Data Journal Case Study 1: Cyclistic “Bike Share”

El objetivo principal de este caso de estudio es dar respuesta a la pregunta planteada por la directora de marketing Lily Moreno: ***Diseñar una estrategia de marketing con el objetivo de convertir a los usuarios casuales del servicio en usuarias con una membresía anual.***

- **Ask**

Durante está primera etapa se busca encontrar el objetivo y el porqué de realizar este análisis. Esto para orientar de mejor manera los resultados.

Las preguntas que se buscan responder con este análisis son:

- Cómo se diferencian los usuarios esporádicos de los usuarios anuales
- Porque los usuarios esporádicos deberían tener membresia
- Cómo se pueden usar los medios digitales para influenciar a los usuarios esporádicos para comprar membresias.

Ahora con el fin de acotar este análisis, la directora de marketing solo dejo la tarea de responder a la primera pregunta: *Cómo se diferencian los usuarios esporádicos de los usuarios anuales.*

Está pregunta y el objetivo planteado permiten decir que el tipo de problema es uno orientado a la realización de predicciones y de realizar conexiones. Desde cierto punto de vista puede verse la búsqueda de patrones entre los usuarios, también con fines de predicción.

Summary Table:

Step	**ASK**
Guiding Questions	**Cual es el problema que se trata de resolver** Se busca encontrar las diferencias entre los consumidores esporádicos y aquellos consumidores con una membresía anual **Cómo pueden sus conocimientos impulsar decisiones comerciales** Pueden impulsarlo ya que permite orientar una campaña de marketing enfocada solo en los consumidores esporádicos para que compren una membresía anual.
Key tasks	Identificar la **business task** Considerar a los **key stakeholders** ● En este caso el principal stakeholder es la directora de marketing Lily Moreno
Deliverables	**A clear statement of the business task** Realizar un análisis cuantitativo sobre la base de datos disponibles para detectar las diferencias de consumo entre los usuarios esporádicos del servicio versus los usuarios con una membresía anual con el fin de generar una campaña de marketing para convencer a los usuarios esporádicos de adquirir una membresía anual

- **Prepare**

En está etapa se procesa la información procedente de la base de datos.
Los datos obtenidos provienen de una base de datos de acceso público y anonimizada. Debido a está última característica se es posible conectar las compras de pases con los registros de tarjetas de crédito para determinar si los compradores ocasionales viven o no en las cercanías de las áreas de servicio o si han realizado múltiples compras de diferentes pases.

La información se descargó y se encuentra almacenada para su procesamiento en un directorio en Google Drive y, además, se almacenó una copia del archivo original en un servidor local.
La veracidad y confiabilidad de la información se sustenta en que la data fue obtenida mediante una base de datos anonimizada y de acceso público verificada. Está cumple con los requisitos **ROCCC (Reliable, Original, Comprehensive, Current, Cited).**

Al revisar la base de datos se detectaron celdas vacías e información repetida en ambas tablas de datos. Para ordenar y filtrar se decidió utilizar Google BigQuery para realizar estas acciones.

- En la tabla de datos “Divvy_Trips_2019_Q1”, existían celdas vacías en la columnas de género (19.711) y año de nacimiento (1). La información contenida en esas filas fue eliminada para que no incidieran en el análisis.
- En la tabla de datos “Divvy_Trips_2020_Q1”, solo existían datos faltantes en la información de llegada de un usuarios. Para mantener la consistencia, está también fue eliminada.
- En el caso de ambas tablas no existía información repetida.
- Se detecto que una key es *usertype* (2019) y *member_casual* (2020)

Summary Table:

Step	**PREPARE**
Guiding Questions	**Donde está localizada la data** La data estaba localizada en una base de datos de público acceso y anonimizada. En ella se podía encontrar registro de diversos años y periodos. Se descargaron los de los años 2019 y 2020. Importados en formatos CSV. **Cómo está organizada la data** Se encuentran organizadas en tablas CSV en formato de columnas. Se puede considerar una base de datos interna, discreta y estructurada. Se pueden apreciar que existen ***primary y foreign key*** lo que permitirá conectar la información de ambas tablas. **Existe algún asunto con respecto a la credibilidad o sesgo** No **Cómo se trabaja con la licencia, privacidad, seguridad y accesibilidad a la data** La base de datos posee una licencia de acceso libre, con la excepción del uso de información sensible sin el permiso de los usuarios (anonimizada). **Cómo se verificó la veracidad de la data** Se verificó la fuente de la información y la veracidad del sitio web de donde se extrajo la información. Se determinó que es un sitio válido y de acceso público.Además, en el sitio web se informa como se obtuvo la información y el nivel de acceso que se puede tener a ella.

