Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Maestría en Ciencia de Datos

Materia: Aprendizaje Automático

Práctica 2

Regresión Lineal

Profesor: José Anastacio Hernández Saldaña

Alumno: Francisco David Treviño Bautista

Matrícula: 562795

Grupo: 1

Fecha de entrega: 20 de julio de 2024

Conjunto de Datos

Se eligió el conjunto de datos con la información de viajes de taxis amarillos reportada por la ciudad de Nueva York TLC mensualmente y que es de carácter público.

Para más información se puede consultar la siguiente página:

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

El conjunto de datos original se descargó en archivo con formato tipo ".parquet", el cuál maneja compresión de datos, y en R Studio se cuenta con una librería para cargar los datos como un dataframe de Pandas.

Variable	Descripción
VendorID	Un código que indica el proveedor de TPEP que proporcionó el registro. 1 = Tecnologías móviles creativas, LLC 2 = VeriFone Inc.
tpep_pickup_datetime	La fecha y hora en que se activó el taxímetro.
tpep_dropoff_datetime	La fecha y hora en que se desconectó el taxímetro.
Passenger_count	El número de pasajeros en el vehículo. Este es un valor ingresado por el conductor del taxi.
Trip_distance	La distancia del viaje transcurrido en millas reportadas por el taxímetro.
PULocationID	Zona de Taxi TLC en la que estaba activado el taxímetro
DOLocationID	Zona de Taxi TLC en la que se desactivó el taxímetro
RateCodeID	El código de tarifa final vigente al finalizar el viaje. 1 = Tarifa estándar, 2 = JFK , 3 = Newark , 4 = Nassau o Westchester 5 = Tarifa negociada, 6 = Viaje en grupo
Store_and_fwd_flag	Este indicador indica si el registro de viaje se mantuvo en la memoria del vehículo antes de enviarlo al proveedor, también conocido como "almacenar y reenviar", porque el vehículo no tenía una conexión con el servidor. Y = viaje de almacenamiento y avance N = no es un viaje de almacenamiento y avance
Payment_type	Un código numérico que indica cómo el pasajero pagó el viaje. 1 = Tarjeta de crédito, 2 = Efectivo, 3 = Sin cargo 4 = Disputa, 5 = Desconocido, 6 = Viaje anulado
Fare_amount	La tarifa de tiempo y distancia calculada por el taxímetro.
Extra	Extras y recargos varios. Actualmente, esto solo incluye los cargos de $0,50 y $1 por hora pico y por noche.
MTA_tax	Impuesto MTA de $0.50 que se activa automáticamente según la tarifa medida en uso.
Improvement_surcharge	Recargo por mejora de $0.30 en viajes evaluados al bajar la bandera. El recargo por mejora comenzó a cobrarse en 2015.
Tip_amount	Monto de la propina: este campo se completa automáticamente para las propinas de tarjetas de crédito. Las propinas en efectivo no están incluidas.
Tolls_amount	Importe total de todos los peajes pagados en el viaje.
Total_amount	El importe total cobrado a los pasajeros. No incluye propinas en efectivo.
Congestion_Surcharge	Monto total cobrado en el viaje por el recargo por congestión del Estado de Nueva York.
Airport_fee	$1.25 para traslados únicamente en los aeropuertos LaGuardia y John F. Kennedy

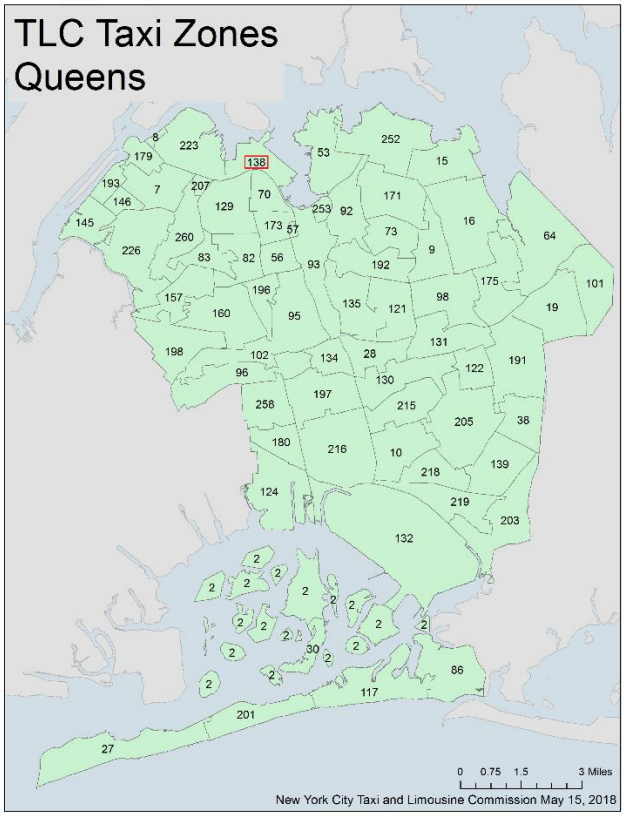
Posteriormente a la carga de la información, se tuvo que realizar un análisis del conjunto de datos para entender su distribución con estadística descriptiva, así como detectar los valores mínimos y/o máximos.

También hubo actividades de limpieza de los datos, porque se detectaron registros con datos inválidos, negativos o fuera de rango. Por ejemplo, había viajes con montos negativos, con una duración mayor a 12 horas, etc.

