**PROYECTO GRUPAL FINAL - DATA SCIENCE**

* **Propuesta de negocio:**

La empresa de servicios de transporte de pasajeros “*Greyhound*” nos contrató como equipo de trabajo de Data Science con el objetivo de analizar su inversión en transportes particulares en Nueva York y revisar su impacto en el medio ambiente.

Para ello, como equipo de trabajo definimos utilizar como referencia del análisis los archivos de trabajo, datasets, de los taxis de la ciudad de Nueva York, junto con los de la calidad del aire y la contaminación acústica de esa misma ciudad.

* **Objetivos:** Las acciones que buscamos lograr en esta propuesta de negocio son las siguientes:
  + Asesorar al cliente **“***Greyhound”* brindándole perspectivas ambientales relevantes para su toma de decisiones.
  + Entregar conclusiones que faciliten el análisis de la rentabilidad del negocio.
  + Elaborar productos como un dashboard y un modelo de Machine Learning con los cuales el cliente podrá identificar la situación actual de la problemática y tener perspectiva a futuro de la misma y tomar decisiones en consecuencia.
* **KPI´s:** Buscamos identificar patrones de viajes, demandas, zonas y mediciones representativas en las métricas establecidas para verificar el cumplimiento de los objetivos definidos y su correspondiente fundamento.

1) **Promedio de Contaminación de Aire y Acústica por viaje.**

* **Específico:** Medir los niveles de contaminación de aire y ruido en cada viaje realizado durante el período 2021-2022.
* **Medible:** Utilizar métricas de calidad del aire y ruido para calcular el promedio en cada viaje.
* **Alcanzable:** La empresa puede recopilar datos de calidad del aire y ruido y analizarlos.
* **Relevante:** Evaluar si la entrada de Greyhound aumenta los niveles de contaminación y ruido, lo que afecta la calidad de vida de los residentes.
* **Tiempo Definido:** Evaluar continuamente a lo largo del período 2021-2022.

Fundamentación: La empresa está interesada en evaluar el impacto de la entrada de Greyhound al mercado en la calidad del aire y el nivel de ruido en la ciudad. El objetivo es asegurarse de que esta entrada no aumente los niveles promedio de contaminación del aire y ruido en más de un 30%.

2) **Ratio de Viajes ~~Personas~~ que Toman Taxi por Viaje: Específico:** Calcular la relación entre el número de personas que toman taxis y el número de viajes en las áreas donde Greyhound opera durante 2021-2022.

* **Medible:** Registrar el número de pasajeros y el número de viajes en esas áreas.
* **Alcanzable:** La empresa puede recopilar datos sobre pasajeros y viajes.
* **Relevante:** Identificar si la presencia de Greyhound influye en el uso de taxis en áreas específicas.
* **Tiempo Definido:** Evaluar el ratio durante el período 2021-2022.

Fundamentación: La empresa busca entender cómo la presencia de Greyhound en ciertas áreas afecta la demanda de taxis. El objetivo es evaluar si la presencia de Greyhound tiene un impacto significativo en el número de personas que optan por tomar un taxi en lugar de utilizar otros medios de transporte.

3) **Precio Promedio de Viajes en Distritos.**

* **Específico:** Medir el precio promedio de los viajes en diferentes distritos durante el último año (2022).
* **Medible:** Calcular el precio promedio en cada distrito.
* **Alcanzable:** La empresa puede recopilar y analizar los datos de tarifas.
* **Relevante:** Identificar las zonas con tarifas promedio más altas, lo que puede influir en las estrategias de fijación de precios.
* **Tiempo Definido:** Evaluar el precio promedio durante el año 2022.

Fundamentación: El objetivo es identificar cuáles son los distritos o zonas de la ciudad en los que los precios promedio de los viajes son más altos. Esto puede ayudar a la empresa a ajustar su estrategia de precios y comprender las áreas donde los viajes son más costosos.

