“Dynamic pricing of Taxi Fare”

Project Leadership

Sharat Manikonda
Director at Innodatatics and Sponsor
linkedin.com/in/sharat-chandra

Hrishikesh
From IBM

Team Members

Name: MOHANRAJ S
linkedin.com/in/mohanraj-s-6a59a5142

Name: DINESH T
https://www.linkedin.com/in/dinesh-t-299554172

Contents

Contents
Project Overview and scope
Business Problem
Objective and Constraints
CRISP-ML-(Q) Methodology
Technical Stacks
Project Architecture
Data Collection and Understanding
Data Information
Data Dictionary
System Requirements
Exploratory Data Analysis

Contents

Contents
EDA Description
Missing Values Observation
Data Visualization
Model Building
Model Accuracy Comparison
Best Model
Model Deployment - Strategy
Screenshot of Output
Video of Output
Challenges
Future Scopes

Project Overview and Scope

- Design and develop a system that makes intelligent taxi fare decisions. The organization, driver partners, and passengers should benefit from this system's capacity to predict the right taxi fee in response to shifting driving conditions.

- The scope involves Data collection, Data preprocessing, EDA, supervised models, Deployment and maintenance and monitoring.

Business Problem

- A transportation (taxi) company has approached your team to design and create a system that intelligently decides on the taxi fare.
- This system should help the company, driver partners and customers by predicting the appropriate taxi fare depending on changing driving conditions.

Objectives and Constraints

Objective

- Maximize strategy for dynamic pricing
- Maximize profits
- Maximize accuracy

Constraints

- Minimize Customer churn
- Minimize Computational Burden

CRISP-ML(Q) Methodology

1.Business Understanding:

- Business problem: A transportation (taxi) company has approached your team to design and create a system that intelligently decides on the taxi fare.
- This system should help the company, driver partners and customers by predicting the appropriate taxi fare depending on changing driving conditions.
- Business objective: Maximize strategy for dynamic pricing, Maximize profits, Maximize accuracy.
- Business constraints: Minimize customer churn, Minimize computational burden.

2.Data Understanding:

- Collect and analyze data.
- Identify relevant features for dynamic pricing.
- Target variable is Total amount and it is continuous in nature.

3.Data Preparation:

- Clean and preprocess the data.
- Typecasting, Handling Duplicates, Handling Missing Values and Outliers removal.
- Did Feature Engineering and Transformed the data into a suitable format for modeling.

CRISP-ML(Q) Methodology

4.Modeling:

- Used machine learning algorithm OLS, Lasso, Ridge, Elastic net, decision tree, ANN and compare those models to find the best fit.

5.Evaluation:

- Compare model performance with existing approaches.
- Cross validation, Hyper tune parameters are used to check the robustness and scalability of the data.
- Evaluate model performance using appropriate metrics such as R square, Adjusted R square, RMSE, pvalue, MAE.

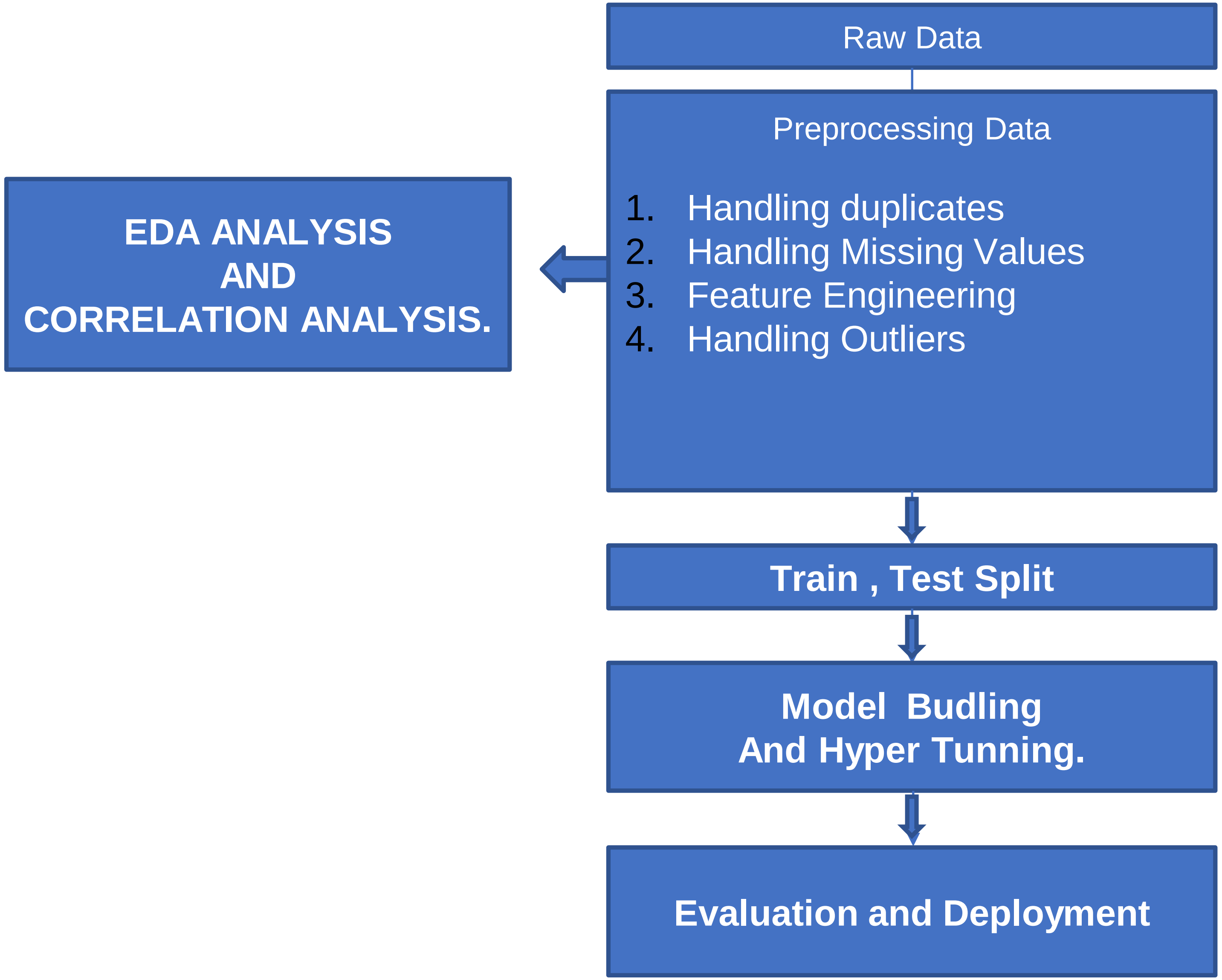
6.Deployment:

- Deployment done using flask and created webpage using HTML,CSS.
- Monitor model performance and update as necessary to ensure continued effectiveness.

Technical Stacks

- Anaconda IDE: Spyder
- Database Management: MySQL
- Python Libraries: NumPy, pandas, matplotlib, seaborn, feature_engine, sklearn, joblib, pickle, sweetviz, model selection, sqlalchemy, mysql, statsmodels and keras.

