

**Proyecto 1 - 82.01**

**Trabajo Práctico Final**

Segundo cuatrimestre de 2023

**Grupo “Subte boys”**

**Integrantes:**

Agustín Chaud Olivieri, 63253

Ignacio Gallo, 62089

**Introducción**

El objetivo de este informe es realizar un profundo análisis descriptivo sobre el dataset elegido. Para esto fue necesario un proceso de limpieza y manejo de los datos, mediante consultas SQL.

Para extraer conclusiones certeras sobre los resultados obtenidos, complementamos el análisis con noticias e informes para darle mayor contexto.

Finalmente exploramos el potencial del dataset y como este se podría mejorar.

**Ficha Descriptiva**

**Nombre de base:** Viajes en Subte 2023

**URL: Fuente:** https://data.buenosaires.gob.ar/dataset/subte-viajes-molinetes

**Periodo de análisis:** 01/01/2020 - 31/09/2023

**Frecuencia de actualización:** mensual

Se unificaron las bases de datos del 2020 al 2023, totalizando 44 CSVs.

**Cantidad total de Filas:** 32.882.180

**Columnas:** fecha, desde, hasta, línea, molinete, estación, pax\_pagos, pax\_pases\_pagos, pax\_franq, pax\_totales

**Preparación de los datos**

Al bajar la base de datos en SQL, la fecha tenía formato “integer”. Para evitar que esto genere errores, cambiamos el formato de la fecha a “date”. Repetimos el mismo proceso en los campos “desde” y “hasta”, que representan horarios.

El dataset tiene inconsistencias en el campo fecha. Hay faltantes, errores en la fecha y CSVs con columnas adicionales o cambios de formato de algunas variables, algunos tienen comillas dobles al principio y al final de cada fila.

Para pasar a un solo tipo de fecha implementamos las siguientes sentencias SQL:

**SELECT** fecha **FROM** (**SELECT** fecha **FROM** "2020\_modificado" m

**UNION** **ALL**

**SELECT** fecha **FROM** historico h

**UNION** **ALL**

**SELECT** fecha **FROM** "2022\_modificado" m2

**UNION** **ALL**

**SELECT** fecha **FROM** "2023\_modificado" m3) **AS** subquery

**WHERE** (fecha !~ '^\\d{1,2}/\\d{1,2}/(2020|2021|2022|2023)$')

**OR** (**TO\_DATE**(fecha, 'DD/MM/YYYY') **IS** **NULL**)

**OR** (**EXTRACT**(**YEAR** **FROM** **TO\_DATE**(fecha, 'DD/MM/YYYY')) **NOT** **IN** (2020, 2021, 2022, 2023));

**UPDATE** historico

**SET** fecha = **TO\_CHAR**(**TO\_DATE**(**trim**(fecha), 'YYYY-MM-DD'), 'DD/MM/YYYY')

**WHERE** **trim**(fecha) ~ '^\\d{4}-\\d{2}-\\d{2}$';

**ALTER** **TABLE** "2023\_modificado"

**ALTER** **COLUMN** fecha **TYPE** **DATE** **USING** **TO\_DATE**(fecha, 'DD/MM/YYYY');

**ALTER** **TABLE** historico

**ALTER** **COLUMN** fecha **TYPE** **DATE** **USING**

**CASE**

**WHEN** fecha ~ '^\d{1,2}/\d{1,2}/\d{4}$' **THEN** **TO\_DATE**(fecha, 'DD/MM/YYYY')

**WHEN** fecha ~ '^\d{4}-\d{2}-\d{2}$' **THEN** **TO\_DATE**(fecha, 'YYYY-MM-DD')

**ELSE** **NULL** -- Si no se reconoce el formato, se podría establecer como NULL o manejar de otra manera

**END**;

**CREATE** **TABLE** subte **AS** **TABLE** "2020\_modificado" ;

**INSERT** **INTO** subte **SELECT** \* **FROM** "2023\_modificado";

Estas sentencias fueron utilizadas para resolver los problemas de formato de la fecha, casos de dd/MM/yyyy, yyyy-mm-dd, en formato char y en algunos casos expresando día y/o mes con un solo dígito.