	Como ayuda a responder la pregunta planteada Es de gran ayuda ya que permite generar patrones de consumo entre los diferentes usuarios y ver las tendencias de ellos. **Existe algún problema en la data** Si, existen vacíos de información que deben ser filtrados para un análisis más expedito.
Key tasks	Descargar la data y almacenarla correctamente. Identificar cómo se encuentra organizada. Ordenar y filtrar la data. ● Se encontraron vacíos de información y se eliminaron. ● No se encontró información repetida. Determinar el nivel de credibilidad de ella.
Deliverables	**A description of all data sources used** Las fuentes de información que se utilizaron fue la base de datos de carácter público y anonimizados

- **Process**

En está etapa del análisis el enfoque se centrará en el de procesar la data con el fin de filtrar y ordenar más en detalle en comparación con el filtrado en la etapa de preparación.

Para la realización de está etapa se continúa utilizando la herramienta de Google BigQuery, esto se justifica debido a la enorme cantidad de información contenida en la base de datos de ambos años, y utilizar SQL facilita el análisis de los datos.

Se realizó una revisión más completa de la información en la búsqueda de datos repetidos o información que no estuviera en línea con lo esperado o que estuviera fuera de lugar bajo el contexto que se está trabajando.

Las acciones realizadas para la limpieza de la base de datos fueron:

- ❖ Se determinó la duración de cada viaje realizado independiente del tipo de membresía.
- ❖ Utilizando está información se detectaron viajes que tenían duraciones inferiores a 60 segundos llegando a durar menos de 10 segundos.
- ❖ También se detectaron viajes de corta duración y donde las estaciones de inicio y fin eran las mismas.
- ❖ Con el fin de optimizar el análisis y que este fuera considerado significativo se eliminaron aquellos viajes de corta duración y aquellos con estaciones repetidas:
  - ➢ Divvy_Trips_2019_Q1:
    - ■ Estaciones repetidas: 24.190
    - ■ Viajes de corta duración: 0
  - ➢ Divvy_Trips_2020_Q1:
    - ■ Estaciones repetidas: 20.295
    - ■ Viajes de corta duración: 120

Cabe mencionar que como etapa de la limpieza de la base de datos se agregaron 2 columnas de información, una correspondiente a la duración de cada viaje y el día de la semana donde inició.

STEP	PROCESS
Guiding Questions	**¿Qué herramienta se usó y por qué?** Se utilizó Google BigQuery como herramienta para la limpieza y filtrado de la base de datos. Esto se justificó debido a la gran cantidad de información contenida en ellas que de haber sido procesada en Google SpreadSheet o Microsoft Excel hubiera tardado más tiempo y a costa de una mayor cantidad de recursos digitales. **Se aseguró la integridad de la data?** Si, esto mediante el filtrado y limpieza de la data **¿Qué pasos se tomaron para asegurar la limpieza de la data?** Una re-revisión de los datos repetidos y los datos faltantes, los cuales no serán considerados para el análisis. Además, no será considerada aquella información donde las estaciones de inicio y fin sean las mismas y la duración del viaje sea menor a un minuto. **¿Cómo se verificó que la data está limpia y lista para su uso?** Lo primero es verificar si existen términos *null,* y ordenar la información para determinar si las etapas de filtrados fueron útiles. También se utiliza la data original con el fin de comparar el número de filas de información original y las filas de información luego del filtrado, viendo una reducción significativa en el número de ellas. Este filtrado se realizó debido a que esa información no tiene sentido lógico cuando se le compara en el contexto general del análisis que se quiere realizar. **¿Se aseguro de documentar su proceso de limpieza para que sea revisado y compartido?** Si, todos los procesamientos y cambios realizados se documentaron y se encuentran comentados en el código de cada una de las tablas de información.
Key tasks	Revisar la data en busca de errores Elegir la herramienta correcta para el procesamiento Transformar la data para que el trabajo sea eficiente ● Aquí en lugar de transformar en si la data se utilizaron funciones específicas para cada tipo de dato específico. En particular para los datos en formato de *TIMESTAMP* Documentar el proceso de limpieza
Deliverables	**Documentation of any cleaning or manipulation of data** Todo lo realizado fue descrito en términos generales al inicio de este encabezado y las funciones utilizadas así como los pasos realizados, se encuentran comentados en el código correspondiente a cada tabla de datos.

