Proyecto Final
Data Science

Alumnos: D'Alessio - Pampillón - Sosa Pivatto

CONTENIDO:

1. Abstract
2. Preguntas de interés
3. EDA
4. Entrenamiento y testeo
5. Selección del modelo final
6. Análisis de resultados
7. Conclusión

1. ABSTRACT

Citi Bike es el programa de alquiler de bicicletas compartidas más grande de Estados Unidos. El sistema es pago y está disponible para su uso 24/24. La data del uso del sistema es pública y se encuentra disponible en su sitio web.

El dataset seleccionado está conformado por un archivo CSV, de **15 columnas y 404.947 registros**, que contienen información del servicio de alquiler de bicicletas durante el año 2019 en la ciudad de New Jersey.

El análisis tiene como **objetivo** predecir la cantidad de rentas de bicicletas en la estación de mayor uso (Grove St PATH) durante el mes de septiembre.

La **audiencia** de este análisis son los directivos del servicio **Citi Bike**.

2. PREGUNTAS DE INTERÉS

Pregunta principal:

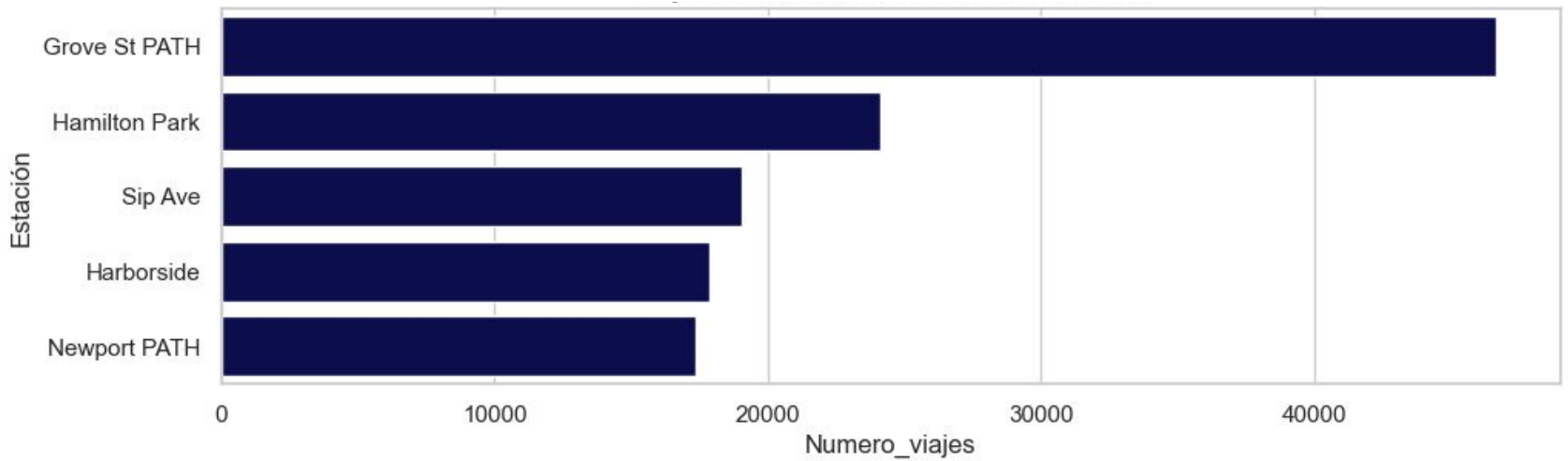
- ¿Cuál será la cantidad de usuarios que alquilarán bicicletas del sistema Citi Bike en la estación Grove St PATH durante el mes de septiembre?

Preguntas secundarias:

- ¿Cuál es el top de estaciones más usadas?
- ¿Cuáles son los meses con mayor uso del sistema?
- ¿Cuáles son los horarios de mayor uso del sistema?
- ¿Cuál es la distribución por género y tipo de usuario del sistema?
- ¿Cuál es la edad promedio de los usuarios del sistema?