Además exclusivamente en junio de 2022 se agregaron más campos al dataset pero luego no se mantuvo esa estructura, generandose una inconsistencia en los datos de ese año. Cometieron el mismo error en septiembre del mismo año, sólo con los datos de 3 líneas. Solucionamos dicho inconveniente eliminando esa columna, trabajando los CSV con pandas en python.

import pandas as pd

from datetime import datetime

# Cargar el archivo CSV

df = pd.read\_csv('Fix/202009\_DEH.csv', delimiter=';', encoding='ISO-8859-1')

# Eliminar las columnas que no se necesitan

columnas\_a\_mantener = ['FECHA', 'DESDE', 'HASTA', 'LINEA', 'MOLINETE', 'ESTACION', 'pax\_pagos', 'pax\_TOTAL']

df = df[columnas\_a\_mantener]

# Convertir el formato de la fecha

# Suponiendo que 'FECHA' es la columna que contiene las fechas en el formato incorrecto

df['FECHA'] = pd.to\_datetime(df['FECHA'], format='%d/%m/%Y').dt.strftime('%Y-%m-%d')

# Guardar el DataFrame modificado a un nuevo archivo CSV para su importación

df.to\_csv('Fix/09DEH.csv', index=False, sep=';', encoding='ISO-8859-1')

Otro problema que nos surgió fue que cargamos mal y duplicamos los datos pertenecientes a julio 2023. A causa de eso ejecutamos un script en toda la base de datos para arreglar datos repetidos y encontramos datos repetidos en el mismo molinete y exactamente el mismo horario. De esta forma un error nuestro nos permitió hallar y corregir datos duplicados del dataset.

**SELECT** \*

**FROM** subte

**WHERE** ctid **IN** (

**SELECT** ctid

**FROM** (

**SELECT** ctid,

**ROW\_NUMBER**() **OVER** (**PARTITION** **BY** fecha, desde, hasta, linea, molinete, estacion, pax\_pagos, pax\_total **ORDER** **BY** ctid) **as** rn

**FROM** subte

**WHERE** fecha >= '2023-07-01' **AND** fecha < '2023-08-01'

) t

**WHERE** t.rn > 1

);

Finalmente unificamos todos los registros en una tabla nueva, copiando registros de a batch para evitar colapsar el motor de base de datos. Luego conectamos el postgres a tableau.

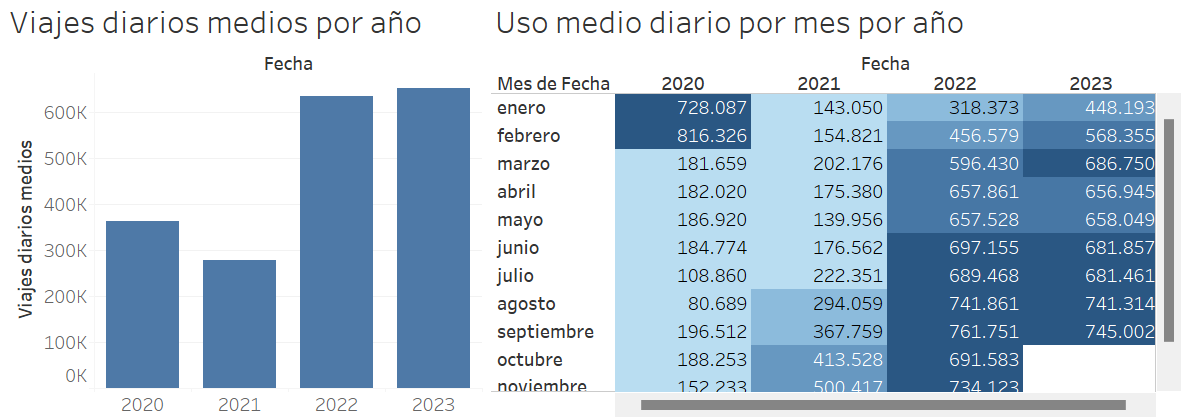
**Preguntas**

Nos planteamos responder a las siguientes preguntas

* ¿ Cómo impactó la pandemia en el flujo de personas del subte?
* ¿Cómo se encuentran distribuidos los viajes temporal y geográficamente?
* ¿ Es necesario ampliar el horario de servicio?
* Exploramos la eficacia del Acuerdo de operación del servicio público con respecto al mantenimiento de los molinetes

Partiendo de estos puntos fuimos generando consultas de SQL para agrupar y filtrar los datos en busca de respuestas y gráficos en tableau que faciliten la interpretación de las mismas. Luego integramos todos estos gráficos en un dashboard para construir el storytelling.

