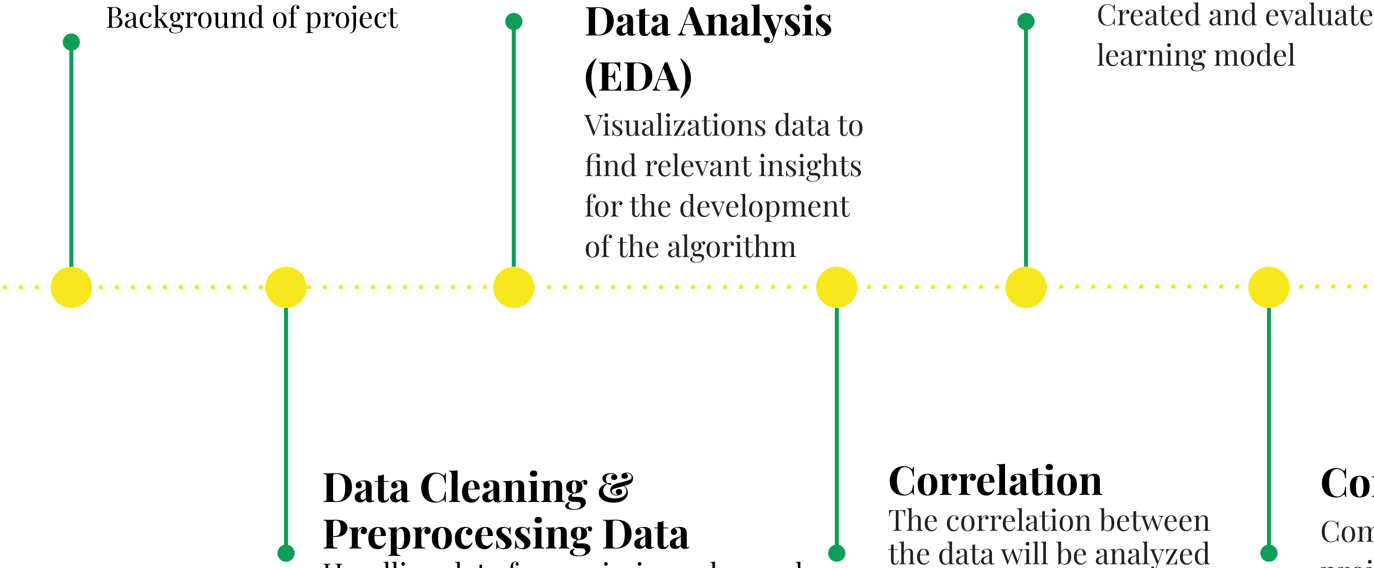
FINAL PROJECT

NYC Taxi Trip Duration

By Aditya Nur Ilman Wibowo
Batch 17

Outline



Introduction

Background of project

Data Cleaning & Preprocessing Data

Handling data from missing value and inconsistent data

Exploratory Data Analysis (EDA)

Visualizations data to find relevant insights for the development of the algorithm

Correlation

The correlation between the data will be analyzed in order to avoid possible multicollinearity problems.

Model Implementation

Created and evaluated machine learning model

Conclusion

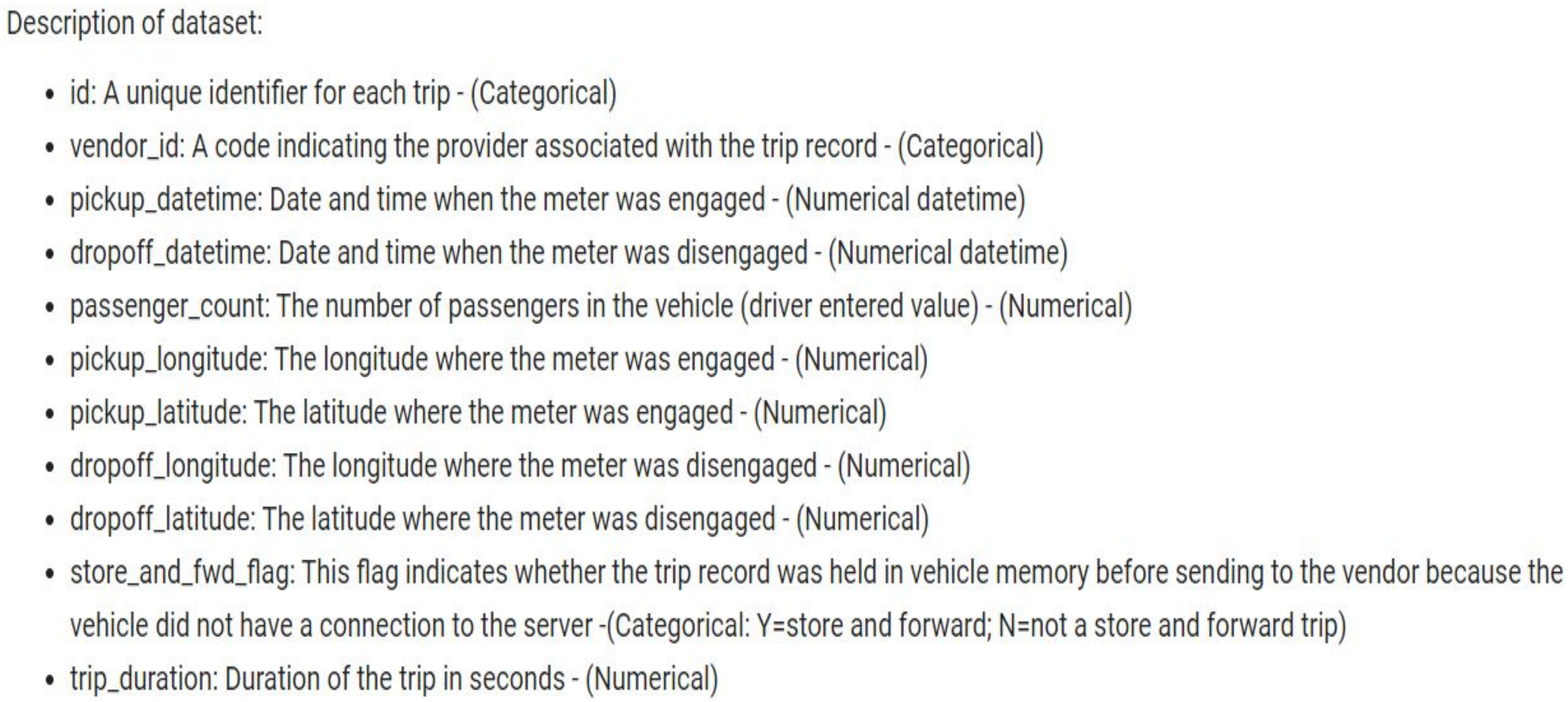
Comment of over all project and future improvment