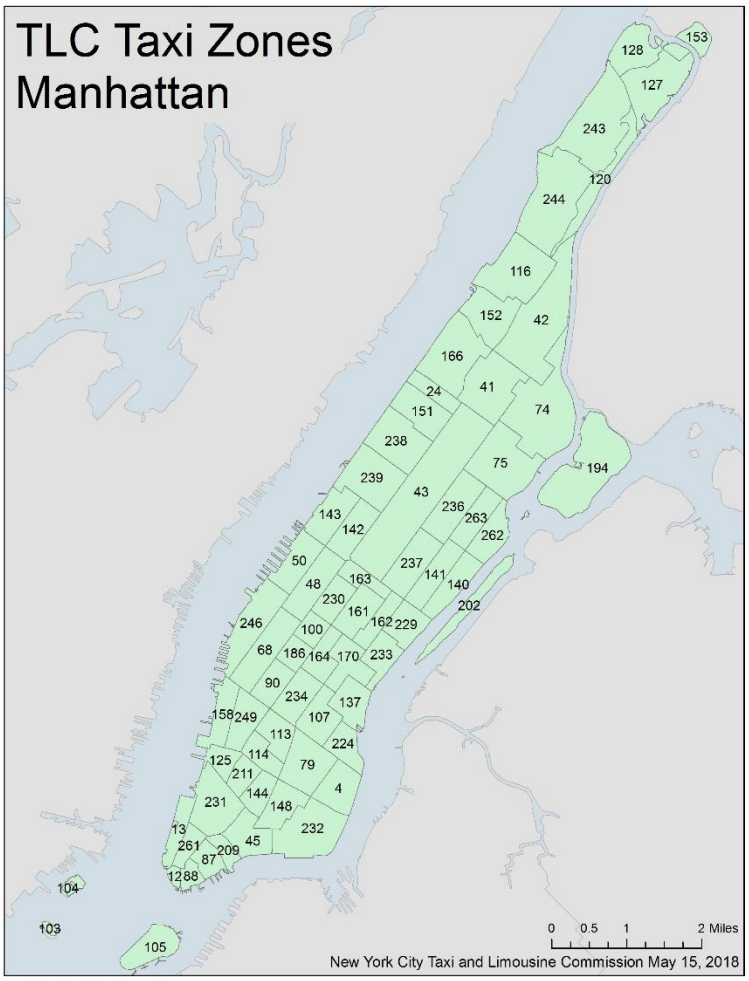
El conjunto de datos tiene las siguientes características:

- Viajes de taxi que salen del aeropuerto LaGuardia.
- Tienen como destino cualquier ubicación dentro de Manhattan.
- El pasajero paga una propina.
- Los viajes corresponden a todos los lunes del mes.

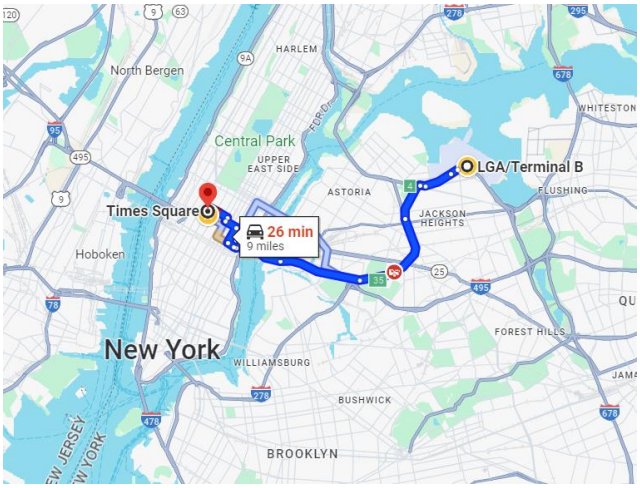
Mapa de la zona de taxis de donde se encuentra el aeropuerto LaGuardia (LocationID = 138):



Mapa de la zona de taxis de Manhattan:



Ejemplo de las 2 rutas para viajar desde el aeropuerto al centro de Manhattan:



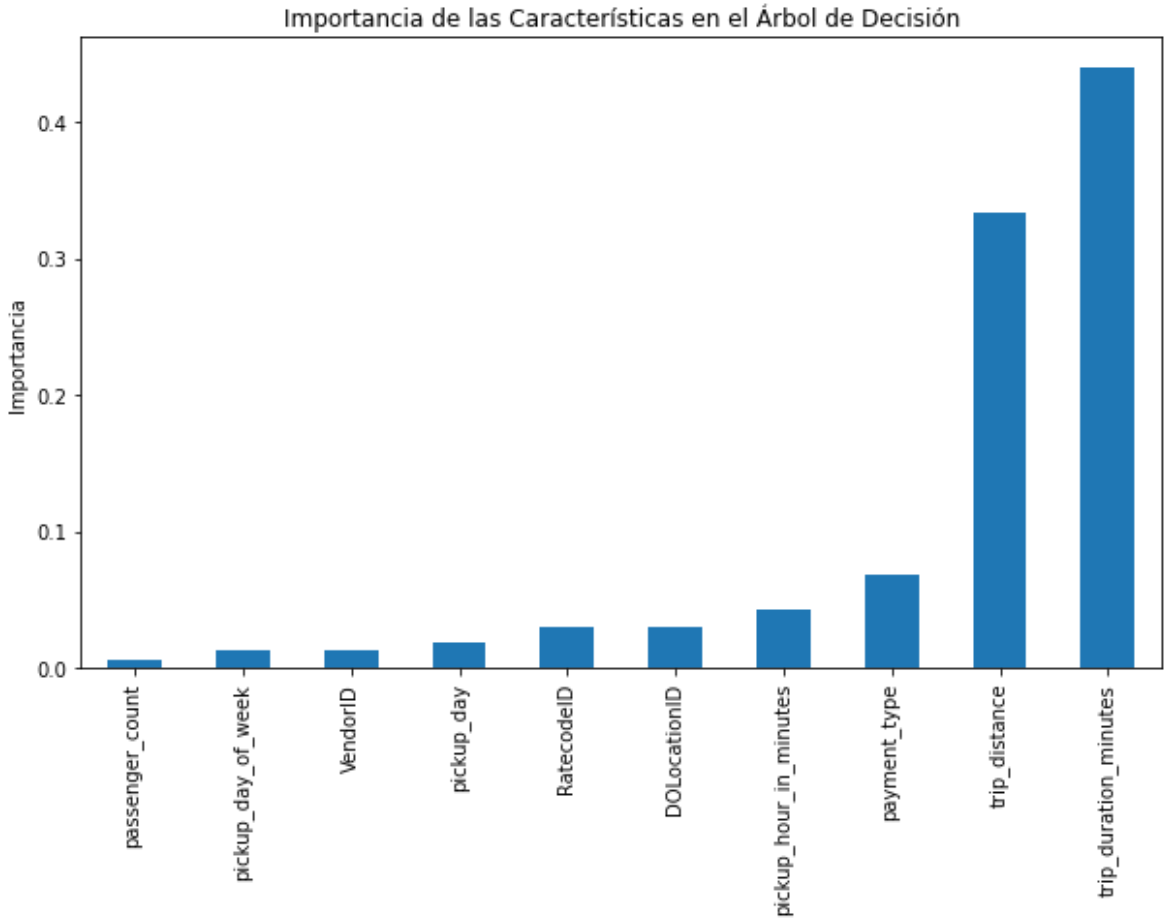
Elección de Variables

Seleccionar las variables independientes y la variable dependiente para el modelo de regresión.

Tipo de Variable	Variable	Descripción
Independiente	trip_distance	La distancia del viaje transcurrido en millas reportadas por el taxímetro.
Independiente	trip_duration_minutes	Duración del viaje en minutos. Campo calculado en función a la duración del viaje en segundos.
Independiente	pickup_hour_in_minutes	La hora y minuto del día en que se inició el viaje en taxi. Campo calculado en función a la hora, minuto y segundo en que inició el viaje.
Independiente	pickup_day	Día del mes en que se inició el viaje en taxi.
Independiente	pickup_day_of_week	Día de la semana en que se inició el viaje en taxi.
Independiente	DOLocationID	Zona de Taxi TLC en la que se desactivó el taxímetro y/o terminó el viaje.
Independiente	VendorID	Un código que indica el proveedor de TPEP que proporcionó el registro. 1 = Tecnologías móviles creativas, LLC 2 = VeriFone Inc.
Independiente	payment_type	Un código numérico que indica cómo el pasajero pagó el viaje. 1 = Tarjeta de crédito, 2 = Efectivo, 3 = Sin cargo 4 = Disputa, 5 = Desconocido, 6 = Viaje anulado
Independiente	RatecodeID	El código de tarifa final vigente al finalizar el viaje. 1 = Tarifa estándar, 2 = JFK, 3 = Newark, 4 = Nassau o Westchester 5 = Tarifa negociada, 6 = Viaje en grupo
Independiente	passenger_count	El número de pasajeros en el vehículo. Este es un valor ingresado por el conductor del taxi.
Dependiente (Y)	total_amount_calc	El importe total cobrado a los pasajeros. No incluye propinas en efectivo.

Modelos de Regresión

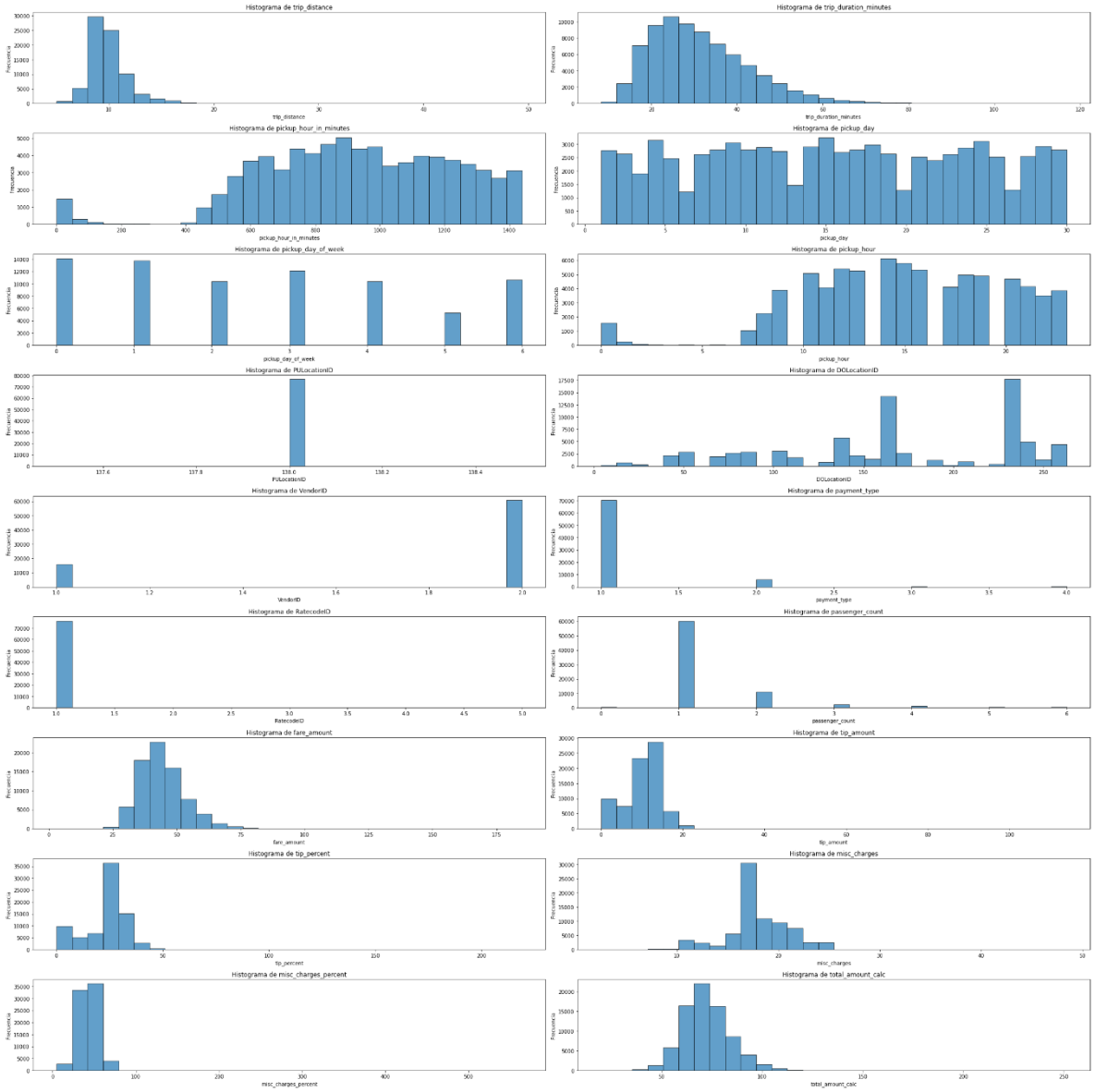
Se desarrolló código adicional para realizar las pruebas del conjunto de variables independientes, de tal forma que sea una regresión múltiple. Se observa que las siguientes variables son las que más impactan al modelo de regresión.