Project Architecture



Raw Data

Preprocessing Data

1. Handling duplicates
2. Handling Missing Values
3. Feature Engineering
4. Handling Outliers

EDA ANALYSIS
AND
CORRELATION ANALYSIS.

Train , Test Split

Model Budling
And Hyper Tunning.

Evaluation and Deployment

Supervised Learning

Data Collection and Understanding

- Data collected from data. World.
- Did Typecasting, Handling Duplicates, Handling Missing Values and Outliers removal.
- Did feature engineering and feature selection (Relevant inputs to the current business problem).

Preprocessing:

- Checked for duplicates, 0.2% of duplicate records found then, removed those duplicate values.

- Checked for missing values, 0.8% of missing values found then, removed those missing values.

- Passenger_count column contains a value of 0.0 and it is invalid if 0 passengers have booked the taxi.1.8% of records got dropped.

- Ratecode_ID contains category of code from 1 to 6. It also contains 99.0 it doesn’t make any sense. 0.004% of records got dropped.

Data Collection and Understanding

Feature Engineering:

- ➢ Calculated time difference in minutes by using trip pickup time and drop off time columns. Therefore, we found a new feature time difference(Minutes) and dropped trip pickup time and drop off time.
- ➢ Converted trip distance in miles to meter because trip distance in miles approaching near zero value.

- ➢ Making new feature New_total_amount by adding features like total_amount and congestion_surcharge(congestion surcharge is not included in the total amount before. It is also paid by customer. Hence we are including). Since total _amount is a sum of fare amount, extra, mta tax, trip amount, tolls amount, improvement surcharge. Hence, dropping these columns. In this way we combined columns together and we made 8 columns into single column New_total_amount. The shape of the final data is (6339567 x 8)

- ➢ Outliers found in trip_distance, time_difference_minutes, New_total_amount and those are removed using winsorization and built in a pipeline.

Data Information

- The dataset tells about yellow taxi trip data in the U.S
- The dataset consists of 6,405,008 records and 18 columns.
- Features involved in the dataset VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge.
- Did Exploratory Data Analysis and the insights are documented.

Data Dictionary

Name of Feature	Description	Data Type	Relevance
VendorID	A code indicating the TPEP provider that provided the record. **1 = Creative Mobile Technologies, LLC; 2 = VeriFone Inc.**	**Numerical(float)**	**Relevant**
tpep_pickup_datetime	The date and time when the meter was engaged.	**Object**	**Relevant**
tpep_dropoff_datetime	The date and time when the meter was disengaged.	**Object**	**Relevant**
passenger_count	The Number of passengers in the vehicle. This is a driver entered value	**Numerical(float)**	**Relevant**
trip_distance	The elapsed trip distance in miles reported by the taximeter.	**Numerical(float)**	**Relevant**
RatecodeID	The final rate code in effect at the end of the trip. **1 = Standard rate** **2= JKF** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare** **6=Group ride**	**Numerical(float)**	**Relevant**

Data Dictionary

Name of Feature	Description	Data Type	Relevance
store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**	**Object**	Relevant but No variance in the dataset
PULocationID	TLC Taxi Zone in which the taximeter was engaged	**Numerical(integer)**	**Relevant**
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged	**Numerical(integer)**	**Relevant**
payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**	**Numerical(float)**	**Relevant**
fare_amount	The time-and-distance fare calculated by the meter.	**Numerical(float)**	**Relevant**
extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.	**Numerical(float)**	**Relevant**

Data Dictionary

Name of Feature	Description	Data Type	Relevance
mta_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.	**Numerical(float)**	**Relevant**
tip_amount	This field is automatically populated for credit card tips. Cash tips are not included.	**Numerical(float)**	**Relevant**
tolls_amount	Total amount of all tolls paid in trip.	**Numerical(float)**	**Relevant**
improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015	**Numerical(float)**	**Relevant**
total_amount	The total amount charged to passengers. Does not include cash tips.	**Numerical(float)**	**Relevant**
congestion_surcharge	Total amount collected in trip for NYS congestion surcharge.	**Numerical(float)**	**Relevant**

System Requirements

- Operating system: Windows 10 or higher, macOS 10.14 or higher
- Processor: Intel Core i3 or higher
- Memory (RAM): 4 GB or higher
- Storage: At least 250 GB free hard disk space
- Software: Anaconda, MYSQL
- Internet connection: Mobile internet connection with computer for accessing the internet

Exploratory Data Analysis [EDA]

- Data visualization: Using charts, graphs, histograms, and other visual tools to explore the data and identify patterns and trends.
- Summary statistics: Calculating descriptive statistics such as mean, median, mode, variance, and standard deviation to summarize the data.
- Correlation analysis: Examining the relationship between two or more variables in the data.

EDA Description

EDA is a crucial step in the data analysis process where analysts examine and visualize the data to gain insights and understand the underlying patterns, trends, relationships, and anomalies in the data. EDA helps us to understand the data and identify potential problems or errors that may impact the accuracy and validity of the results.

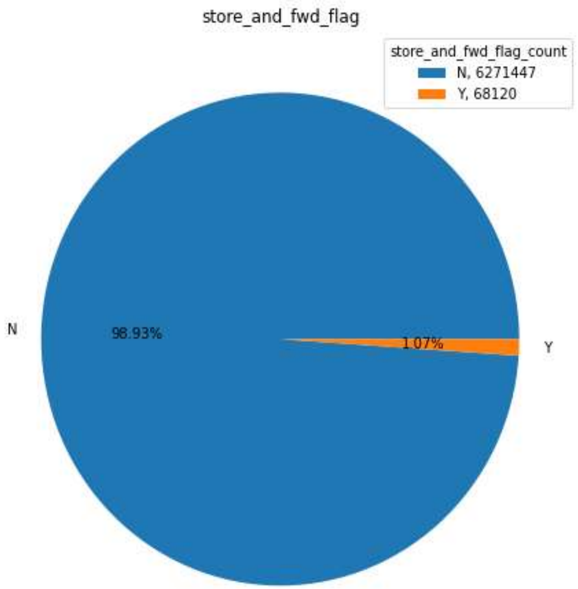
EDA involves various techniques and tools such as data visualization, summary statistics, correlation analysis and data cleaning. The objective of EDA is to answer questions about the data and generate new questions that can be investigated using more advanced statistical techniques.

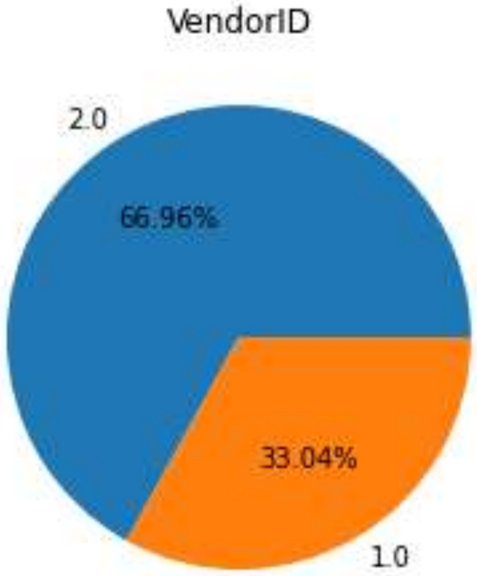
Missing Values Observation

```
VendorID                    52492
tpep_pickup_datetime     0
tpep_dropoff_datetime   0
passenger_count           52492
trip_distance                0
RatecodeID                 52492
store_and_fwd_flag       52492
PULocationID               0
DOLocationID               0
payment_type               52492
fare_amount                 0
extra                           0
mta_tax                       0
tip_amount                   0
tolls_amount                 0
improvement_surcharge  0
total_amount                 0
congestion_surcharge     0
```

There are missing values in the category type of data . If we do imputation the correlation will change accordingly. Hence, we simply drop this missing values. We are dropping 0.8% of the over all data.

