# DS Test

**Josué de Jesús Juárez Vidales**

# DS Test.
# Descripción.

- En Los Ángeles existe un **sistema compartido de bicicletas** que brinda datos anónimos acerca del uso del servicio que ofrecen.
- La tabla que se proporciona contiene el histórico de viajes que se han realizado durante cerca de 9 meses.
- **Se desea saber si es posible inferir si el tipo de pase es “Monthly Pass”**

# DS Test. **Descripción.**

- **Features**
- trip_id: identificador único para el viaje
- duration: duración del viaje en minutos
- start_time: dia/hora donde en viaje inicia en formato ISO 8601 tiempo local
- end_time: dia/hora donde el viaje termina en formato ISO 8601 tiempo local
- start_station: la estación donde el viaje inició
- start_lat: la latitud de la estación donde el viaje se originó
- start_lon: la longitud de la estación donde el viaje se originó
- end_station: la estación donde el viaje terminó
- end_lat: la latitud de la estación donde terminó el viaje
- end_lon: la longitud de la estación donde terminó el viaje
- bike_id: un entero único que identifica la bicicleta
- plan_duration: número de dias que el usuario tendrá el paso. 0 significa un viaje único (Walk-up plan)
- trip*route*category: “Round trip” son viajes que empiezan y terminan en la misma estación
- **Target**
- passholder_type: El nombre del plan de passholder

# DS Test. **Análisis exploratorio de datos**

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | trip_id | duration | start_time | end_time | start_lat | start_lon | end_lat | end_lon | bike_id | plan_duration | trip_route_category | passholder_type | start_station | end_station |
| 2 | 101750280 | 35 | 8/7/2018 11:20 | 8/7/2018 11:55 | 33.74892 | -118.275192 | 33.74892 | -118.275192 | 6530 | 1 | Round Trip | Walk-up | 4127 | 4127 |
| 3 | 46560345 | 32 | 9/17/2017 17:51 | 9/17/2017 18:23 | 34.035679 | -118.270813 | 34.047749 | -118.243172 | 6683 | 0 | One Way | Walk-up | 3057 | 3062 |
| 4 | 120016336 | 6 | 4/22/2019 9:22 | 4/22/2019 9:28 | 34.04607 | -118.233093 | 34.047749 | -118.243172 | 6710 | 30 | One Way | Monthly Pass | 3022 | 3062 |
| 5 | 129547190 | 138 | 9/22/2019 11:27 | 9/22/2019 13:45 | 34.06258 | -118.290092 | 34.059689 | -118.294662 | 17068 | 1 | One Way | One Day Pass | 4304 | 4311 |
| 6 | 136619463 | 14 | 1/31/2020 17:11 | 1/31/2020 17:25 | 34.026291 | -118.277687 | 34.02166 | -118.278687 | 18841 | 30 | One Way | Monthly Pass | 4266 | 4443 |
| 7 | 63406498 | 30 | 12/16/2017 15:18 | 12/16/2017 15:48 | 34.13525 | -118.13237 | 34.13525 | -118.13237 | 5768 | 0 | Round Trip | Walk-up | 4158 | 4158 |
| 8 | 25033469 | 11 | 4/15/2017 22:02 | 4/15/2017 22:13 | 34.045181 | -118.250237 | 34.05357 | -118.266357 | 6527 | 30 | One Way | Monthly Pass | 3067 | 3040 |
| 9 | 107479459 | 15 | 10/16/2018 17:27 | 10/16/2018 17:42 | 34.04113 | -118.267982 | 34.045422 | -118.253517 | 6333 | 1 | One Way | Walk-up | 3011 | 3051 |
| 10 | 132750788 | 19 | 11/16/2019 11:24 | 11/16/2019 11:43 | 34.046822 | -118.248352 | 34.046822 | -118.248352 | 19855 | 30 | Round Trip | Monthly Pass | 3038 | 3038 |
| 11 | 107465757 | 8 | 10/16/2018 12:05 | 10/16/2018 12:13 | 34.052872 | -118.24749 | 34.04607 | -118.233093 | 5926 | 1 | One Way | Walk-up | 3046 | 3022 |
| 12 | 43527123 | 24 | 8/31/2017 23:22 | 8/31/2017 23:46 | 34.0485 | -118.258537 | 34.0485 | -118.258537 | 6482 | 0 | Round Trip | Walk-up | 3005 | 3005 |
| 13 | 59951416 | 27 | 11/25/2017 15:46 | 11/25/2017 16:13 | 34.049301 | -118.2388 | 34.049198 | -118.25283 | 6066 | 1 | One Way | One Day Pass | 3042 | 3063 |
| 14 | 136555586 | 73 | 1/30/2020 11:55 | 1/30/2020 13:08 | 34.06258 | -118.290092 | 34.049301 | -118.238808 | 12244 | 1 | One Way | Walk-up | 4304 | 3042 |
| 15 | 137129003 | 24 | 2/9/2020 11:38 | 2/9/2020 12:02 | 34.029121 | -118.403168 | 34.029121 | -118.403168 | 15688 | 1 | Round Trip | Walk-up | 4329 | 4329 |
| 16 | 116292309 | 8 | 2/26/2019 7:54 | 2/26/2019 8:02 | 34.05661 | -118.237213 | 34.047749 | -118.243172 | 6188 | 30 | One Way | Monthly Pass | 3014 | 3062 |
| 17 | 53930793 | 6 | 10/24/2017 18:25 | 10/24/2017 18:31 | 34.049889 | -118.25588 | 34.042061 | -118.26338 | 6728 | 0 | One Way | Walk-up | 3032 | 3034 |
| 18 | 28587664 | 5 | 5/23/2017 10:16 | 5/23/2017 10:21 | 34.056969 | -118.253593 | 34.05088 | -118.248253 | 6685 | 30 | One Way | Monthly Pass | 3049 | 3069 |
| 19 | 76916100 | 13 | 3/15/2018 9:14 | 3/15/2018 9:27 | 34.058319 | -118.246094 | 34.039871 | -118.250038 | 12128 | 30 | One Way | Monthly Pass | 3028 | 3077 |
| 20 | 79014486 | 25 | 3/28/2018 11:09 | 3/28/2018 11:34 | 34.063179 | -118.24588 | 34.04681 | -118.256981 | 6526 | 0 | One Way | Walk-up | 3026 | 3064 |
| 21 | 108346057 | 12 | 10/29/2018 16:58 | 10/29/2018 17:10 | 34.031052 | -118.26709 | 34.044701 | -118.252441 | 6441 | 30 | One Way | Monthly Pass | 3020 | 3031 |
| 22 | 165364244 | 7 | 7/2/2021 10:28 | 7/2/2021 10:35 | 33.97366 | -118.422859 | 33.980991 | -118.414879 | 18833 | 30 | One Way | Monthly Pass | 4255 | 4582 |
| 23 | 103556763 | 360 | 8/19/2018 4:38 | 8/19/2018 10:38 | 34.046822 | -118.248352 | 34.046822 | -118.248352 | 12407 | 30 | Round Trip | Monthly Pass | 3038 | 3038 |
| 24 | 153108231 | 21 | 2/5/2021 16:50 | 2/5/2021 17:11 | 33.984341 | -118.47155 | 33.99556 | -118.481552 | 15680 | 1 | One Way | Walk-up | 4210 | 4214 |
| 25 | 104470685 | 12 | 9/1/2018 16:00 | 9/1/2018 16:12 | 34.058319 | -118.246094 | 34.05661 | -118.237213 | 12024 | 1 | One Way | One Day Pass | 3028 | 3014 |