4) **Cantidad de Viajes en Horas Pico**

* **Específico:** Contar la cantidad de viajes realizados en horas pico durante el último año (2022).
* **Medible:** Registrar la cantidad de viajes en las horas específicas definidas como horas pico.
* **Alcanzable:** La empresa puede recopilar datos de horarios de viajes.
* **Relevante:** Evaluar si la empresa es capaz de gestionar la demanda en las horas pico.
* **Tiempo Definido:** Evaluar la cantidad de viajes en horas pico durante el año 2022

Fundamentación: La empresa necesita asegurarse de que puede manejar eficazmente la demanda durante las horas pico. El objetivo es definir un mínimo de viajes que la empresa debe ser capaz de gestionar durante estas horas para garantizar un servicio de calidad.

* **Alcance:** El trabajo lo haremos definiendo primero los archivos a analizar, y los años que contemplaremos en los mismos para poder ser representativos en el informe. Para ello establecimos:

Usar como análisis:

1) Dataset “Taxis de la ciudad de Nueva york”.

2) Dataset “Calidad del aire en la ciudad de Nueva York”.

3) Dataset “Contaminación acústica en la ciudad de Nueva york”.

Ambos serán analizados desde una:

* **Perspectiva Ambiental:** En el dataset “Calidad de aire” elegimos tomar dos años desde el 2021 al 2022 inclusive. Y del dataset “Contaminación acústica” el año 2019.
* **Perspectiva de Negocio:** En el dataset de Taxis en la Ciudad de N.Y. elegimos tomar los años desde el 2020 al 2022 inclusive.

Posteriormente, realizamos un cronograma con las tareas que definimos llevar a cabo para ejecutar la propuesta y les asignamos roles de todos los colaboradores del equipo en el software Jira.

Gráfico, Gráfico de barras

Descripción generada automáticamente

De nuestra planificación surgieron las siguientes tareas:

* MARCO DE REFERENCIA DEL PROYECTO: Desarrollo de la estructuración y planeación inicial del proyecto. Definición de objetivos y alcance del proyecto.
* OBJETIVOS ESPECIFICOS DEL PROYECTO: Definir los objetivos que queremos trabajar.
* ETL: Realizar la transformación y limpieza de los datos de los diferentes archivos definidos para que sean de buena calidad y representativos. Para esta actividad hemos definido utilizar los roles de Data Engeneering así garantizábamos la correcta realización. Se Definió también utilizar diccionarios para los datasets para disponer de una visibilización de los datos. También se ha hecho un análisis preliminar de las bases realizadas. Estos estarán alojados en el repositorio compartido.
* DATASETS: Definición de los datasets que se emplearán durante el proyecto, diccionarios de datos, descripción y documentación de las fuentes.
* EDA: Análisis exploratorio de los datos. Inicialmente haremos el recorrido de los datos alojados en los 3 datasets. El EDA proporciona una visión integral del impacto de la inversión de la empresa de transporte de pasajeros en la calidad del aire y la contaminación sonora en Nueva York. Los KPI´s propuestos permiten evaluar el alcance y los efectos de la inversión, lo que puede ayudar a la empresa a tomar decisiones informadas para minimizar su impacto ambiental y mejorar la calidad de vida en la ciudad. El análisis geoespacial es esencial para identificar áreas críticas y dirigir las acciones de mitigación de manera eficiente.
* DATA ENGENEERING: Generar los pasos finales del ETL, documentos, diccionarios, Pipeline, diseño de modelo Entidad relacional, Pipelines para alimentar el DW y validación de los datos. Dataware house, workflow. Análisis de datos de la muestra realizada. MVP.
* DASHBOARD Y DATA ANALYTICS: Una vez definido el ETL y los datasets disponibles, se procederá a realizar el dashboard interactivo de la propuesta con los análisis de los objetivos propuestos. Este incluirá los Kpi´s definidos y su resultado.
* KPI´S: Plantear qué Kpi´s encontramos relevantes para asegurar los objetivos de la propuesta.
* MODELO DE MACHINE LEARNING: una vez definido el análisis se define el modelo a implementar y se pondrá en funcionamiento.
* **Stack Tecnológico:** El stack tecnológico que decidimos utilizar es mayormente gratuito, de fácil acceso y el conjunto de personas quienes formamos parte de esta consultora tenemos larga experiencia con ellos.