Data Visualization





TPEP Provider VeriFone Inc has give majority of Records 66.96%

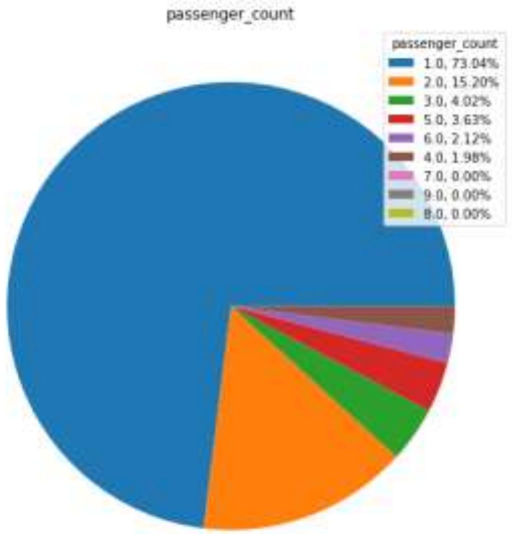
The rest provider is Creative Mobile Technologies 33.04%

The majority values in the column is No.

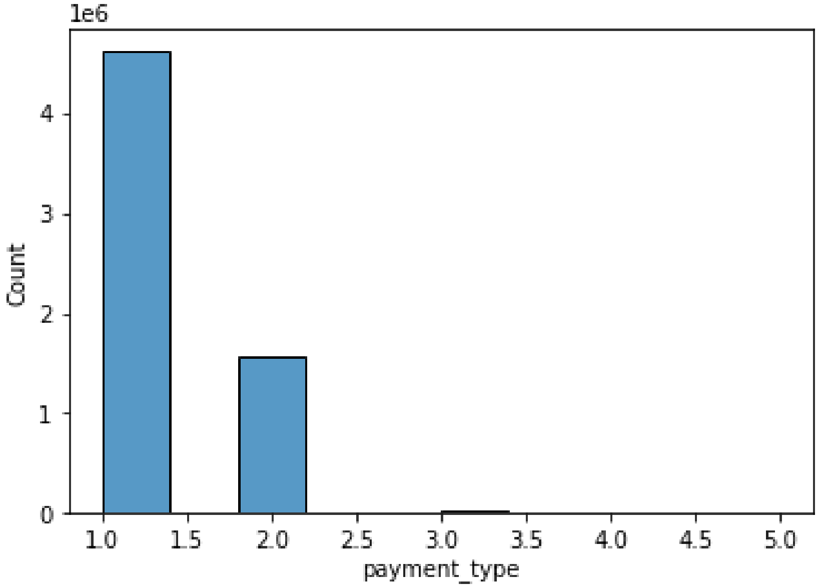
- The connection to server working fine. So, automatically trip data's are recorded.

The majority of entries are same so we remove the column.

Data Visualization



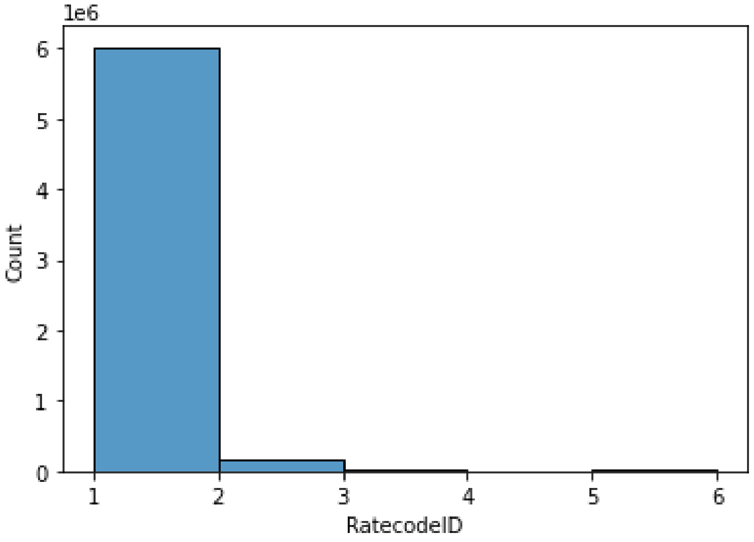
In passenger_count the proportion of single Customer booking a cab is 73.04%.
Two Customers travelling percentage is 15.20%
More than 2 customers travelling together is rare

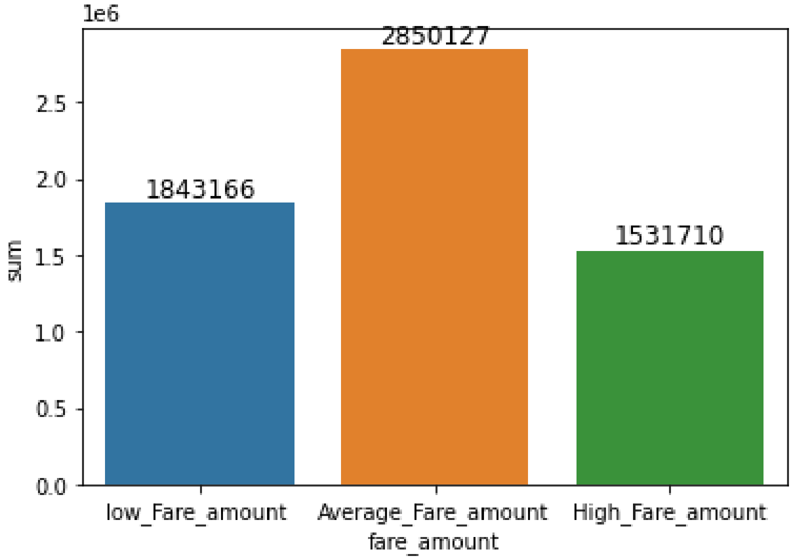


payment_type Credit card and Cash are used most often by customers.