Introduction

This primary data set was releas by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. This company wants to **develop an application to predict the duration of trips** that will be made.

Based on this, my final project will develop a **machine learning regression** algorithm capable of predicting the duration of a trip based on the variables provided by the application user and the driver

Data Understanding



Description of dataset:

- id: A unique identifier for each trip - (Categorical)
- vendor_id: A code indicating the provider associated with the trip record - (Categorical)
- pickup_datetime: Date and time when the meter was engaged - (Numerical datetime)
- dropoff_datetime: Date and time when the meter was disengaged - (Numerical datetime)
- passenger_count: The number of passengers in the vehicle (driver entered value) - (Numerical)
- pickup_longitude: The longitude where the meter was engaged - (Numerical)
- pickup_latitude: The latitude where the meter was engaged - (Numerical)
- dropoff_longitude: The longitude where the meter was disengaged - (Numerical)
- dropoff_latitude: The latitude where the meter was disengaged - (Numerical)
- store_and_fwd_flag: This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server -(Categorical: Y=store and forward; N=not a store and forward trip)
- trip_duration: Duration of the trip in seconds - (Numerical)

Data Cleaning & Preprocessing Data

Cek missing value →

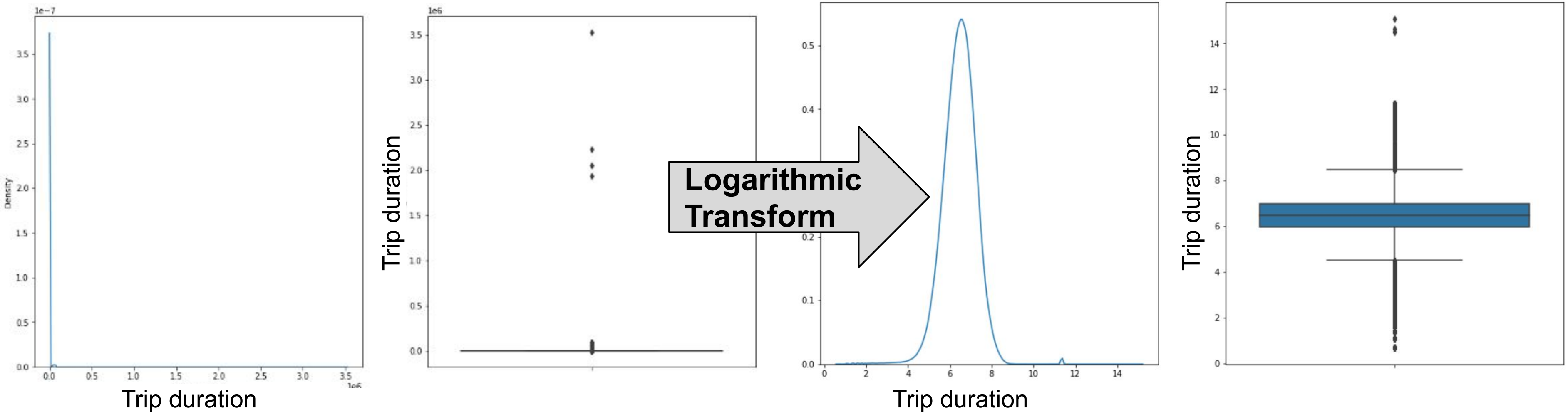
- Convert data type date time
- Creating feature distance from latitude and longitude

Resume Analyzing :