**Impacto de la pandemia en el flujo de personas**



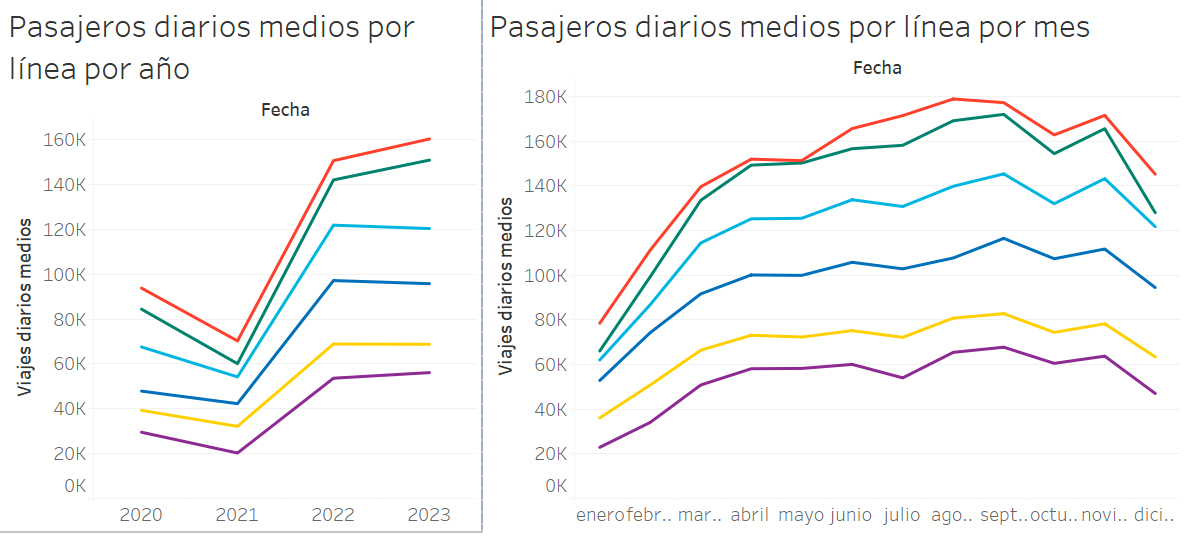
*Gráfico 1*

Viendo las variaciones anuales se identifica inmediatamente una caída fuerte del nivel de viajes diarios medios del 2020 al 2021, con una marcada recuperación que inicia en 2022 y continua 2023. Vemos que los niveles pre-pandemia, enero y febrero de 2020, de 728.000 y 816.000 viajes diarios medios respectivamente, no se recuperarán hasta agosto del 2022.

**Distribución temporal** **y geográfica de los viajes**

En el gráfico 2 identificamos que la pandemia afectó de forma similar a todas las líneas.

También destacamos que las líneas B (roja) y D (verde) son las más utilizadas y las líneas H (amarilla) y E (violeta) son las menos transitadas.