The other methods like no charge, dispute, unknown and void trip are least used

Data Visualization

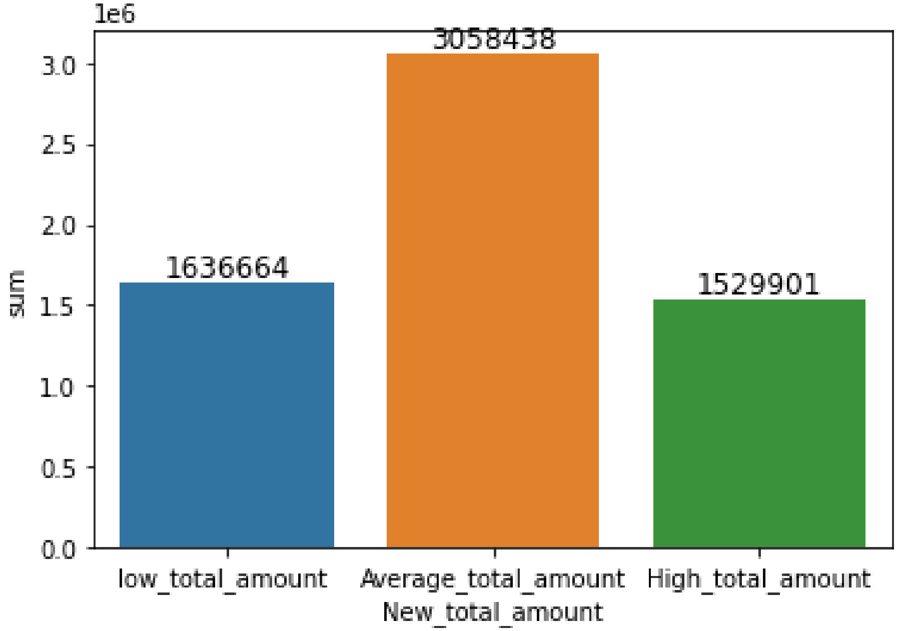


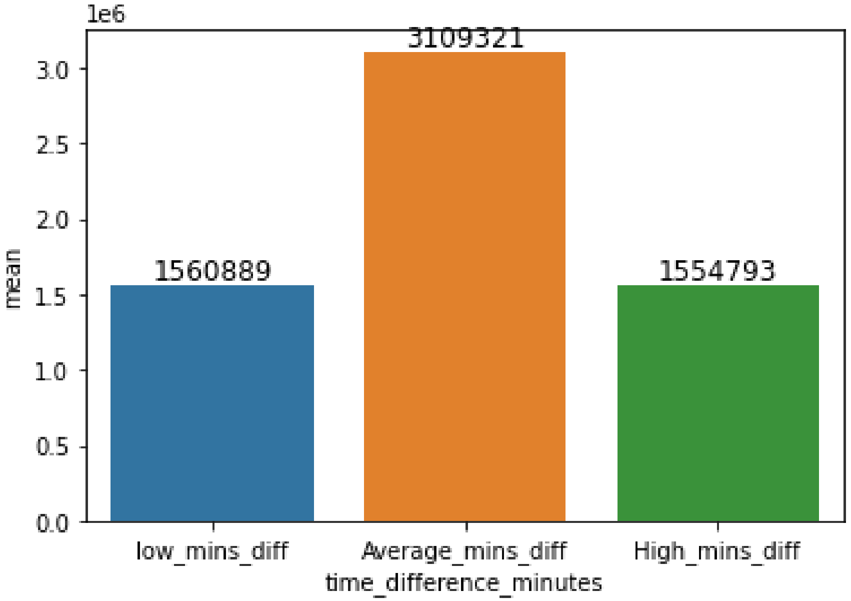


- Most of the people prefer credit Card over other Mode of payment.

- Average Fare amount generated by taxi company is high.

Data Visualization

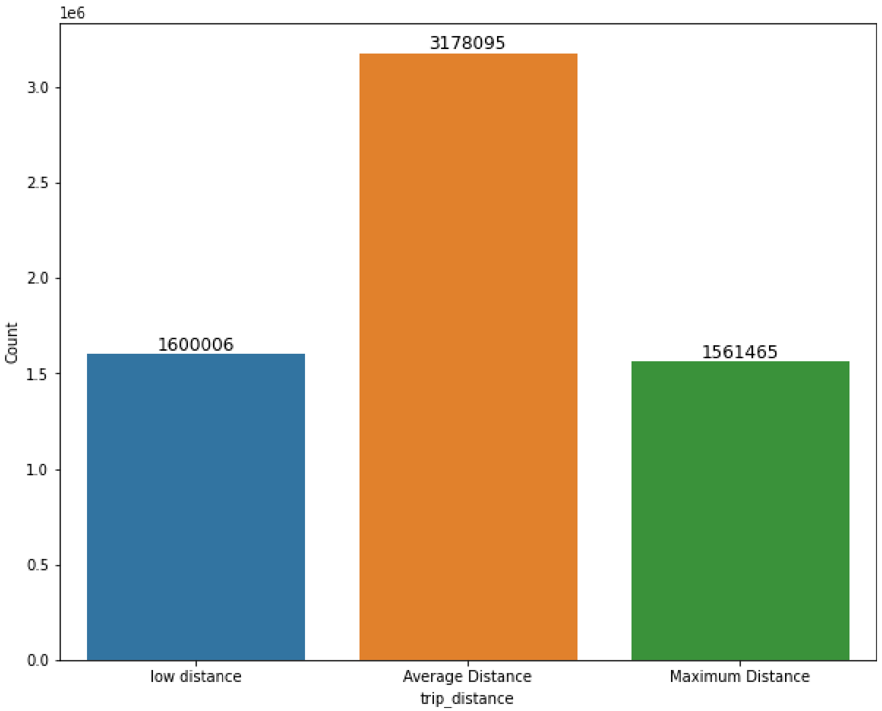




- Average Total amount which includes all kind of taxes generated by taxi company is high.