- On **average** trips have 1 to 2 **passengers**.
- The **Smallest number of passengers** is **0**, so there are some records that are missing or filled out incorrectly.
- The **highest number** of passengers carried at the same time was **9 people**, indicating that these were limousine trips with a high number of passengers.
- On **average** a **trip duration** lasts **959.49 seconds**, which is approximately **16 minutes**.
- The **smallest value** of the **duration of the trip**, the value of **1 second** is found, being an inconsistent value for a trip.
- The **longest time** a **trip** took was **3.526*10^6 seconds**, which is approximately **41 days**, representing inconsistent data.
- On **average the trip** has a distance of **3.94 km**.
- The **Nearest travel distance** values equal to **0**, indicating that a trip did not occur or that the destination of the trip was the same as the departure.
- The **longest trip** had a distance of **1243.4 km**, which is a very high value, which could indicate a recording error.
- The **highest value** of 'month_pickup' is **6** and the **lowest** is **1**, indicating that the data are from the months between **January** and **June** of the analyzed year.

Data Cleaning & Preprocessing Data

Trip Duration Analysis

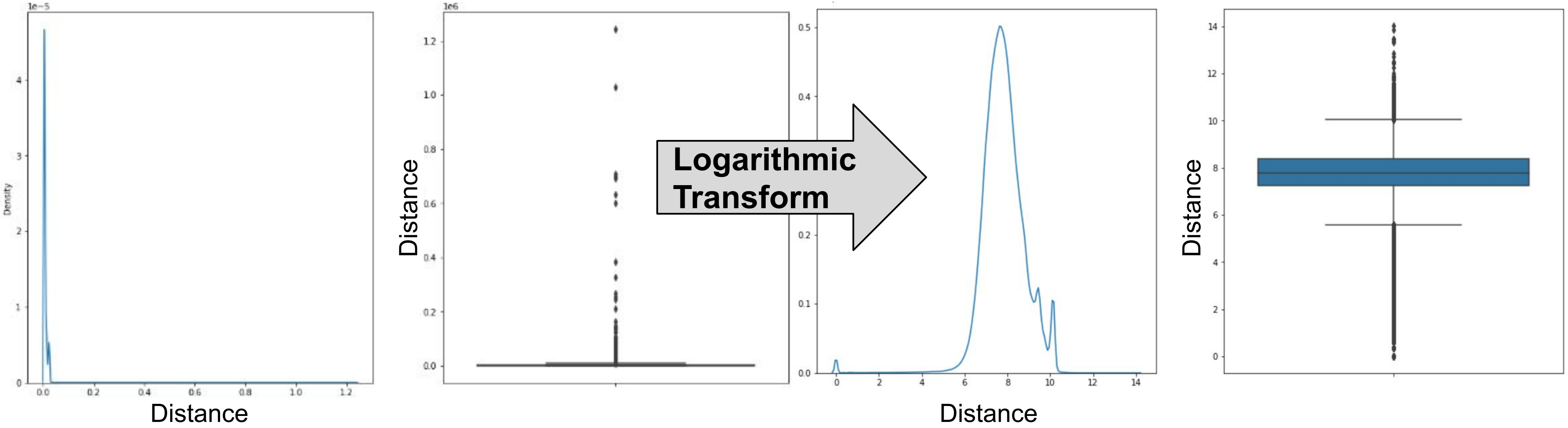


Through the logarithmic representation, the graph starts to have a format more similar to a normal distribution, having the **highest concentration** of data between the values of **4 and 9** (representing values between **55 and 8100 seconds**). There is also a **small density peak** between values of **11 and 12** on the logarithmic scale (values between **59900 and 162800 seconds**), which are intuitively very high values for a trip. There are **small ripples** for values less than **3** on the logarithmic scale (travel times less than **20 seconds**), which are inconsistent values of travel time.

In order to build a more robust model, data with trip_duration **above 59900 seconds** and **less than 20 seconds** will be **removed**.

Data Cleaning & Preprocessing Data

Distance Analysis

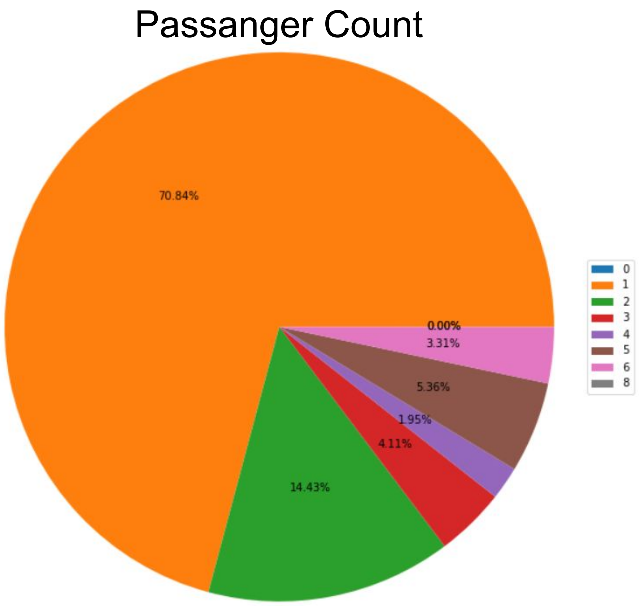


In the density graph, a certain similarity to a normal distribution can be noted, being a little different for logarithmic distance values between **9 and 10**, where there is a peak of concentrations at the same distance.