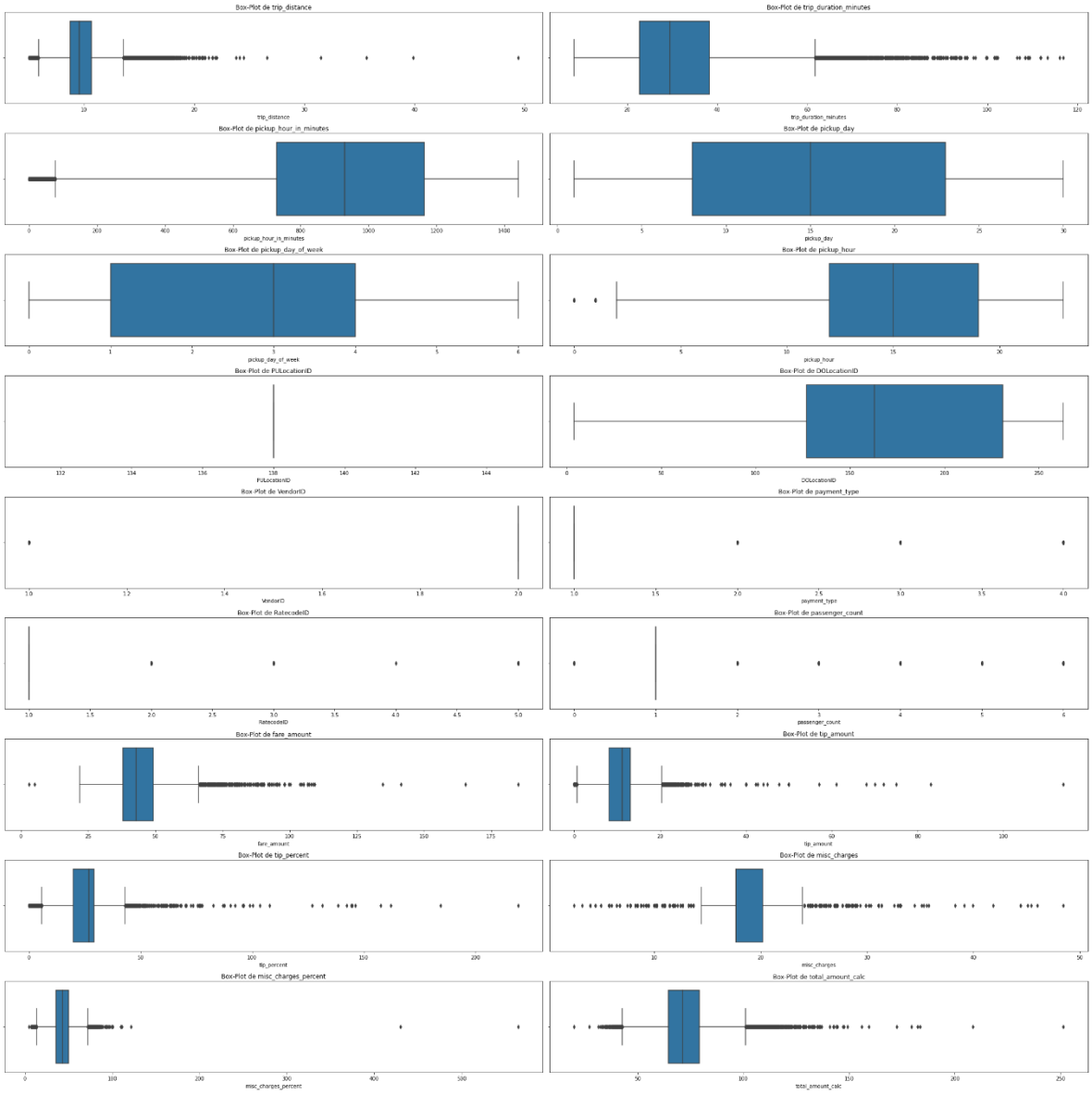
Por otro lado, se realizaron pruebas usando de referencia el código del repositorio del profesor, que se revisó en clase, que evalúa para cada modelo una variable independiente vs la variable dependiente.

