skills.tech

Education Is Transformation

Diplomado

Data Scientist

Proyecto Python

skills.tech

Proyecto Movilidad

Tema: Transporte Público CDMX
Herramienta: Python

Alfredo Jimenez

skills

Objetivo:

En este proyecto, el alumno aplicará las habilidades obtenidas en el curso para analizar y evaluar el transporte público de la CDMX, basado en datos y estadística básica

Escenario

Eres un consultor privado y tienes como tarea explorar y analizar el uso del transporte público en la Ciudad de México a través del tiempo para todos sus distintos medios de transporte, así como evaluar posibles impactos que pueda tener el cierre de la línea 1 del metro.

Materiales:

Ambiente de desarrollo de Python:

- Google colab
- Ó Jupyter notebook en local

Datos:

https://datos.cdmx.gob.mx/dataset/da3fcf80-f15f-4478-9795-26eddaa6fe86/resource/5d33f9c7-e033-4676-a02d-9e2129017acf/download/afluencia-preliminar-en-transporte-publico.xlsx-afluencia_diaria.csv

Instrucciones:

Sigue las diapositivas y presenta tus resultados,
respuestas e insights en las diapositivas designadas.

Entregable:

1. **Esta presentación:** con tu nombre en la portada y con las respuestas en las diapositivas designadas.

2. **EL Jupyter Notebook** con tus métodos y procesos de dónde sacaste las respuestas.

Práctica

Sección 1: Obtención de datos

Instrucciones 1:

Descargar los datos:
https://datos.cdmx.gob.mx/dataset/da3fcf80-f15f-4478-9795-26eddaa6fe86/resource/5d33f9c7-e033-4676-a02d-9e2129017acf/download/afluencia-preliminar-en-transporte-publico.xlsx-afluencia_diaria.csv

Los datos contendrán varias columnas:

- **id:** ID del registro o número de registro
- **organismo:** organismo utilizado de transporte (ecobici, metrobus, etc.)
- **linea_servicio:** línea de servicio, en caso de aplicar (línea 1, por ejemplo)
- **dia:** de la semana, domingo, lunes, etc.
- **fecha:** fecha del registro
- **afluencia_tarjeta:** Pago por tarjeta
- **afluencia_boleto:** pago por boleto
- **afluencia_total_preliminar:** pago total

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones :
En esta sección lo más importante es hacer las preguntas correctas para despertar interés e iniciar con la exploración de los datos.

Contesta las siguientes pregunta:

Preguntas	Respuestas
¿Cuál es la columna con más datos no nulos?	id, organismo, dia, fecha
¿Son todos los tipos de datos para las columnas correctos?	no, hay que modificar fecha a datetime y afluencia total preliminar a int
¿Cuál es el organismo más utilizado?	stc

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones :
En esta sección lo más importante es hacer las preguntas correctas para despertar interés e iniciar con la exploración de los datos.

Contesta las siguientes pregunta:

Preguntas	Respuestas
¿Cuál es la línea de servicio más utilizada?	L1
¿Qué día de la semana tiene más uso de transporte público? ¿Por cuánto?	Martes con 2667 registros de afluencias totales preliminares

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones

Responde a las siguientes preguntas:

¿Son los tipos de datos para cada columna correctos? ¿Cuáles sí y cuáles no?

Notarás que las columnas de afluencias son de tipo objeto. Cambialas a valores numéricos.

Hint: encontrarás errores al momento de convertirlas, debido a que contienen el string “,”. Retira este string antes de convertirlos utilizando **.str.replace,** y después conviértelos a datos numéricos.

