Case study: How Does a Bike-Share Navigate Speedy Success?

1 Introducción

Este es un proyecto para finalizar el curso "Google Data Analytics" por Coursera. En este proyecto se busca poner en marcha los diferentes puntos aprendidos en el curso , así cómo el uso de las diferentes herramientas para el análisis y visualización de datos. Para este proyecto utilizamos el Case Study: How does a bike-share navigate speedy success?, el caso contiene instrucciones las cuales nos guían en como solucionar este proyecto.De acuerdo a la metodología del curso una de las mejores estrategias para obtener soluciones guiadas por datos se sigue con el proceso ask,prepare,process,analyze,share and act. El cual seguiremos en este proyecto.

1.1 Escenario

El director de marketing de Cyclistic, una compañía de bicicletas en Chicago, cree que el éxito de la compañía depende en maximizar el número de membership anuales, por tanto el equipo quiere entender cómo los usuarios casuales y los usuarios con membresía utilizan las bicicletas. De acuerdo a este análisis el quipo puede diseñar una nueva estrategia para convertir a los usuarios casuales en miembros.

2 Ask

Preguntas clave para el equipo de marketing en un futuro:

- ¿Cómo los usuarios miembros y los usuarios casuales utilizan las bicicletas de distinta manera ?
- ¿Por qué los usuarios casuales comprarían la membresía casual?
- Cómo Cyclistic puede utilizar las redes sociales para influenciar a los usuarios casuales a convertirse en miembros?

Puntos clave para nuestro análisis.

- ¿Cómo los usuarios miembros y los usuarios casuales utilizan las bicicletas de distinta manera?
- Tener claro el objetivo de la tarea.
- Una descripción de la fuente de datos a usar.
- Documentación del proceso de la limpieza y manipulación de datos.
- Un resumen del análisis hecho.
- Visualizaciones del análisis.
- Recomendaciones con base en tu análisis.

Identificando estas preguntas podemos decir que tendremos que analizar las principales diferencias entre los usuarios miembros y los usuarios casuales en orden de encontrar insights e información valiosa para entregar conclusiones que ayuden a la creación de estrategias para convertir a los usuarios casuales en usuarios miembros.

3 Prepare

Los datos se encuentran en una página web , estos datos son de uso libre y se pueden acceder a ellos mediante el enlace a dicha página. Los datos están organizados en archivos zip por mes y tenemos datos desde 2020-08 hasta 2021-08 , es decir tenemos un año de datos recabados. También se incluyen datos de diferentes años separados por trimestres, tenemos datos desde 2013 hasta el 2020. Cada archivo zip incluye una tabla con diferentes campos y una tabla con la descripción de cada campo. Procederemos a realizar un Análisis ROCCC de los datos :

- Reliable ; Debido a que los datos son propuestos por el curso y son datos no reales, ahora no podemos saber si los datos son sesgados o no pero por la fuente de los datos entonces podemos confiar en estos datos.
- Original: Por las mismas cuestiones anteriores podemos decir que son originales.
- Comprehensive: Los datos contienen la información necesaria para responder a la pregunta y objetivos de este análisis.
- Current: La fecha de los datos nos ayudan al análisis.
- Cited: Estos datos fueron revisados por profesionales que cubren el curso, entonces es información confiable.

Un problema con estos datos, es un problema típico pues hay espacios en blanco en estas tablas de datos lo cual es un problema que se puede solucionar conforme al análisis.

4 Process

Algunos errores de estos archivos son que hacen falta datos en algunas filas, también tenemos que el número de miembros es más grande que el número de causales, entonces en ese sentido este hecho puede sesgar algunas conclusiones , por tanto debemos de tener cuidado con este punto. Las herramientas que utilizamos fueron microsoft excel y R. Los datos de cada tabla son más de 100k datos entonces son tablas bastante pesadas. El formato de cada campo está bien establecidos entonces no existe mayor problema para trabajar con ellos. Para los datos en blanco , como representaban menos del 5% de la cantidad de datos total decidimos eliminarlos pues para las operaciones que se realizaron no nos guían a valores perdidos. Ahora existe una cuestión importante pues en ciertos campos, en particular la hora de partida y la hora de llegada , la hora de llegada es menor que la hora de partida entonces leyendo un poco , esto tiene que ver por el mantenimiento de las bicicletas principalmente. Por tanto también podemos eliminar estos datos.