Se hicieron pruebas separadas de cada variable de las 2 más importantes del modelo de regresión (trip_duration_minutes, trip_distance).

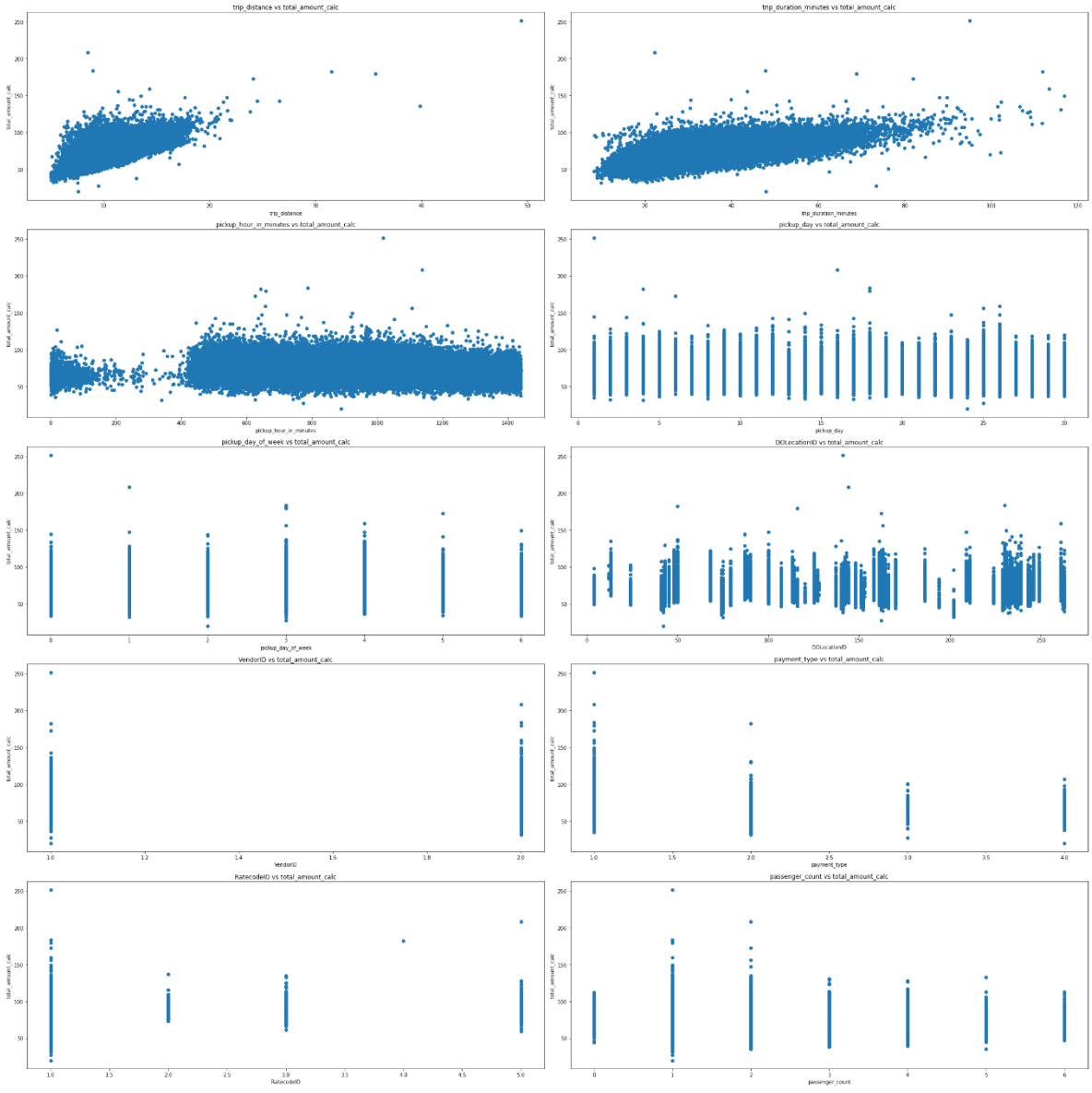
Análisis Exploratorio



Gráficas Tipo Box-Plot

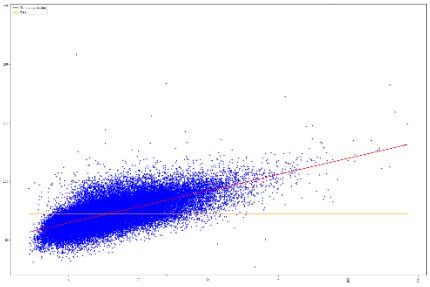


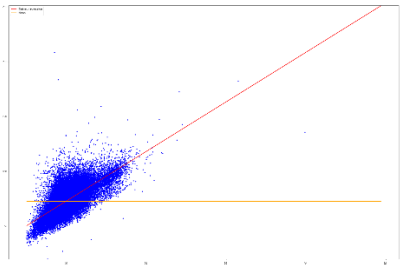
Gráficas por Variable Independiente vs Variable Dependiente