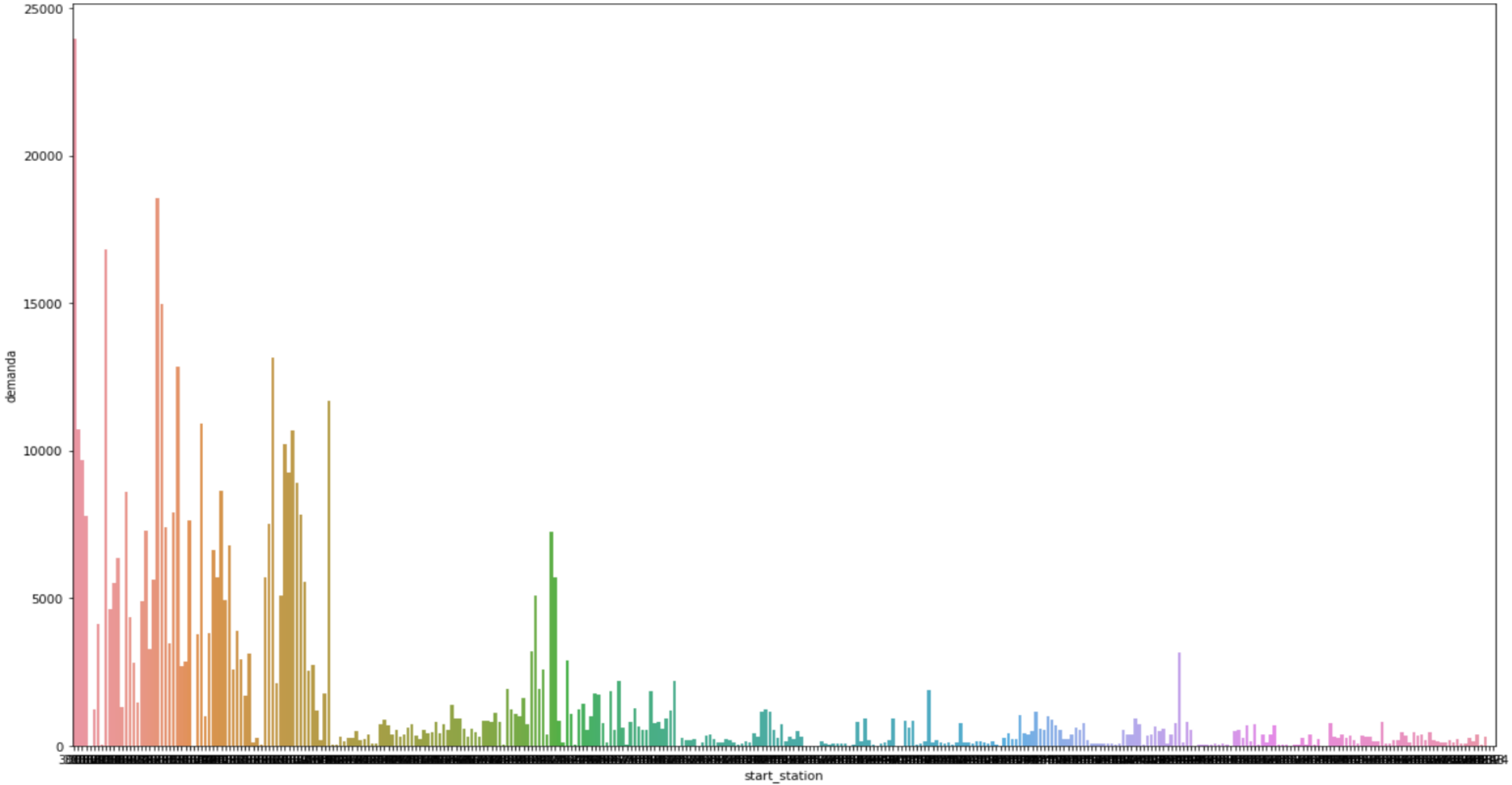
# DS Test. **Análisis exploratorio de datos**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| trip_id | 700000.0 | 1.069468e+08 | 4.497342e+07 | 8.369648e+06 | 7.538051e+07 | 1.179410e+08 | 1.404178e+08 | 1.794831e+08 |
| duration | 700000.0 | 3.708498e+01 | 1.253025e+02 | 1.000000e+00 | 7.000000e+00 | 1.300000e+01 | 2.600000e+01 | 1.440000e+03 |
| start_lat | 694437.0 | 3.404495e+01 | 3.252554e-01 | 3.371098e+01 | 3.403746e+01 | 3.404661e+01 | 3.405194e+01 | 5.570553e+01 |
| start_lon | 694437.0 | -1.182538e+02 | 2.332640e+00 | -1.184954e+02 | -1.182810e+02 | -1.182570e+02 | -1.182472e+02 | 1.182383e+02 |
| end_lat | 681426.0 | 3.404417e+01 | 2.997205e-01 | 3.371098e+01 | 3.403705e+01 | 3.404652e+01 | 3.405091e+01 | 5.570553e+01 |
| end_lon | 681426.0 | -1.182592e+02 | 2.129781e+00 | -1.184954e+02 | -1.182810e+02 | -1.182570e+02 | -1.182464e+02 | 3.760654e+01 |
| plan_duration | 699792.0 | 4.492870e+01 | 9.281630e+01 | 0.000000e+00 | 1.000000e+00 | 3.000000e+01 | 3.000000e+01 | 9.990000e+02 |
| start_station | 700000.0 | 3.499720e+03 | 6.159188e+02 | 3.000000e+03 | 3.031000e+03 | 3.064000e+03 | 4.214000e+03 | 4.594000e+03 |
| end_station | 700000.0 | 3.489727e+03 | 6.130408e+02 | 3.000000e+03 | 3.030000e+03 | 3.064000e+03 | 4.214000e+03 | 4.594000e+03 |

# DS Test. **Análisis exploratorio de datos**

- Extracción de características: year, month, day, hour, weekday, **distance.**
- Exploración de Estaciones
- Crecimiento de planes