5 Analyze

En esta sección creamos pivot tables en excel para obtener resultados en cada tabla mensual, los resultados que queremos es el promedio de duración de los viajes , el máximo de duración y el número de viajes realizados , esto separado por días de la semana y por miembros o casuales. Esto se hizo para cada mes, después se creó una nueva tabla en la cual se agregaron estos resultados . Para los datos trimestrales , se decidió trabajar bajo el software R, teniendo un problema con el primer trimestre del año 2019, ya que no pudimos abrir el archivo pues se desconocía. El script de este proceso se agrega como archivo aparte de este documento.

6 Share

Tenemos dos puntos de vista uno local y otro más general , empezando por el punto de vista local , tenemos los resultados de la duración promedio de viajes por semana y el número de viajes por semana, Esto separado por categorías de casuales y de miembros.

Para la figura 1 notamos algo evidente, la duración promedio por día es muchísimo más grande en la categoría de causales que en la categoría de miembros , algo interesante es que esta duración no cambia a través de los días , es decir se podría esperar que exista un cambio en los fines de semana pero parece ser algo constante entre los días de la semana. El tiempo promedio de los casuales ronda los 4000 segundos, esto es aproximadamente una hora mientras que los miembros ronda la media hora. Ahora comparemos el número de viajes por día , en la gráfica 2 notamos que los números de viajes por día son más los usuarios miembros que los casuales , aquí sí podemos notar que los fines de semana los miembros suelen ocupar menos el servicio y los usuarios casuales suelen aumentar. Entonces hay una tendencia clara que los miembros ocupan más el servicio en los días entre semana , es decir de lunes a viernes , mientras que los fines de semana los usuarios casuales ocupan más el servicio a comparación de los días entre semana . Procereremos entonces a ver cómo es el comportamiento mensual de todo el año recorrido , recordando que tenemos

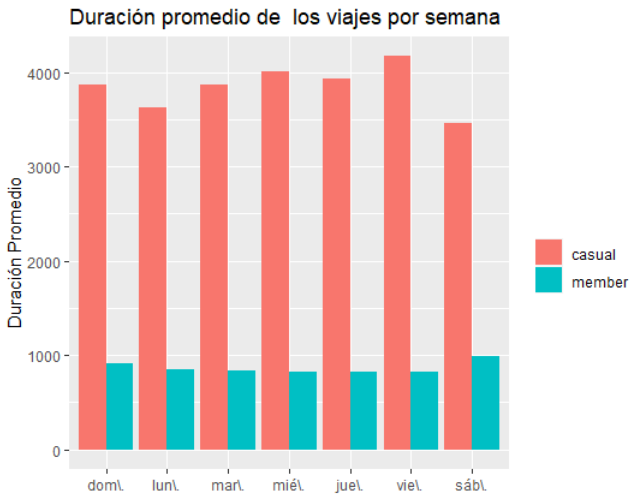

Figure 1: Duración prodemio de viajes semanales.

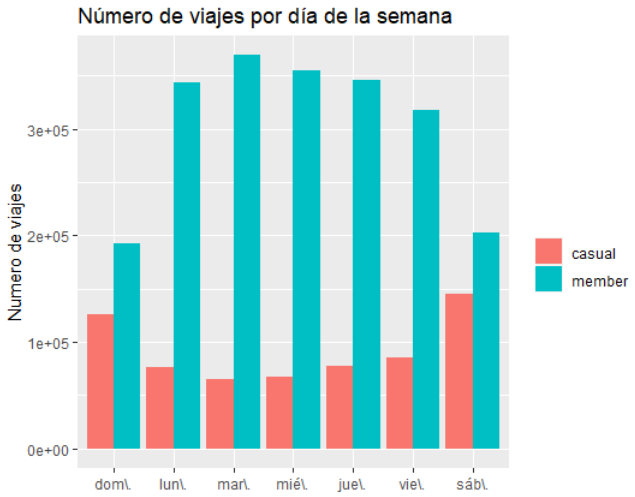

Figure 2: Numero de viajes por semana.

datos para un año de estudio, en la gráfica 3 podemos ver la duración promedio de viajes entre casuales y miembros , notamos la misma tendencia que por día de la semana , la duración de los viajes de los miembros son menores que los viajes de los casuales, ahora lo interesante es que el promedio de los casuales es muy variable e interesantemente estos tiempos van bajando conforme avanza el tiempo. Ahora analizando la gráfica 4 notamos que el número de viajes por mes tiene una tendencia bastante peculiar , esto es que asemeja una tendencia periódica es decir cada seis meses existe una tendencia alcista y bajista , siendo los primeros meses del año entrante los valores más bajos , algo interesante en el último semestre es que el número de usuarios casuales aumentó más que el número de miembros .