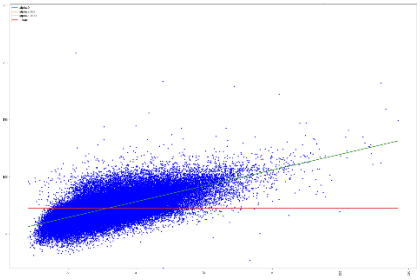
Gráficas de Modelos

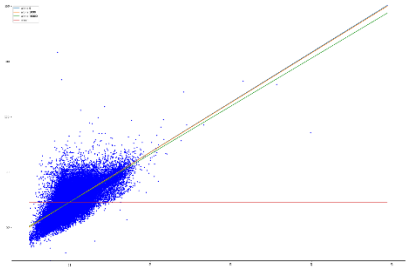
a) Linear Regression



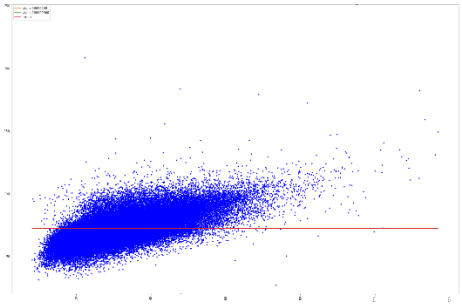


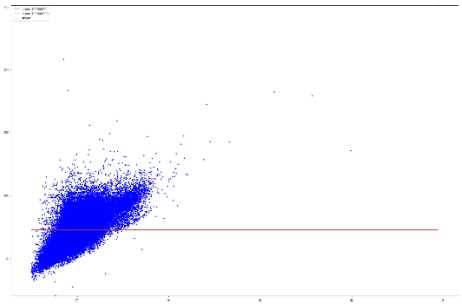
b) Ridge Regression



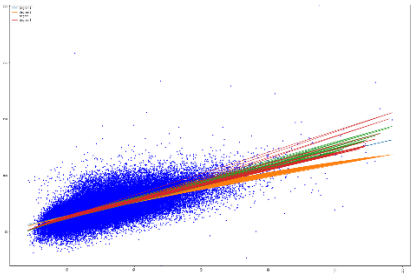


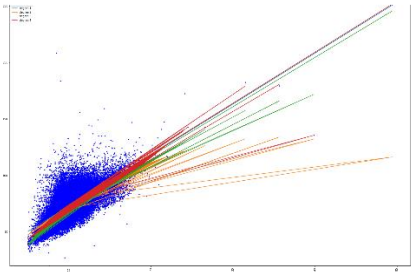
c) Lasso Regression



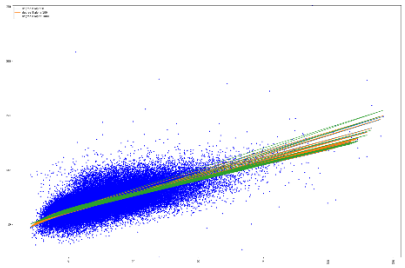


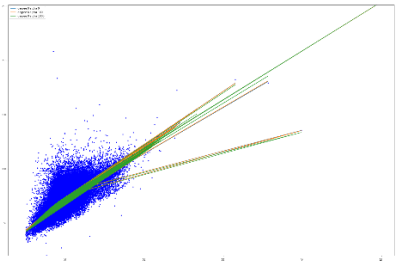
d) Polynomial Regression



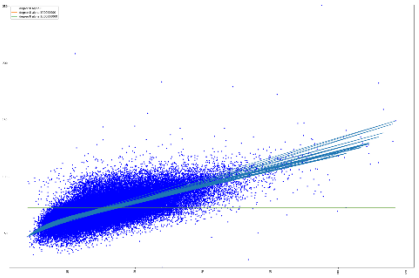


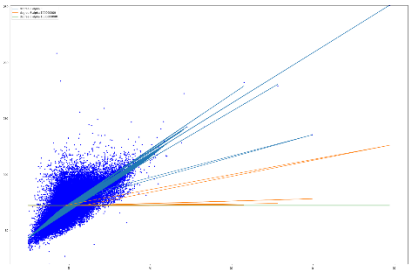
e) Polynomial Ridge Regression



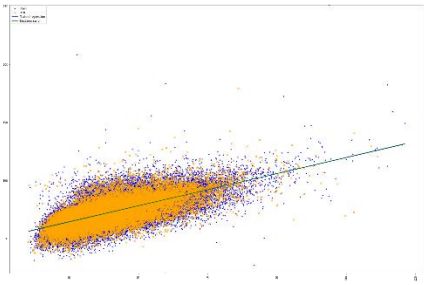


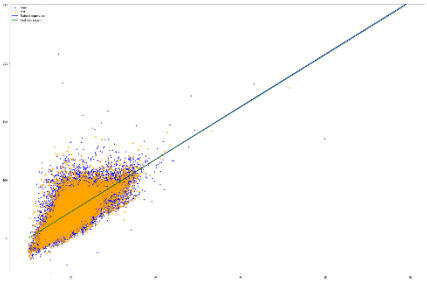
f) Polynomial Lasso Regression



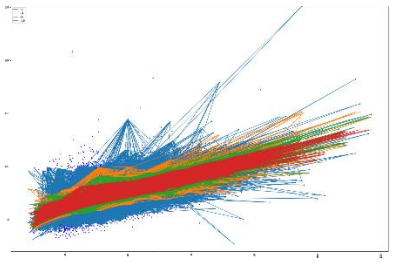


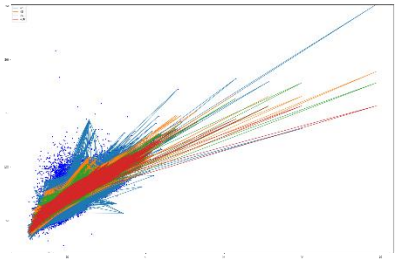
g) Linear Regression Model Predict



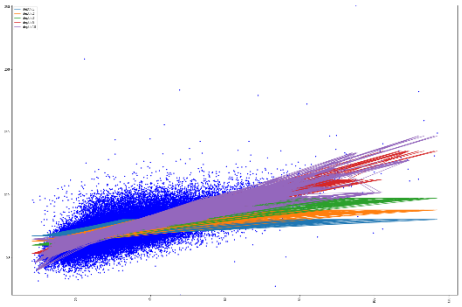


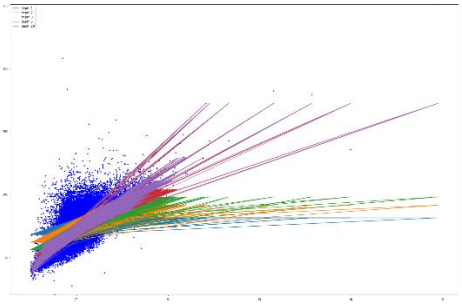
h) KNN Regression

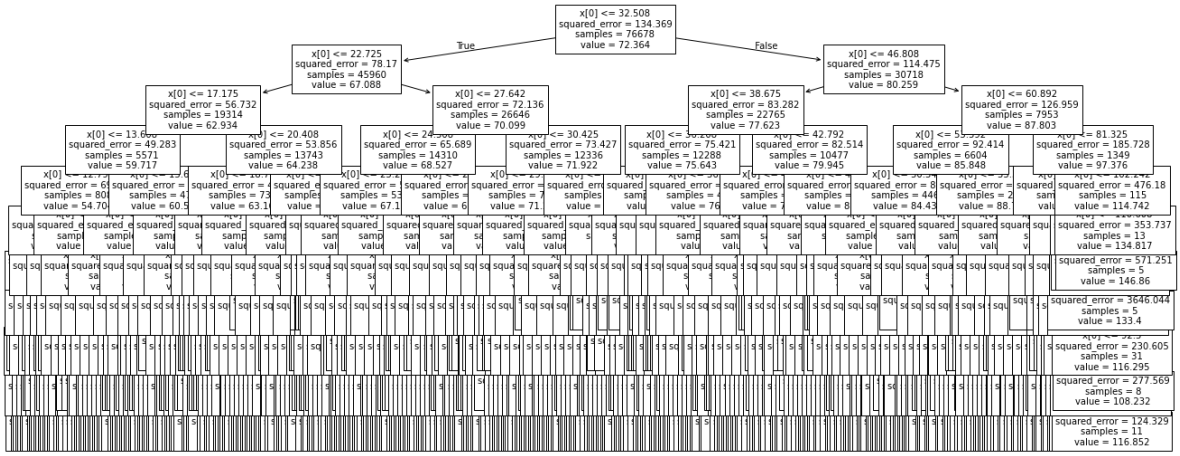


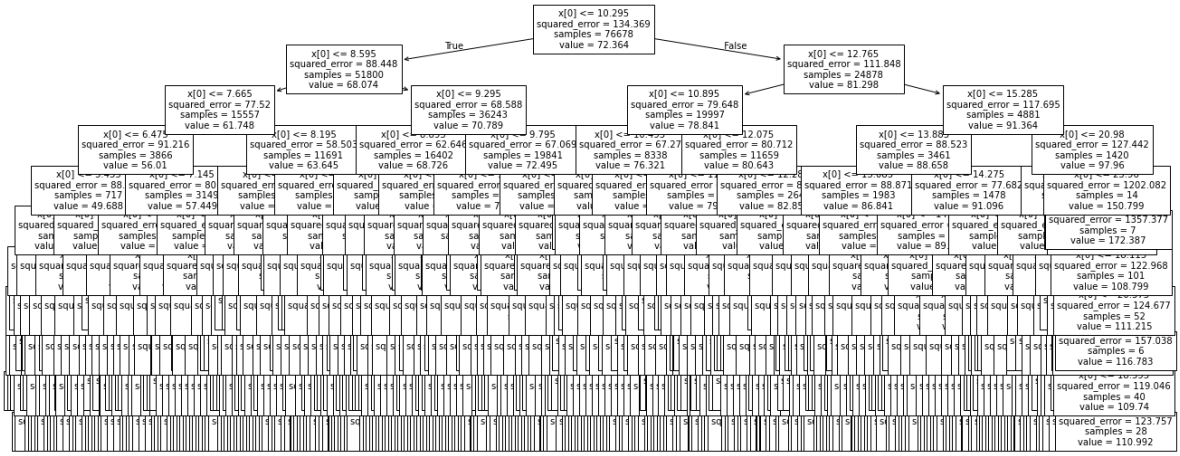


i) Decision Tree Regression





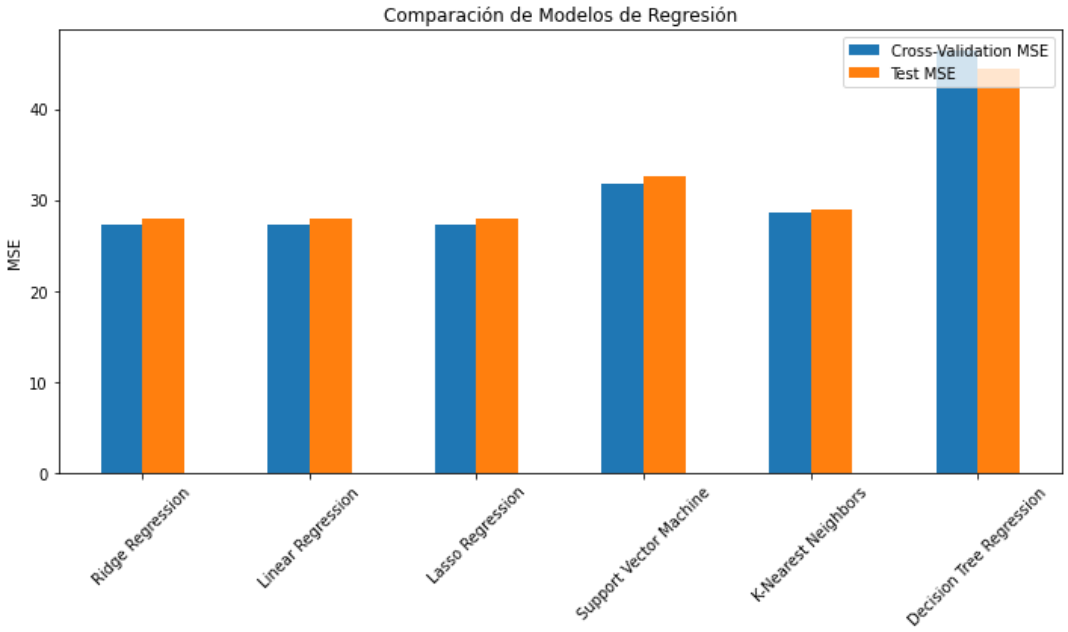

x[0] <= 32.508
squared_error = 134.369
samples = 76678
value = 72.364
True
False
x[0] <= 22.725
squared_error = 78.17
samples = 45960
value = 67.088
x[0] <= 46.808
squared_error = 114.475
samples = 30718
value = 80.259
x[0] <= 17.175
squared_error = 56.732
samples = 19314
value = 62.934
x[0] <= 27.642
squared_error = 72.136
samples = 26646
value = 70.099
x[0] <= 38.675
squared_error = 83.282
samples = 22765
value = 77.623
x[0] <= 60.892
squared_error = 126.959
samples = 7953
value = 87.803
x[0] <= 13.625
squared_error = 49.283
samples = 5571
value = 59.717
x[0] <= 20.408
squared_error = 53.856