# DS Test. **Análisis exploratorio de datos**



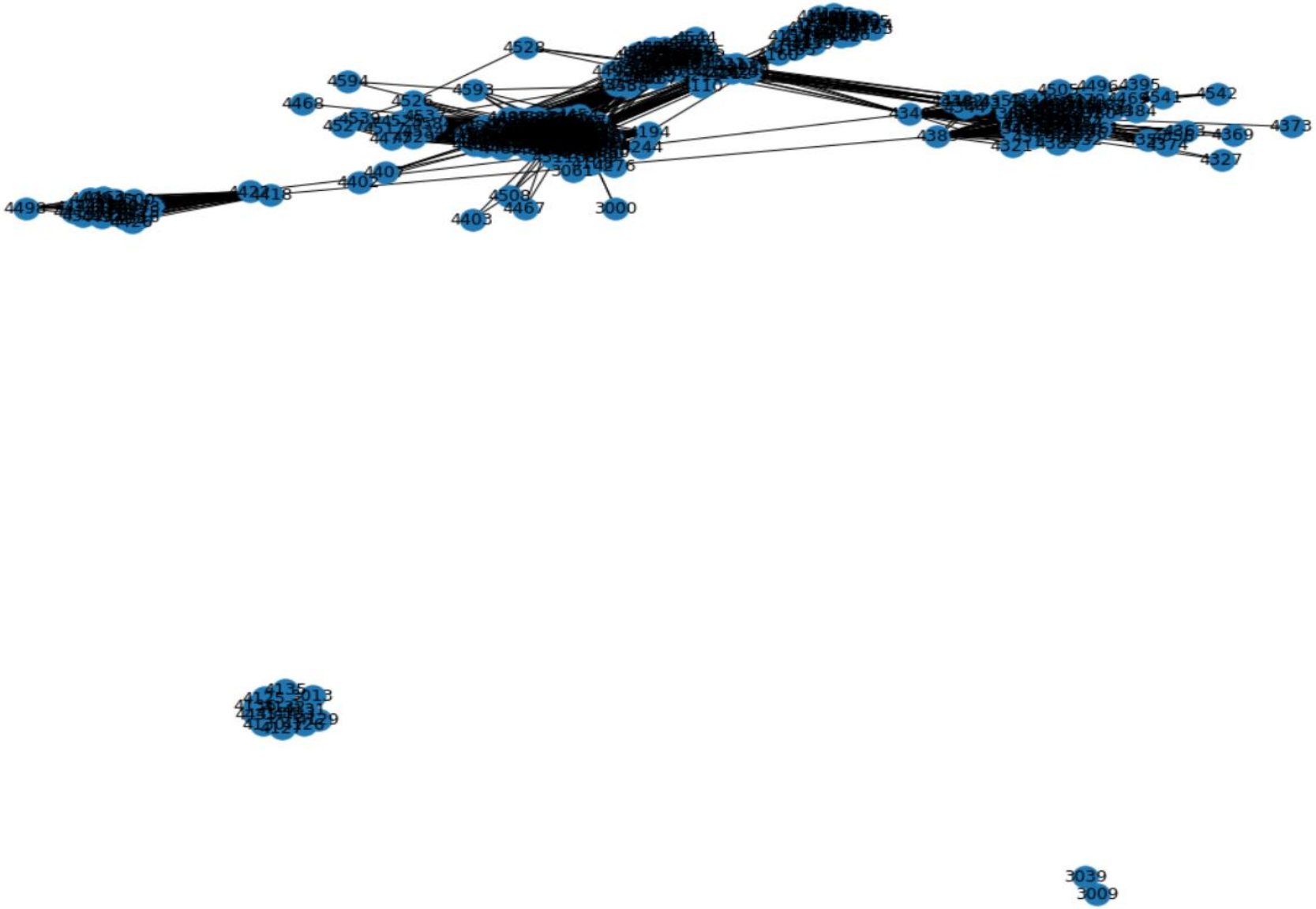
# DS Test. **Análisis exploratorio de datos**

- Estaciones con mayor demanda (percentil 90): 36 estaciones de 358.

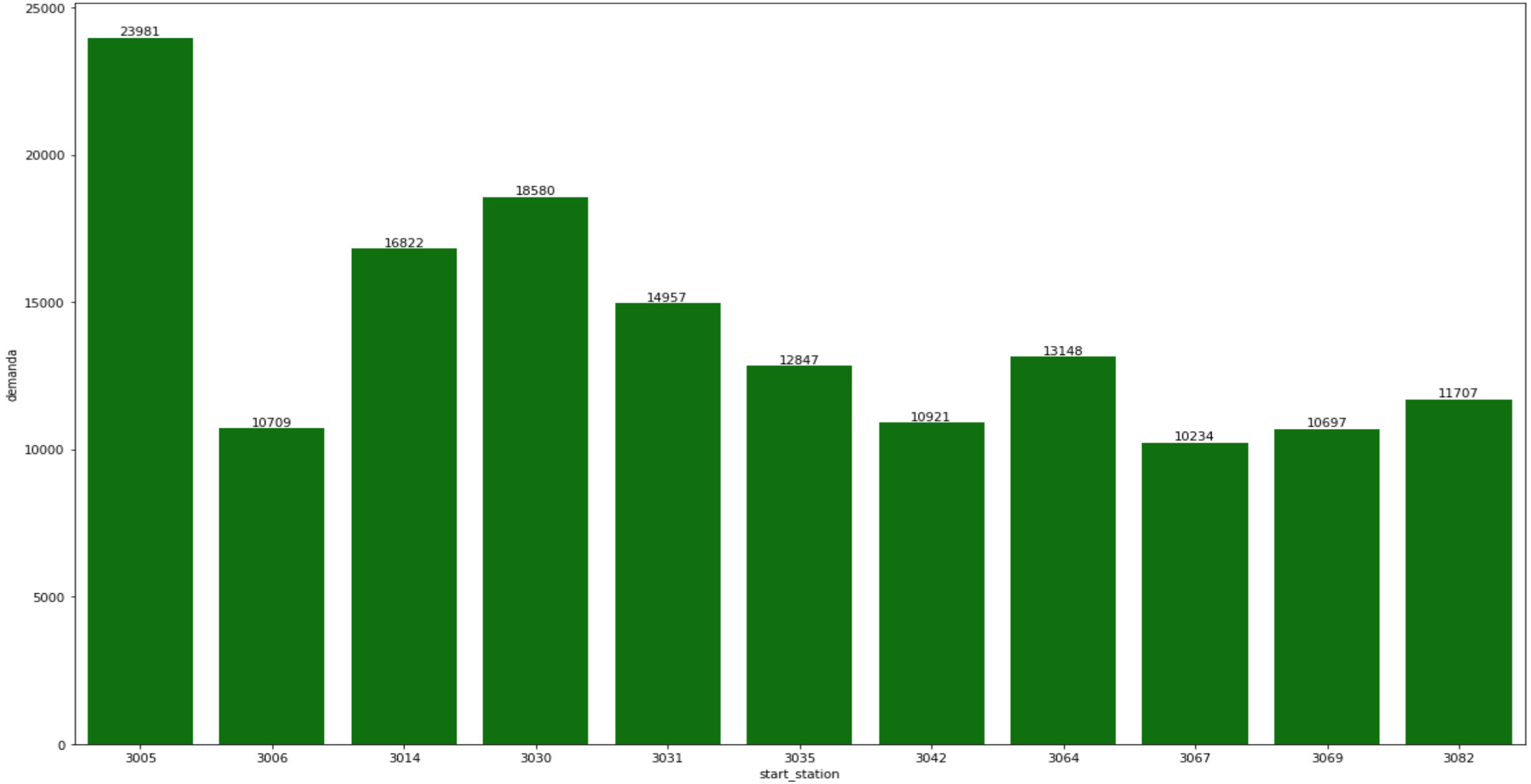
| | start_station | demanda |
|---|---|---|
| 0 | 3005 | 23981 |
| 21 | 3030 | 18580 |
| 8 | 3014 | 16822 |
| 22 | 3031 | 14957 |
| 50 | 3064 | 13148 |
| 26 | 3035 | 12847 |
| 64 | 3082 | 11707 |
| 32 | 3042 | 10921 |
| 1 | 3006 | 10709 |
| 55 | 3069 | 10697 |

| start_time_hour | start_station |
|---|---|
| 6 | 3014 |
| 7 | 3014 |
| 8 | 3014 |
| 9 | 3014 |
| 10 | 3005 |
| 11 | 3005 |
| 12 | 3005 |
| 13 | 3005 |
| 14 | 3030 |
| 15 | 3030 |
| 16 | 3030 |
| 17 | 3030 |
| 18 | 3005 |
| 19 | 3005 |
| 20 | 3005 |
| 21 | 3005 |
| 22 | 3005 |

