Wheelie Wonka Bike Station: Predicciones de viaje

MAYO DE 2021

Fidel González, Jesús Hernández, David Rojas y Mónica Contreras

Contenido

1. Introducción
2. Objetivo: Disponibilidad de bicicletas en tiempo real
3. Patrones de consumo
   - Geográficos/Meteorológicos
   - Consumidor
4. Duración de un viaje promedio
   - Análisis de predicción
5. Número de bicicletas por estación cada 10 min

1. Introducción

Introducción

- **Wheelie Wonka Bike** es una empresa que ofrece el servicio de renta de bicicletas en la ciudad de Boston, actualmente con 145 estaciones y aproximadamente 1,164 bicicletas en su inventario

- Este servicio se ofrece a través de una suscripción anual o de uso no cotidiano

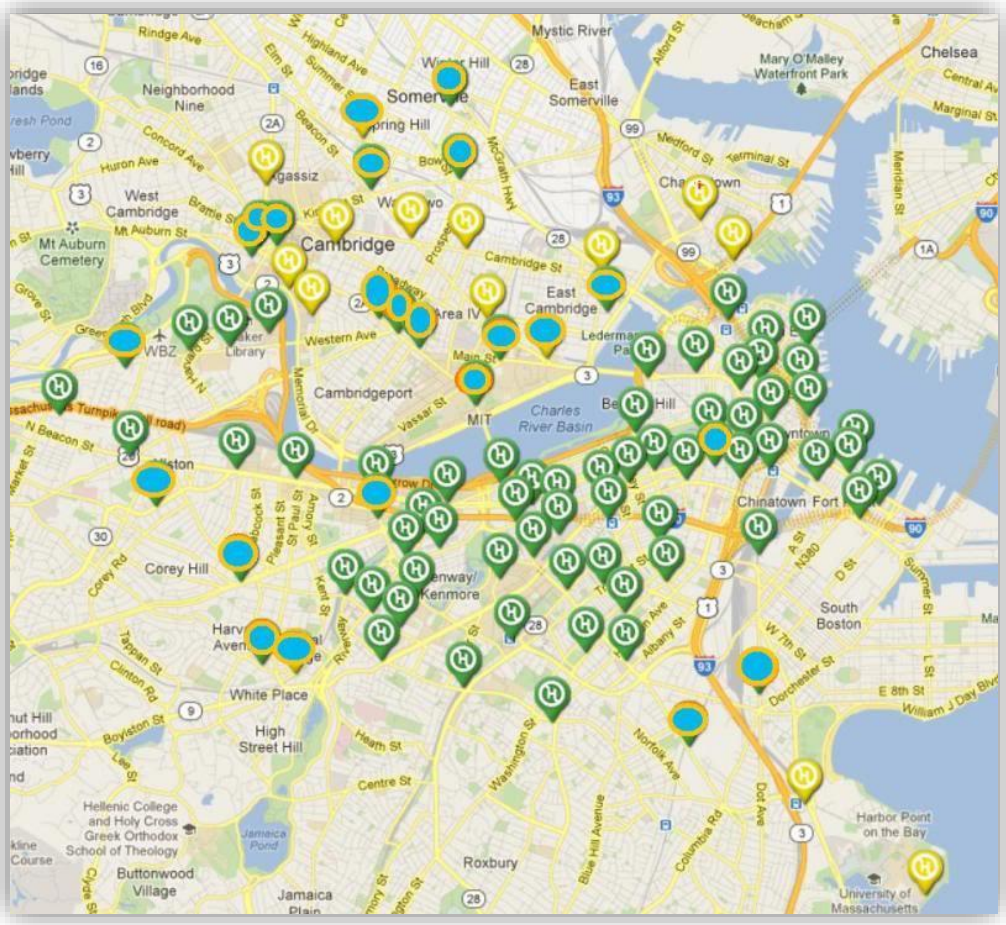
- La accesibilidad de las bicicletas y el poder dejar la bicicleta en un punto distinto al de inicio del viaje son características que los usuarios toman en cuenta al momento de contratar el servicio

2. Objetivo: Disponibilidad de bicicletas en tiempo real

Objetivo

- Con el objetivo de que aumente el número de usuarios del servicio, a través de una mejor experiencia, se pretende que los usuarios puedan identificar en tiempo real el número de bicicletas disponibles por estación

- Para ello, se realizó una análisis incorporando variables de características de cada individuo, condiciones climatológicas, de tiempo y sobre el servicio de bicicletas (ej. Suscripción, ubicación estación, duración del viaje, número de bicicleta, etc.)