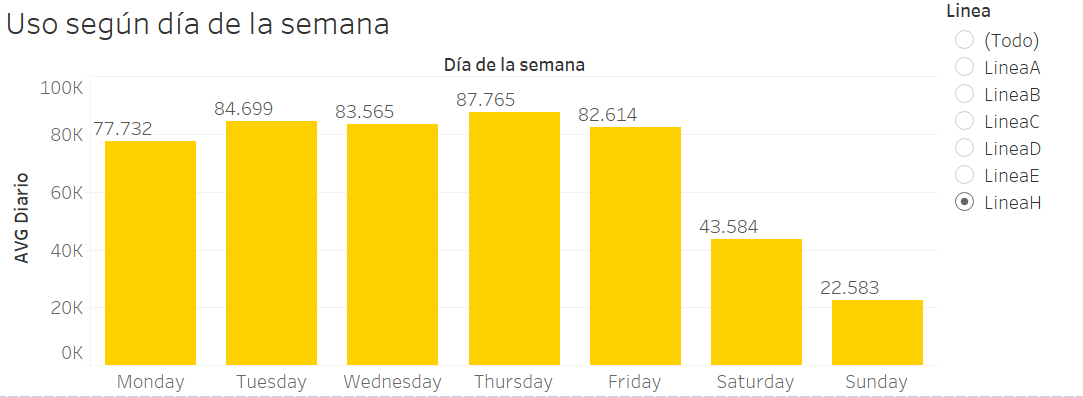
****

*Gráfico 2*

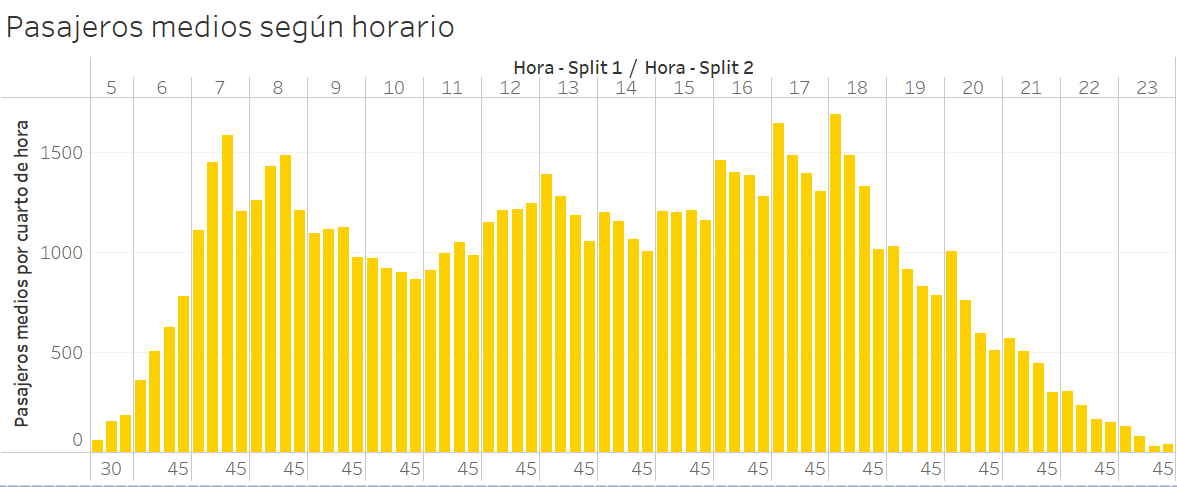
A los gráficos 3 y 4 se les aplica un filtro de línea,en este ejemplo está seleccionada la línea H.

En el gráfico 3 se puede observar como va variando el promedio de viajes diarios según el día de la semana y la línea elegida, siendo este relativamente homogéneo durante los días de la semana y sufriendo una gran disminución los fines de semana. El gráfico 4 refleja cómo cada línea tiene dos picos, uno por la mañana y otro por la tarde.

Más adelante corroboramos este patrón con una noticia.