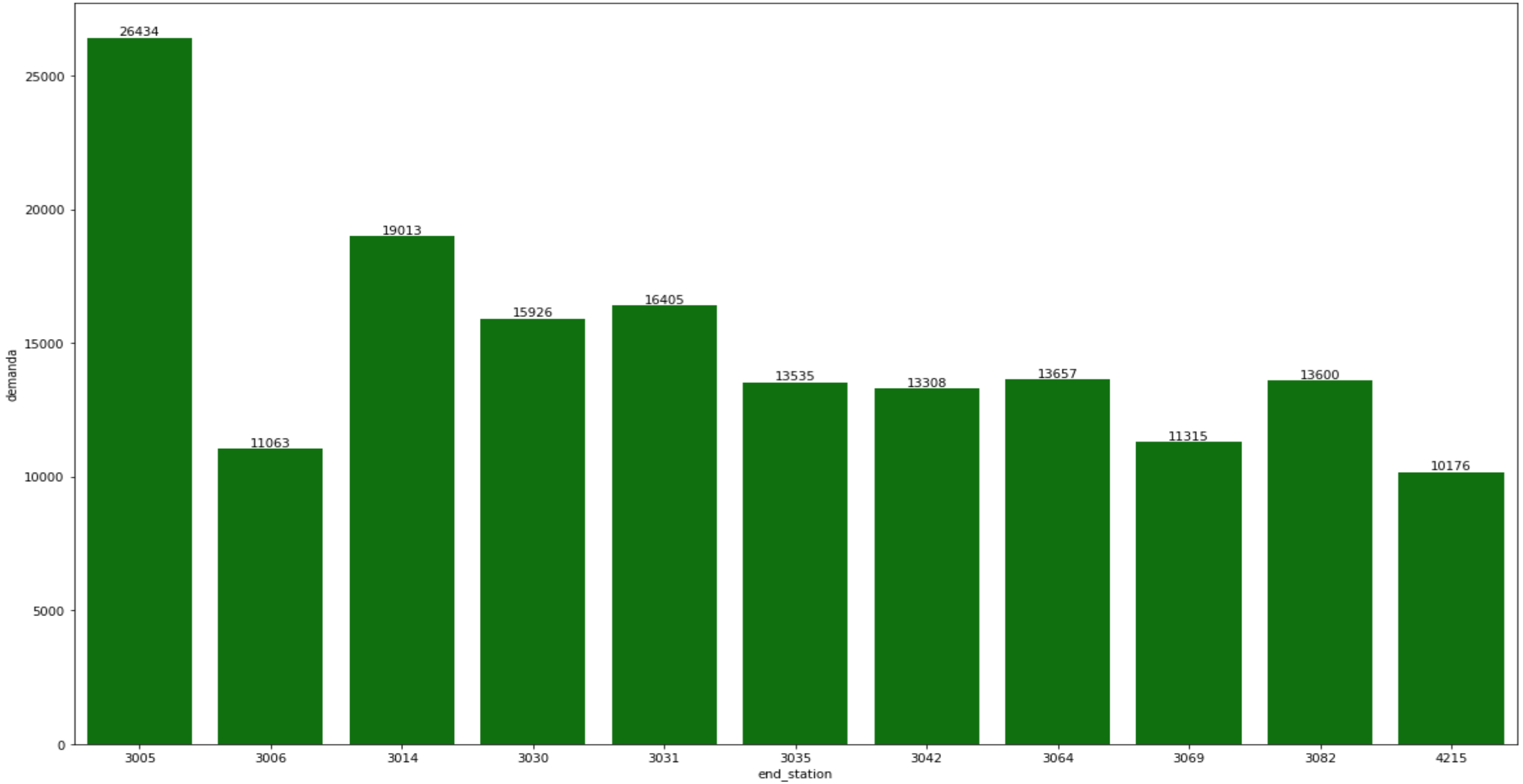
# DS Test. **Análisis exploratorio de datos**



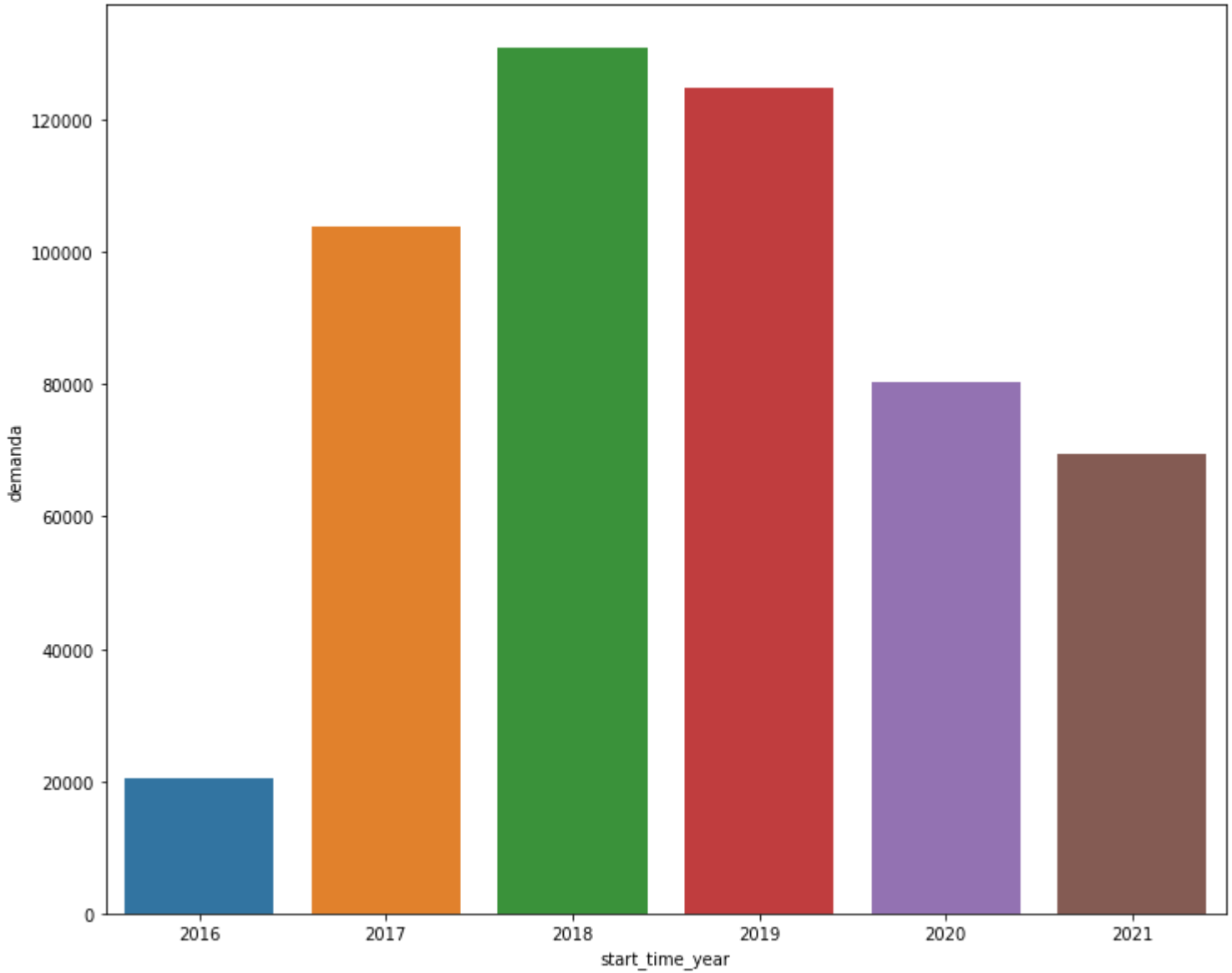
# DS Test. **Análisis exploratorio de datos**