Pega las celdas de código que utilizaste para esto

```
#fecha a datetime
df['fecha'] = pd.to_datetime(df['fecha'])

#afluencias a numericos
df['afluencia_total_preliminar'] = df['afluencia_total_preliminar'].str.replace(',','')
df['afluencia_boleto'] = df['afluencia_boleto'].str.replace(',','')
df['afluencia_tarjeta'] = df['afluencia_tarjeta'].str.replace(',','')

df[['afluencia_total_preliminar', 'afluencia_boleto', 'afluencia_tarjeta']] = df[['afluencia_total_preliminar', 'afluencia_boleto',
'afluencia_tarjeta']].astype('float')

df.info()
```

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
¿Cuáles son las columnas con mayor cantidad de datos nulos? Crea el siguiente **dataframe** y pega las líneas de código que utilizaste en donde se muestre el porcentaje de datos nulos **por columna**.

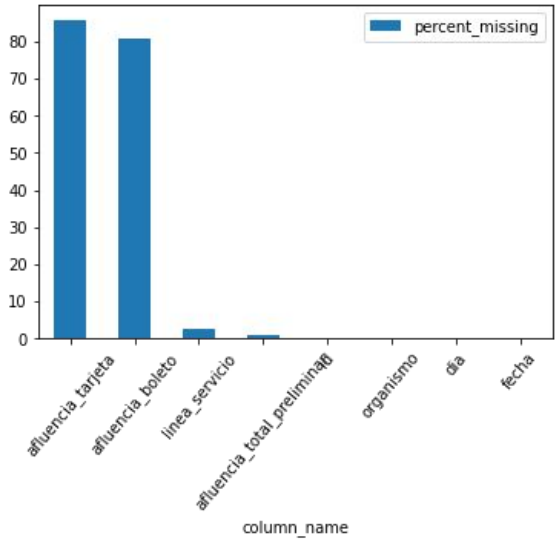
	columnas	porcentaje
0	id	
1	organismo	
2	linea_servicio	
3	dia	
4	fecha	
5	afluencia_tarjeta	
6	afluencia_boleto	
7	afluencia_total_preliminar	

```
percent_missing = df.isnull().sum() * 100 /
len(df)
missing_value_df = pd.DataFrame({'column_name':
df.columns,'percent_missing': percent_missing})
missing_value_df
```

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
Una vez que hayas creado el DataFrame, crea una **gráfica de barras** visualizando el **porcentaje** de datos nulos utilizando la **librería plotly.** En el eje x estará la columna, (id, organismo, etc.) y en el eje y estará el porcentaje. Pega aquí tu gráfica.



Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
Notarás que las columnas de boletos tienen una gran cantidad de valores nulos. ¿Por qué? ¿Se pueden eliminar todos los registros que contengan datos nulos? ¿Por qué sí o por qué no? Contesta

La columna de boletos posee una gran cantidad de datos nulos ya que puede presentar registros para un organismo o línea de servicio específica. Solo un organismo es el que recibe boletos para contabilizar la afluencia de usuarios

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Haciendo relación con la pregunta anterior, piensa en diferentes situaciones del registro de datos y por qué sí o por qué no estas tendrían valores nulos. Por ejemplo, esperaríamos que la columna de línea de servicio esté vacía cuando se usa ecobici, pues esta no tiene líneas. Crea una función que, dado un organismo, grafique la distribución de valores nulos cuando se tiene ése organismo, y pega la distribución de valores nulos para las demás columnas para el caso de **ecobici.**

Hint: crea una función que seleccione el dataframe cuando en organismo se tiene un valor específico, y después grafica los valores nulos. Esperarás obtener ahora un 100% de datos nulos en las columnas de boletos, tarjeta y línea de servicio.

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Reto

Crea la misma gráfica, que anteriormente, pero en lugar de revisar los datos nulos cuando se utiliza un organismo en específico, revisa la distribución de organismos, dada una columna es nula. Por ejemplo: para los valores nulos de la columna de linea_servicio, ¿qué porción de estos son debido a ecobici? Los que no, ¿a qué se le pueden atribuir? Pega la distribución para la columna de **afluencia_tarjeta**.