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

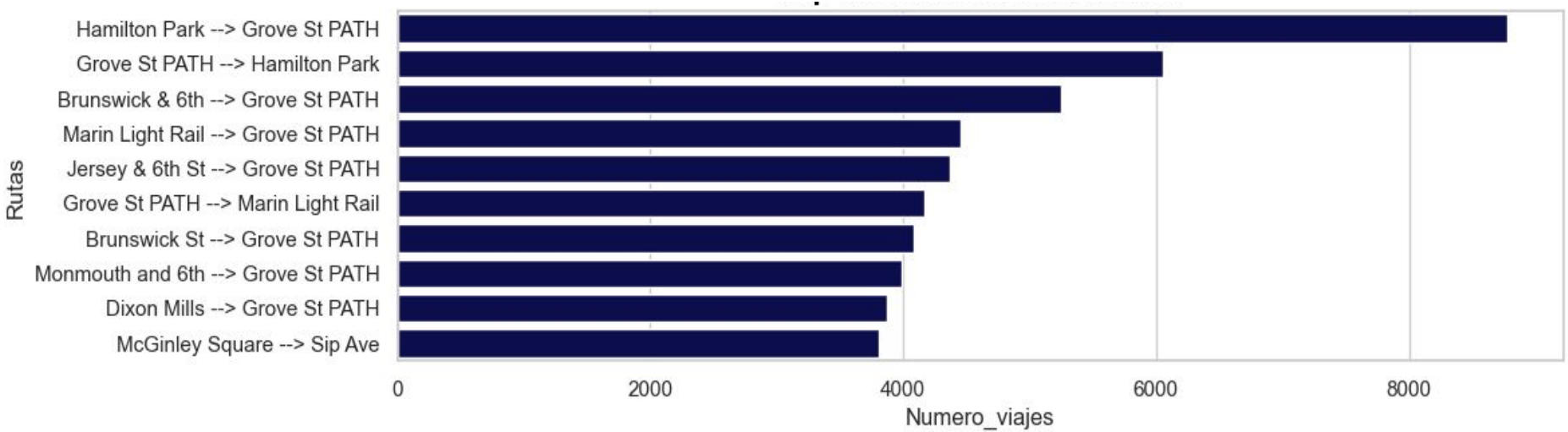
Top 5 de estaciones de salida más utilizadas



La estación más utilizada es **Grove St PATH**

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Top 10 de recorridos realizados

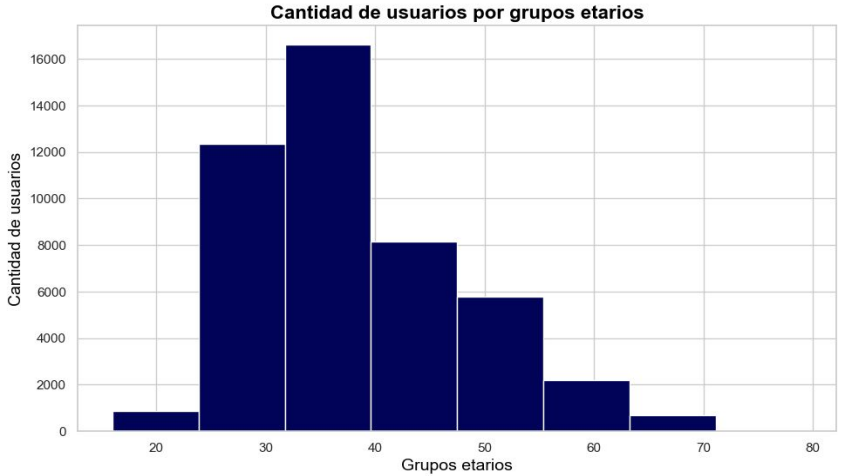


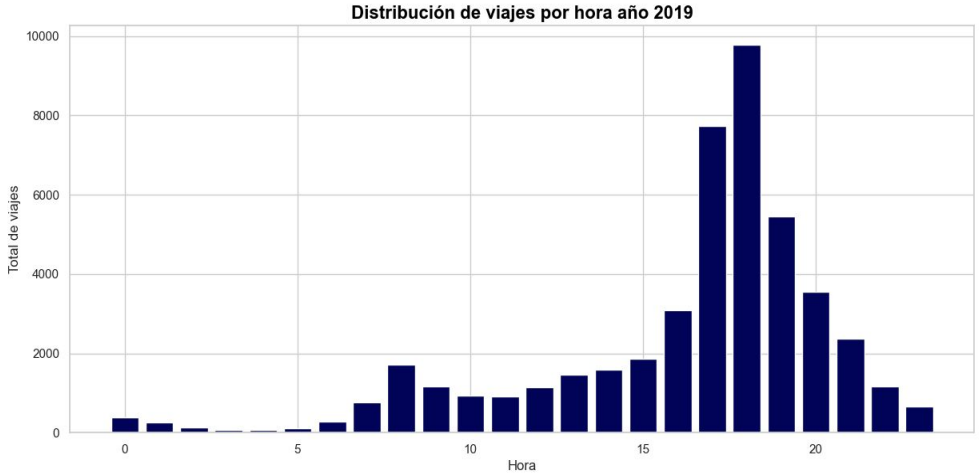
El **90%** de los recorridos tiene como origen o destino la estación **Grove St PATH**