7 Act

Con este análisis rápido sobre estas variables podemos dar algunos insights.

- existe una tendencia de disminución de la duración temporal de los viajes en los usuarios casuales en el último semestre.
- existe un aumento en el número de usuarios casuales en el último semestre.
- Los usuarios casuales son más propensos a utilizar el servicio los fines de semana.

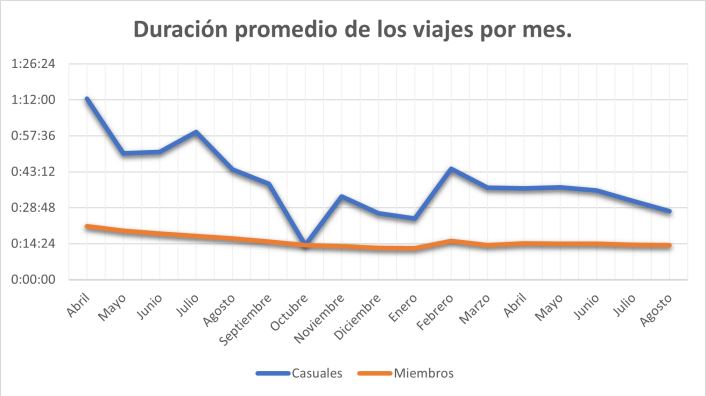

Figure 3: Duración promedio de viaje anual.

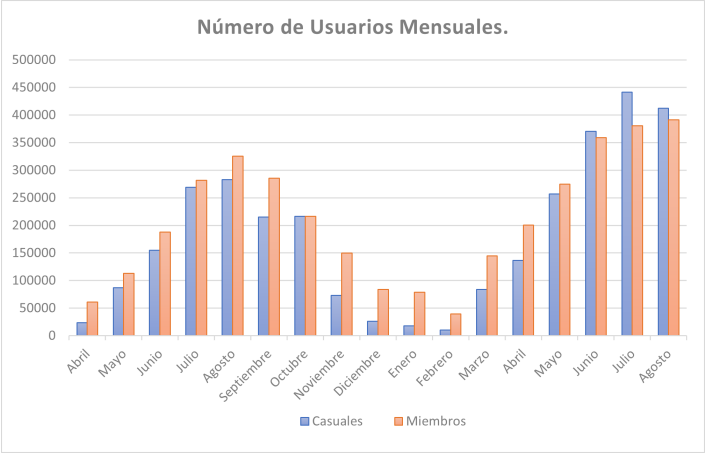

Figure 4: Número de viajes mensuales.

Con esto en mente se puede construir las bases para un desarrollo de un programa de acción para que los usuarios que son casuales , adquieran la membresía. Existen cuestiones que también quedarían por resolver cómo ¿Por qué los números de viajes por mes toman esta forma periódica , siendo el principio de año el valor más pequeño de todos?, ¿Si la duración del tiempo promedio por viaje tiene una tendencia de disminución en el último semestre , esto a que es debido?, ¿Cuales son los puntos en los cuales hay más viajes en el último semestre?. Existen más preguntas que estos datos no son suficientes o se podrían construir nuevas entradas las cuales podremos tratar de resolver estas preguntas para así tener mejores insights .

8 Anexo

8.1 Documentación del proceso de limpieza

Después de interactuar un poco con las tablas, la primera decisión es eliminar los NA de nuestras tablas, para esto el software r es más rápido con este proceso entonces lo que se hizo fué

```
library(tidyverse)
library(lubridate)

df <- read_csv("Divvy_Trips_2020_09.csv")
df <- drop_na(df)
write_csv(df, "2020_09.csv")
```

Esto para cada uno de los meses, después se decidió a trabajar con Excel, utilizando este programa decidimos crear dos campos extras "ride_length" y "day_of_the_week" , el primero consiste en la diferencia de de la fecha de llegada y la fecha de partida, mientras que la segunda es el día de la semana en el que se hace el viaje. El primero se calculó de forma $= D2 - C2$ y el segundo con $= TEXT(C2, "dddd")$ donde el campo "ride_length" se transformó en entrada

$HH:DD:SS$, aquí nos encontramos con el problema de que algunas fechas de llegada eran menores que las fechas de partida , entonces en este caso se pueden hacer dos cuestiones eliminarlas o tomarlas con valor absoluto. En R esto se puede hacer de manera más directa

```
df <- df[!(df$start_station_name == "HQ_QR"
| df$ride_length<0),]
```

mientras que en Excel requiere un poco más de pasos, lo primero es filtrar los datos y después eliminarlos a mano, esta solo es una manera .