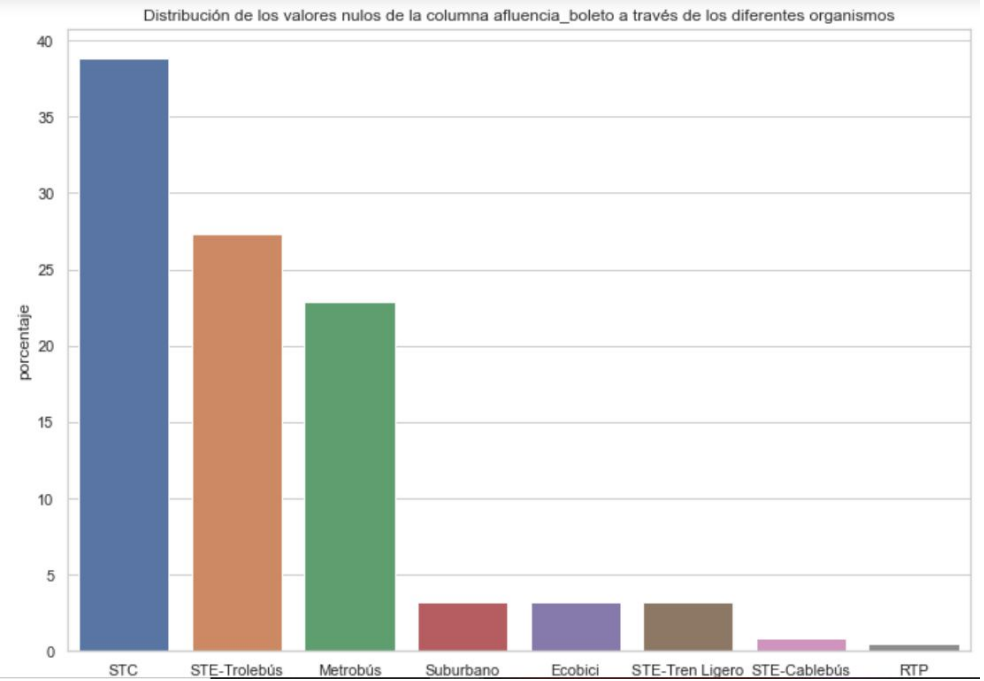
Análisis como estos son los que se deben de tomar en consideración a la hora de limpiar datos. Regresando al ejemplo, si se tienen datos nulos en la columna de linea_servicio que NO se les pueden dar una explicación o justificación, se consideran errores en la captura de datos, y se eliminan.

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Reto

Ejemplo: esta es la gráfica que se genera con la columna de afluencia_boleto



Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
¿Cómo se distribuyen las columnas de afluencia_tarjeta, afluencia_boleto y afluencia_total_preliminar? ¿En dónde está el promedio? ¿Se concentran más en los valores altos o bajos? Crea un histograma con 40 bins para cada columna, uno encima del otro (3 renglones, 1 columna).
hint: px.histogram o go.Histogram

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
Para continuar con el análisis de estas columnas, crea un boxplot de dichas columnas.

Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

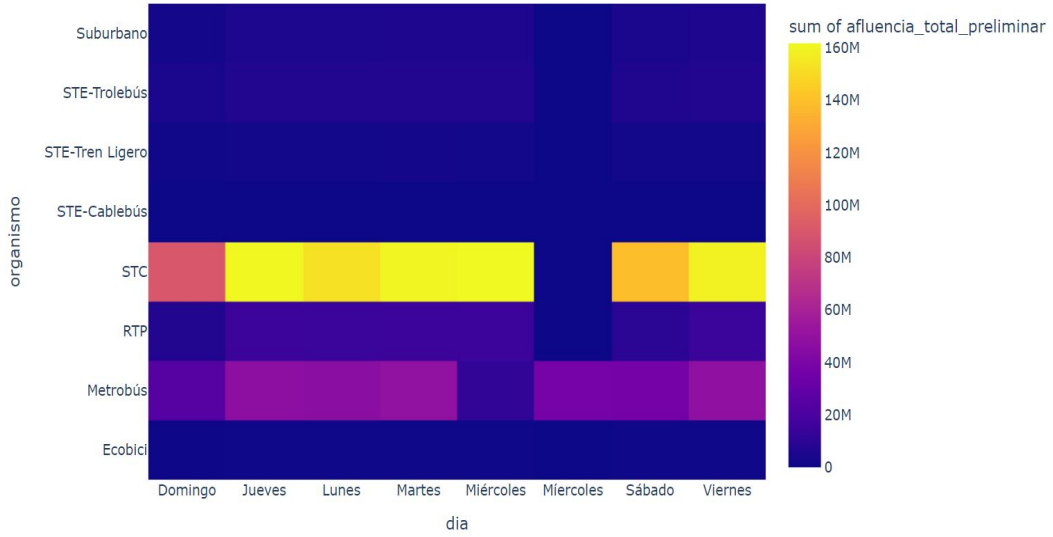
Instrucciones
Responde a las siguientes preguntas:
¿Cuál es el día con más usuarios del transporte público? Pega aquí el número de usuarios de transporte público (en total) por día

Dia	Numero de Usuarios
Lunes	770843
Martes	820067
Miercoles	807048
Jueves	802291
Viernes	874324
Sabado	747252
Domingo	454374

Diapositiva de Respuestas

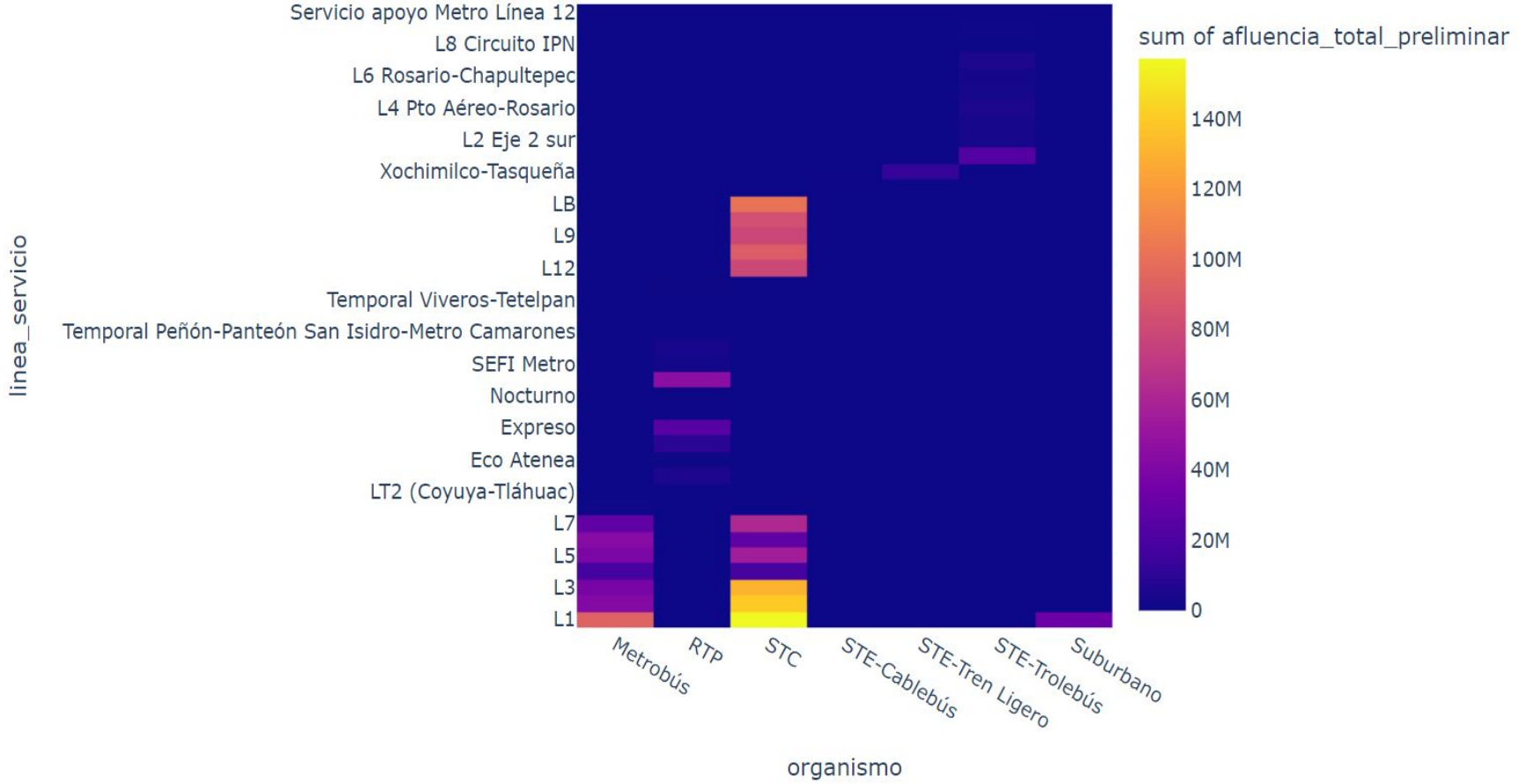
Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
¿Cuál es la combinación de día y método de transporte público con más usuarios en promedio? Visualizalo en un heatmap, como el siguiente visualizando **la suma** de afluencia_total_preliminar y pega el código para generarlo en esta diapositiva y responde a la pregunta inicial