A continuación, los detallamos con su respectiva fundamentación:

* + Python**:** Es uno de los lenguajes de programación más populares al día de hoy para el análisis de datos debido a su simplicidad y legibilidad, además de contar con una amplia gama de bibliotecas especializadas.
  + Pandas: Proporciona estructuras de datos y funciones esenciales para realizar análisis de datos rápidos y fáciles en Python. Es especialmente útil para la manipulación y análisis de datos estructurados.
  + Numpy: Biblioteca fundamental para cálculos científicos en Python. Proporciona objetos de matriz multidimensional y herramientas para trabajar con estas matrices.
  + Matplotlib: Una de las bibliotecas más populares para visualización en Python. Permite crear una gran variedad de gráficos estáticos, animados e interactivos.
  + Seaborn: Basado en Matplotlib, Seaborn ayuda a crear visualizaciones estadísticas atractivas y ofrece una interfaz de alto nivel para dibujar gráficos estadísticos informativos.
  + Scikit-learn: Biblioteca para aprendizaje automático (Machine Learning) en Python. Proporciona herramientas simples y eficientes para análisis predictivo de datos.
  + Requests: Esencial para trabajar con solicitudes HTTP. Facilita el proceso de hacer peticiones a servidores web y manejar respuestas.
  + Jupyter Notebook: Un entorno de desarrollo interactivo que permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto. Es ampliamente utilizado en la comunidad de ciencia de datos para investigación y enseñanza. Permite realizar el código en partes, lo cual es altamente funcional para poder detectar errores.
  + Visual Studio Code: Editor de código versátil con soporte para múltiples lenguajes y herramientas de desarrollo integradas.
  + MySQL: Uno de los sistemas de gestión de bases de datos relacionales más populares. Utilizado para almacenar, recuperar y gestionar datos en aplicaciones. Es gratuito a diferencia de otros softwares similares.
  + PowerBI: Herramienta de visualización de datos y elaboración de informes que permite convertir datos sin procesar en información significativa a través de paneles interactivos. Es gratuito a diferencia de otros softwares similares para usuarios de Microsoft.
  + GitHub: Plataforma de control de versiones y colaboración que permite trabajar conjuntamente con otros desarrolladores en proyectos.
  + Jira: Plataforma de gestión de proyectos que facilita la planificación, seguimiento, gestión de errores y trabajo en equipo.
  + OneDrive: Servicio de almacenamiento en la nube que garantiza que los archivos estén respaldados y puedan ser compartidos y accedidos desde cualquier lugar.
  + Discord: Plataforma de comunicación que facilita la colaboración y el trabajo en tiempo real.
  + WhatsApp: Herramienta de comunicación rápida y directa en asuntos urgentes.
  + PySpark: Interfaz para Apache Spark y Python. Permite el procesamiento de datos a gran escala en clústeres.
  + Google Colab: Entorno de cuaderno Jupyter que no requiere configuración y se ejecuta completamente en la nube. Tiene la particularidad de brindarle al usuario herramientas computaciones como GPUS, RAM.
  + Paquete Office: Conjunto de herramientas de productividad que incluye programas como Word, Excel y PowerPoint, esenciales para la creación y presentación de documentos.
  + Gmail: Servicio de correo electrónico que facilita la comunicación y el intercambio de información.  
    Diagrama

    Descripción generada automáticamente
* **Repositorio:** La información está disponible en GitHub. Allí trabajamos conjuntamente. El link es: <https://github.com/MrNobody-505/ProyectoGrupalHENRY.git>

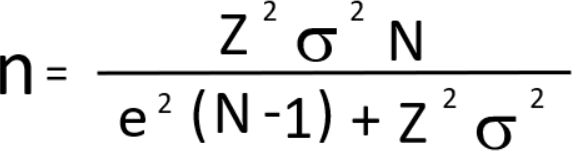
Sprint 2:

**Muestra de los Datos:**

* + DATASETS DE TAXIS

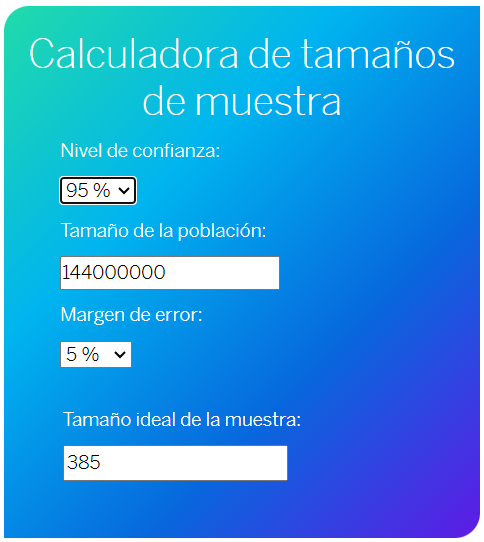
En lugar de trabajar con los 144 millones de registros completos, considera realizar un muestreo aleatorio de los datos. Puedes seleccionar una muestra representativa de los datos que sea significativa para tu análisis. Esto te permitirá trabajar con un conjunto de datos más manejable sin perder de vista la representatividad.

# Definir el tamaño de la muestra:



Donde:

N = 144 millones Z = 1.96 (nivel de confianza 95%) e= 0.05(error=5%) σ= 0.5 lo más seguro es optar por una desviación estándar de 0,5, que garantizará que el tamaño de la muestra sea lo suficientemente grande.



TOTAL de registros mínimos= 385 REGISTROS UTILIZADOS = 2.443

Marzo: En marzo, generalmente se espera una rutina normal de actividades escolares y laborales en Nueva York, sin eventos de vacaciones importantes ni días festivos que alteren significativamente la demanda de servicios de taxis. Esto lo convierte en un mes relativamente estable en términos de demanda de taxis.

Agosto: Durante agosto, es común que se produzcan vacaciones de verano, tanto para las escuelas como para muchas empresas. Como resultado, la ciudad puede experimentar una disminución en la demanda de taxis, especialmente en áreas donde las personas se van de vacaciones. Sin embargo, la ciudad aún alberga a turistas durante el verano, lo que puede equilibrar esta disminución en la demanda.

Diciembre: Diciembre es un mes festivo en el que se celebran varias festividades, incluyendo la Navidad y el Año Nuevo. Durante este período, es común que las personas viajen, hagan compras navideñas y asistan a eventos y celebraciones. La demanda de servicios de taxis suele ser más alta en áreas comerciales y de entretenimiento, especialmente cerca de eventos navideños y lugares de compras.

Teniendo en cuenta estos patrones, tus selecciones de semanas en marzo, agosto y diciembre parecen apropiadas para capturar una variedad de condiciones estacionales en la ciudad de Nueva York. Como mencioné anteriormente, documenta claramente tus criterios de selección y asegúrate de que la muestra sea lo suficientemente grande para tus análisis. Si necesitas realizar ajustes a lo largo de tu proyecto, puedes hacerlo de manera informada.

**DataWarehouse**

Para crear un Data Warehouse en Google Cloud, utilizamos BigQuery, que es un servicio de almacenamiento de datos y análisis en la nube altamente escalable y completamente administrado.

**Cloud**

Para este trabajo decidimos utilizar un servicio en la nube como GCP principalmente por su gratuidad y sencilles y además por ventajas tales como:

Elasticidad y Escalabilidad: Los recursos en la nube se pueden escalar hacia arriba o hacia abajo fácilmente para adaptarse a las necesidades cambiantes, lo que permite a las empresas manejar aumentos de carga sin la necesidad de invertir en hardware físico.