# DS Test. **Análisis exploratorio de datos**



# DS Test. **Análisis exploratorio de datos**



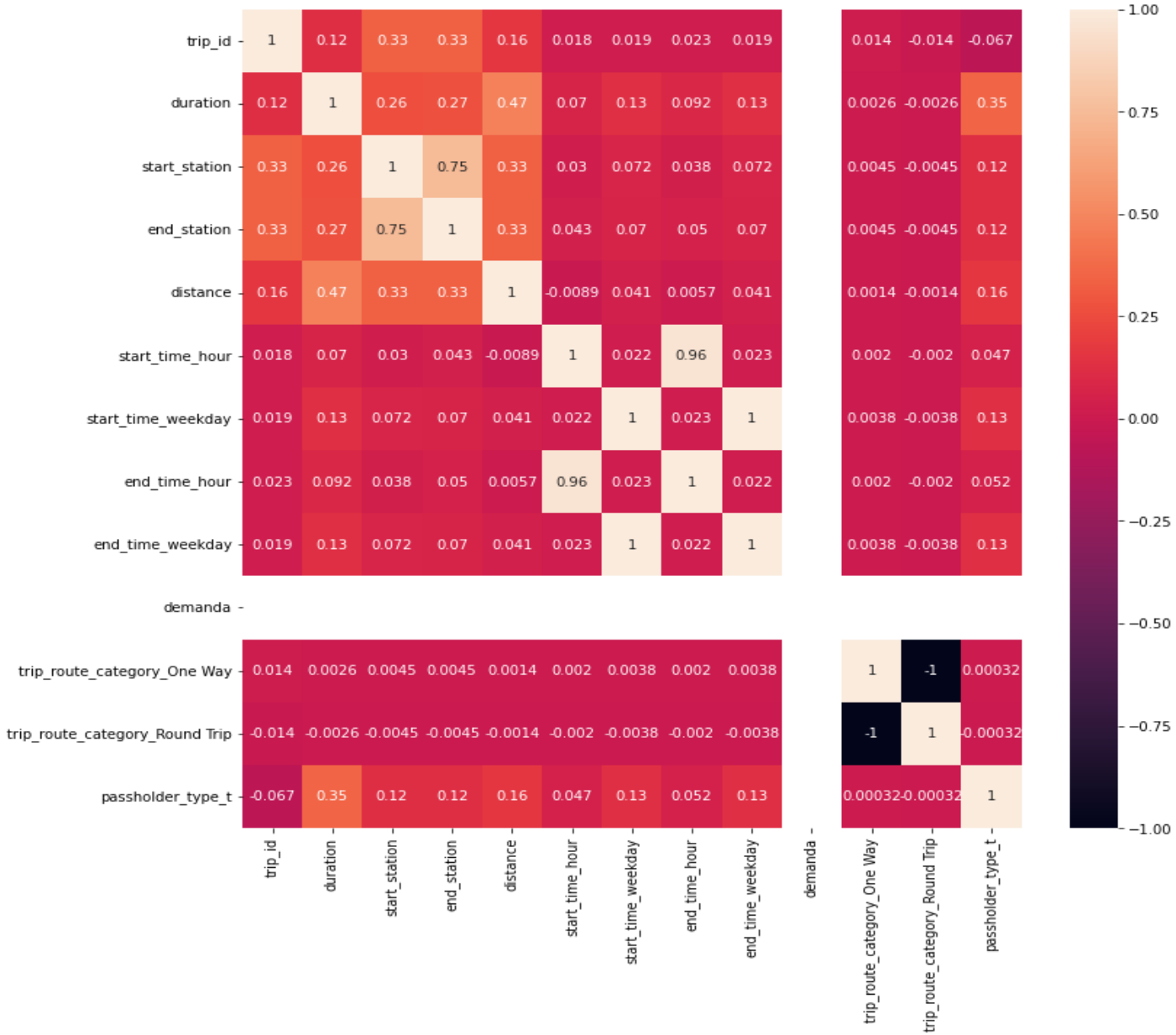
# DS Test. Análisis exploratorio de datos. Conclusiones

- Demanda de servicios en las distintas estaciones:
  - 11 estaciones concentran el 97% de la demanda de servicio (parten de 11 estaciones principalmente), se trasladan y finalizan el servicio en 11 estaciones (1 estación de diferencia entre las estaciones donde iniciaron el servicio)
- Crecimiento de planes:
  - Si hay un crecimiento en la demanda, se incorporaron nuevas estaciones.
  - El “virus” posiblemente afectó la baja en demanda durante 2020 y 2021.

# DS Test. Modelo de datos

- **Hipótesis**:
- La distancia recorrida será o no una característica que define el tipo de pase
- La duración del recorrido será o no una característica que define el tipo de pase
- La hora en que se hace uso del servicio será o no una característica que define el tipo de pase