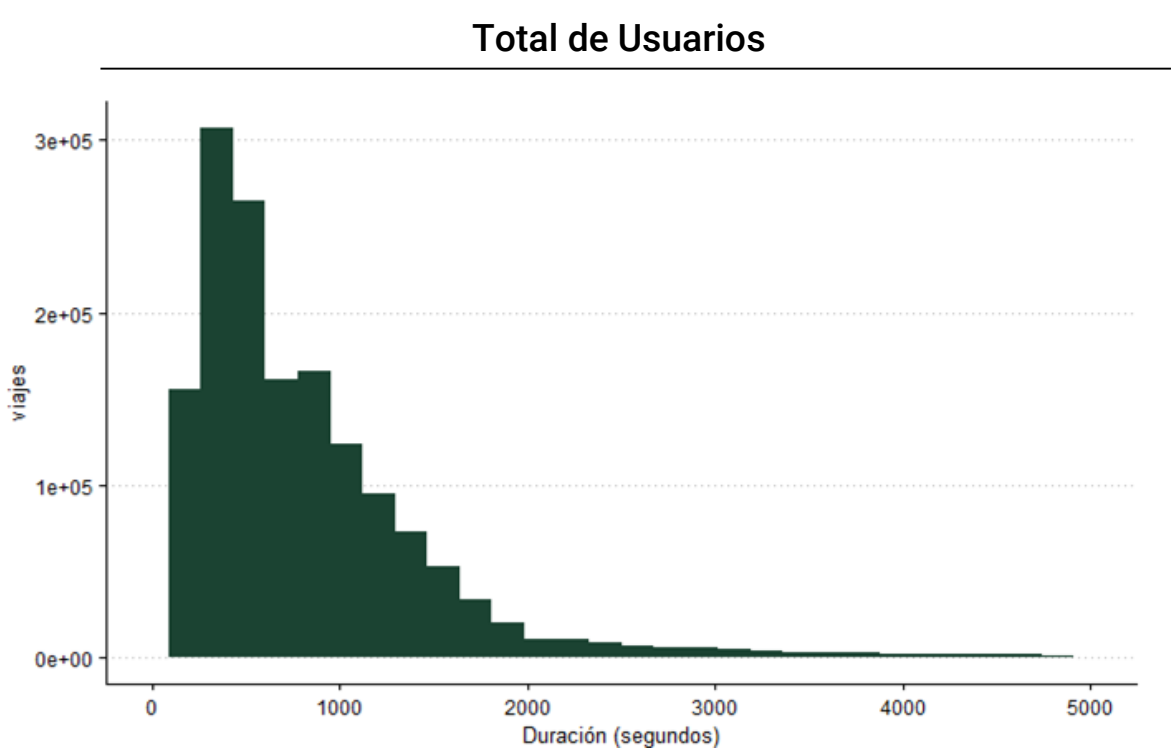


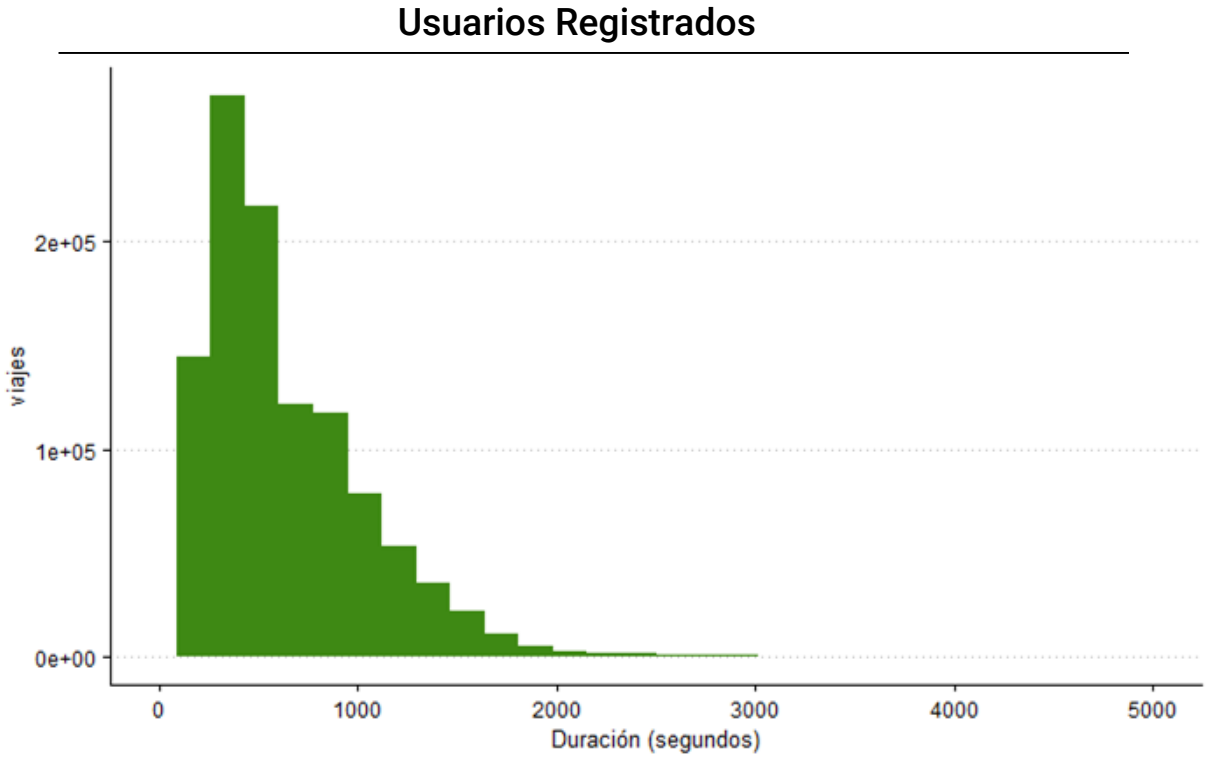
3. Patrones de consumo

Geográficos/Meteorológicos
Consumidor

Estadística descriptiva: duración de viajes

- Se observa que la distribución total está muy centrada en 18.5 minutos y la media de la duración es de 19.16 minutos por viaje. La mayoría de los viajes son de menos de 30 minutos

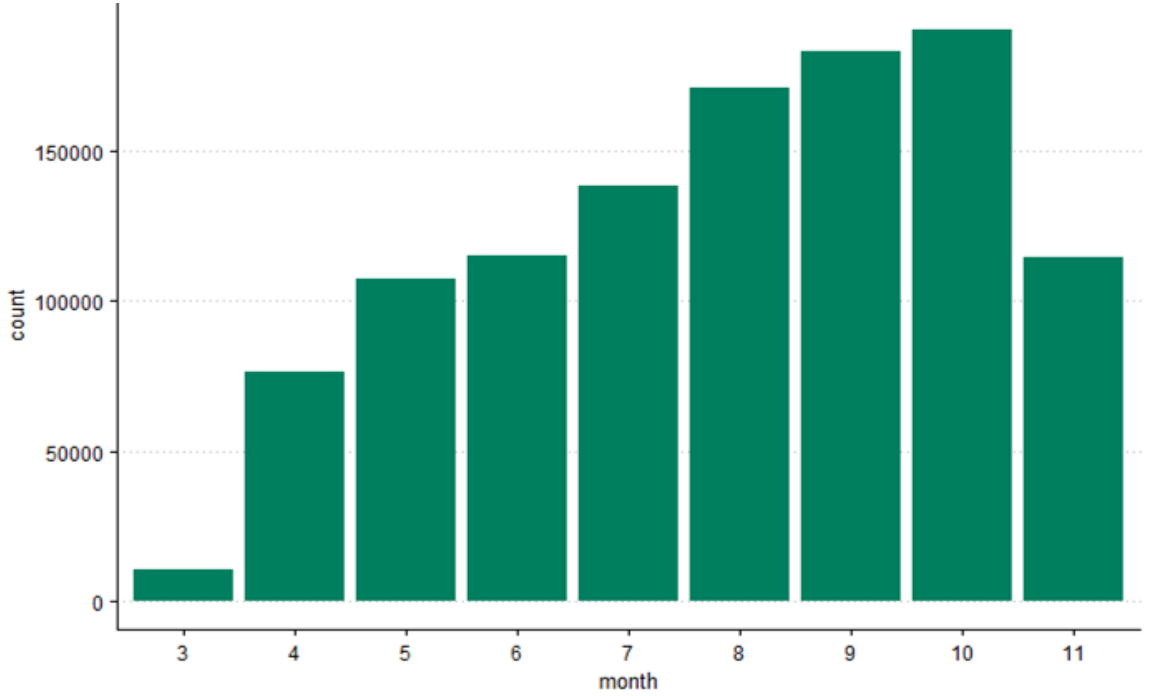




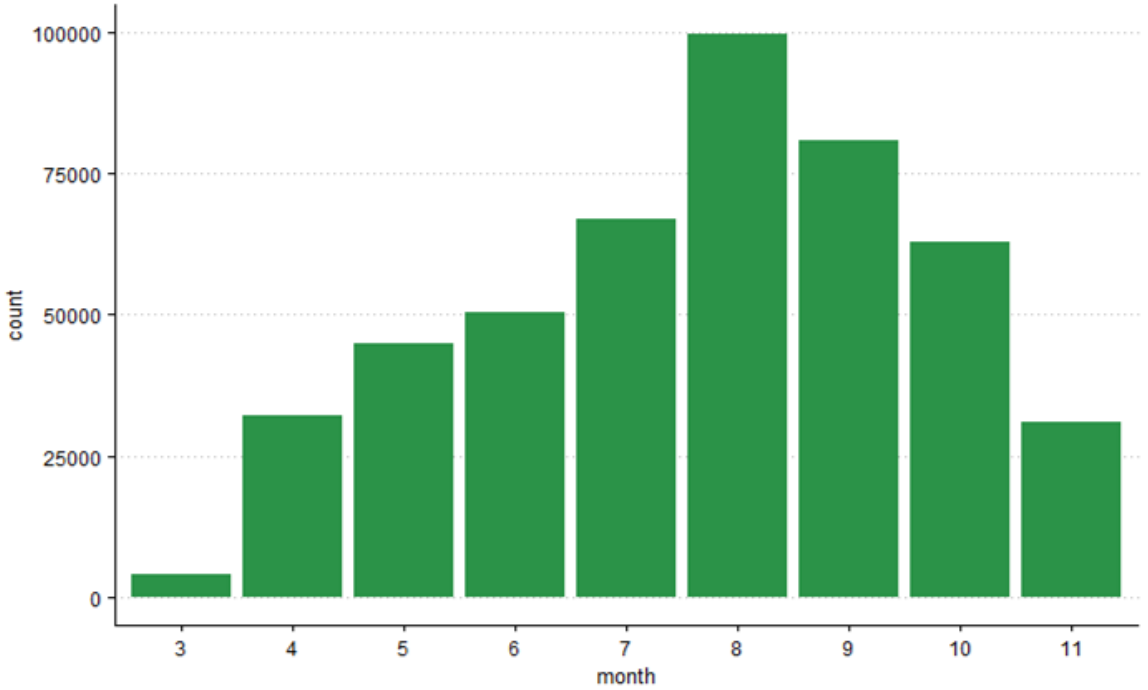
Tipo de Usuario

- Podemos observar que los usuarios utilizan en mayor medida el servicio de bicicletas en el 3er trimestre de cada por posibles temas de clima y aquellos que se encuentran registrados su variación de uso entre meses es menor a diferencia de los casuales