samples = 13743
value = 64.238
x[0] <= 24.308
squared_error = 65.689
samples = 14310
value = 68.527
x[0] <= 30.425
squared_error = 73.427
samples = 12336
value = 71.922
x[0] <= 33.188
squared_error = 75.421
samples = 12288
value = 75.643
x[0] <= 42.792
squared_error = 82.514
samples = 10477
value = 79.945
x[0] <= 51.356
squared_error = 92.414
samples = 6604
value = 85.848
x[0] <= 81.325
squared_error = 185.728
samples = 1349
value = 97.376
squared_error = 571.251
samples = 5
value = 146.86
squared_error = 3646.044
samples = 5
value = 133.4
squared_error = 230.605
samples = 31
value = 116.295
squared_error = 277.569
samples = 8
value = 108.232
squared_error = 124.329
samples = 11
value = 116.852


x[0] <= 10.295
squared_error = 134.369
samples = 76678
value = 72.364
True
False
x[0] <= 8.595
squared_error = 88.448
samples = 51800
value = 68.074
x[0] <= 12.765
squared_error = 111.848
samples = 24878
value = 81.298
x[0] <= 7.665
squared_error = 77.52
samples = 15557
value = 61.748
x[0] <= 9.295
squared_error = 68.588
samples = 36243
value = 70.789
x[0] <= 10.895
squared_error = 79.648
samples = 19997
value = 78.841
x[0] <= 15.285
squared_error = 117.695
samples = 4881
value = 91.364
x[0] <= 6.475
squared_error = 91.216
samples = 3866
value = 56.01
x[0] <= 8.195
squared_error = 58.503
samples = 11691
value = 63.645
x[0] <= 9.795
squared_error = 67.069
samples = 19841
value = 72.495
x[0] <= 12.075
squared_error = 80.712
samples = 11659
value = 80.643
x[0] <= 13.885
squared_error = 88.523
samples = 3461
value = 88.658
x[0] <= 20.98
squared_error = 127.442
samples = 1420
value = 97.96
squared_error = 1357.377
samples = 7
value = 172.387
squared_error = 122.968