# DS Test. **Modelo de datos**

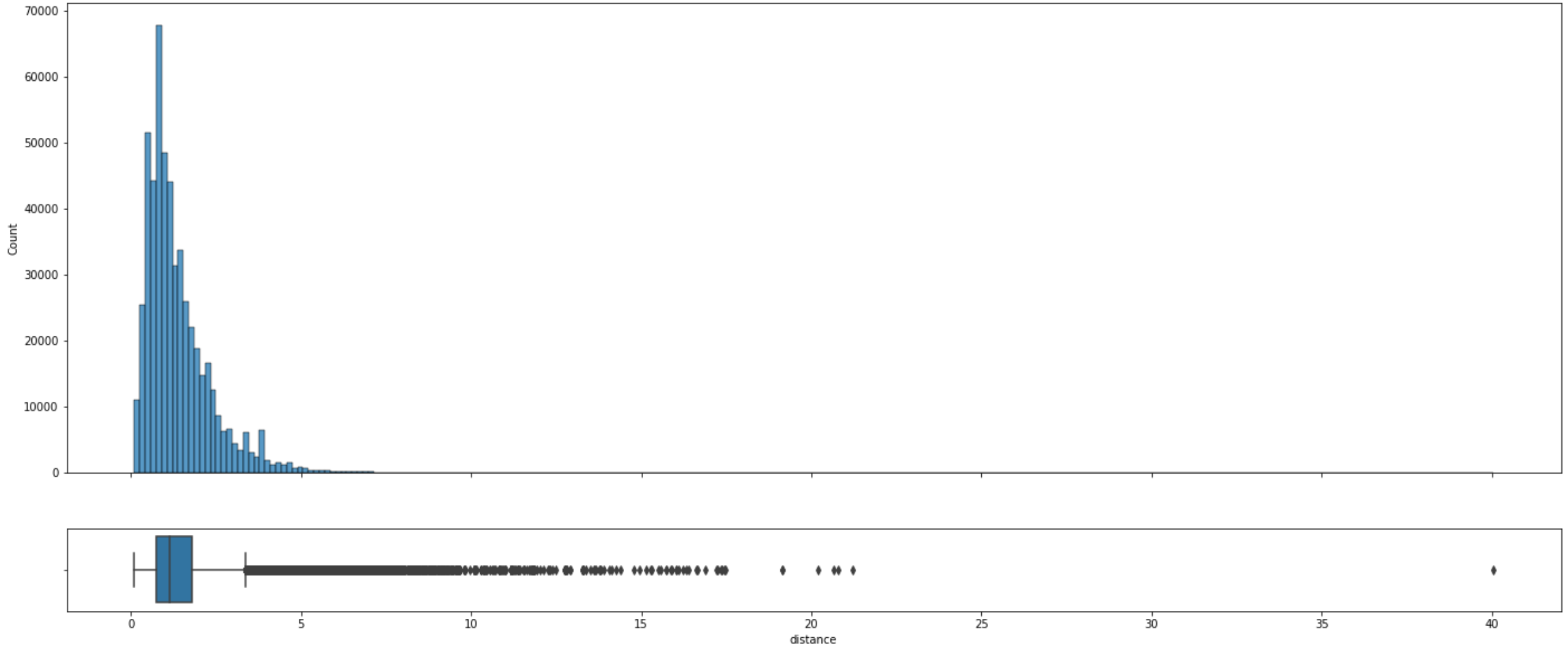


# DS Test. Modelo de datos

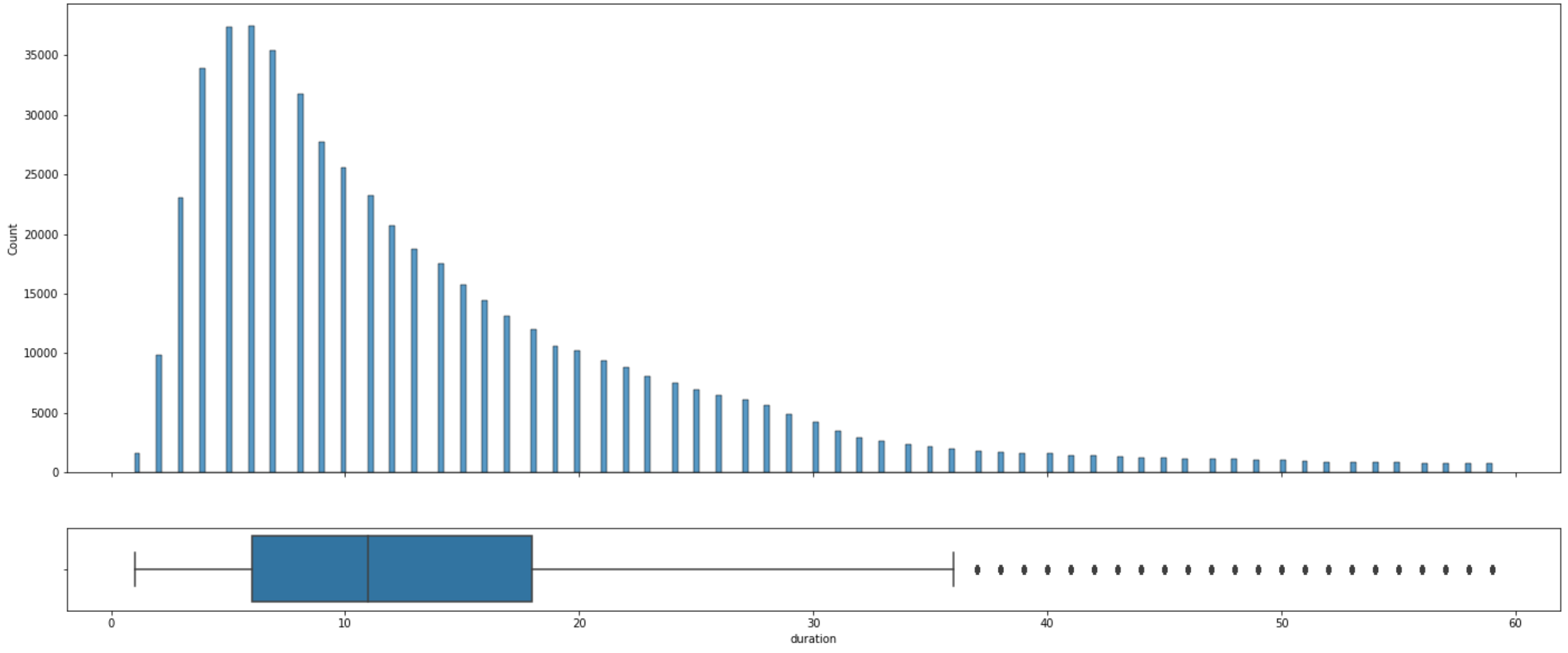
- **Procesamiento (transformación) de datos:**
  - Eliminación de datos nulos
  - Extracción de características:
    - Creación de nuevos atributos: year, month, day, hour, weekday, **distance.**
  - Eliminación de outliers
  - Codificación de categorías
  - Eliminación de características correlacionadas