Usuario Registrado



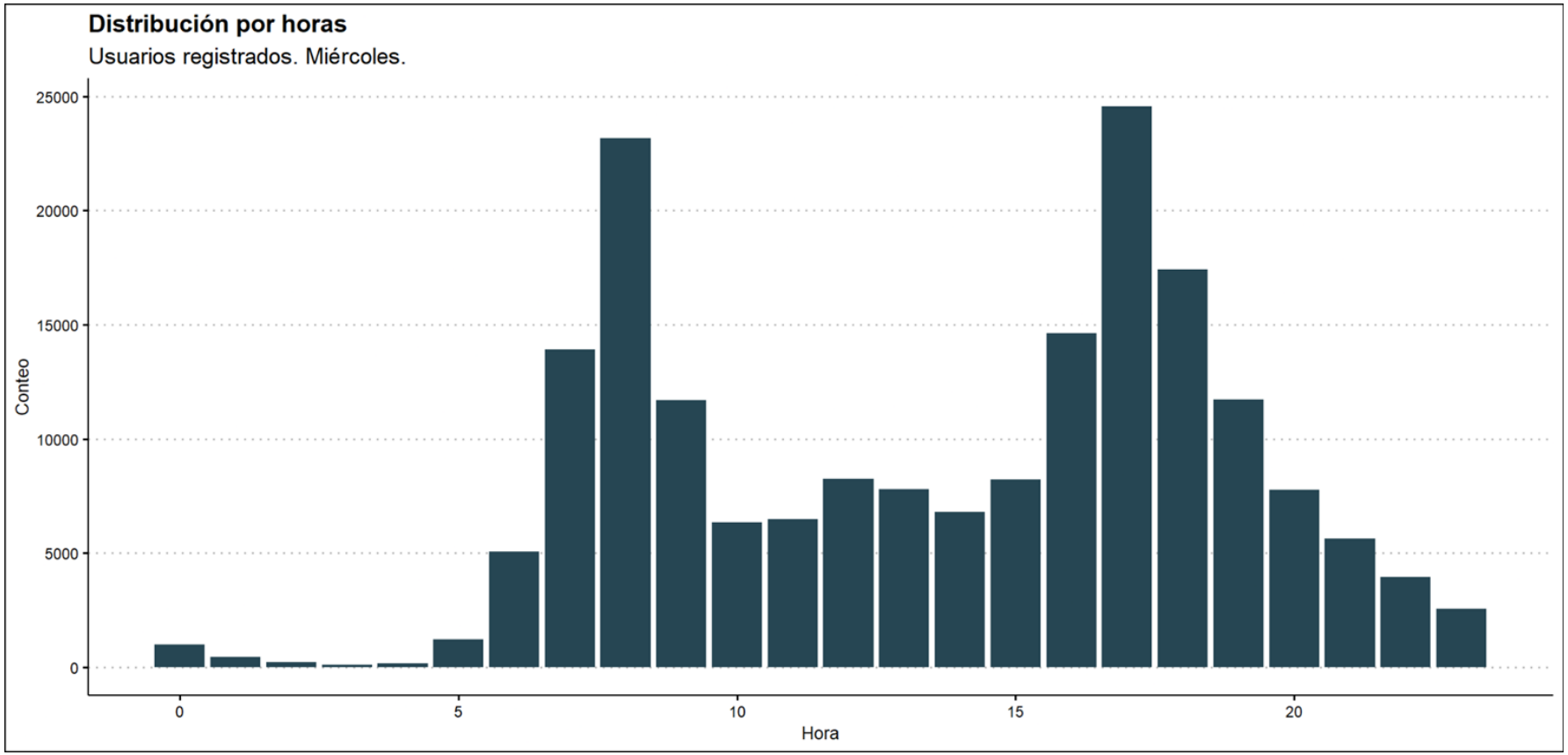
Usuario Casual



Usuarios registrados

- Los usuarios registrados tienen picos de uso en horarios de entrada y salida de oficinas.