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

A continuación se procede a realizar un análisis más específico para nuestra estación de estudio:

GROVE ST PATH





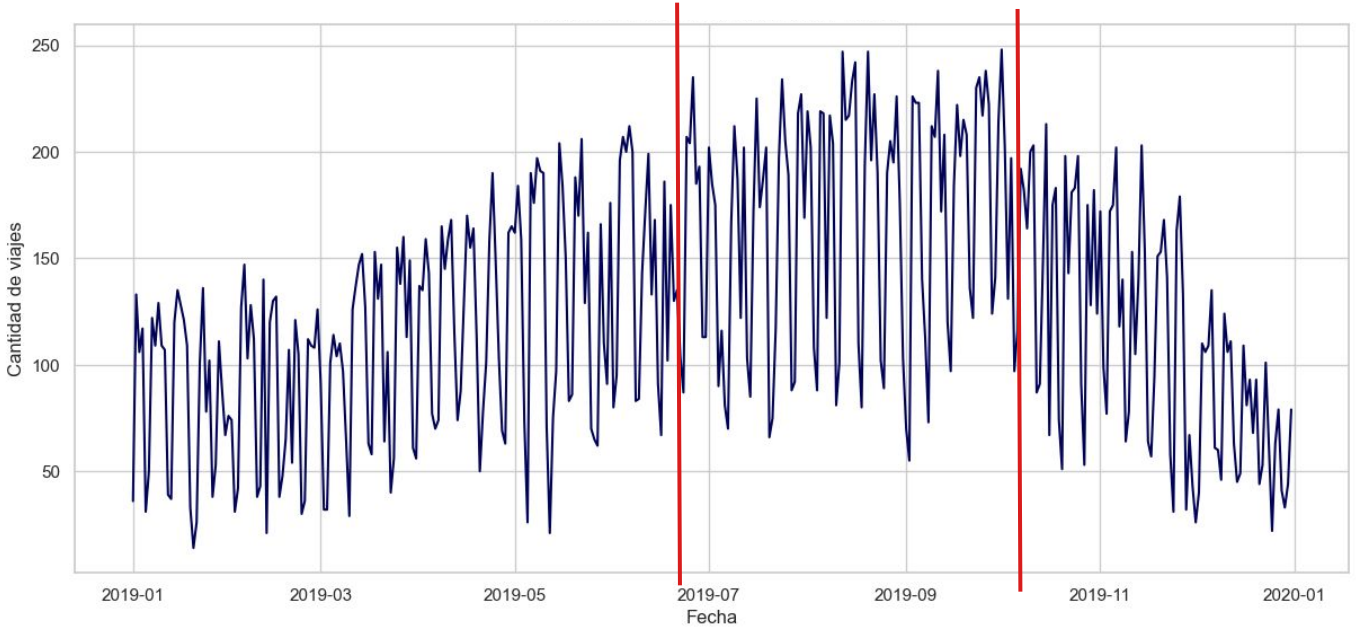
La mayoría de los usuarios tienen **35 años**

El **90%** tienen menos de 51 años.

Mayor demanda: **entre las 17 - 19 hs**

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Distribución de viajes por día