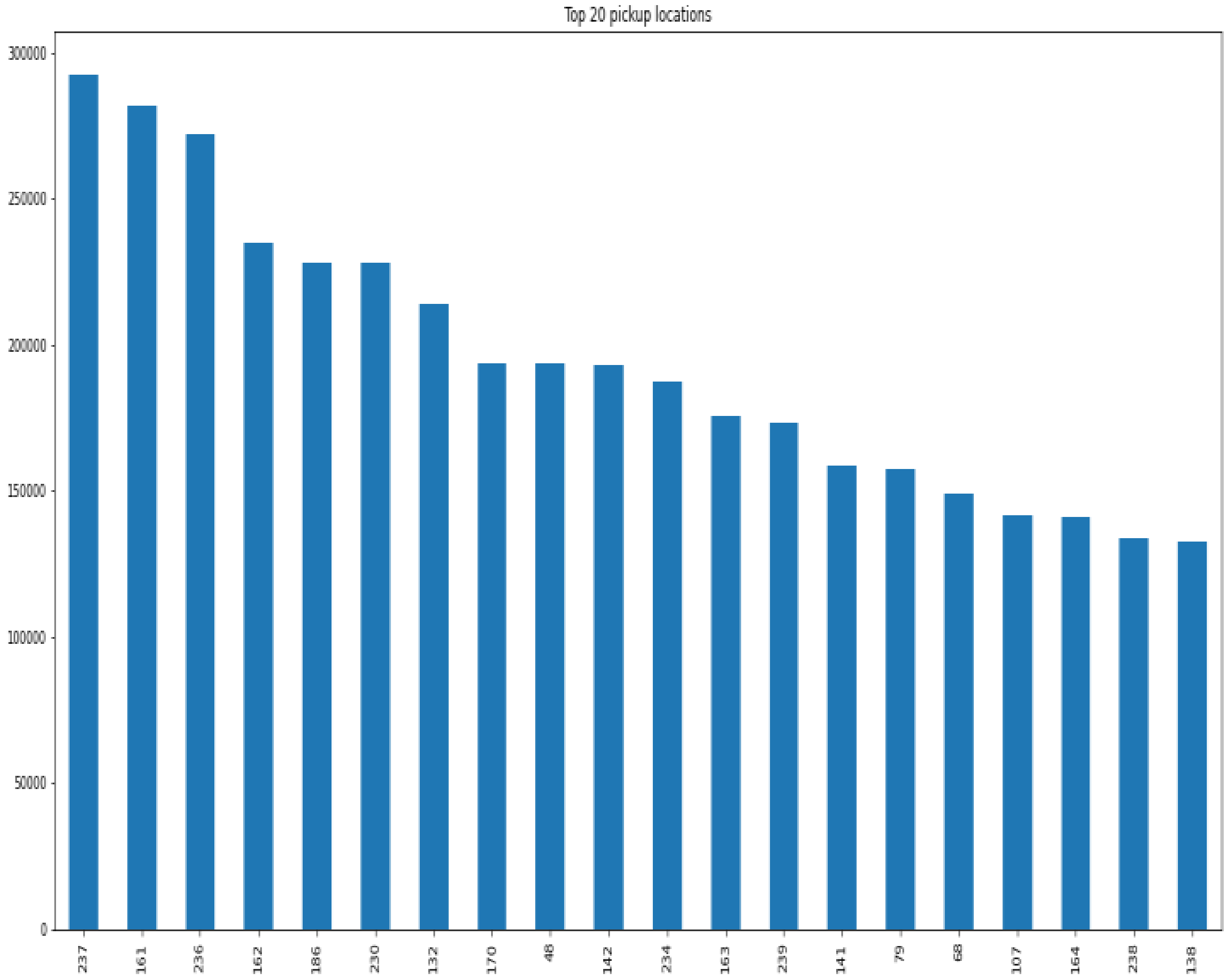
- Average Minutes spend by customer while traveling is high.

Data Visualization



- Average Distance Travelled by taxi company is high.

Data Visualization

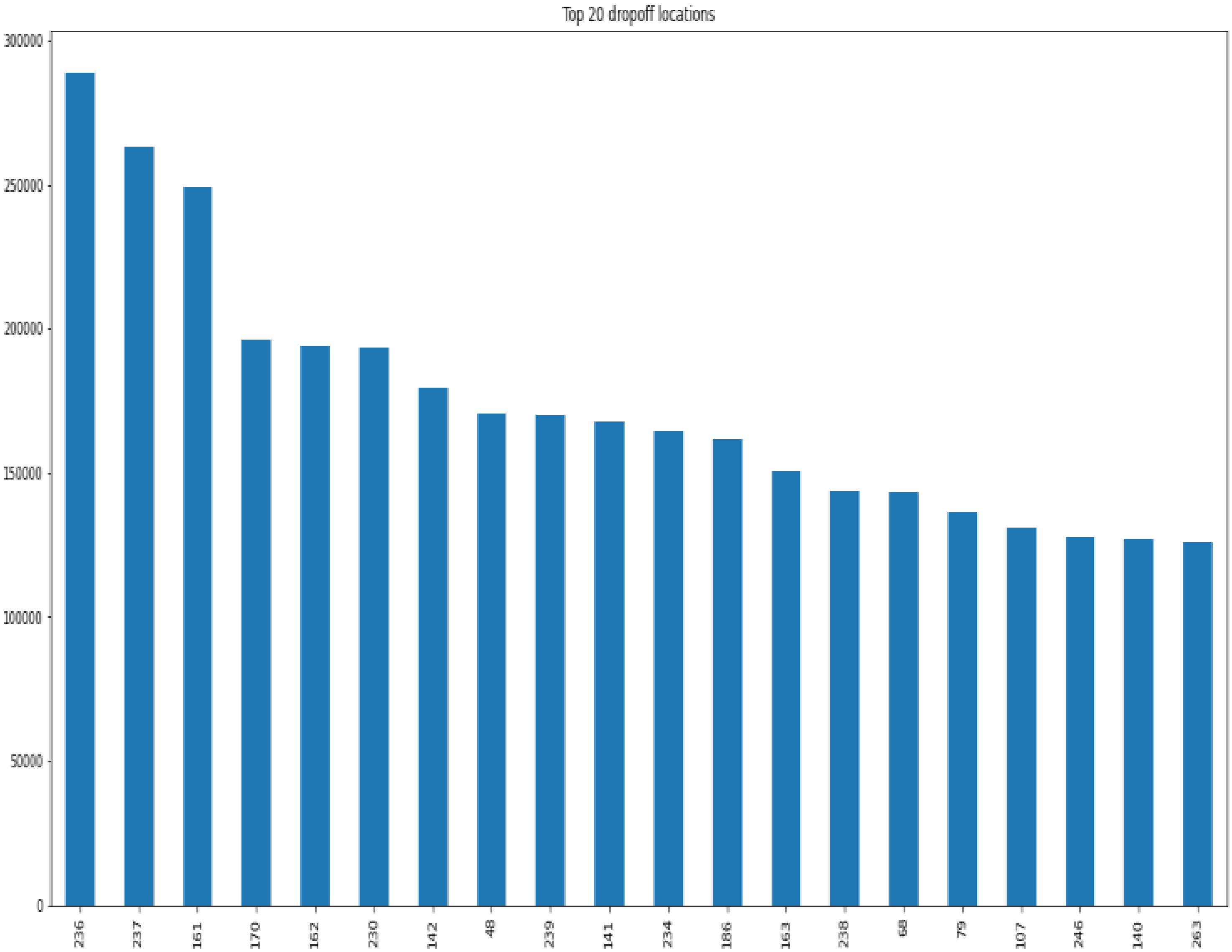


For sample:
According to TLC's documentation, PULocationID 237 corresponds to the "Upper East Side North" taxi zone,
which includes the area bounded by East 96th Street to the north, Central Park to the west, East 59th Street to the south, and the East River to the east.

161 refers to the "Crown Heights North" taxi zone.
This zone is located in Brooklyn and is roughly bounded by Atlantic Avenue to the north,
Utica Avenue to the east, Eastern Parkway to the south, and Washington Avenue to the west.
It includes several notable landmarks such as the Brooklyn Museum, the Brooklyn Botanic Garden, and the Brooklyn Children's Museum.