samples = 101
value = 108.799
squared_error = 124.677
samples = 52
value = 111.215
squared_error = 157.038
samples = 6
value = 116.783
squared_error = 119.046
samples = 40
value = 109.74
squared_error = 123.757
samples = 28
value = 110.992

Evaluación del Modelo

En base al código generado en Python y al conjunto de datos, variables independientes, se obtienen los siguientes resultados para cada modelo de regresión:

Modelo	R-squared	Cross-Validation MSE	Test MSE
Ridge Regression	0.794221000349588	27.228010155429900	28.021607147320700
Linear Regression	0.794220632314303	27.228010430591900	28.021657263903100
Lasso Regression	0.794216903553377	27.320025375346600	28.022165021613900
K-Nearest Neighbors	0.787545778997191	28.637928061549900	28.930594122065700
Support Vector Machine	0.760356068755520	31.732741470593300	32.633106915577000
Decision Tree Regression	0.674175563039869	46.374475172167200	44.368591484089700



Proceso de selección del modelo

En base a la tabla comparativa, se elegir el modelo Ridge porque tenemos como criterio usar la validación cruzada con el menor valor.

Conclusiones

Se encuentra que las 5 variables que más afectan al modelo en orden de importancia son:

- trip_duration_minutes: la duración del viaje es de alrededor de 45% en importancia.
- trip_distance: la distancia en millas para llegar al destino con 35% aproximadamente.
- payment_type: el tipo o forma de pago, con menos del 10%.
- pickup_hour_in_minutes: la hora en que inicia el viaje, tiene relación con las horas pico y afecta la duración principalmente, con alrededor del 5%.
- DOLocationID: es la ubicación del destino, que tienen diferentes distancia para llegar.