skills.tech

Diapositiva de Respuestas



Diapositiva de Respuestas

Sección 2: Análisis Exploratorio

Instrucciones
Responde a las siguientes preguntas:
¿Cuál es la línea de servicio más utilizada? ¿Qué porcentaje de todas las líneas constituye esta línea? ¿El cierre de la línea 1 del Metro afectará fuertemente a los usuarios?

La línea de Servicio más Usada es L1, representa el 8,05% de todas las líneas, su cierre podría afectar notablemente, ya que constituye la línea con mayor tránsito de personas y al no encontrarse disponible, podrían llegar a colapsar otros medios de transporte que cubren la ruta L1.

Diapositiva de Respuestas

Sección 3: Transformación de datos

Instrucciones :
Regresando al tratamiento de valores nulos, como un experimento, elimina **todos** los registros / renglones que tengan al menos **un** valor nulo en alguna columna. Si hacemos esto, ¿qué porcentaje de datos limpios tendríamos con respecto al original?
Hint: utiliza df.dropna(how = “any”)

Se tendria un 14,35% de datos limpios respecto a la base de datos original

Diapositiva de Respuestas

Sección 3: Transformación de datos

Instrucciones :
¿Cuántos valores duplicados hay? Después, elimina todos los valores duplicados.
Hint: utiliza .duplicated() para revisar cuántos duplicados hay, y después, drop_duplicates(subset=<una lista con las columnas que quieres verificar que no sedupliquen>) para eliminarlos

No hay valores duplicados

Diapositiva de Respuestas

Sección 3: Transformación de datos

Instrucciones :
Cambia el formato de la columna “Fecha”. Recuerda utilizar el formato adecuado para la fecha. Una vez hecho éso, pega aquí los días de cada registro y el día de la semana
Hint: utiliza .dt.day y dt.weekday