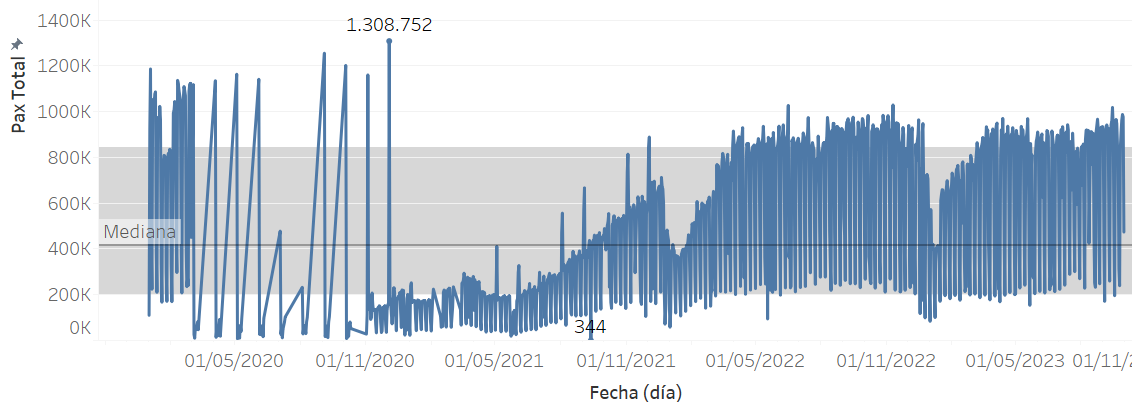


*Gráfico 3*



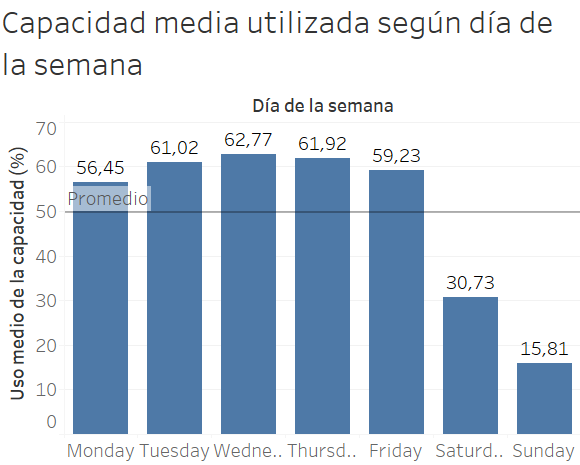
*Gráfico 4*

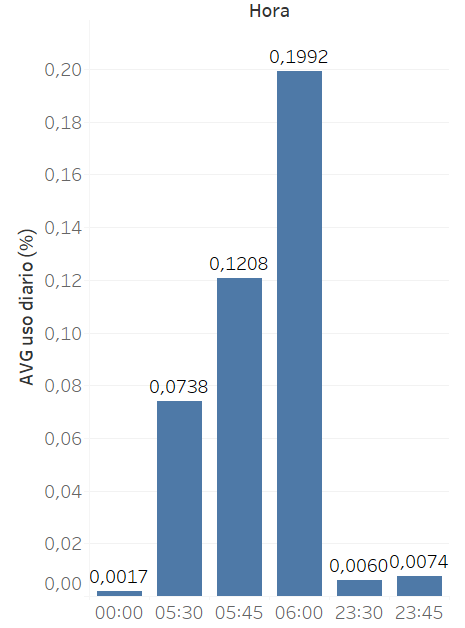
**¿ Es necesario ampliar el horario de servicio?**



*Gráfico 5*

Tomamos el máximo diario de los últimos cuatro años, 1.308.752, para establecer la capacidad máxima demostrada por el sistema. En base a eso calculamos el porcentaje de uso de la capacidad de todas las líneas, según el día de la semana. Su promedio nos dió 50%.