  - Conjunto de datos no balanceado

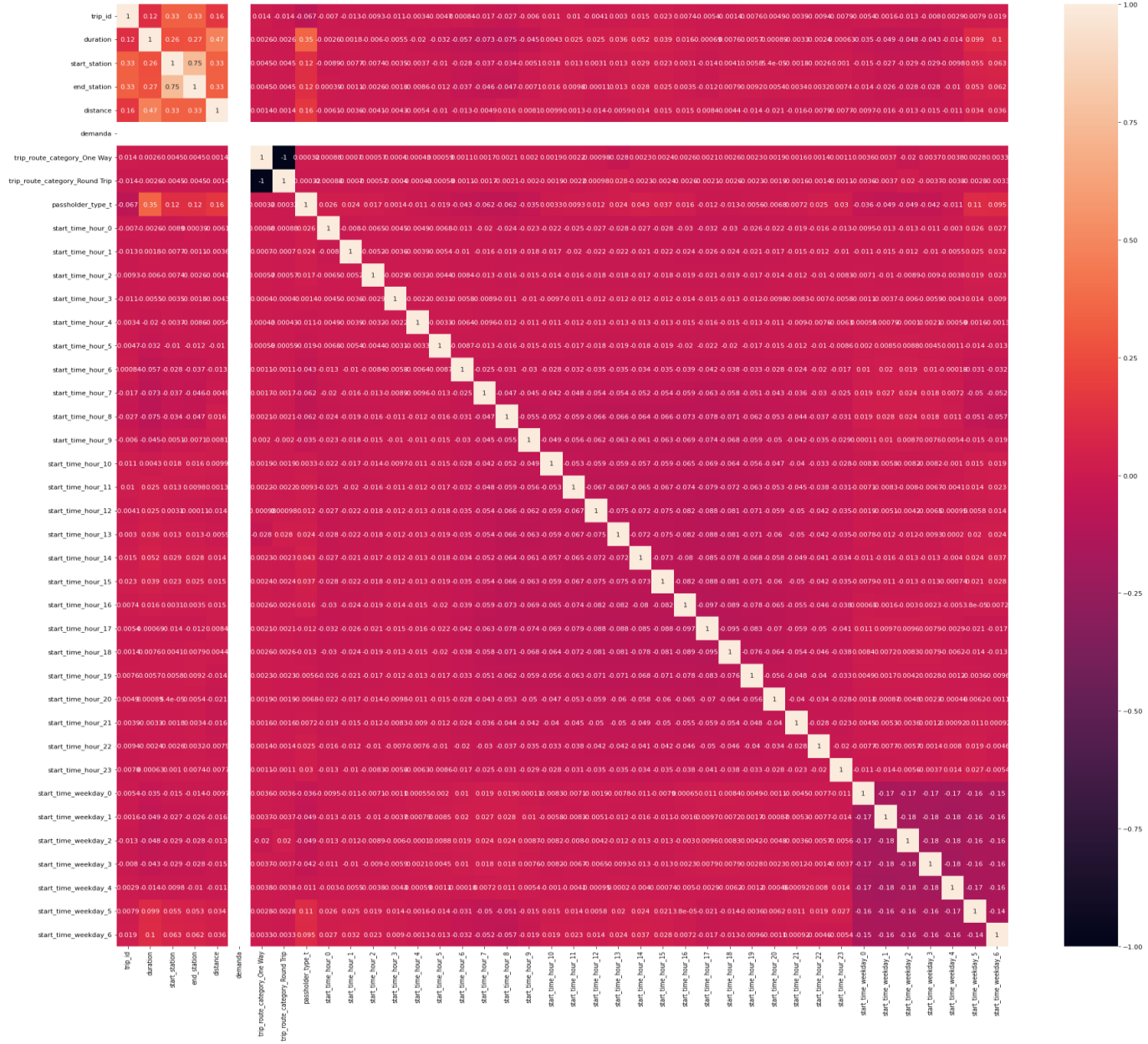
# DS Test. **Modelo de datos**



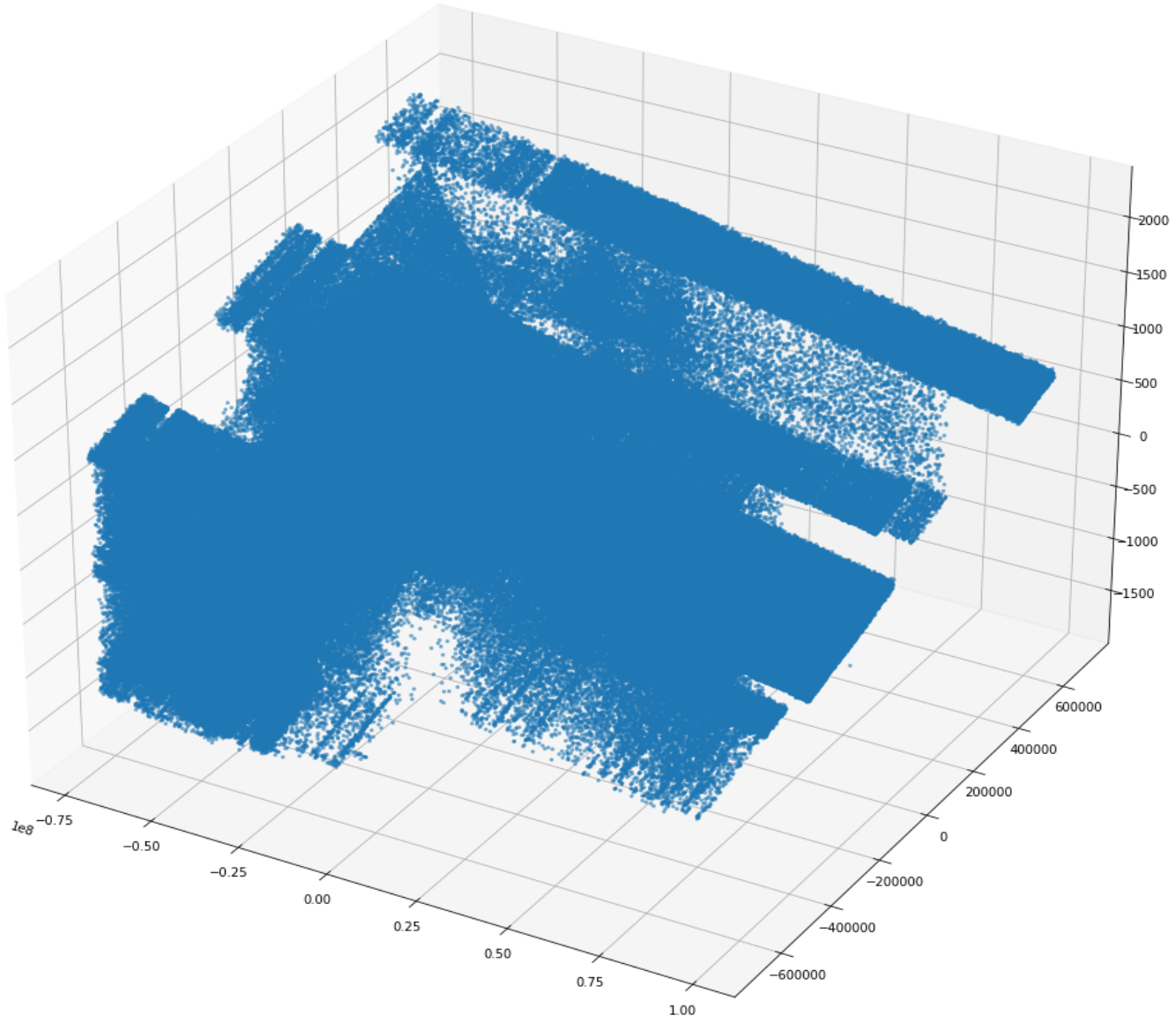
# DS Test. **Modelo de datos**



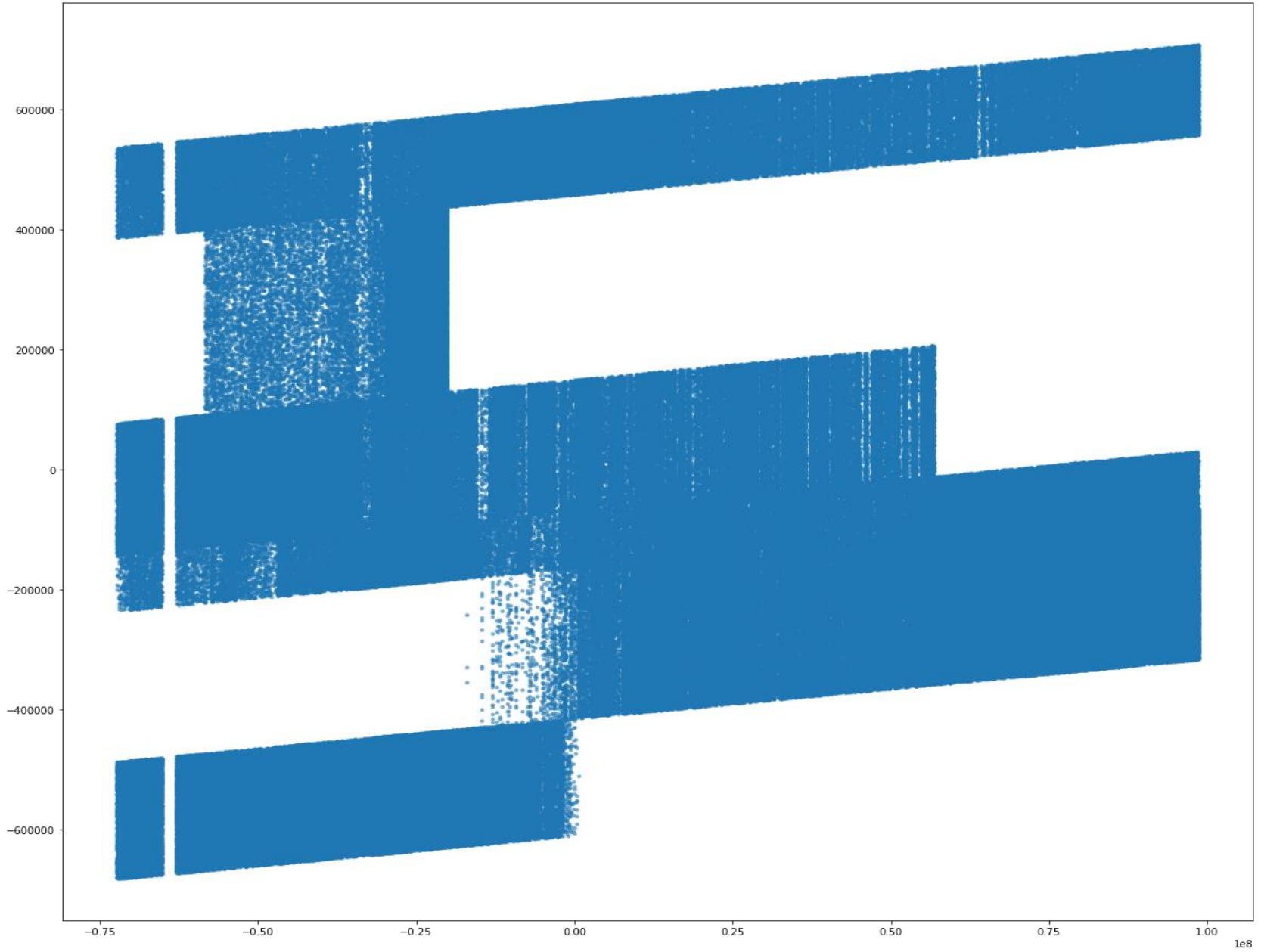
# DS Test. Modelo de datos



# DS Test. **Modelo de datos**

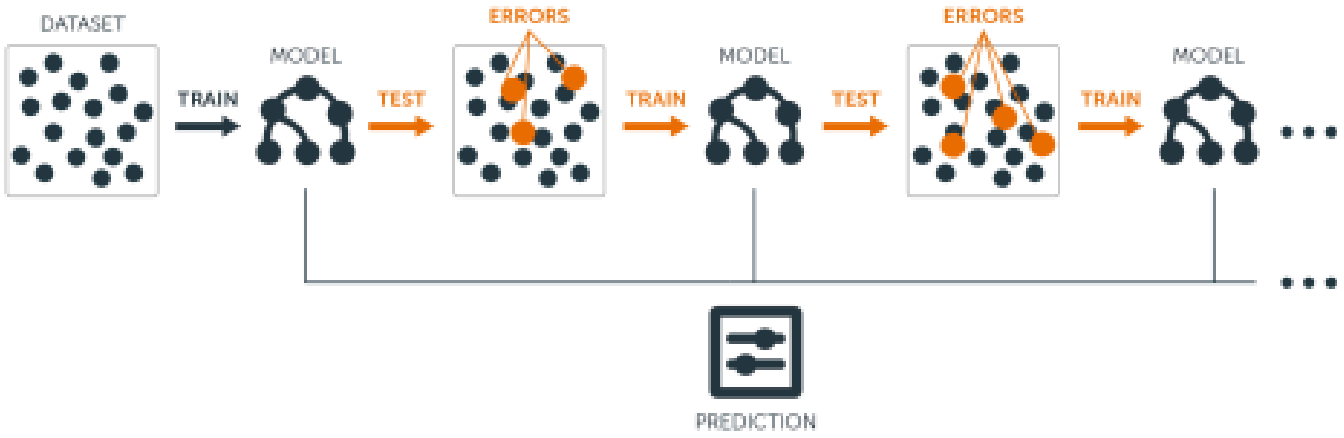


# DS Test. **Modelo de datos**



# DS Test. **Modelo de datos**

- Clasificador CatBoost
- Algoritmo basado en gradientes en árboles de decisión.
- Funciona ajustando árboles de decisión a residuos de árboles de decisión previos, transformando lentamente las predicciones débiles en fuertes.



# DS Test. Modelo de datos

- Ajuste de hiperparámetros:
- Iteraciones
- Tasa de aprendizaje
- Outlier removal

Community Prediction Competition

# DS Test

Data Science Programming Test

26 teams · a month to go

Overview Data Code Discussion Leaderboard Rules Team Submissions **Submit Predictions** ···

## Leaderboard

**Raw Data** **Refresh**

YOUR RECENT SUBMISSION

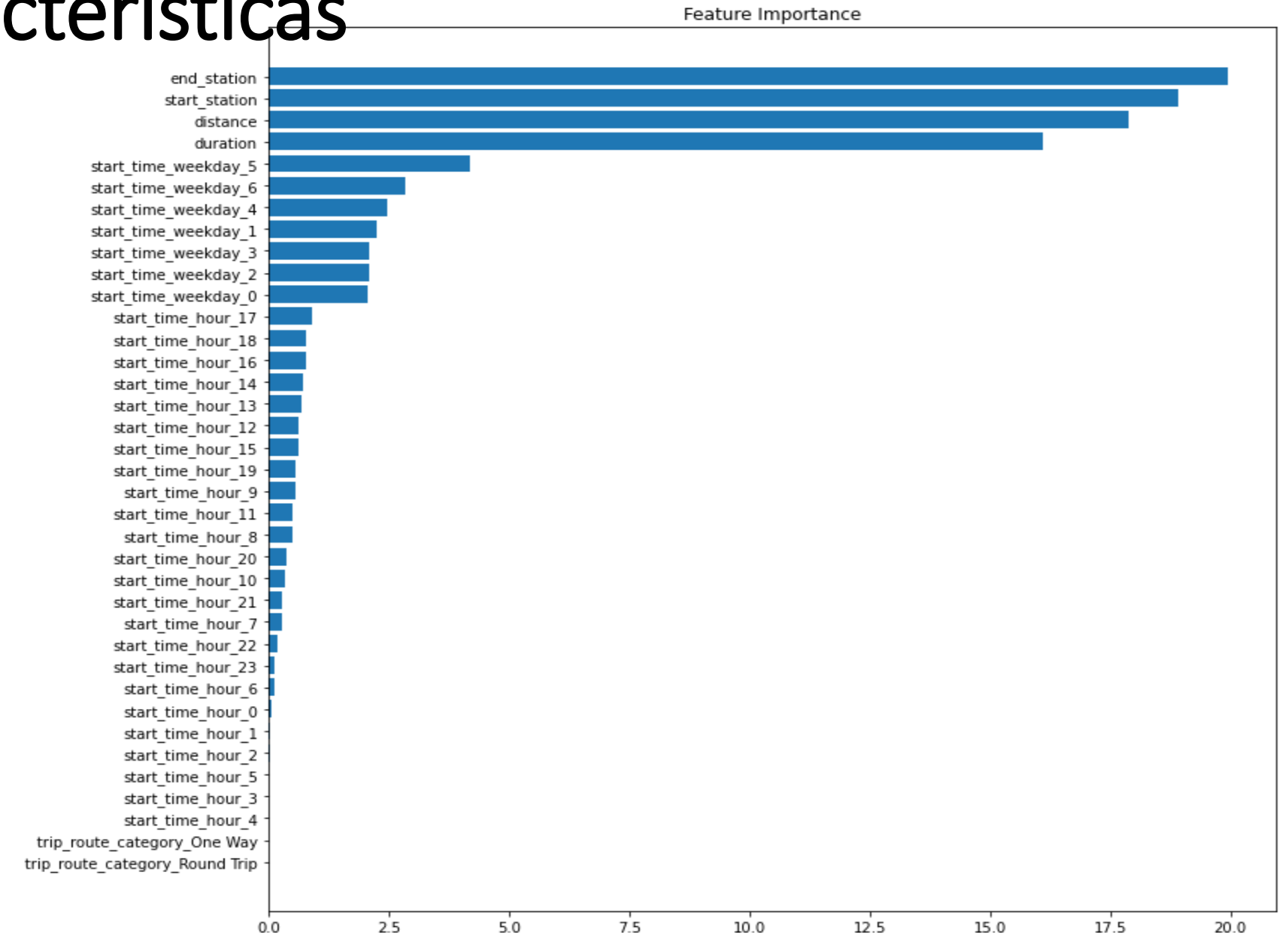
**jcbc3_model_onehot_1.csv**

Submitted by Josué de Jesús Juárez Vidales · Submitted 2 days ago

**Score: 0.56760**

Private score:

# DS Test. Modelo de datos. Importancia de características



# DS Test. **Modelo de datos**

- **Hipótesis**:
- La **distancia** recorrida será o no una característica que define el tipo de pase. **SI**
- La **duración** del recorrido será o no una característica que define el tipo de pase. **SI**
- La **hora** en que se hace uso del servicio será o no una característica que define el tipo de pase. **NO**

- Tanto la **estación de llegada** como la **estación de salida** resultaron ser importantes para el modelo

# DS Test. Modelo de datos

- **Conclusiones**:
- Tanto la **estación de llegada** como la **estación de salida** resultaron ser importantes para el modelo.
- También la distancia y duración del uso del servicio es relevante para el modelo de predicción.
- Aunque el modelo no tuvo un óptimo desempeño, si es posible emplearlo para obtener información del comportamiento de los usuarios.
- Es posible mejorar el modelo procesando los datos de fecha y generando nuevas características de las estaciones de servicio.

# DS Test. Modelo de datos

- **Otro modelos usados:**
- Clasificador bosque de árboles aleatorios
- Classificador XGBoost

**Diagrama de flujo.** Framework CRISP-DM (Cross Industry Standard Process for Data Mining)

# Gracias