fue debido al desconocimiento de algunas de las utilidades que nos proveía Impala SQL.

Una vez estandarizados los nombres de las tablas, se procedía a hacer una combinación de estas con los datos que habían sido anteriormente cargados dentro de la plataforma, es decir, aquellos que contenían la información geográfica de escuelas y hospitales. Esta combinación la realizábamos mediante un LEFT JOIN en SQL, usando como columna en común el centroide generado una vez que se realizó sectorización del país en cuestión.

Esta combinación generaba una nueva tabla con un mayor número de registros que la tabla original de escuelas y hospitales, pero a su vez, un menor número que la tabla que contenía datos

estadísticos. De esta forma, podíamos tener para cada uno de los registros de la tabla, un dato relacionado respecto a una característica asociada con su valor máximo, mínimo y promedio.

De la misma forma, una vez generada esta primera tabla, se procedía a combinar una segunda tabla con datos similares a los anteriores, pero de un tipo distinto. Generando así una tabla que contenía cada uno de los establecimientos de interés, con repetición en sus datos por tipo de servicio, por cada uno de los distintos proveedores de este, y, además, un máximo, mínimo y promedio por cada uno.

En este punto es importante mencionar que generalmente y debido a la naturaleza de las tablas que contenían los datos estadísticos, las tablas que resultaban de ambas operaciones de LEFT JOIN, solían tener entre 50,000 registros, hasta unos pocos millones, dependiendo del número de establecimientos localizados para el país en cuestión, la cantidad de proveedores de servicio y el número de sectores creados al principio del proceso.

Para finalizar este paso, procedía a exportar y realizar una breve descripción acerca de las tablas creadas por si es que estas combinaciones podían llegar a ser de utilidad a otros empleados. Además de eso último, la razón principal por la que la exportación de las tablas creadas era necesaria era debido a que, en el siguiente paso, durante el uso de la aplicación de visualización, únicamente podía hacer uso de las tablas que se encontraban exportadas.

Huawei's Visualization Tool

Figura 34: Analogía del procesamiento de datos: Visualización (Imagen generada con Inteligencia Artificial)

El último paso de trabajo con los datos recabados era el correspondiente a su visualización, la cual era llevada a cabo mediante una aplicación web muy similar a Tableau y Power BI.

Mediante el uso de esta herramienta, se podía hacer despliegue de gráficas de distintos tipos, tales como barras, líneas, circulares, de caja, entre muchos otros. Además de esto, se contaban con herramientas de despliegue geográfico, con las cuales era posible mostrar mapas para los distintos países investigados, con distintas características en ellos como, por ejemplo, su densidad poblacional.

Dentro del aspecto de generación de información con el uso de esta herramienta, también nos era posible formular datos nuevos haciendo uso de los existentes. Esto era posible mediante funciones de agregación, lógicas o matemáticas que nos permitían agregar información valiosa para el despliegue o bien, modificar la existente a modo de que esta pudiese ser más entendible.

Dentro de un panel específico, a nuestro equipo nos era requerido diseñar una interfaz entendible para el usuario, con la cual pudiese interactuar cambiando parámetros y que, a su vez, estos cambios generasen resultados en la visualización que estábamos mostrando.

Siguiendo esta idea, se formaron distintos gráficos con estadísticas de interés y números aproximados correspondientes tanto a la tabla de datos estadísticos, como a los datos correspondientes a los establecimientos, entre los cuales se podían incluir desde número de profesores, alumnos o personal, el número de aulas o camas de hospital, si un establecimiento se catalogaba como rural o urbano e incluso el porcentaje de establecimientos públicos y privados. Para el caso particular en donde se trabajaba con locaciones bancarias, únicamente se desplegaba su ubicación y una medida estadística para cada una de ellas.

Dependiendo de las características de cada gráfico, los datos desplegados podían ser filtrados para mostrar resultados de un estado o municipio en específico, o bien acotar los resultados por la magnitud de algún parámetro seleccionado.