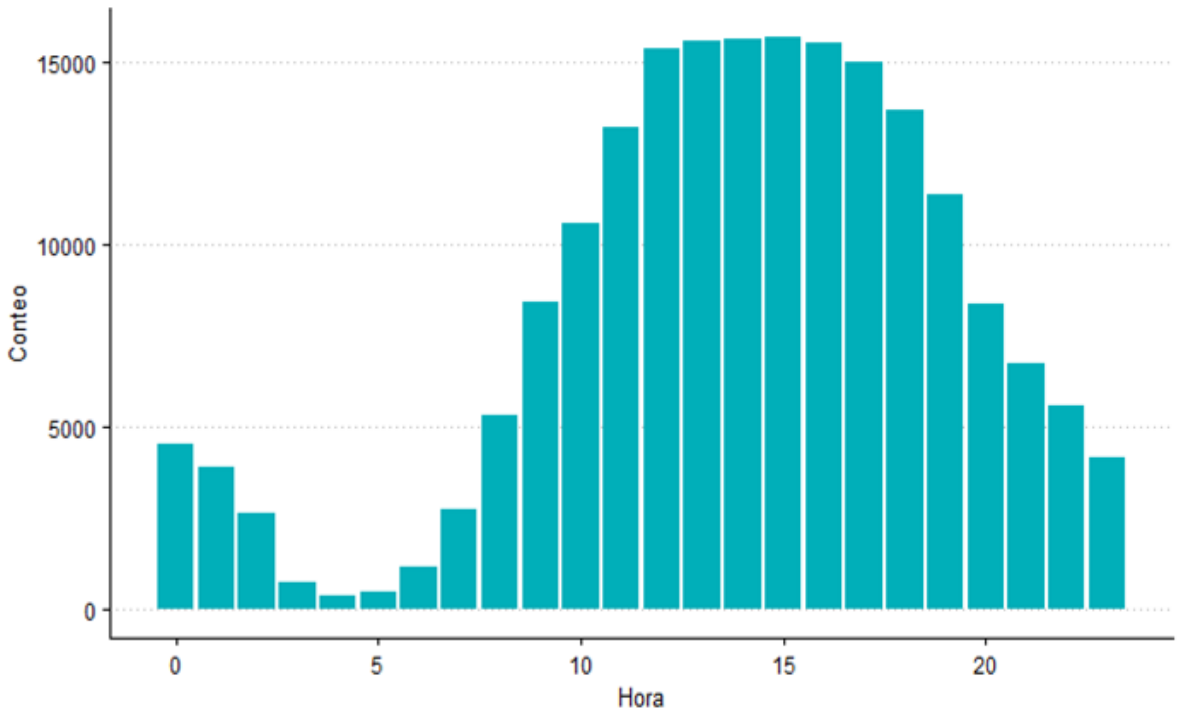
Usuario Registrado



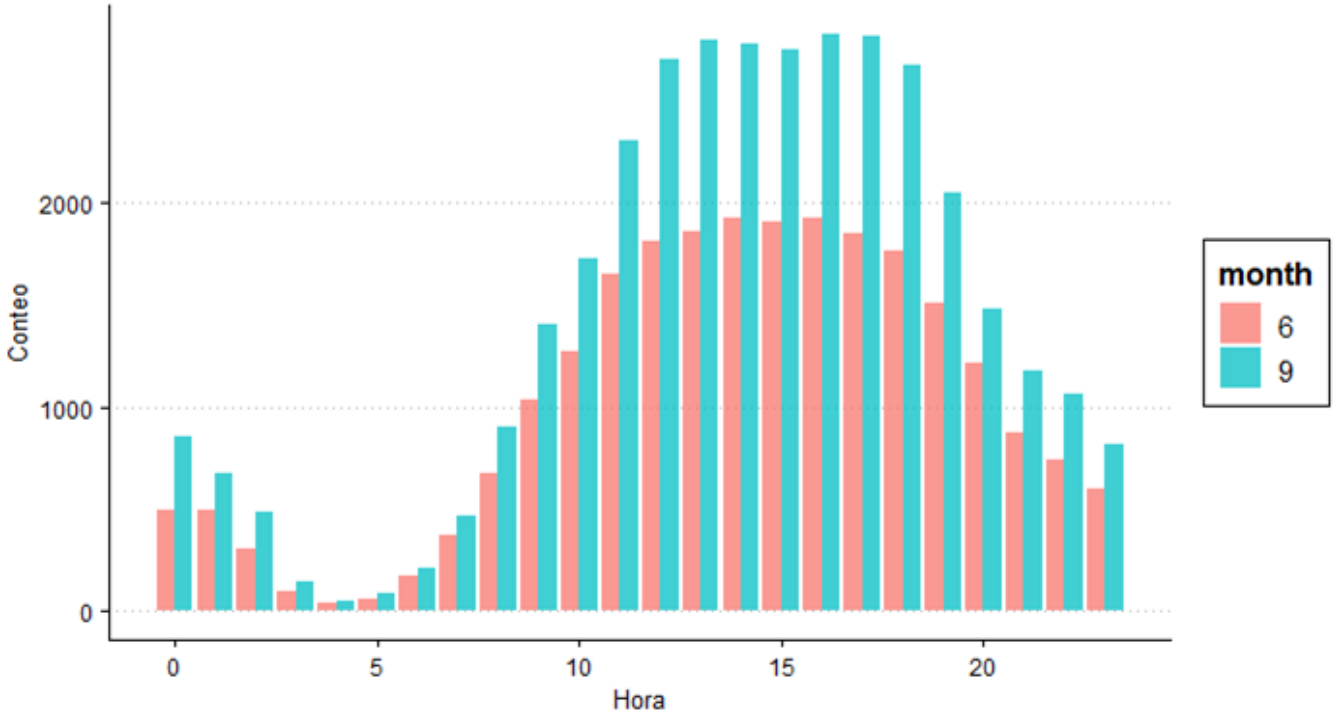
Estacionalidad de horario de viajes: usuarios registrados

- Se puede observar que los usuarios registrado utilizan menos las bicicletas en sábados y domingos que los usuarios casuales
- De igual forma se puede observar cierta estacionalidad en el mes de septiembre manteniendo el mismo horario de mayor uso pero aumentando significativamente

Distribución por horas (sábados y domingos)



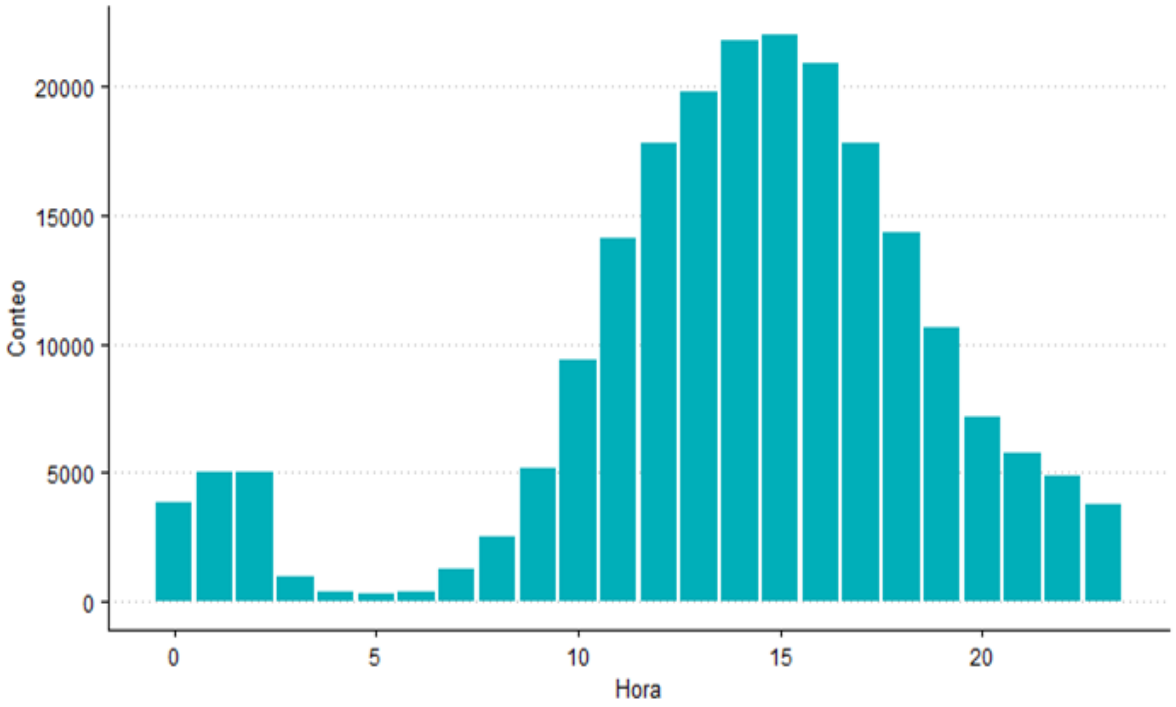
Distribución por horas (domingos de sep-jul)



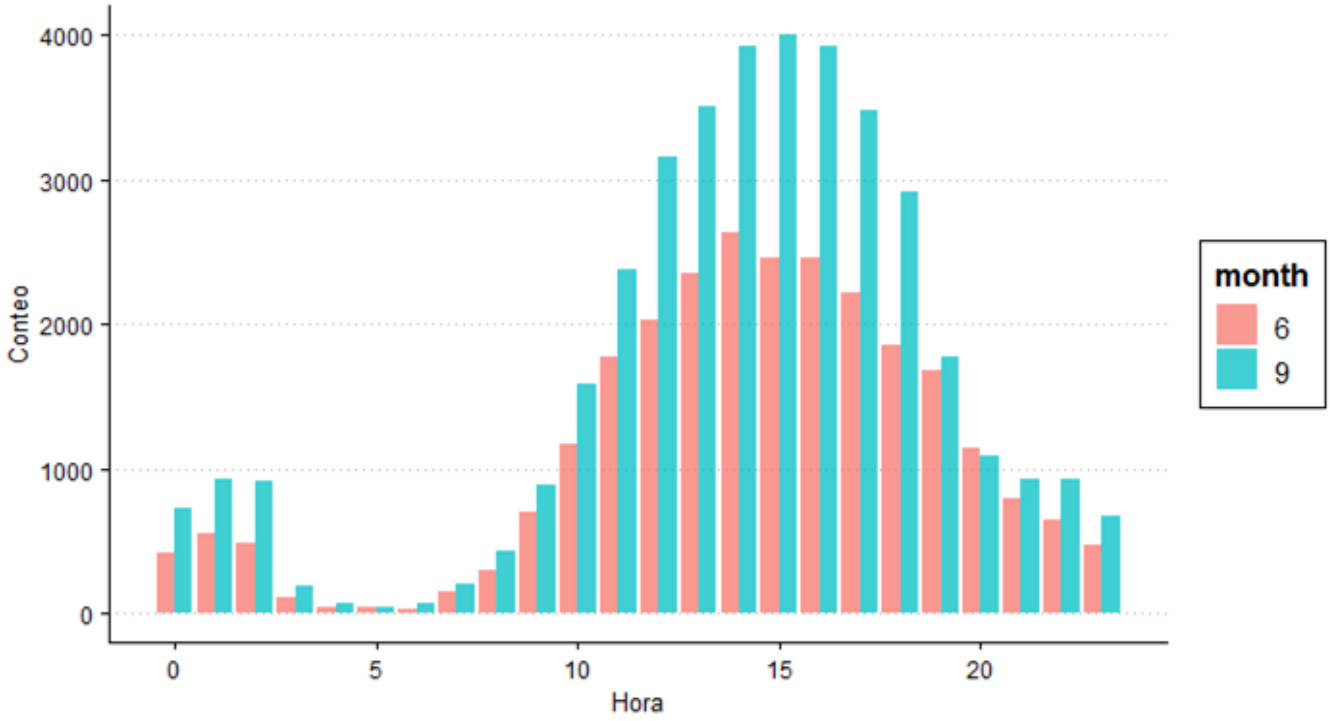
Estacionalidad de horario de viajes: usuarios casuales

- Se puede observar que los usuarios casuales utilizan más las bicicletas en sábados y domingos que los usuarios registrados
- Adicional se puede observar cierta estacionalidad en el mes de septiembre manteniendo el mismo horario de mayor uso pero aumentando significativamente