There are distance data close to or **equal to 0** (approximately 1 meter away). These are very low values for a trip by car, so they probably must have been storage or typing errors by the user when requesting the trip, and can be considered outliers.

Based on this, distance values less than 20 meters will be removed.

Data Cleaning & Preprocessing Data

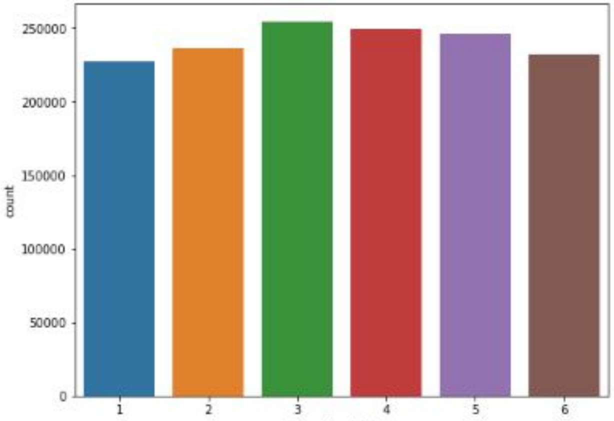


It can be seen from the graph that **more than 90%** of the data are from trips with **up to 3 passengers**, with **almost 71%** of trips made with **one passenger**.

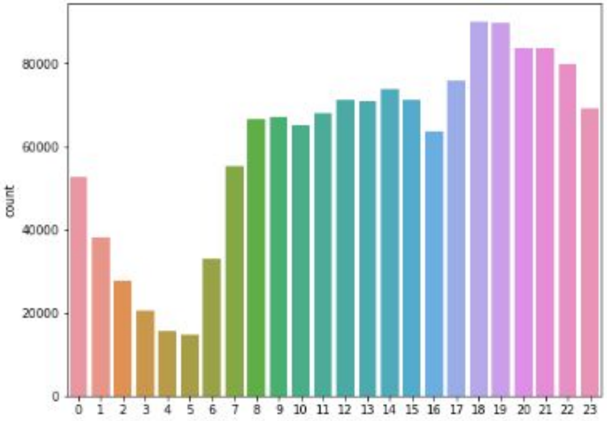
The data where initially there were **9 passengers** per trip were **removed** in previous processes.

It is observed the presence of data with number of **passengers equal to 0**, but in a very small amount. As the amount of data is small and it is an inconsistent value, this data will be **removed**.

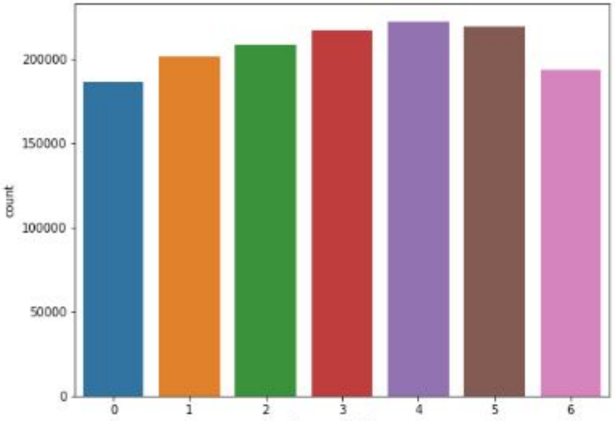
Exploratory Data Analysis



month_pickup



hour_pickup



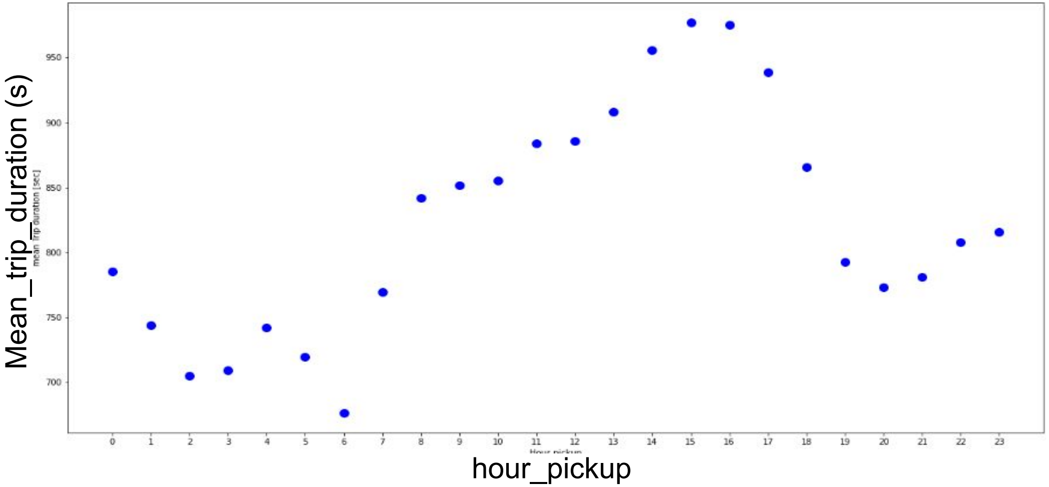
day_week_pickup

From the graph of the **months**, it can be seen that all months have values close to the amount of data, with the **lowest value** in **January** and **the highest** in **March**.