- **Analyze**

Es en está etapa donde se realizan los análisis propiamente tal. Aquí se realizan cálculos basados en la información ordenada y recolectada y los primeros cálculos realizados en las etapas anteriores (duración del viaje y los días de la semana con mayor coincidencia).

Se continúa el trabajo de la información en Google BigQuery.

El primer paso es la realización de algunos cálculos básicos, como lo son el conocer el número total de viajes realizados luego del proceso de filtrado y ordenado, esto a través del número de filas. También se utilizan valores distintos para conocer los valores máximos, mínimos y la duración promedio de los viajes.

Las acciones realizadas para el cálculo de estos parámetros básicos fueron:

- La determinación de la duración de cada uno de los viajes, utilizando las funciones TIMESTAMP_DIFF, que entrega la diferencia entre el tiempo de inicio y el tiempo final en el formato TIMESTAMP, y la función MOD, que entrega el residuo de la división entre el resultado de la diferencia y un valor entero).
- Luego mediante las mismas funciones se determina la duración de cada viaje pero en segundo, esto con el propósito de estandarizar y poder tener una forma sencilla de comparación.
- Posteriormente se genera una nueva tabla donde se muestran estos resultados extraídos mediante las funciones MIN, MAX y AVG.
- Los resultados fueron organizados de dos formas diferentes. La primera fue según los días de la semana, se utilizó este parámetro para ver si existe alguna tendencia de los días de la semana donde se utiliza con mayor frecuencia el servicio de bicicletas. La otra forma fue basado únicamente en la cantidad de viajes respecto de las diferentes rutas, está última basada en las estaciones donde se iniciaban los viajes

Para una correcta lectura de estos resultados, cree una tabla con toda la información original más los nuevos cálculos realizados añadiendoles como nuevas columnas a los valores anteriores.

A continuación, para obtener resultados orientados a la business question, se crean dos tablas separadas con los datos de los años 2019 y 2020 pero solo extrayendo los datos que se consideran más relevantes para el análisis.

Las acciones realizadas para la construcción de está tabla fueron las siguientes:

- Con el fin de obtener una selección de los datos ordenados se utilizó la función CREATE TABLE IF NOT EXIST para crear tablas donde se mostrarán los datos característicos de ambos años en estudios.
- Los datos característicos para la toma de decisiones orientada al business task que se consideraron son:
  - Las estaciones de inicio y fin y el conteo asociado a cada una de ellas.
  - La duración de los viajes. Viajes mínimos, máximos y promedio para cada ruta.
  - El día de la semana con mayor cantidad de rutas realizadas.

- Estas tablas fueron separadas considerando los dos tipos de usuarios que utilizan el servicio de bicicletas, miembros y casuales. Ellas fueron realizadas durante dos años en estudio.