Distribución por horas (sábados y domingos)



Distribución por horas (domingos de sep-jul)



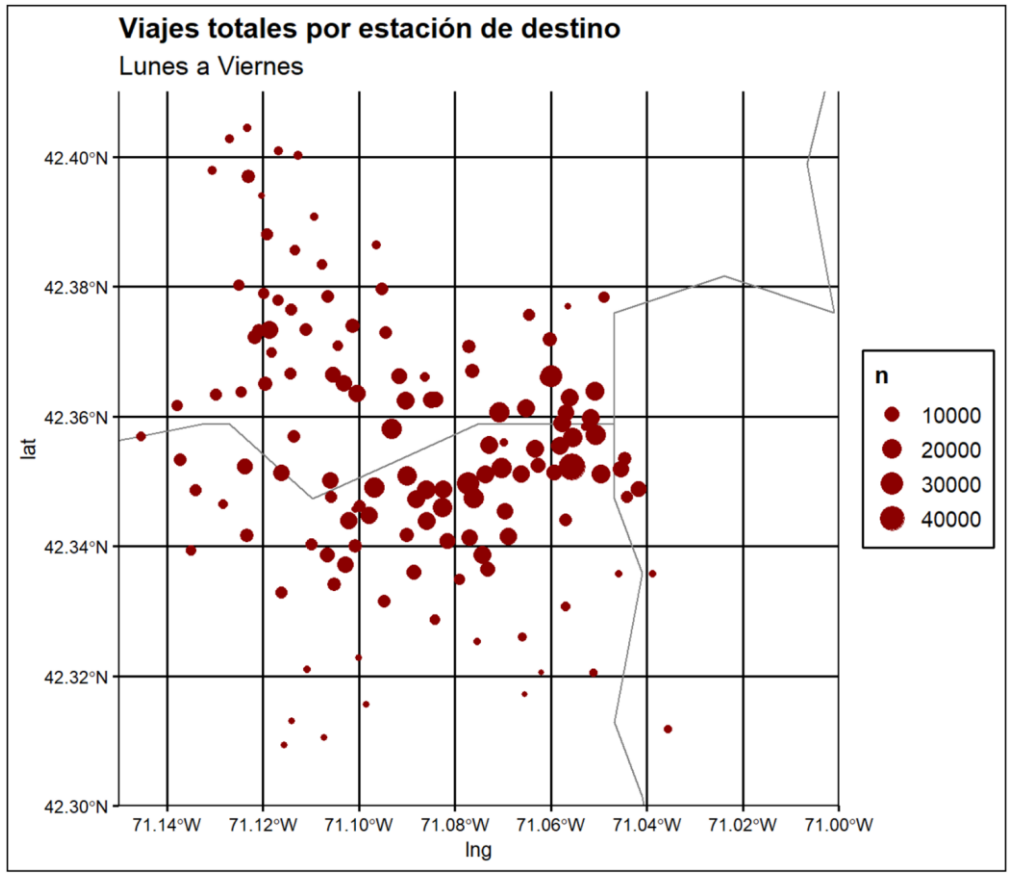
Análisis geográfico: rutas

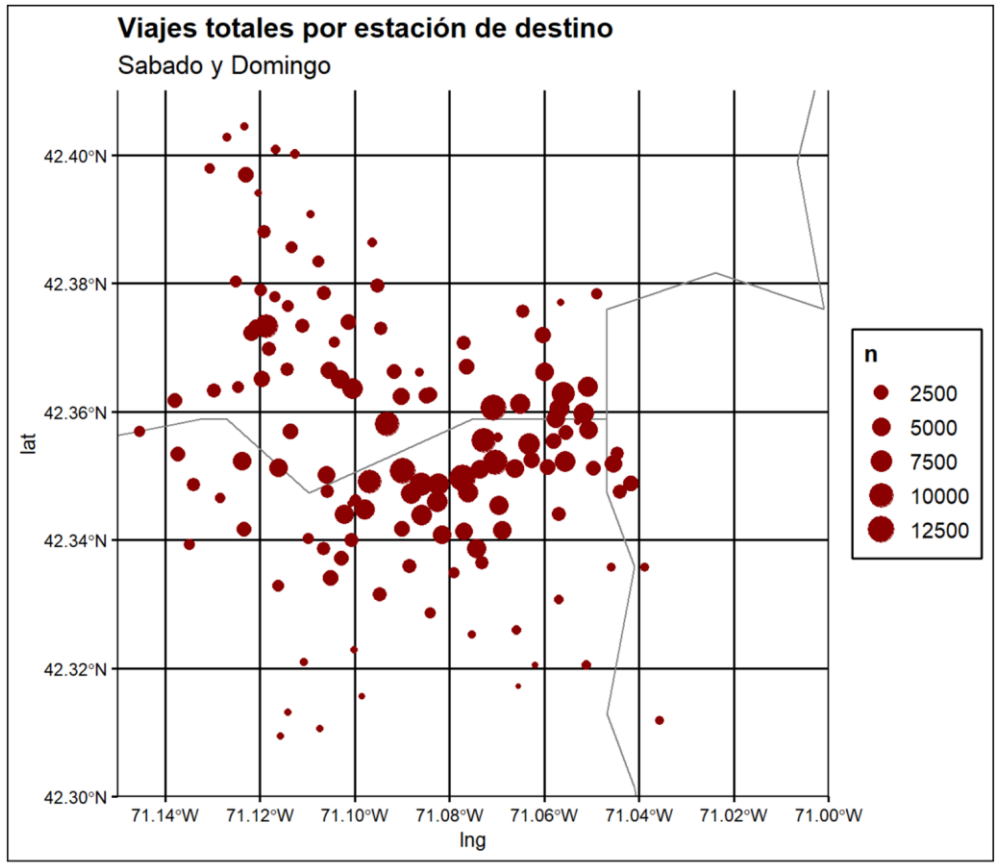
- En el 96% de los viajes, los usuarios toman la bicicleta en una estación y la dejan en otra estación. Casi el 70% de los viajes son dentro del municipio de Boston.

Municipios	Viajes	Proporción
Boston_Boston	1076844	0.69
Cambridge_Cambridge	159949	0.10
Boston_Cambridge	111089	0.07
Cambridge_Boston	110185	0.07
Cambridge_Somerville	21016	0.01

Estaciones más utilizadas entre semana y fin de semana

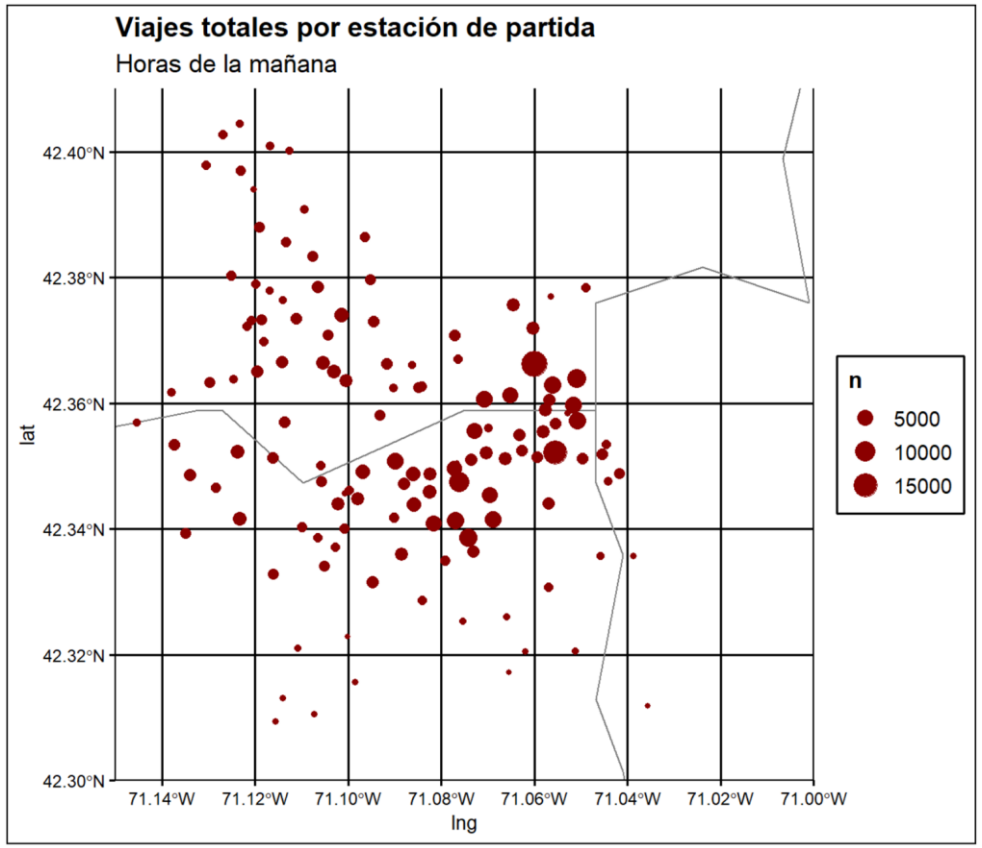
- **Entre semana los viajes se concentran en el centro de la ciudad, mientras que en el fin de semana hay una distribución un poco más uniforme.**