From the **boarding time chart**, it can be seen that **few trips** take place between **0:00 am and 5:00 am** and that **many trips** take place between **6:00 pm and 10:00 pm**.

From the **day week pickup graph**, it is possible to observe that the days of the week do not have a relevant difference in the amount of data. **Monday and Sunday** are the days with the **fewest trips** and **Saturday and Friday** are the days with the **most trips**.

Exploratory Data Analysis



From the graph it is possible to see that **the trips made** between **1 pm and 5 pm** have an **average duration** greater than **900 seconds**, being the period that has an average greater than all other times. This may indicate that this period is the time of greatest traffic in this city.

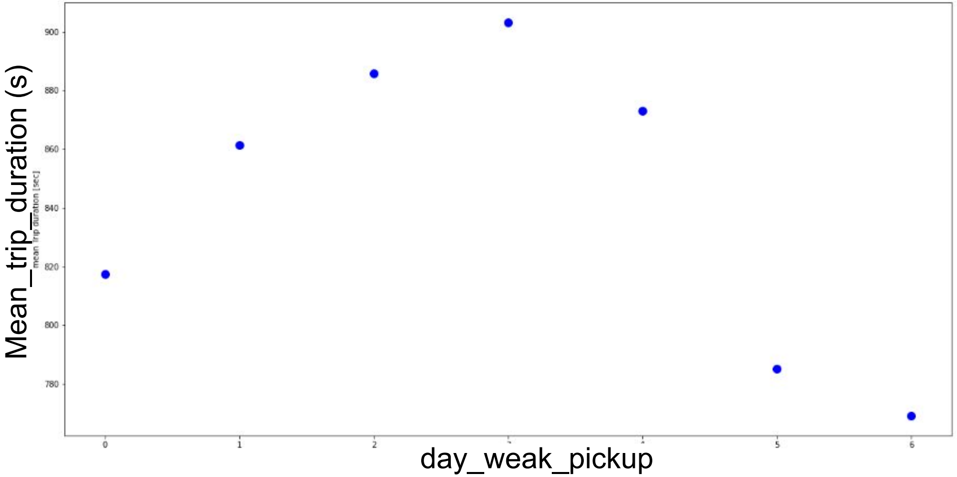
Trips made between **2 am and 6 am** have the **lowest average trip duration**.

Trips made between **23:00 and 05:00** seem to have a **decreasing relation of trip duration** with the passage of time.

Between the period from **6:00 am to 1:00 pm** there is a **gradual increase** in the **average duration** of the trip.

Between the period from **17:00 to 19:00** there is a **gradual decrease** in the **average duration** of the trip.

Exploratory Data Analysis



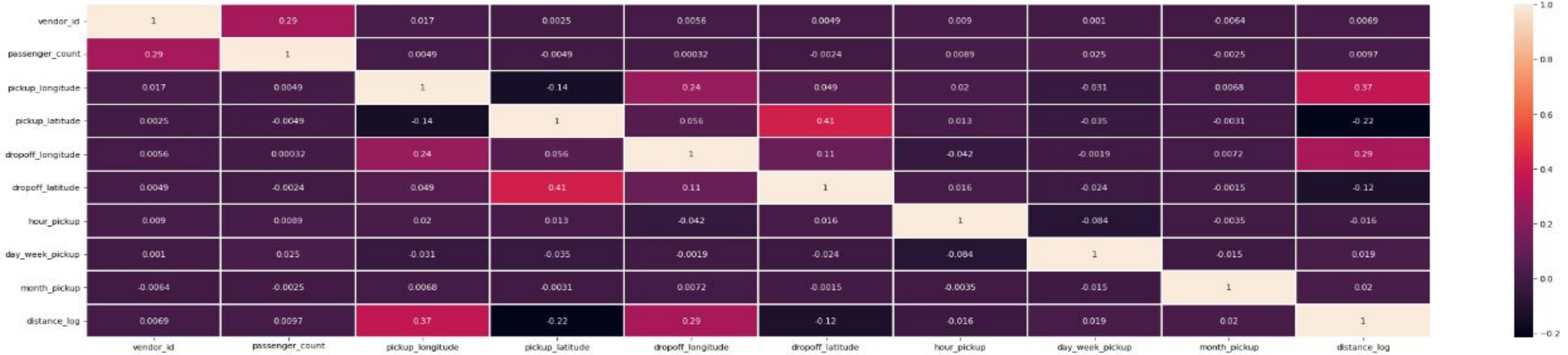
It is possible to identify an **increasing relationship** between **the trip duration average** and **the days from Monday to Thursday**.

It is possible to identify a **decreasing relationship** between **the trip duration average** and **the days from Thursday to Sunday**.