Costo-efectividad: Con los servicios en la nube, pagas solo por los recursos que usas, lo que puede reducir significativamente los costos operativos y de capital, al no tener que invertir en infraestructura física y su mantenimiento.

Agilidad y Rapidez de Implementación: La nube permite desplegar rápidamente aplicaciones y servicios, lo que facilita la experimentación y la innovación al reducir el tiempo de lanzamiento de nuevos productos y servicios.

Acceso Global: Los servicios en la nube están accesibles desde cualquier lugar con conexión a Internet, lo que permite trabajar y colaborar de forma remota y global sin restricciones de ubicación.

Seguridad y Conformidad: Proveedores como GCP ofrecen sólidas capacidades de seguridad y cumplen con muchos estándares de regulación, reduciendo la carga de las empresas para mantener la seguridad y la conformidad.

Recuperación ante Desastres y Continuidad del Negocio: Los servicios en la nube suelen incluir opciones para la copia de seguridad y la recuperación de datos, lo que ayuda a garantizar la continuidad del negocio en caso de fallos o desastres.

Servicios y Herramientas Avanzadas: La nube proporciona acceso a una amplia gama de servicios y herramientas avanzadas, incluyendo potentes capacidades de IA, análisis de datos y machine learning, que de otra manera podrían ser demasiado costosas o complejas para implementar de forma local.

Actualizaciones Automáticas: Los servicios en la nube se actualizan automáticamente con las últimas características y parches de seguridad, lo que reduce la carga del mantenimiento de sistemas.

Ecosistema de Servicios Integrados: Plataformas como GCP ofrecen un conjunto de servicios integrados que trabajan juntos, como computación, almacenamiento, bases de datos, y servicios de machine learning, facilitando la creación de soluciones completas en la nube.

Sostenibilidad: Al optimizar el uso de recursos y operar en centros de datos eficientes, los servicios en la nube pueden contribuir a reducir la huella de carbono de las operaciones de TI.

**Automatización**

A) La automatización completa se realizó sobre un solo archivo “borough.csv”:  
1. ETL y posterior guardado del archivo “borough.csv” en Cloud Storage:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

2. pasar el archivo “borough.csv” del Cloud Storage a Big Query(DataWarehouse):

Texto

Descripción generada automáticamente

Estas dos funciones se ejecutan automáticamente cada un año.

B) La automatización parcial del resto de archivos.

1. ETL realizado de manera local.

Lo más relevante de esta sección es que el ETL de Taxis\_NY fue realizado con pyspark que es una herramienta de big data.

Todos los archivos se suben al Cloud Storage.

2. Pasar todos los archivos del Cloud Storage a Big Query(DataWarehouse) de manera automática, replicando la función “A) 2.” (la función de la imagen de arriba)

**Modelo ER**

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Validación de datos:**

Verificar la integridad y la calidad de los datos antes de cargarlos en el Data Warehouse. Esto implica identificar y corregir posibles problemas, como errores de formato, valores faltantes o incoherencias en los datos. Es crucial para garantizar que los datos en el DW sean precisos y confiables, ya que los análisis y las decisiones empresariales se basan en estos datos. Algunos aspectos clave de la validación de datos incluyen:

* Comprobamos la integridad referencial: Aseguramos que las relaciones entre las tablas se mantengan y que no haya datos huérfanos.
* Validar valores y tipos de datos: Verificar que los valores estén dentro de los rangos esperados y que los tipos de datos sean coherentes.
* Eliminamos duplicados: Identificamos y eliminamos registros duplicados.
* Garantizar la coherencia: Aseguramos que los datos sean consistentes en todo el DW.

La validación de datos esta realizada en el ETL de cada archivo.