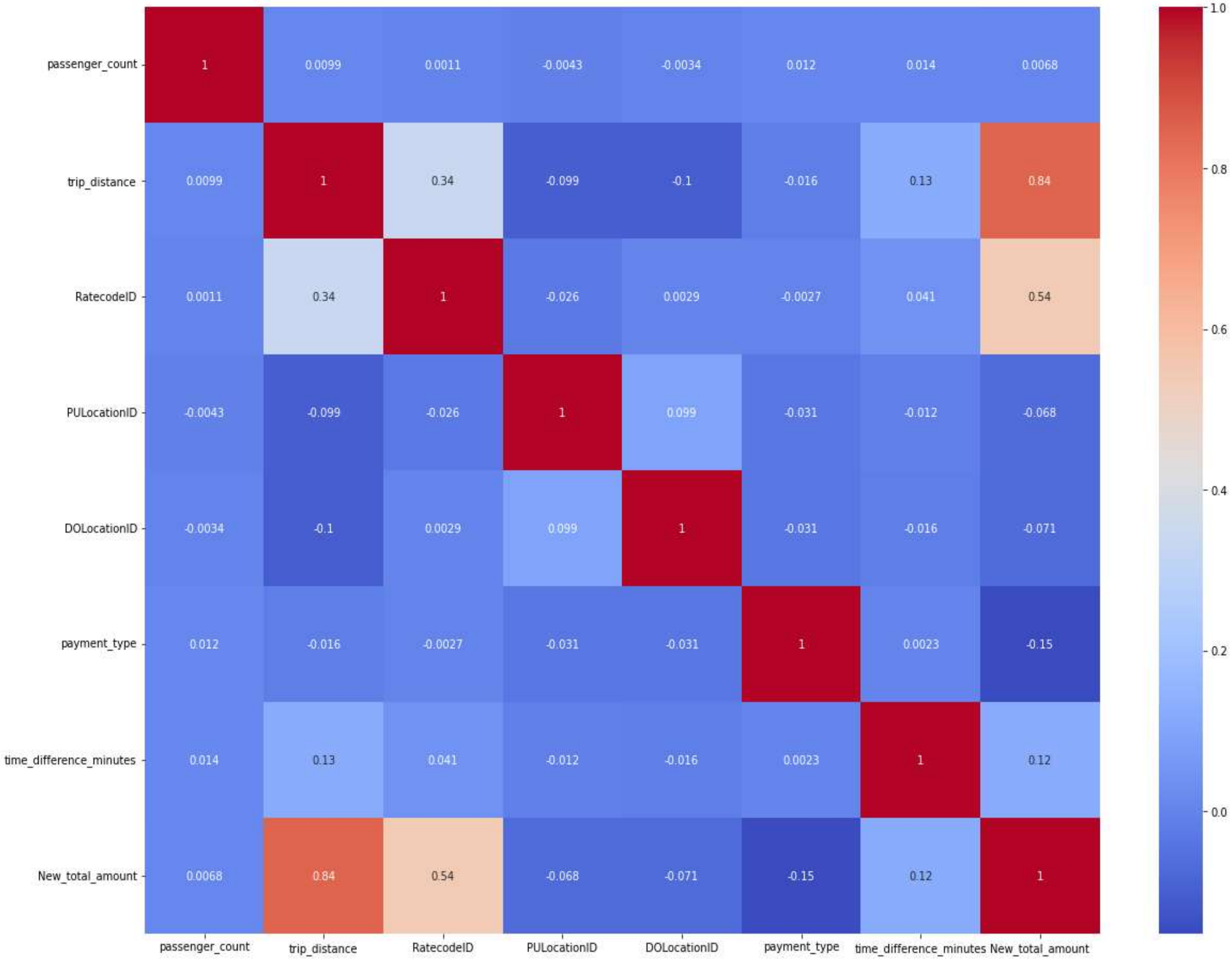
INNODATATICS
Innovation • Data • Analytics

Data Visualization



For sample:
236 refers to the "Upper East Side South" taxi zone. This zone is located in Manhattan and is roughly bounded by East 59th Street to the north,
Fifth Avenue to the west, East 42nd Street to the south, and the East River to the east. It includes several notable landmarks such as the Plaza Hotel, Bloomingdale's, and Central Park.

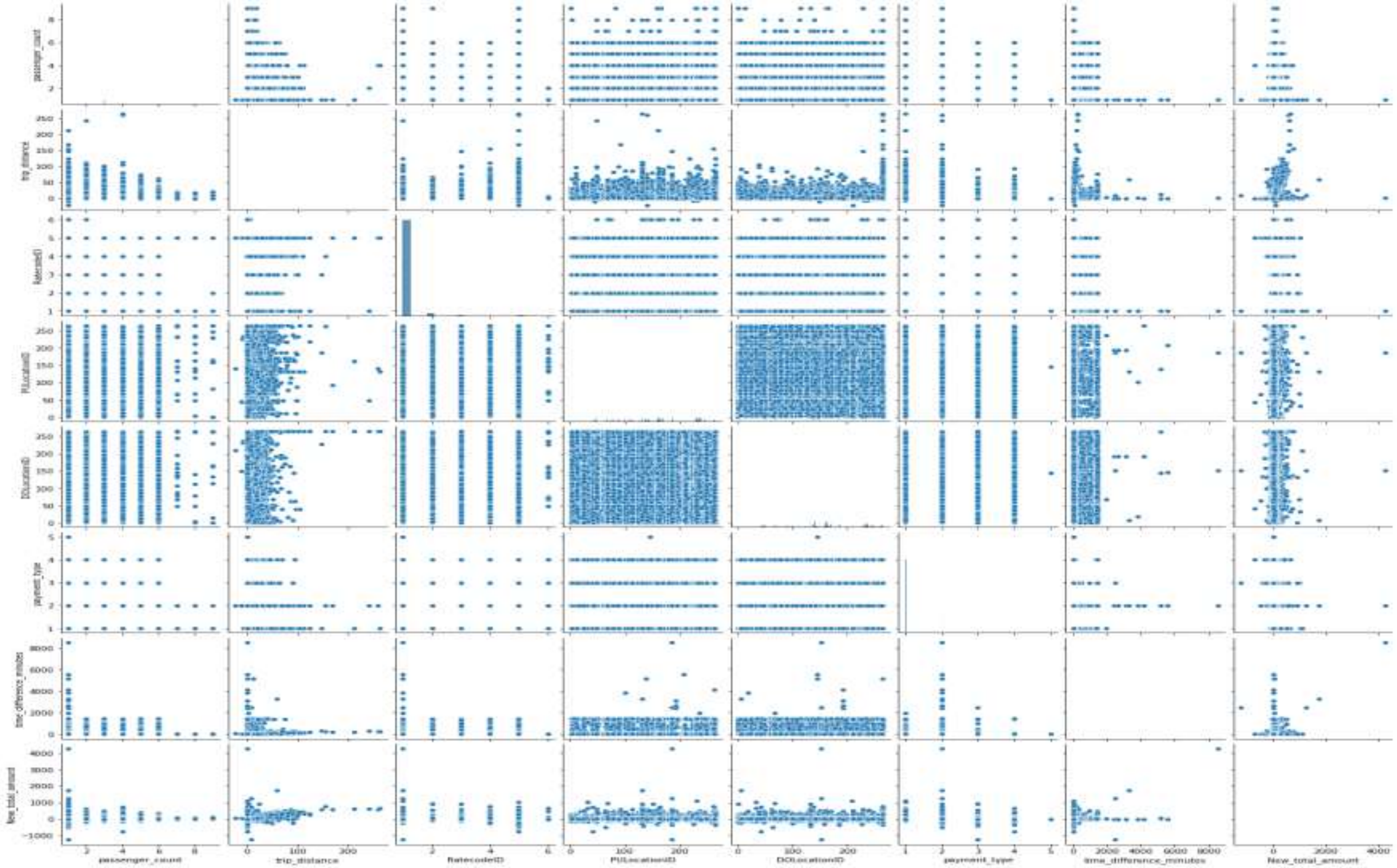
Data Visualization



- New total amount and trip distance having correlation of 0.84.The correlation is strong or positive correlation.
- New total amount and rateID code having correlation of 0.54. The correlation is moderate.
- Rate ID code and trip distance having slight correlation 0.34.