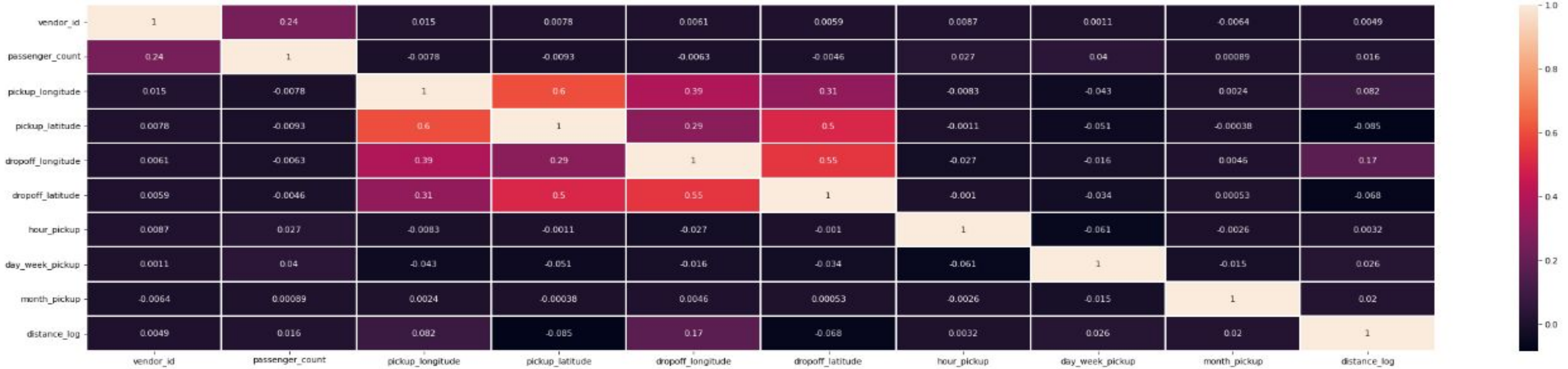
The days with the **lowest average trip duration** values are the **weekends**. This could be because the weekends are less busy or because on weekends the taxi and limousine systems are more used at off-peak times.

Correlation

In order to **avoid multicollinearity problems**, two types of correlation will be evaluated for each feature pair: **Pearson's correlation** and **Spearman's rank correlation**. As there is not a very clear literature for **'acceptable' correlation values** that avoid multicollinearity, values **greater than 0.8 of correlation** will be defined as **cutoff values to be removed**



	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	hour_pickup	day_week_pickup	month_pickup	distance_log
vendor_id	1	0.29	0.017	0.0025	0.0056	0.0049	0.009	0.001	-0.0064	0.0069
passenger_count	0.29	1	0.0049	-0.0049	0.00032	-0.0024	0.0089	0.025	-0.0025	0.0097
pickup_longitude	0.017	0.0049	1	-0.14	0.24	0.049	0.02	-0.031	0.0068	0.37
pickup_latitude	0.0025	-0.0049	-0.14	1	0.056	0.41	0.013	-0.035	-0.0031	-0.22
dropoff_longitude	0.0056	0.00032	0.24	0.056	1	0.11	-0.042	-0.0019	0.0072	0.29
dropoff_latitude	0.0049	-0.0024	0.049	0.41	0.11	1	0.016	-0.024	-0.0015	-0.12
hour_pickup	0.009	0.0089	0.02	0.013	-0.042	0.016	1	-0.084	-0.0035	-0.016
day_week_pickup	0.001	0.025	-0.031	-0.035	-0.0019	-0.024	-0.084	1	-0.015	0.019
month_pickup	-0.0064	-0.0025	0.0068	-0.0031	0.0072	-0.0015	-0.0035	-0.015	1	0.02
distance_log	0.0069	0.0097	0.37	-0.22	0.29	-0.12	-0.016	0.019	0.02	1