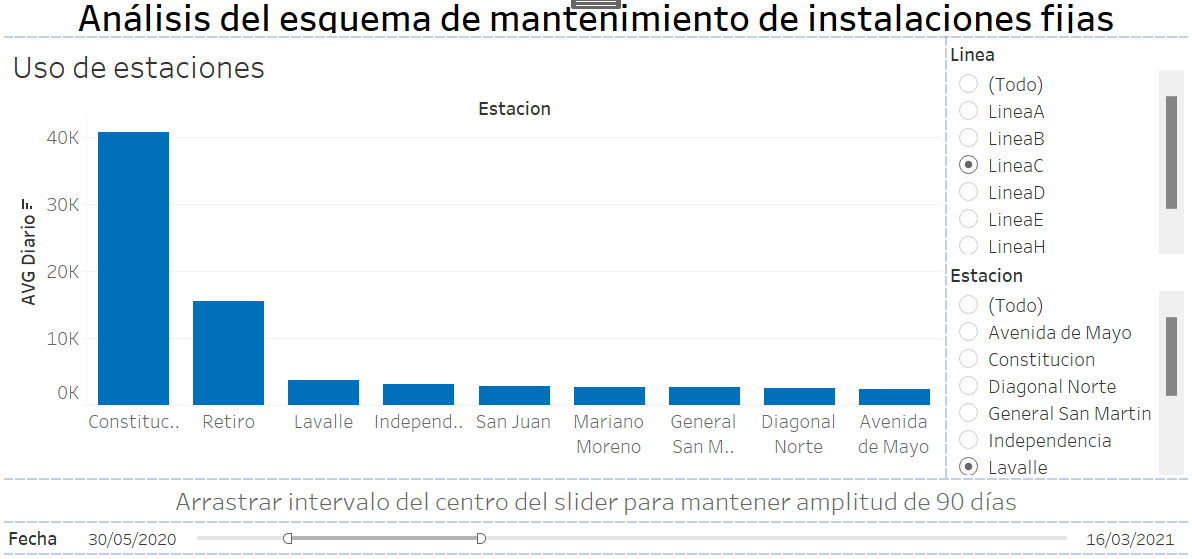
*Gráfico 6*



*Gráfico 7*

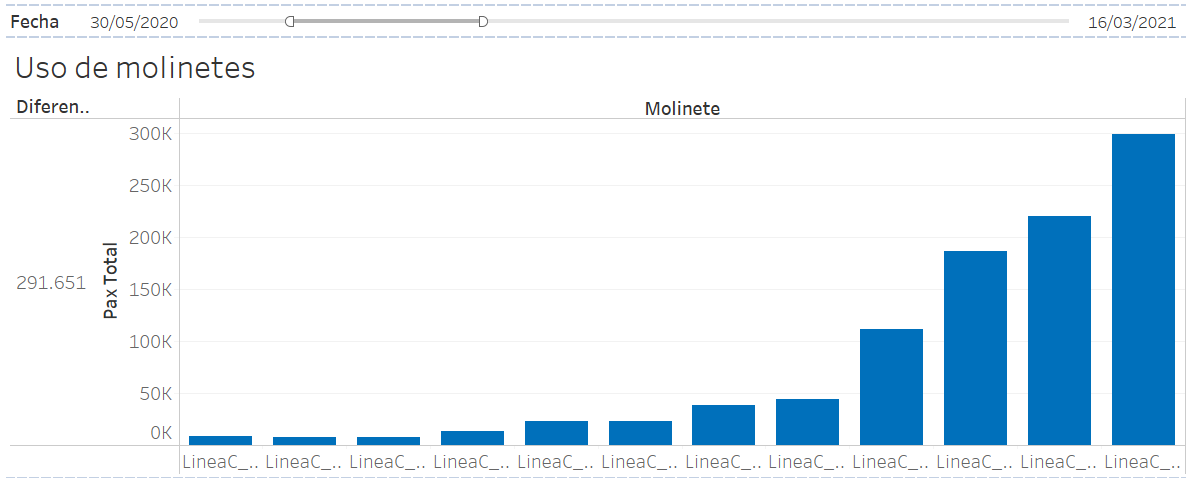
En el gráfico 7 vemos que en los horarios pertenecientes al primer subte(varían desde 5:30 a 6:00 según la línea y día) y el último (varían desde 23:30 a 00:00) nunca se supera el 0,2% de ocupación, por lo que no hay ninguna necesidad de aumentar la franja horaria de servicio.

**Operación del servicio y mantenimiento de los molinetes**

****

*Gráfico 8*

Los gráficos interactivos 8 y 9 permiten filtrar por línea y por estación, representando el flujo de pasajeros por estación y el uso de cada molinete. En todas las líneas se repite el patrón de que las estaciones terminales y aquellas con cercanía a terminales del servicio de ferrocarriles tienen mayor afluencia de pasajeros. Estas diferencias de uso reflejan necesidades distintas de mantenimiento y limpieza, las cuales no se ven reflejadas en el AOM.



*Gráfico 9*

El gráfico 9 refleja la gran diferencia de uso entre los molinetes de una misma estación, debido principalmente a su posición dentro de la misma. De esta forma el periodo de 90 días establecido,en muchos molinetes es excesivo mientras que no resulta suficiente para otros.

**Dashboard completo**

Por la cantidad de registros no se puede subir el dashboard a tableau public.

Este es el link para descargarlo:

<https://drive.google.com/file/d/1ZdXqJe6KSdIk-sujIrXYXRLfLNsPySTJ/view?usp=sharing>

\*\*Nota al lector:

Algunos monitores recortan el dashboard y no permiten ver los textos o generan que algunos gráficos tengan que ser scrolleados. En caso de que los textos estén recortados realizar doble click para ver el resto del contenido.