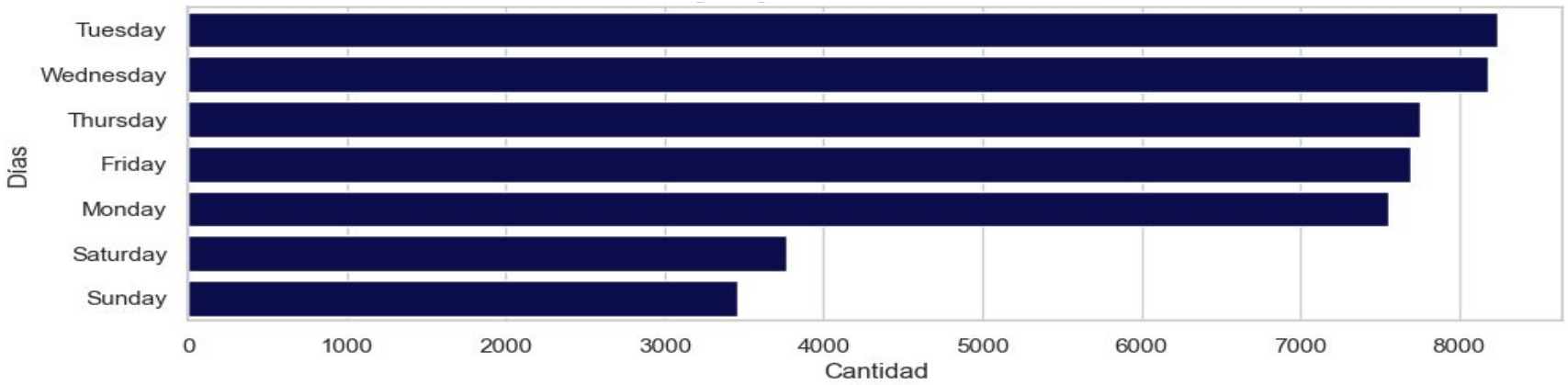


Mes	Cantidad
Agosto	5.463
Septiembre	5.282
Julio	4.833
Octubre	4.678
Junio	4.499
Mayo	4.114
Abril	3.727
Noviembre	3.585
Marzo	3.113
Enero	2.661
Febrero	2.447
Diciembre	2.256

En los meses de **verano (julio-octubre)** es donde se visualiza la **mayor cantidad de viajes**, alrededor del **45% del total anual**.

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Distribución de viajes por día de la semana



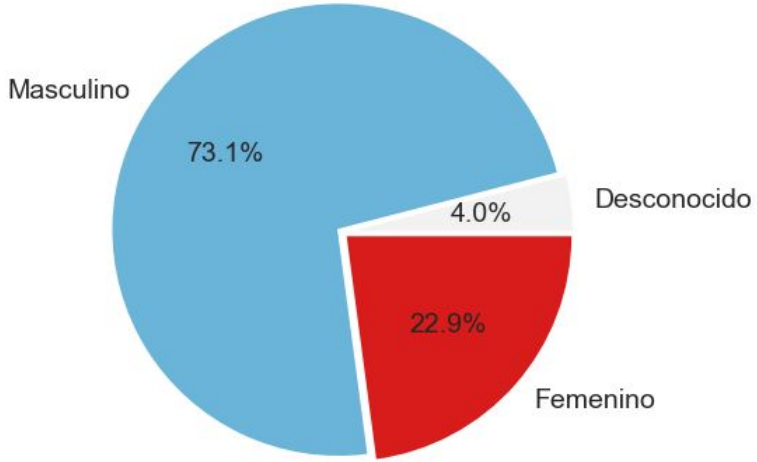
Durante los días de semana se registran la mayor cantidad de rentas (**lunes a viernes**).

Luego de analizar rango etario, horario pico y los días de mayor renta; podemos concretar que:

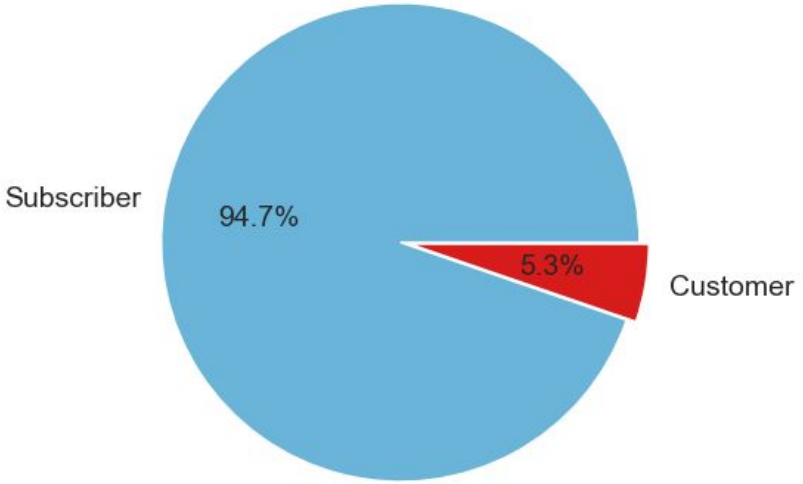
La gran mayoría de los usuarios de nuestra estación es gente en edad laboral, que alquilan bicicletas una vez finalizada la jornada de trabajo para regresar a sus hogares.

3. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Distribución de usuarios por género



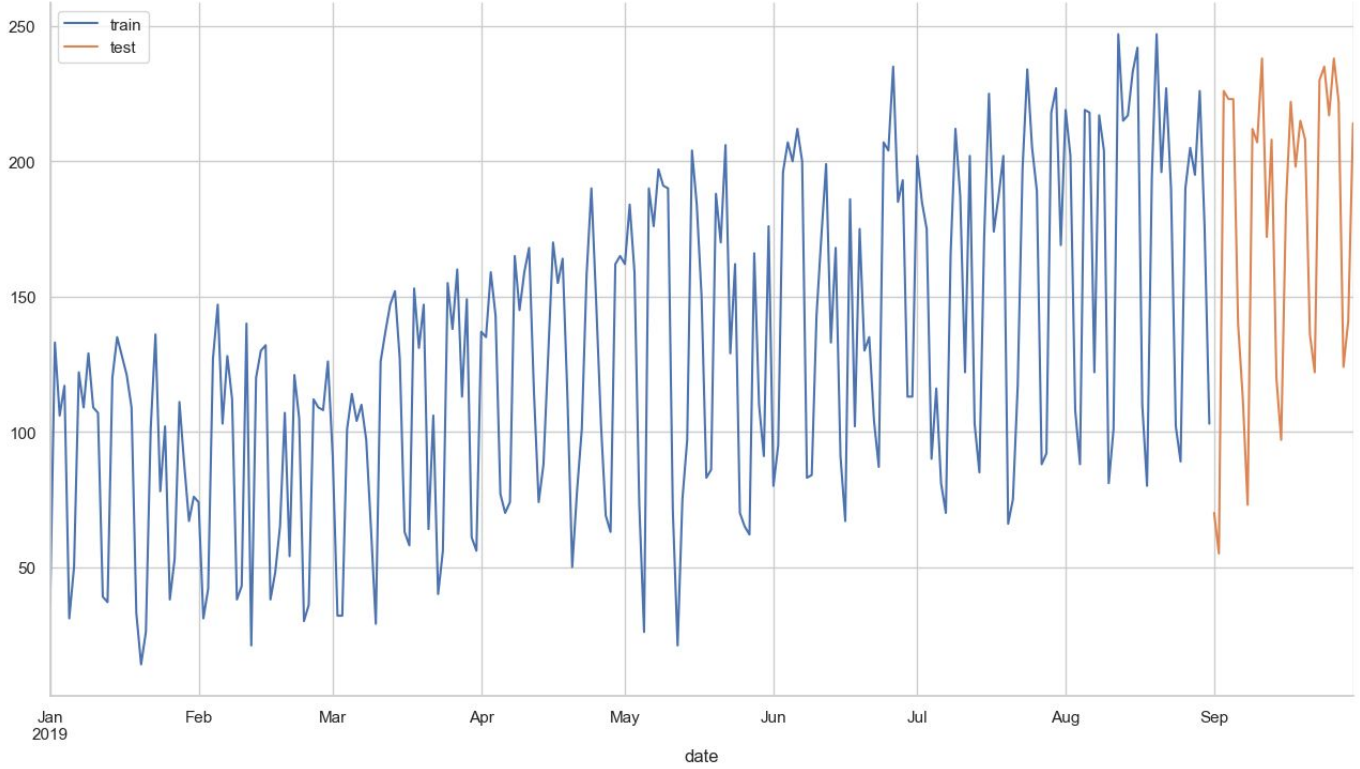
Distribución de usuarios por tipo de suscripción



También sabemos que en su mayoría son **clientes fieles** ya que alrededor del **95%** están **suscriptos** al servicio de bicicletas. Siendo principalmente **hombres (73%).**

4. ENTRENAMIENTO Y TESTEO

Para el modelado dividiremos el dataset en dos partes: **train** y **test**. Siendo test 30 días que corresponden al mes de septiembre, sobre el cual realizaremos nuestra predicción. Tomando los meses anteriores (enero-agosto) como train.