Otro tipo de despliegue geográfico que se utilizó, aunque en pocas ocasiones, fue del trazado de rutas, esto para visualizar estrategias de conexión entre locaciones, o bien para observar claramente una ruta existente de interés y formular la estrategia de aproximación más conveniente a algún punto de esta.

Por último, en esta parte de visualización de datos también era agregada una tabla dinámica que, mediante la variación de los parámetros, era capaz de generar un monto aproximado del costo que tendría la implementación de tecnología Huawei en zonas de relevancia. Dando así al cliente, una respuesta inmediata acerca de la posible solución para la problemática planteada.

StoryTelling

Figura 35: Analogía del procesamiento de datos: Storytelling (Imagen generada con Inteligencia Artificial)

Para finalizar, la última tarea que llevaba a cabo era el estructurar un diálogo, generalmente conocido como "historia" para explicar el significado y profundizar en el valor de los datos que estábamos mostrando.

Esta parte era de suma importancia al momento de presentar nuestra aplicación ante un cliente, por lo que debíamos tener una preparación previa antes de estructurar nuestra historia. Esto lo lográbamos haciendo una breve investigación acerca del cliente al que nos indicaban que se tenía que presentar el producto, tratando de considerar cuáles podrían ser los usos que le diera a la analítica que estábamos brindando de acuerdo con el giro de trabajo, los datos mostrados, un monto de inversión estimado, etc.

En las distintas presentaciones del producto que realicé, me dirigí tanto a empleados de gobierno, como de empresas privadas de distintos países latinoamericanos tanto en inglés como en español. Durante estas, siempre me concentré en proveer un enfoque distinto al uso de la herramienta para los distintos intereses del negocio. Mientras que al presentar ante organismos gubernamentales me enfocaba en el cómo el uso de analítica de datos estadísticos de conexión, en conjunto con datos públicos podrían llegar a ser de gran utilidad en el trabajo por reducir la brecha tecnológica en zonas de alta marginación, en las empresas privadas, procuraba enfocarme en cómo esta misma información era capaz de facilitar la localización de oportunidades de negocio a nivel nacional, tanto para los sectores de educación como salud.

Influencia de mi carrera universitaria durante la labor profesional

Durante mi desempeño profesional en Huawei Technologies de México, tuve diversas oportunidades para hacer uso de mis conocimientos y habilidades desarrolladas a lo largo de mis estudios universitarios como Ingeniero Mecatrónico, no únicamente en un ámbito correspondiente al área lógica-matemática, sino también aquellos desarrollados en las materias del área de Ciencias Sociales y Humanidades.

Al haber trabajado durante poco más de un año en el proyecto anteriormente presentado, puedo concluir que los conocimientos que me fueron de mayor utilidad para la creación de soluciones en este ámbito son aquellos obtenidos en las materias donde se hacía uso de la programación y ciencias básicas. Bajo esta premisa, quiero aclarar que, en mayor o menor medida, todas las materias que cursé durante mi carrera universitaria forjaron mi estructura del pensamiento, logrando de esta forma tener ideas creativas e innovadoras basadas en el conocimiento adquirido y, en consecuencia, tomar decisiones basadas en evidencia.

Algunas de las materias, cuyo contenido me fue de mayor utilidad durante este proyecto fueron:

- Técnicas de programación: El uso de los conocimientos adquiridos durante este curso es bastante obvio, ya que la mayor parte del trabajo desempeñado se hizo con el uso de algoritmos que me ayudasen al momento de procesar, filtrar o transformar los datos crudos obtenidos, en información tangible.
- Inteligencia Artificial: Una de las más populares vertientes de la inteligencia artificial se encuentra ampliamente basada en los datos, el aprendizaje automático. Fue en esta materia donde tuve un primer acercamiento a algunos conceptos como Big Data, o ETL.
- Desarrollo de aplicaciones Móviles: Durante esta materia, tuve por primera vez interacción con una base de datos y con el lenguaje SQL. Además, durante este curso tuve una primera aproximación al diseño de interfaces de usuario, teniendo de esta forma, cierta familiaridad con los requisitos que estas deban tener para agrado del cliente objetivo.
- Diseño Mecatrónico: Aunque en primera instancia pueda parecer osado el afirmar que los conocimientos adquiridos en este curso son de utilidad en el análisis de datos. He de decir que, los ejercicios realizados para conocer las necesidades de un usuario, el proceso de