Data Visualization

Pair plot



Innovation • Data • Analytics

Model Building –

ORDINARY LEAST SQUARE (OLS):

- OLS helps us to find the relationship between dependent and independent variable.

- The OLS method estimates the coefficients of the regression equation by minimizing the sum of squared differences between the observed values and the predicted values of the dependent variable.

- The resulting coefficients provide estimates of the strength and direction of the relationship between the dependent variable and each independent variable, as well as the overall predictive power of the model.
- P_values are 0 for all the features which is less than 0.05 (Industrial Standard).
- R2 value and Adjusted R2 value are equal which is 0.98 in summary().
- After transformation Model is not working well.

Model Building –

Lasso:

- Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization method used in linear regression analysis to prevent overfitting and improve the accuracy of the model.
- The purpose of Lasso is to select the subset of predictor variables that are most relevant to the response variable, by penalizing the absolute size of the coefficients of the linear regression model.
- This penalty shrinks the coefficients towards zero, and can effectively set some coefficients to exactly zero, leading to a simpler and more interpretable model that only includes the most important predictor variables.
- Performed Gridsearch CV for hypertuning and obtained best estimator for prediction

Model Building –

Ridge:

- The purpose of Ridge regression is similar to that of Lasso regression, which is to prevent overfitting by shrinking the coefficients of the linear regression model. However, Ridge regression shrinks the coefficients by adding a penalty term to the sum of squared errors, which is proportional to the square of the magnitude of the coefficients.

- This penalty term is known as the L2 norm and effectively reduces the size of the coefficients, but does not set any of them to zero.

- By shrinking the coefficients towards zero, Ridge regression reduces the variance of the model and improves its stability, especially when the number of predictor variables is large and the correlation among the predictors is high

- Performed Gridsearch CV for hypertuning and obtained best estimator for prediction

Model Building –

Ridge:

- The purpose of Ridge regression is similar to that of Lasso regression, which is to prevent overfitting by shrinking the coefficients of the linear regression model. However, Ridge regression shrinks the coefficients by adding a penalty term to the sum of squared errors, which is proportional to the square of the magnitude of the coefficients.

- This penalty term is known as the L2 norm and effectively reduces the size of the coefficients, but does not set any of them to zero.

- By shrinking the coefficients towards zero, Ridge regression reduces the variance of the model and improves its stability, especially when the number of predictor variables is large and the correlation among the predictors is high

- Performed Gridsearch CV for hypertuning and obtained best estimator for prediction

Model Building –

Elastic Net:

- Elastic net is a combination of Lasso (L1 Norm) and Ridge (L2 Norm)

- Performed Gridsearch CV for hypertuning and obtained best estimator for prediction

Decision Tree:

- The decision tree is constructed by recursively partitioning the feature space into smaller regions based on the values of the predictor variables, with the goal of minimizing the residual sum of squares.

- Each terminal node of the tree represents a prediction of the response variable based on the predictor variables within that region. The resulting model is interpretable and can be used for prediction, inference, and feature selection.

- We used this hyper parameters (min_samples_split=60, min_samples_leaf=3)

Model Building –

Artificial Neural Network(ANN) :

- ANNs is to learn from data through a process of training, in which the algorithm adjusts its weights and biases to minimize the difference between its predictions and the actual values.
- This training process allows ANNs to model complex relationships between input data and output predictions, and to generalize to new, unseen data.
- The main advantage of ANNs is their ability to model complex, nonlinear relationships between input and output data, making them well-suited for problems with large and high-dimensional datasets.
- Activation function used relu and linear.
- Loss function used mean square error and optimizer used adam

Model Building –

K Nearest Neighbour (KNN) :

- The purpose of KNN Regressor is to predict the numerical value of a new data point based on the average value of the k nearest neighbors in the training dataset.

- KNN Regressor is a supervised machine learning algorithm that operates by calculating the distance between the new data point and each point in the training dataset. It then selects the k closest neighbors based on the distance metric, and computes the average of their target values.

- This average value is then assigned as the predicted value for the new data point.

Model Accuracy Comparison

Models	Test r2_score	Train r2_score	Test mae	Train mae	Test rmse	Train rmse
OLS	0.84	0.84	1.62	1.62	2.65	2.65
Lasso	0.88	0.88	1.25	1.25	2.28	2.28
Ridge	0.88	0.88	1.25	1.25	2.28	2.28
Elastic net	0.88	0.88	1.25	1.25	2.28	2.28
Decision Tree	0.93	0.95	0.92	0.8	1.67	1.46
ANN	0.88	0.89	1.21	1.2	2.38	2.43
KNN	0.88	0.92	5.11	3.3	2.26	1.81

Best Model –

- After comparing metrics of the models such as Ordinary Least Square, Lasso, ridge, Elastic net, Decision tree, Artificial Neural Network and K- Nearest Neighbours.

- After hype tuning the respective parameters for the respective models, we obtained decision tree as the right fit model.

- We obtained best **R2 score(0.93), Mean Absolute Error(0.92) and Root Mean Squared Error(1.67)** for decision tree among those models.