**Pipeline: en esta gráfica podemos ver el flujo de los datos y las tareas que se aplican desde la transformación hasta su ingesta en el Data Warehouse(BigQuery).**

**Tenemos dos formas de tratar los datos:**

**Una de ellas es el proceso de ETL de manera local en los archivos: 1)Noise.csv, 2)Parameter\_index.csv,3)Taxis\_NY.csv, 4)Queens\_air\_quality.csv, 5) year.csv. Estos mismos se guardan en Cloud Storage una vez finalizada la transformación y limpieza.**

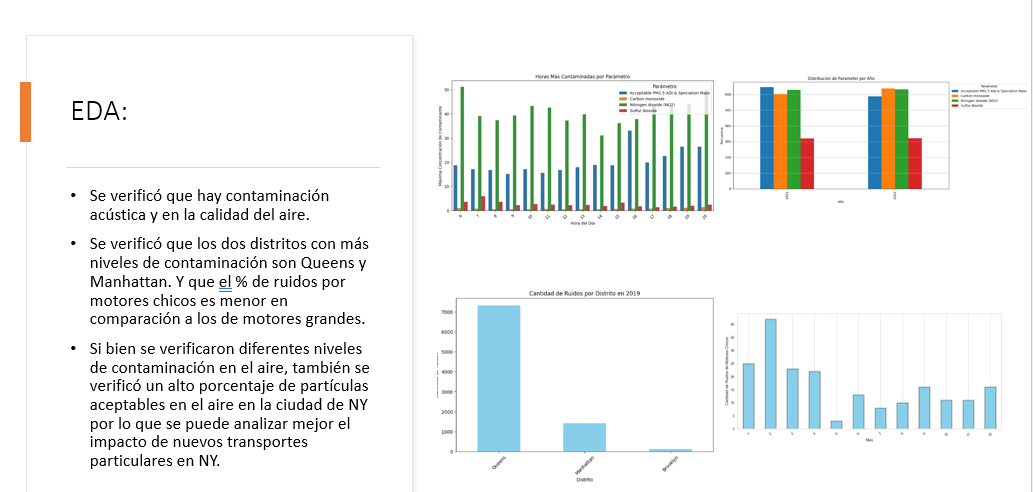
**Por su parte el archivo borough.csv es el resultado de un proceso de ETL automatizado realizado en Google Cloud y activado mediante una tarea planificada en Google Scheduler. Este archivo se guarda con los anteriores en Cloud Storage.**

**Por ultimo la totalidad de los archivos se ingestan automáticamente en BigQuery reflejándose como una colección de tablas contenidas en una base de datos llamadas taxis.**

Diagrama

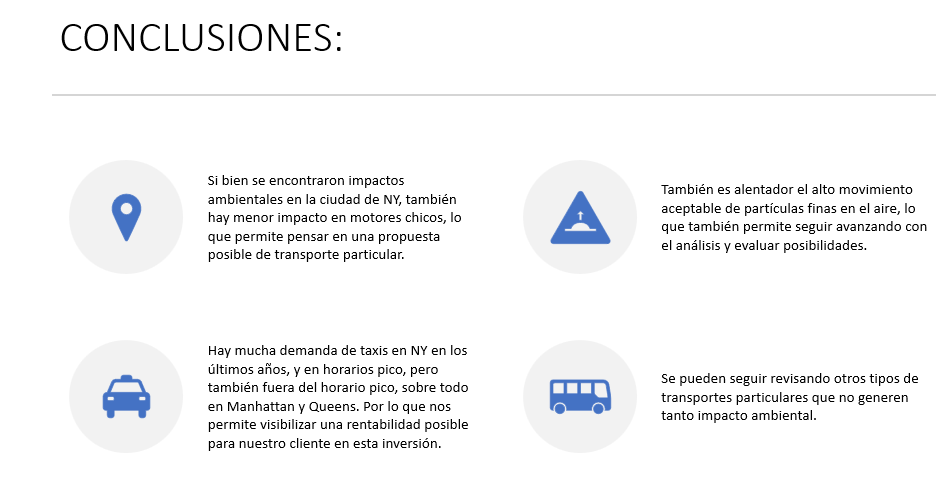
Descripción generada automáticamente

**EDA:**

****

**Imagen que contiene Tabla

Descripción generada automáticamente**

****