diseño de una solución y el cómo adaptar los recursos disponibles para la misma, fueron imprescindibles durante una gran variedad tareas.

- Estadística: Esta materia también me fue primordial para poder realizar conclusiones sobre los datos que eran obtenidos y posteriormente presentados ante clientes. Cosas tan sencillas como la obtención de la media, las diversas formas de visualizar datos y cuáles son los gráficos correctos para representar cierta información, son de alta importancia al momento de mostrar una idea.
- Redacción y exposición de temas de Ingeniería: Esta materia en particular, fue un tanto tormentosa para mí cuando la cursé en mi primer semestre de la universidad. Sin embargo, reconozco ampliamente que los conocimientos impartidos en ella me ayudaron a estructurar mi diálogo de una forma que fuese entendible y convincente para las personas a las que se les presentaba el producto.
- Creatividad e innovación: Esta materia, en ocasiones infravalorada en ingeniería, es de una gran utilidad para resolver problemas cuando se necesita de un enfoque distinto al tradicional. En distintos momentos, principalmente cuando trataba de crear un panel que pudiese ser de utilidad para el usuario, usaba técnicas basadas en lo aprendido a lo largo de este curso, con las cuales podía identificar más fácilmente las necesidades del cliente y así identificar las potenciales soluciones.
- Cálculo, Algebra y Geometría Analítica: Este bloque de materias lo utilicé durante una tarea en específico, y es que hubo un momento en el que se me encargó realizar un programa que pudiese realizar peticiones a la API de elevación de Google Maps de forma dinámica.

  Tomando dos puntos A y B, la API de forma nativa puede dividir en N partes el segmento y realizar el número correspondiente de peticiones para saber la altitud en cada uno de ellos. Sin embargo, mi tarea era realizar una petición cada 100 metros de distancia, por lo que me fue necesario calcular el número de segmentos necesario para alcanzar este objetivo. Además de esto, tuve que realizar un cálculo para tomar en cuenta la curvatura de la tierra al momento de realizar la petición, ya que lo requerido era determinar si entre el punto A y B había “Línea de vista” o bien, si es que se podía trazar una línea recta entre ambos sin encontrar obstáculo alguno.

Conclusiones

El análisis e ingeniería de datos son disciplinas que día con día se hacen más necesarias en el mundo laboral moderno. Las exorbitantes cantidades de datos generados anualmente, gracias al creciente número de dispositivos conectados a internet, generan posibilidades nunca vistas en empresas de cualquier sector para tener una introspectiva más cercana al comportamiento real del producto o servicio que ofrecen al público. Sin embargo, su procesamiento y transformación en información relevante es un reto complicado, debido al enorme poder de cómputo que esto implica y la poca homogeneidad con la que la mayoría de estos son generados y/u obtenidos.

En particular para este proyecto, se hizo uso de las muchas bondades que ofrecen las plataformas de Huawei, para procesar y transformar datos de forma masiva, para posteriormente crear una herramienta de Inteligencia Empresarial que facilitara la visualización y entendimiento de las estadísticas creadas. La información obtenida a través de esta analítica de datos tuvo la finalidad de generar una introspectiva certera y confiable acerca de la infraestructura en telecomunicaciones para diversos países en Latinoamérica, ayudando al cliente a plantear proyectos que generarían impacto en una comunidad objetivo.