STEP	ANALYZE
Guiding Questions	**¿Cómo deberías organizar tus datos para realizar análisis sobre ellos?** Todos los datos utilizados se organizaron siempre manteniendo la business task en mente, es decir, sacar insight para transformar a los conductores esporádicos en miembros. En este caso particular, el conocer las duraciones de los viajes, los días y las estaciones más frecuentes y las rutas con mayor tránsito nos permiten extraer estos insights. **¿Tus datos están correctamente formateados?** En general si tenían el formato correcto, pero como se utilizó el tiempo fue necesario cambiar el formato de este para que fuera más trabajable y que al momento de analizar y presentar los datos estos tuvieran sentido y fueran más sencillos de ver, el formato utilizó fue el de mantener la cantidad de segundos de duración de cada viaje **¿Qué sorpresas descubriste en los datos?** Fue una sorpresa encontrar que el número de conductores miembros es bastante mayor al número de miembros casuales, independiente del año en estudio. Eso sí la tendencia a las estaciones con mayor flujo es casi siempre la misma independiente de está condición, favoreciendo el objetivo del negocio y mejorando la eficiencia de recursos, porque no se tendrá que enfocar en zonas donde la presencia de estos dos grupos sea tan dispar. Eso sí, dado que no se tiene la información si cada ruta fue realizada por el mismo usuario no es posible decir una diferencia exacta entre el número de miembros y no miembros. **¿Qué tendencias o relaciones encontró en los datos?** Una de las relaciones esperadas es que los conductores casuales tienden a aumentar los fines de semana mientras que los conductores miembros suelen tener los mayores números los días de semana. Otra tendencia es que las rutas más comunes entre miembros se encuentran en zonas más comerciales de la ciudad, se puede suponer edificios de oficinas y puntos de conexión con el transporte público, mientras que las rutas de los usuarios no miembros tienden a estar concentradas en zonas de recreación como parques y costaneras, eso sí también existe frecuenci con puntos de conexión con el transporte público. **¿Cómo le ayudarán estos conocimientos a responder las preguntas de su empresa?** Está información permitirá enfocar los recursos, para convencer a los no miembros a pagar membresias y además fortalecer a los que ya son miembros a continuar con sus planes, también permite tener un enfoque en que lugares geográficos de la ciudad colocar los recursos. Por último, permite tener un panorama general del uso de los servicios por parte de los usuarios.
Key tasks	Agregue sus datos para que sean útiles y accesibles Organiza y formatea la data Realizar cálculos o procedimientos Identificar tendencias y relaciones

Deliverables	Resumen del análisis En términos generales, se obtuvieron varios resultados a los cálculos realizados, se generó información útil y eficaz para la toma de decisiones y la información obtenida se considera valiosa para la toma de ellas.

- **Share**

En está etapa se generan las visualizaciones necesarias para la correcta y eficiente muestra de resultados del análisis realizado previamente.

Con el fin de mejorar la forma en que se muestran los datos se decidió utilizar Microsoft Power Bi para la muestra de resultados. Dada la gran cantidad de información de las tablas generadas en la etapa de análisis, no es eficiente la descarga y posterior importación de ellas a Power Bi por lo que fue necesario cargar directamente las tablas desde Google BigQuery a Microsoft Power Bi.

Inicialmente, se generaron 3 reportes en Power Bi, uno para cada año y un tercero que compara las métricas entre ambos años. El propósito es exponer el reporte con la comparación de métricas y tener los otros dos como respaldo en caso de ser necesario mostrar la información más detallada.

Se utilizan etiquetas, gráficos de barra y diagramas de Sankey para mostrar las principales conclusiones. Las etiquetas y gráficos de barra se utilizan para mostrar las frecuencias en estaciones de inicio y fin y la popularidad de los días de las semana, como también la duración de los viajes realizados. Por otro lado el diagrama de Sankey se utilizó como una forma de representar las rutas con mayores tránsitos sin la necesidad de mostrarlas en un mapa.

Comentario: Cabe señalar que dado que la cuenta con la que se crearon las visualizaciones de Power BI es una cuenta básica no es posible compartir completamente la visualización en Github por lo que solo se puede subir una parte de está visualización en formato pdf.

STEP	SHARE
Guiding questions:	**¿Pudo responder a la pregunta de cómo los miembros anuales y los ciclistas ocasionales usan las bicicletas de Cyclistic de manera diferente?** Considerando la business question ha responder, considero que la información proporcionada en este análisis ayudará a responder está pregunta. Esto lo podemos destacar sobre todo en los patrones de uso, ya sean los días o las estaciones con mayores frecuencias por los usuarios **¿Qué historia cuentan sus datos?** Una de las primeras conclusiones objetivas a tener en cuenta es que en general existe un aumento en el uso de los servicios por parte de los usuarios entre los años 2019 y 2020, independiente de su categoría. Algo interesante que se muestra en la información es que los usuarios miembros son en gran mayoría los usuarios del servicio y los usuarios casuales representan un porcentaje menor al 10% del total de paseos realizados en cada uno de los años aproximadamente. Además es interesante destacar que este explosivo aumento en el uso de los servicios pudo deberse al aumento de las medidas restrictivas debido a la pandemia de COVID-19.