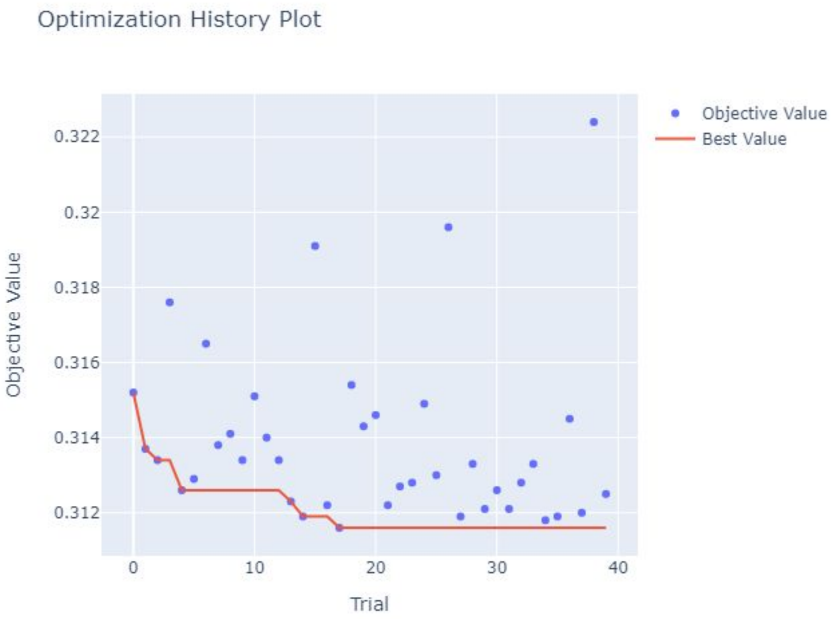


	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	hour_pickup	day_week_pickup	month_pickup	distance_log
vendor_id	1	0.24	0.015	0.0078	0.0061	0.0059	0.0087	0.0011	-0.0064	0.0049
passenger_count	0.24	1	-0.0078	-0.0093	-0.0063	-0.0046	0.027	0.04	0.00089	0.016
pickup_longitude	0.015	-0.0078	1	0.6	0.39	0.31	-0.0083	-0.043	0.0024	0.082
pickup_latitude	0.0078	-0.0093	0.6	1	0.29	0.5	-0.0011	-0.051	-0.00038	-0.085
dropoff_longitude	0.0061	-0.0063	0.39	0.29	1	0.55	-0.027	-0.016	0.0046	0.17
dropoff_latitude	0.0059	-0.0046	0.31	0.5	0.55	1	-0.001	-0.034	0.00053	-0.068
hour_pickup	0.0087	0.027	-0.0083	-0.0011	-0.027	-0.001	1	-0.061	-0.0026	0.0032
day_week_pickup	0.0011	0.04	-0.043	-0.051	-0.016	-0.034	-0.061	1	-0.015	0.026
month_pickup	-0.0064	0.00089	0.0024	-0.00038	0.0046	0.00053	-0.0026	-0.015	1	0.02
distance_log	0.0049	0.016	0.082	-0.085	0.17	-0.068	0.0032	0.026	0.02	1

Model Implementation [LGBM Regressor]

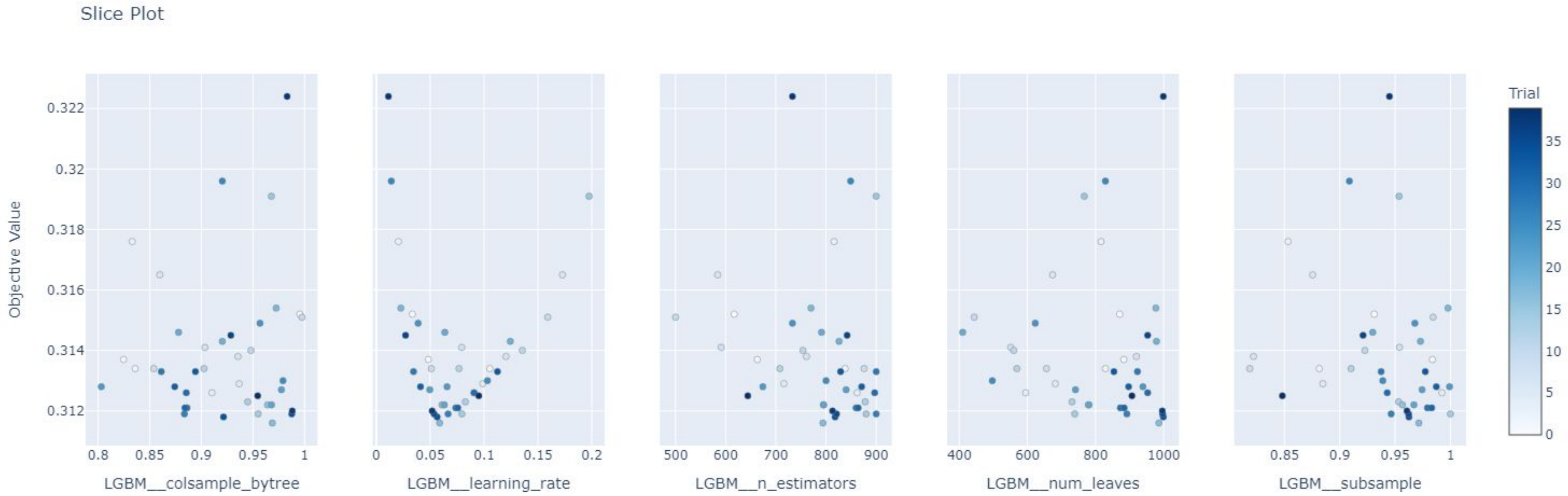
As a regression model, **LightGBM** will be used, being a machine learning model that has a great performance for solving regression problems. Two evaluations will be made in it, one with the **standard model** and another improving the **hyperparameters** of this model, aiming to **increase the accuracy** of the model.

RMSE mean : 455.96
RMSLE mean: 0.35493



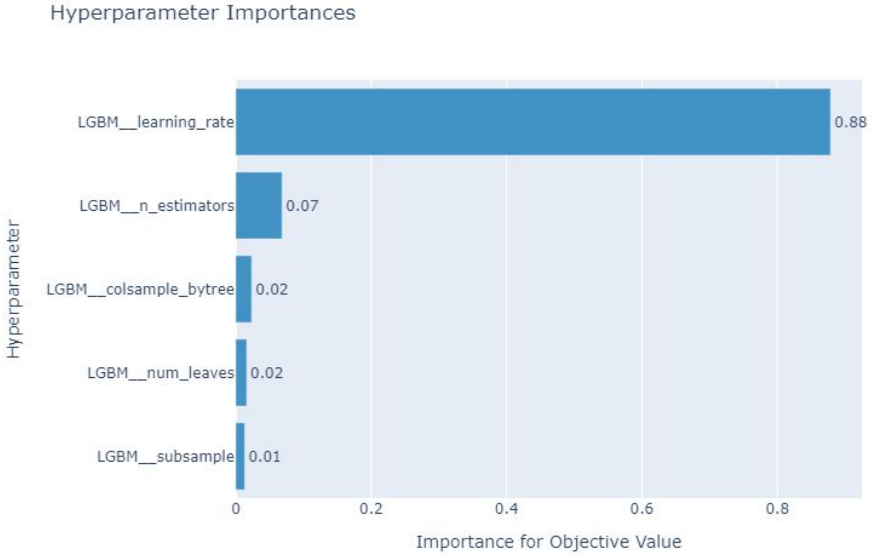
With the optimization process, an improvement of approximately **0.04 in the RMSLE value is observed**, showing that the optimization was significant. It is also observed that the final result obtained was **0.3117**, which is a little higher than what was initially defined.