**Contexto:**

Realizamos una búsqueda de información mediante noticias y papers,para complementar nuestro análisis y enriquecer nuestras conclusiones.Utilizamos una tesis de la universidad Torcuato Di Tella, sobre la aplicación de métodos cuantitativos en la hora pico del subte de Buenos Aires.

La misma puede encontrarse en el siguiente link

<https://drive.google.com/file/d/1E0lBc95splrASsXNbdA04_JJyiWo2zm9/view?usp=sharing>

Se aplicaron una gran cantidad de modelos y fórmulas a lo largo de toda la tesis, pero lo que nos interesa es que realizaron un relevamiento de demanda real *in situ* durante la hora pico de la mañana (7:30 a 8:30) en las 6 estaciones iniciales de la línea D. De esta forma llegaron a la conclusión de que en el horario de mayor afluencia los subtes en promedio viajan al 70% de su capacidad máxima cuando logran mantener una frecuencia de 2 minutos y 45 segundos.

Según sus frecuencias observadas *in situ* más de una vez se verificaron intervalos cercanos y hasta inferiores a los 2 minutos,pero también mayores a 3 minutos. Esto se debe a los altos desvíos estándar entre dichos intervalos.

También complementamos nuestro análisis con esta noticia del diario Perfil

<https://www.perfil.com/noticias/sociedad/subte-se-adelanta-la-hora-pico-y-lo-usa-mas-gente-martes-y-jueves.phtml>

Si bien esta noticia es del 2017, los patrones mencionados siguen vigentes: la hora pico de la tarde registra su mayor afluencia entre las 17 y las 19, y en todas las líneas hay mayor demanda a las 14 hs que a las 19 hs. También se cumple que entre los días de la semana el lunes es el día con menor flujo de viajes, en cada una de las líneas. También en todas las líneas se cumple que el día de la semana con mayor flujo de viajes es el jueves.

Las líneas B y D son las más utilizadas y las menos usadas son la E y H.

**Potencial a futuro:**

El tema del flujo de personas dentro de una ciudad siempre será un tópico de estudio central.

Una modificación que le haríamos a la base de datos consiste en agregar el destino en el que la persona finaliza su viaje. Esto serviría para tener una mirada más integral sobre los destinos más concurridos en días y horarios específicos. Podrían usarse estos datos para armar mapas de calor de los viajes en distintos horarios. Esto sería útil ya que partiendo de datos verídicos se pueden generar reglas y leyes que sean más eficientes para incrementar la seguridad y condiciones de vida en la ciudad. Además, conociendo los puntos de destino, sería más sencillo analizar el diseño del servicio de subte y medir si se ajusta a las necesidades de los ciudadanos. Finalmente creemos que se podrían desbloquear un sin fin de posibilidades analíticas integrando y cruzando los datos de uso de Subte, Colectivo y Tren. Mediante el análisis conjunto podremos entender como la carga de uno alivia al otro, cómo podemos incentivar el uso de uno por sobre otro, como se complementan, cuales son los principales puntos de transbordo y en qué horarios se da la máxima concurrencia. Finalmente, con el dato de donde termina el viaje en subterráneo y empieza el de colectivo y viendo a través de un mapa de calor el trayecto que realizan los mismos podríamos evaluar la necesidad de extender las líneas existentes.

**Conclusiones:**

Al analizar los datos concluimos que a grandes rasgos la pandemia afectó similarmente a todas las líneas. También hallamos un patrón de uso relativamente homogéneo entre los días lunes y viernes, con una disminución muy marcada el día sábado y un decremento adicional el domingo.

Dicho patrón de uso tiene las mismas características generales que el del 2017.

Por otro lado,llegamos a la conclusión de que actualmente no existe una necesidad genuina de extender el servicio de subte a las 24 hs como proponía el candidato a jefe de gobierno Santoro. Esto se debe a que en los horarios de partida del primer y último subte nunca se supera el 0,2% de su capacidad máxima de uso.

También concluimos que el período de mantenimiento de molinetes establecido en el AOM, de 90 días, en muchos casos es excesivo para algunos molinetes mientras que no resulta suficiente para otros. Esto se debe a grandes diferencias de uso entre los distintos molinetes de cada estación.