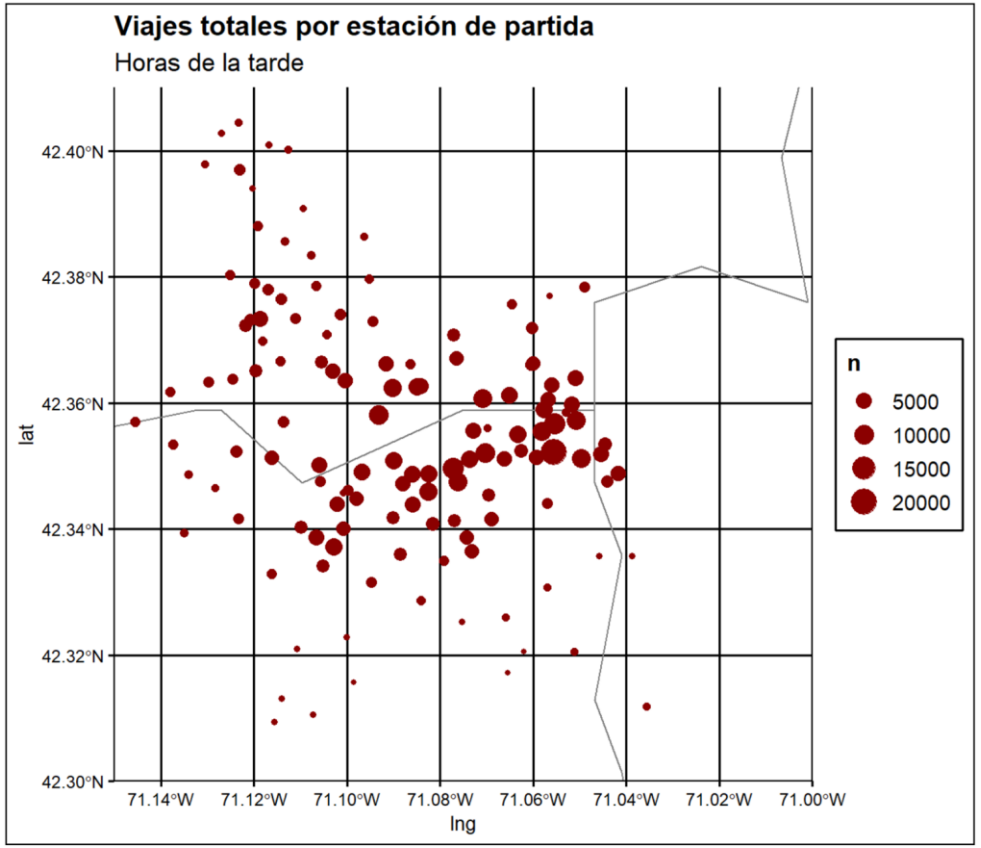




Estaciones más utilizadas en la mañana y en la tarde

- Procedemos a analizar la base de datos con el objetivo de sacar perspectivas en cuanto a la distribución geográfica de las partidas y llegadas en horas pico.