5. SELECCIÓN DEL MODELO FINAL

Se corrió el modelo **FORECAST** con dos regresores diferentes:

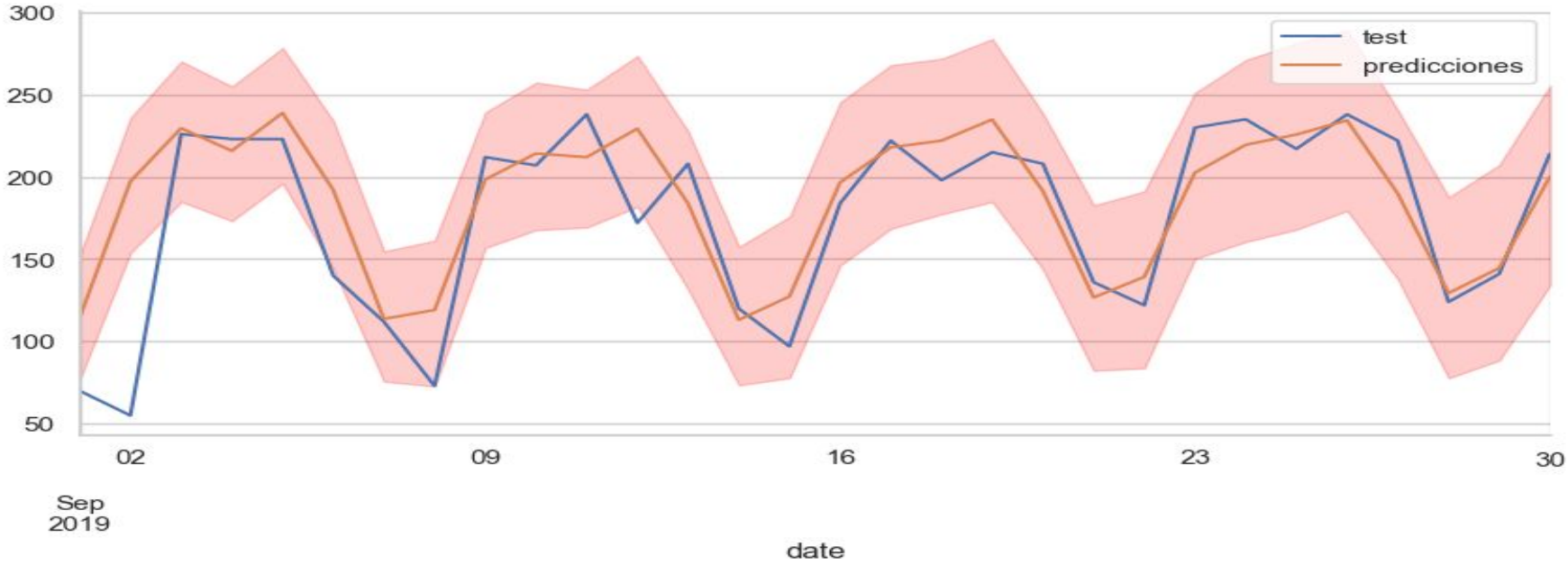
-Linear regression **-Random forest regressor**

Comparando los resultados obtenidos de ambas corridas, decidimos elegir **linear regression** como regressor para el modelo forecast en la etapa de producción; ya que si bien la eficacia es muy similar con ambos, la utilización de recursos y tiempo de ejecución es mucho menor.

	Linear regression	Random forest regressor
Error cuadrático medio	1243	1141
R^2	0.6	0.63
RdR	0.68	0.7
Tiempo búsqueda mejores hiperparametros	**1 seg**	114 seg
Tiempo entrenamiento y ejecución	**2 seg**	453 seg

6. ANÁLISIS RESULTADOS OBTENIDOS

Nuestro modelo nos arroja un valor estimado y un intervalo de alquileres para cada día del mes de septiembre. Si lo comparamos con los valores reales (test) veremos que obtuvimos una **cobertura acertada el 87%** de los días.



6. ANÁLISIS RESULTADOS OBTENIDOS

Además podemos decir que se predijo un **promedio de 186 viajes por día** en la estación **Grove St Path** durante el mes de septiembre.

Podemos medir la precisión de nuestro modelo utilizando diferentes métricas entre las cuales destacan:

- **R^2** = 0.61 Valor aceptable.

- **Error medio absoluto (MAE)** = **23** Nuestros resultados difieren en promedio unas 23 rentas por día respecto al valor real lo cual es un valor aceptable teniendo en cuenta que el promedio de viajes por día real fue de 176.

- **Error absoluto porcentual (MAPE)** = **22%**

- **RdR** = **0.682** Lo cual nos indica que nuestro modelo arroja valores un **68.2%** mejores que si utilizáramos una serie aleatoria o random walk.

7. CONCLUSIÓN

Como conclusión podemos decir que nuestro modelo (**FORECAST** con **Linear regression**) cumple con el objetivo propuesto y podría aplicarse en períodos sucesivos; para predecir de manera acertada cuantas personas alquilarán una bicicleta en nuestra estación de estudio.