	id	organismo	linea_servicio	fecha	dia	Day	Day Week	afluencia_tarjeta	afluencia_boleto	afluencia_total_preliminar
0	1	Ecobici	NaN	2020-03-01	Domingo	1	6	NaN	NaN	11238.0
1	2	Ecobici	NaN	2020-03-02	Lunes	2	0	NaN	NaN	29475.0
2	3	Ecobici	NaN	2020-03-03	Martes	3	1	NaN	NaN	31855.0
3	4	Ecobici	NaN	2020-03-04	Miércoles	4	2	NaN	NaN	31477.0
4	5	Ecobici	NaN	2020-03-05	Jueves	5	3	NaN	NaN	31493.0
...	...	...	...	...	...	...	...	...	...	...
18709	18710	Suburbano	L1	2021-06-28	Lunes	28	0	NaN	NaN	NaN
18710	18711	Suburbano	L1	2021-06-29	Martes	29	1	NaN	NaN	NaN
18711	18712	Suburbano	L1	2021-06-30	Miércoles	30	2	NaN	NaN	NaN
18712	18713	Suburbano	L1	2021-07-01	Jueves	1	3	NaN	NaN	NaN
18713	18714	Suburbano	L1	2021-07-02	Viernes	2	4	NaN	NaN	NaN

18714 rows × 10 columns

	Fecha	day	day week
0	2020-03-01	1	6
1	2020-03-02	2	0
2	2020-03-03	3	1
3	2020-03-04	4	2
4	2020-03-05	5	3

Diapositiva de Respuestas

Sección 3: Transformación de datos

Instrucciones :
Ahora, continuando con la limpieza de datos, es posible que se tenga un renglón en donde, por ejemplo, "Metrobús" aparezca duplicado. Es decir, "MetrobúsMetrobús", por error. O, en su contraparte, que queramos seleccionar todas las variantes de Metrobús, en caso de existir. Por ejemplo: "Metrobús-Sur", "Metrobús-Norte" se pueden agrupar en una sola variable llamada "Metrobús". Esto se hace buscando si la celda contiene la palabra "Metrobús". Revisa en la celda de organismo cuáles renglones contienen la palabra "Metrobús" y pega aquí cuántos hay.

Hint: utiliza .str.contains(“Metrobus”)

Hay 3461 registros de la palabra Metrobus

Diapositiva de Respuestas

Sección 3: Transformación de datos

Instrucciones :
Realiza el mismo ejercicio que anteriormente, pero con la palabra “Temporal”.
¿Cuántas líneas temporales diferentes hay?

Hay 155 registros con la palabra Temporal

skills.tech

Diapositiva de Respuestas

Sección 3: Transformación de datos

Instrucciones :
Agrupa por días, y suma la columna de afluencia_total_preliminar para todos los organismos y líneas de servicio. De esta forma tendrás un dataframe con dos columnas: la fecha y la afluencia_total_preliminar de todos los organismos. Grafica estos en una serie de tiempo. **¿Afectó el COVID-19 al uso del transporte público?** ¿Sí o no? ¿Se puede notar algún incremento o decremento en el uso por la salida o introducción de las diferentes olas? **Copia el código para generar esta gráfica aquí.**
Respuesta: El COVID-19 si afecto al uso del transporte publico, en Marzo de 2020 cuando se declaro la pandemia la afluencia de usuarios bajo drasticamente, asi como en enero del 2021 cuando se declaro una nueva variante del virus, pero la afluencia vuelve a crecer parcialmente al 3er mes

Hint: generará la siguiente gráfica


```
from matplotlib import pyplot

df_time =
df.groupby('fecha')['afluencia_total_preli
minar'].sum().fillna(0)
fig = px.line(df_time,)
fig.show()
```

skills

Diapositiva de Respuestas

Sección 4: Conclusiones

Instrucciones :
El transporte publico en Mexico ha tenido una baja de afluencia de usuarios desde los inicios de la pandemia. El organismo RTP es el unico que acepta o genera registros a traves de boletos y tarjetas, generando de esta manera datos acerca de la afluencia de usuarios que lo usan.

El proyecto nos permite encontrar informacion acerca de como es el flujo de usuarios en la red de transporte publico en el tiempo, pudiendo revisar los organismos mas utilizados, las lineas mas transitadas, la cantidad de personas que usan el transporte publico, y los dias en que es mas probable que se encuentren mas saturados.