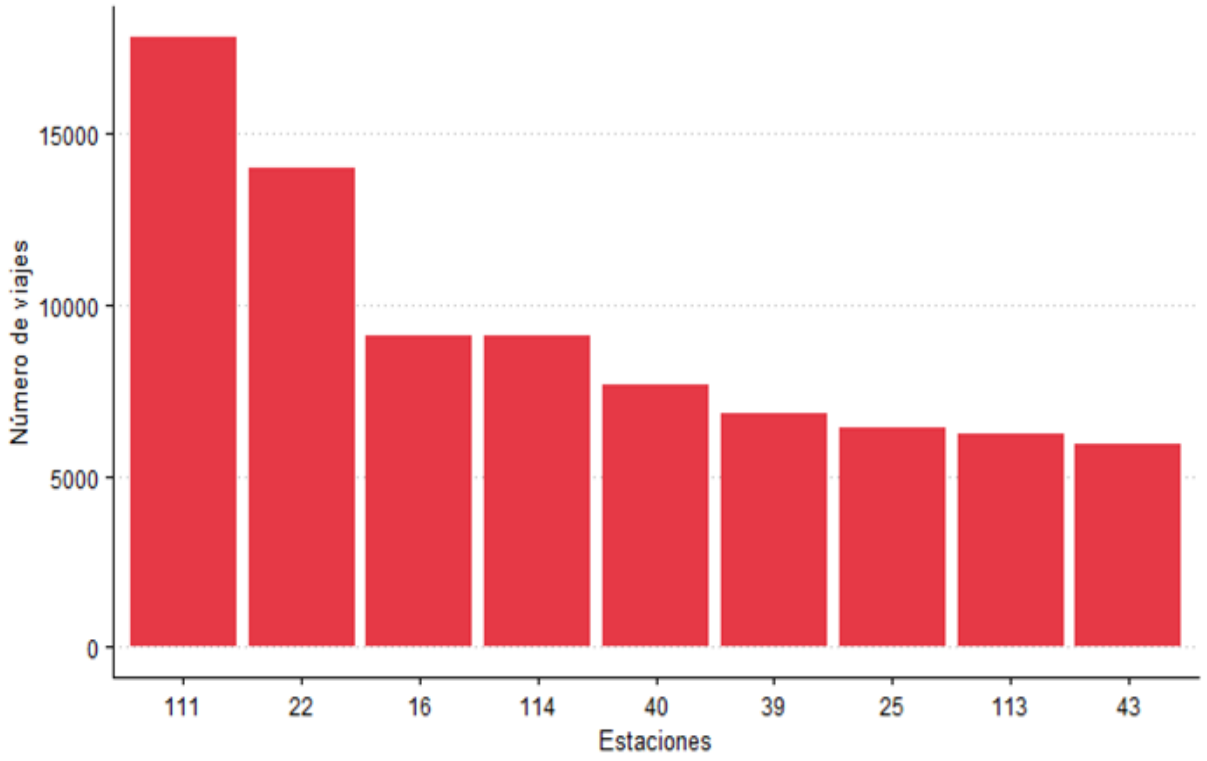




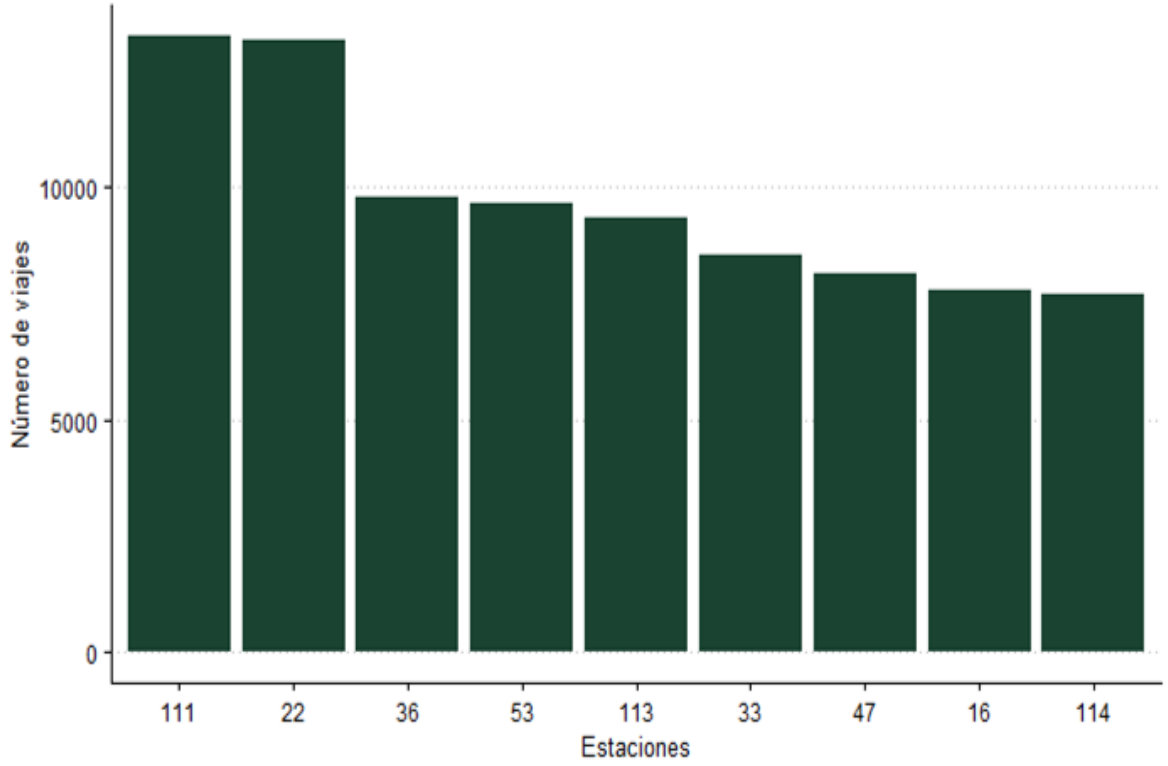
Estaciones más utilizadas para iniciar el viaje

- El principal resultado de estos gráficos es que las estaciones (111,22) que son las que más se utilizan en las mañanas, de igual son las que más se utilizan en las tardes para terminar sus viajes

Viajes por estación de inicio (mañana)

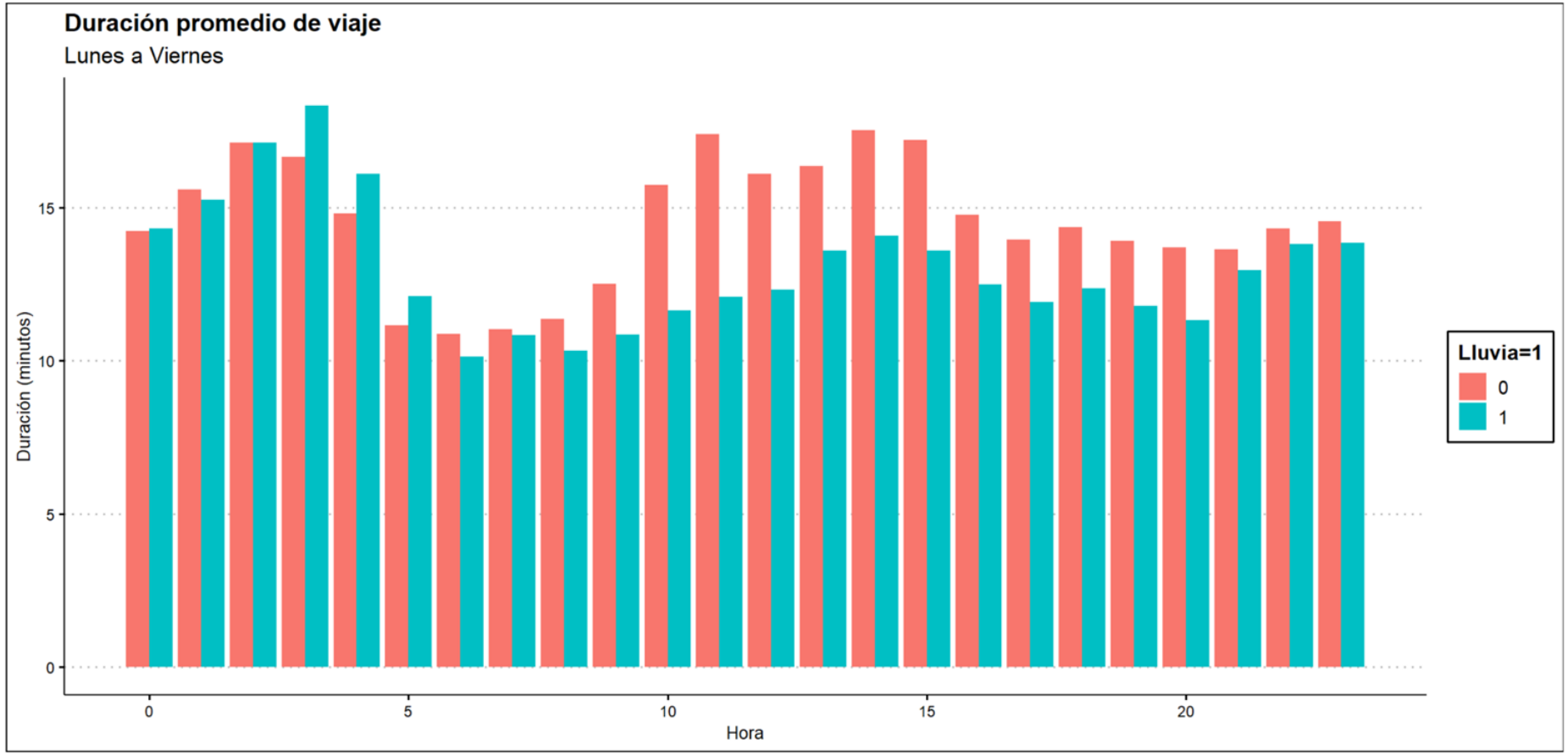


Viajes por estación de fin (tarde)



Duración de viajes con lluvia

- Es de esperarse que cuando llueve la gente use menos las bicicletas. En promedio, los viajes sin lluvia duran alrededor de 3 minutos más en comparación.



4. Duración de los viajes

Análisis de predicción

Elección del Modelo

- Derivado de los análisis realizados sobre las variables que tenemos disponibles, se proponen 3 modelos con las siguientes características para predecir la duración de los viajes:

	Árbol	RF	FDR
Variables explicativas	• Dummy de género • Si es un usuario registrado • Mes (uso de la bicicleta) • Día de la semana • Hora (tiempo) • Dummy que indica si estaba lloviendo o chispeando	• Número de bicicleta • Año de nacimiento del usuario • Mes (uso de la bicicleta) • Día de la semana • Hora • Dummy de género • Si es un usuario registrado • Dummy que indica si está lloviendo • Intensidad de lluvia	• Dummy de género • Si es un usuario registrado • Mes (uso de la bicicleta) • Dummy de día de la semana • Dummy de hora • Dummy de la estación donde inició el viaje • Dummy que indica si estaba lloviendo • Intensidad de lluvia
Variable dependiente	Log de la duración del viaje	Duración del viaje	Log de la duración del viaje

Elección del Modelo: ¿Por qué se eligieron esos modelos?