	¿Cómo se relacionan sus hallazgos con su pregunta original? Se relacionan directamente ya que permiten tener una visión real de los usuarios que principalmente utilizan nuestros servicios y enfocar los recursos en aquellos que queremos que compren la membresía y así aumentar nuestro público **¿Quién es su audiencia? ¿Cuál es la mejor manera de comunicarse con ellos?** La audiencia a la que se le presentaron estos resultados fue Lily Moreno, la directora de Marketing y a su equipo. La mejor manera de presentar los datos es mostrar las estaciones y rutas más populares en un mapa y los datos empíricos que respaldan las conclusiones que se mencionan respecto al enfoque del marketing. **¿Puede la visualización de datos ayudarlo a compartir sus hallazgos?** Si ya que es una presentación intuitiva y fácil de entender. Está posee graficas, datos numericos y representaciones graficas de los datos **¿Su presentación es accesible para su audiencia?** Si, fue construida pensando en los diferentes espectros de color y como se asocian cada uno de ellos.
Key Task	Determinar la mejor manera de compartir lo obtenido Crear visualizaciones efectivas Presentar los hallazgos Asegurarse de que el trabajo es accesible
Deliverable	**Supporting visualizations and key findings** Presentación de los resultados y las principales conclusiones de ellos se muestran al final de la presentación

- **Act**

En está última etapa es donde se presentan las conclusiones finales y se realizan las sugerencias de la implementación de la business task asociada al análisis.

Con fines de facilitar el acceso a la información y a todo el contenido asociado a este análisis, la información, tablas, códigos, visualizaciones, entre otros datos, fueron subidos a GitHub para que puedan ser estudiados, comentados y revisados por otros usuarios.

STEP	**ACT**
Guiding questions:	**¿Cuál es su conclusión final en función de su análisis?** Las conclusiones generales que se pueden obtener del análisis son 3 grandes ideas: la primera es que existe una mucho mayor cantidad de usuarios miembros del servicio que aquellos que lo utilizan de manera casual, la segunda es que las rutas más utilizadas por los usuarios casuales suelen enfocarse en áreas de la ciudad destinadas a la recreación y paseo. Y como tercera ideas, es el aumento considerable en usuarios, miembros y casuales, del año 2019 al 2020, las explicaciones a este fenómeno pueden deberse a la pandemia de COVID-19 que afectó al mundo el 2020.

	¿Cómo podrían su equipo y su empresa aplicar sus conocimientos? La mejor manera de aplicar estos resultados es en una mejor toma de decisiones y la optimización de los recursos de marketing. Esto se debe a que las decisiones son basadas en datos y permiten ser orientadas a un público en particular en lugares más frecuentados por ellos en los tiempos de mayor confluencia. **¿Qué próximos pasos darían usted o sus partes interesadas en función de sus hallazgos?** Los próximos pasos a realizar están dirigidos a basar el marketing de la compañía en las rutas de usuarios casuales con un enfoque en los rangos etarios de los usuarios que mayoritariamente utilizan el servicio y esperar feedback de ellos para ir adaptando está campaña para tornarla siempre más efectiva. **¿Existen datos adicionales que podría utilizar para ampliar sus hallazgos?** Si, otros datos que podrian ser utiles para la mejor toma de decisiones estan las edades de los usuarios del servicio, el propósito con el que ellos lo utilizan, la posibilidad de conocer no necesariamente la identidad pero una forma de reconocer si los usuarios se repiten para conocer con precisión el número de miembros o usuarios casuales ocupan la red de bicicletas.
Key Task	Crear un portafolio Agregar el caso de estudio Practicar la presentación de los resultados obtenidos
Deliverable	**Your top three recommendations based on your analysis** Basado en los datos analizados, mis tres principales recomendaciones son: ● Enfocar el marketing en las 2 o 3 rutas más utilizadas por los usuarios casuales. Tales como la ruta entre las estaciones de *Lake Shore Dr. & Monroe St. y Streeter Dr. & Grand Av*. (ruta más transitada en ambos años), entre otras. ● Si se realiza marketing presencial, enfocar este los fines de semana en está u otras rutas casuales ya que es cuando se concentran los mayores números de usuarios. ● Por último, realizar una campaña de marketing a menor escala con los usuarios miembros para fortalecer permanencia. Una sugerencia es generar un programa de recompensa por recomendación de membresía podría aumentar las inscripciones.