Model Implementation [LGBM Regressor]



From the parameter variation graphs, no variable tended to the limits initially established as the optimization occurred, excluding the variable 'colsample_bytree', whose maximum possible value is 1. This is an **indication that the optimization was carried out correctly and no parameter needs to be expanded** in one new optimization to further improve the performance of the model.

Model Implementation [LGBM Regressor]



It is observed that the 'learning_rate' variable was the variable that added the most value to improve the performance of the model. The other variables like 'subsample', 'n_estimators' and 'num_leaves' had a smaller impact and could be removed in order to reduce the optimization time.

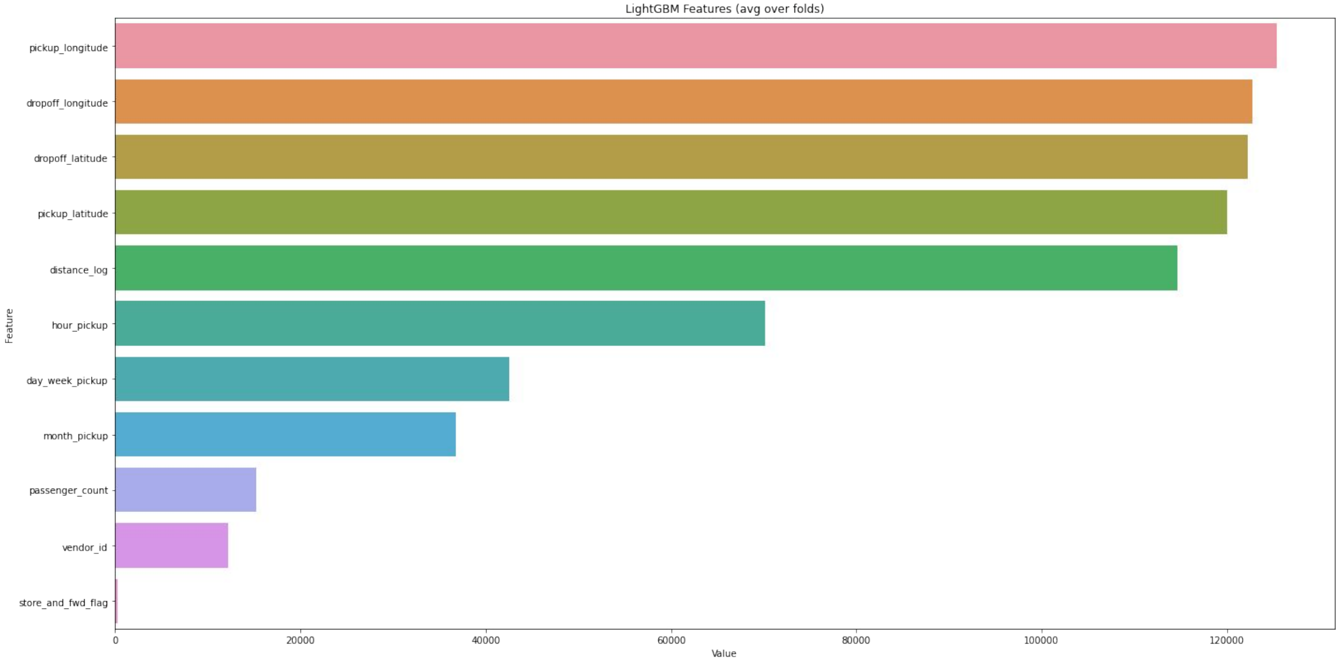
Model Implementation [LGBM Regressor]

RMSE train : 400.61

RMSLE train: 0.24583

In this case, all data for model training is used to increase the performance of the generated algorithm, thus increasing the results with the final test data. Note that there was an improvement of approximately 0.06 in the RMSLE value compared to the results with cross-validation.

Model Implementation [LGBM Regressor]



It is noted by the feature importance that the coordinates of pickup and dropoff together with the distance in log are the most relevant variables for the prediction of the model. The variable store_and_fwd_flag has the least relevance for model prediction, showing that it could be removed without loss of model prediction.

Conclusion

The result obtained from the algorithm was close to the expected result and it is expected that when applied to the test data similar results will be obtained.

In order to improve the project in future work, the following activities can be carried out:

- Removal of Features that add less information to the model, verifying the performance increase.
- Perform a better cleaning on the pickup and dropoff coordinates data, since they have greater predictive power.
- Creation of clusters for pickup and dropoff coordinates, aiming to create alternative variables to increase the predictive power.
- Application of other regression models for performance comparison.