1. **Árbol de Regresión:** Es un método que te permite mapear características de Xi a Yi, en nuestro caso se utiliza un árbol de regresión porque nuestra variable dependiente es continua y la predicción es una media de yi, además de que nos permite ***detectar no-linealidades*** en el modelo. Para evitar caer en un sobreajuste en nuestro caso, se elige estimar el árbol hasta que la disminución de cada rama llegue aun corte de disminución en el deviance mínimo (ej.- min dev = 0.05)

1. **False Discovery Rate:** Este método se utiliza cuando existen muchas columnas, como en nuestro caso, y se corre una regresión donde posiblemente no todas las variables resultan ser significativas. Te ayuda a que tu modelo no tenga demasiados falsos positivos y a elegir qué variables cumplen con esta prueba (ej.- tasa de falsos positivos de 10%)

1. **Random Forest:** Una ventaja de este método comparado contra los árboles es que te ayuda a resolver que lleguen a estar muy correlacionados a pesar de hacer remuestreo. Se elige un parámetro B (número de árboles) a partir del cual ya no ganas mucho mayor predictivo para mejorar las predicciones

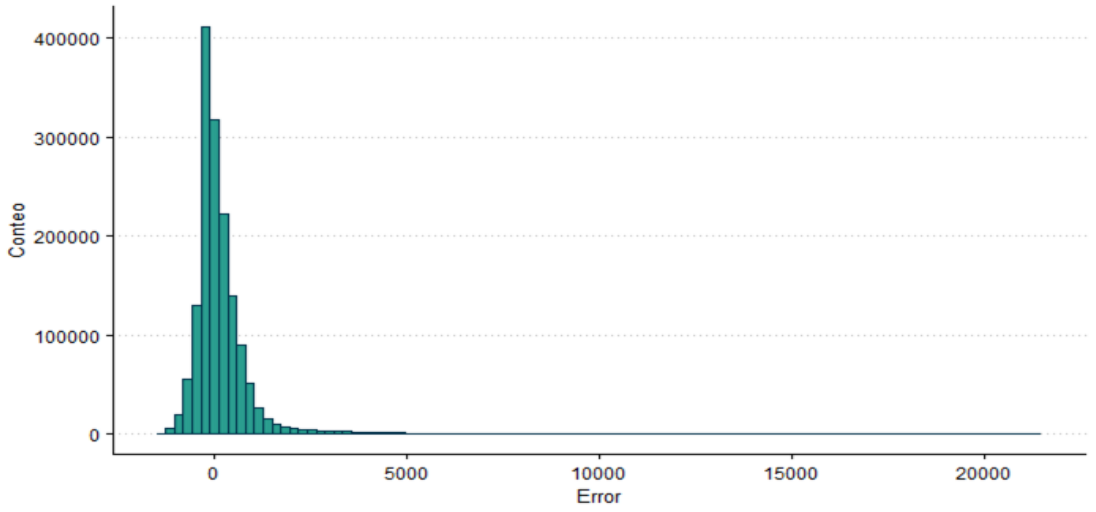
Elección del Modelo: Modelo campeón

- Con base a los resultados de los distintos modelos para predecir el tiempo de duración de los viajes, podemos observar que aquel que tuvo el mejor desempeño basado en la estimación del error medio fue al que se le aplicó el “False Discovery Rate” que obtuvo el valor más bajo de error, seguido del método de Árboles de Regresión y por último el “Random Forest”

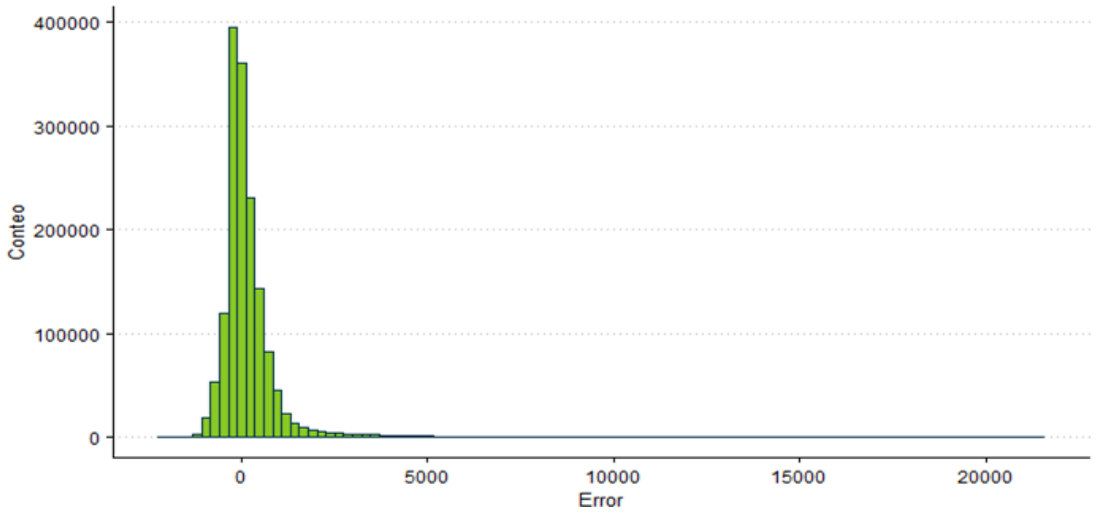
Tabla 1. Estimación de error medio

Entrenamiento	Error medio Árbol 1	Error medio Árbol 2	Error medio FDR	Error medio RF
0	508.64	509.62	502.09	836.77
1	509.61	510.43	502.9	837.04

Distribución de errores de predicción (Árbol con dummy por hora)



Distribución de errores de predicción (FDR)



5. Número de bicicletas por estación cada 10min

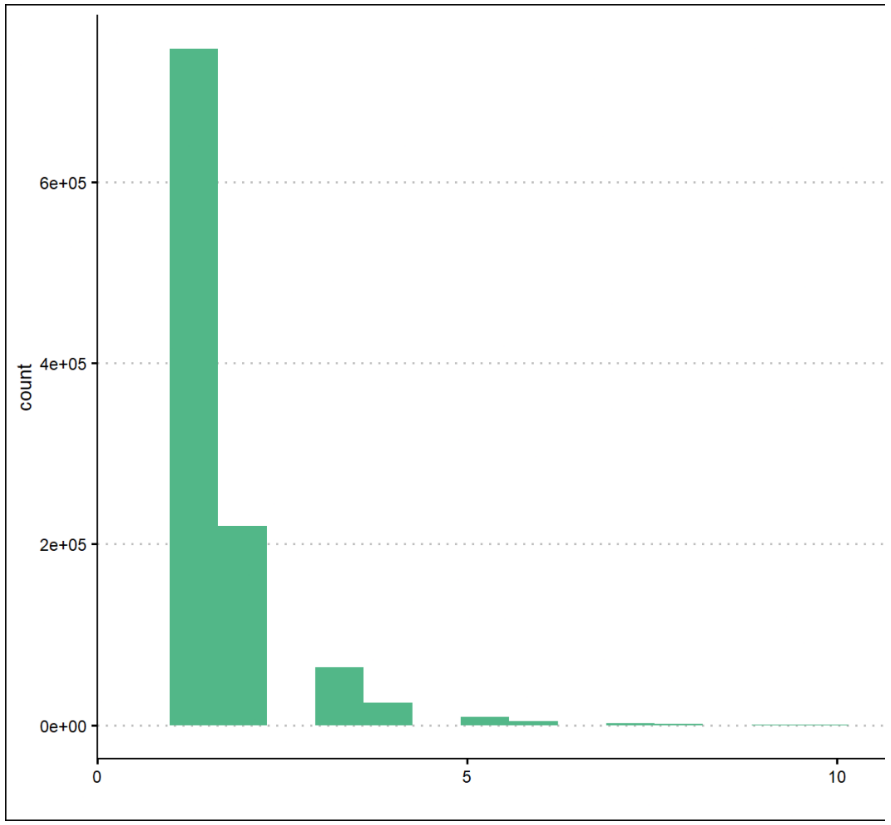
Elección del Modelo

- Se estima un modelo con las siguientes características:

	OLS
Variables explicativas	• Categórica Id Estación • Día de la semana • Intervalo de tiempo (se divide las 24 horas del día en intervalos de 10 mins) • Mes • Dummy que indica si está lloviendo • Intensidad de lluvia
Variable dependiente	Número de llegadas a la estación

Resultados de la estimación

- Se observa que al hacer un conteo del número de bicicletas que llega por estación para la ventana de tiempo observada, la mayoría de los casos es una o dos bicis.



- Si calculamos el promedio del error de predicción de nuestro modelo tenemos lo siguiente.

	Media	Mediana
In sample	0.603	0.4591
Out of sample	0.601	0.4596

Wheelie Wonka Bikes