Si bien, cuando recién ingresé a este proyecto, pensé que mis habilidades como ingeniero mecatrónico no serían del todo útiles en el desempeño de la labor que se me había encomendado, con el tiempo puede darme cuenta de que, trasladar el conocimiento de un área a otra fue más fácil de lo que en un principio imaginé. Haciendo uso de la lógica generada, principalmente en las materias de Ciencias Básicas, Programación y también las habilidades para generar ideas y sintetizar información obtenidas en las materias diseño, pude desarrollar un excelente papel como ingeniero de datos.

Actualmente, habiendo fijado ya mi carrera hacia el mundo de los datos, y tratando de adentrarme en sus distintas disciplinas (Analítica, Ingeniería y Ciencia), aún me encuentro desarrollando habilidades y aprendiendo distintas tecnologías que, en conjunto con mi formación universitaria, me harán un digno profesional egresado de la máxima casa de estudios de nuestra nación.

Referencias

- Microsoft Copilot para generación de Imágenes. Junio de 2024, de https://designer.microsoft.com/image-creator?scenario=texttoimage
- ¿Quiénes somos? (s.f.). *Huawei Technologies Co., Ltd*. Recuperado el 24 de octubre de 2023, de https://www.huawei.com/mx/corporate-information
- Huawei Blog (2024), *Broadening the Innovation Landscape: Huawei Patents & Top Ten Inventions 2021.* Obtenido de https://blog.huawei.com/2022/06/16/broadening-innovation-landscape-huawei-patents-top-ten-inventions-2021/
- World Bank. (2016). World development report 2016: Digital dividends. Washington, DC: World Bank. https://www.worldbank.org/en/publication/wdr2016
- Secretaría de Comunicaciones y Transportes. (s.f.). Programa "Internet para todos". Gobierno de México. Recuperado de https://www.gob.mx/sct/acciones-y-programas/programa-internet-para-todos
- IDC Corporate, 2022: El gasto en TI en América Latina superará el crecimiento del PIB en 2023. Recuperado de: https://www.idc.com/getdoc.jsp?containerId=prLA50040523
- Secretaría de Comunicaciones y Transportes. (s.f.). Programa " Despliegue de infraestructura pasiva de telecomunicaciones ". Gobierno de México. Recuperado de https://www.gob.mx/cms/uploads/attachment/file/566631/Infraestructura_de_Telecom_portal.pdf
- Huawei. (2024). Huawei ayuda a la digitalización y el desarrollo sostenible en América Latina y el Caribe. Recuperado de

https://www.huawei.com/mx/news/mx/2024/huawei-ayuda-a-la-digitalizacion-y-el-desarrollo-sostenible-en-america-latina-y-el-caribe

- Zhao Bo (2017). Web Scraping, *Encyclopedia of Big Data,* DOI 10.1007/978-3-319-32001-4_483-1, https://www.researchgate.net/profile/Bo-Zhao-3/publication/317177787_Web_Scraping/links/5c293f85a6fdccfc7073192f/Web-Scraping.pdf
- Ministerio de Educación de Ecuador. (2023). Sitio de datos abiertos del Ministerio de Educación de Ecuador. Recuperado de https://educacion.gob.ec/base-de-datos
- Muthukadan, Baiju (2006-2018) *Selenium with Python* Recuperado el 30 de octubre de 2023, de https://selenium-python.readthedocs.io/
- Geopy Contributors (2006-2018) *Geopy 2.4.0* Recuperado el 30 de octubre de 2023, de https://pypi.org/project/geopy/
- Jupyter Team (2015) *Project Jupyter Documentation* Recuperado el 30 de octubre de 2023, de https://docs.jupyter.org/en/latest/
- NumFOCUS, Inc. (2023) *pandas documentation* Recuperado el 30 de octubre de 2023, de https://pandas.pydata.org/docs/
- Google, Inc. (2023) *Elevation API documentation* Recuperado el 30 de octubre de 2023, de https://developers.google.com/maps/documentation/elevation/start