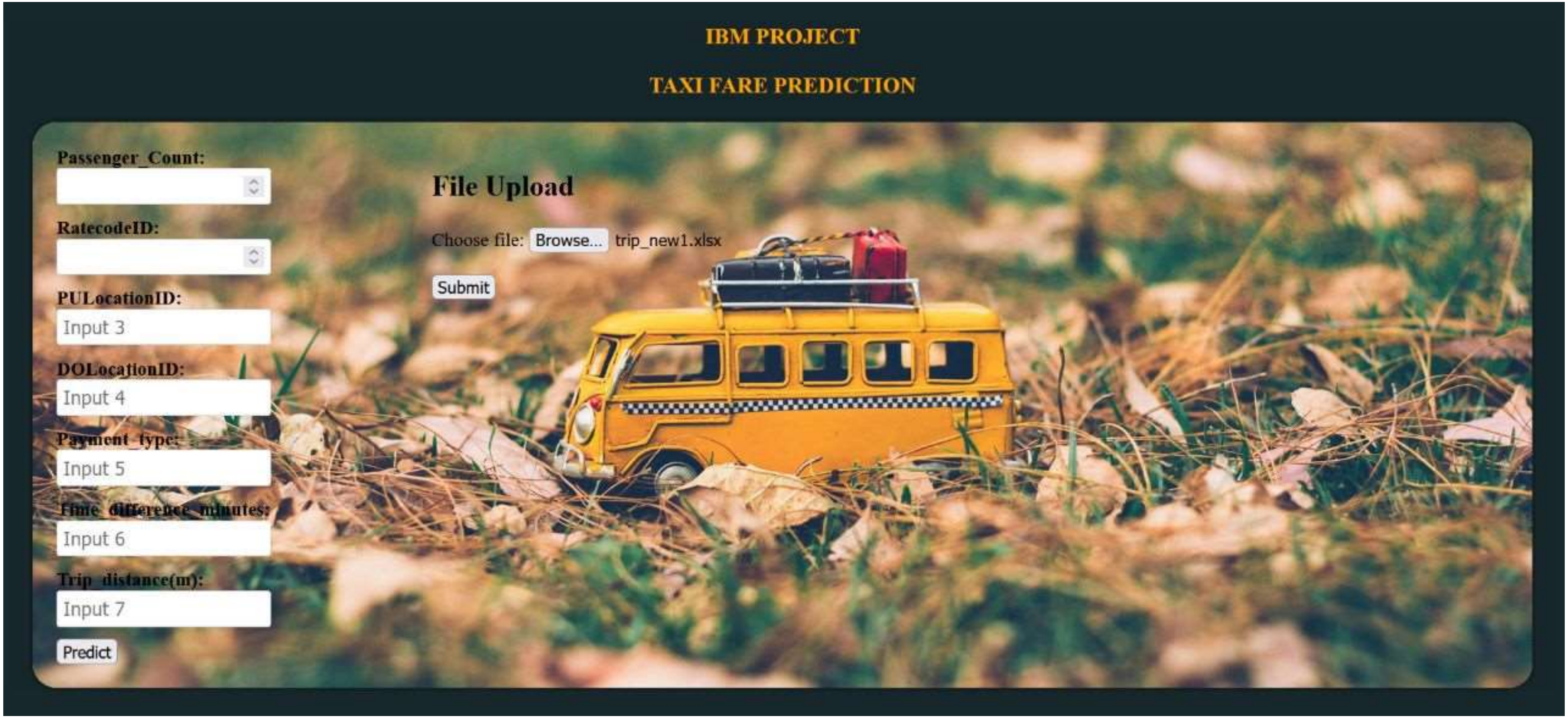
Model Deployment - Strategy

Flask is a Python-based micro framework used for developing small-scale websites. Flask is very easy to make Restful APIs using python. As of now, we have developed a model i.e model.pkl, which can predict a class of the data based on various attributes of the data.

Now we will design a web application where the user will input all the attribute values and the data will be given to the model, based on the training given to the model, the model will predict the drug group based on inputs like Age, Gender, Season and condition of patients.

Create script.py file in the project folder and copy the following code.
Here we import the libraries, then using app=Flask(__name__) we create an instance of flask.
@app.route('/') is used to tell flask what URL should trigger the function index() and in the function index, we use render_template('index.html') to display the script index.html in the browser.
Let's run the application.

Screen shot of output



IBM PROJECT

TAXI FARE PREDICTION

Passenger_Count:

RatecodeID:

PULocationID: Input 3

DOLocationID: Input 4

Payment_type: Input 5

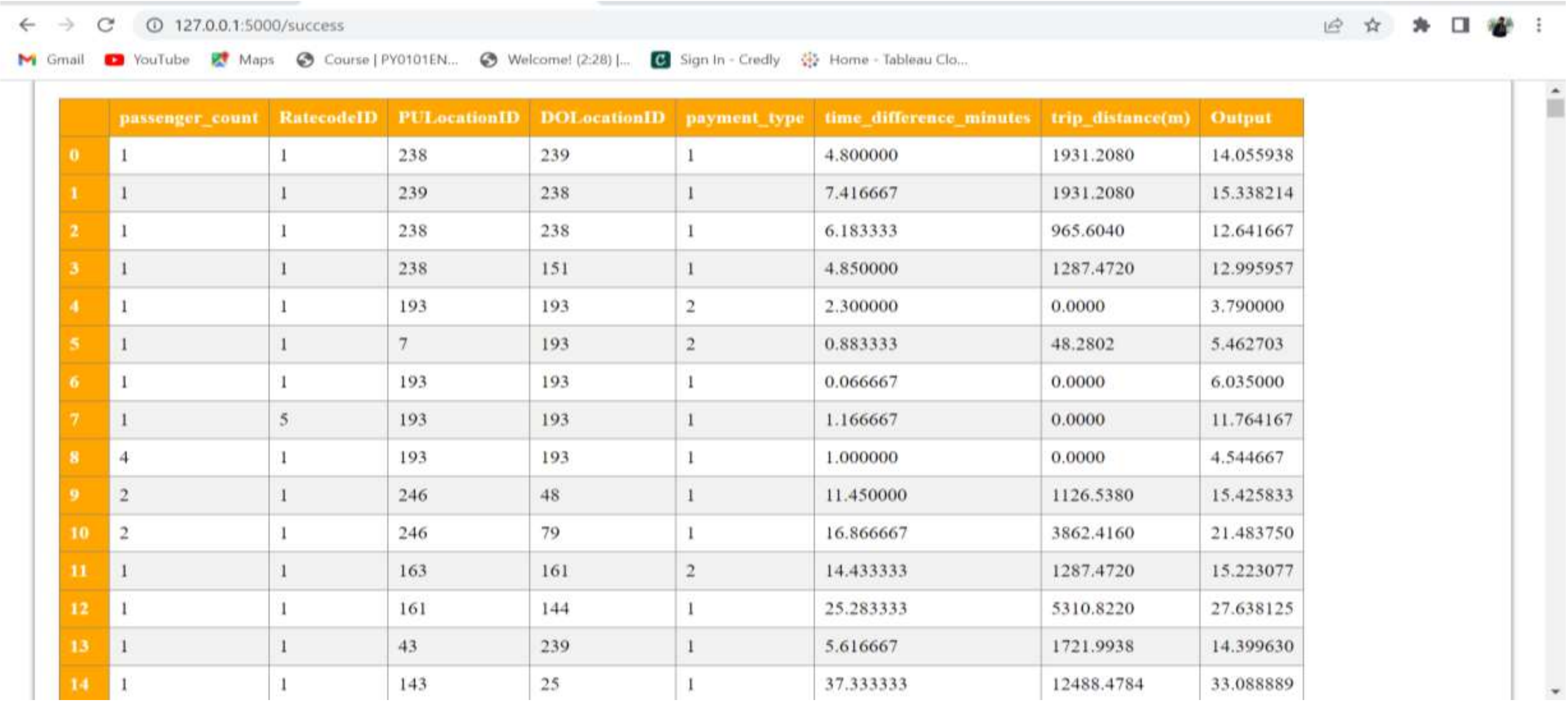
Time_difference_minutes: Input 6

Trip_distance(m): Input 7

File Upload

Choose file: trip_new1.xlsx

Screen shot of output



	passenger_count	RatecodeID	PULocationID	DOLocationID	payment_type	time_difference_minutes	trip_distance(m)	Output
0	1	1	238	239	1	4.800000	1931.2080	14.055938
1	1	1	239	238	1	7.416667	1931.2080	15.338214
2	1	1	238	238	1	6.183333	965.6040	12.641667
3	1	1	238	151	1	4.850000	1287.4720	12.995957
4	1	1	193	193	2	2.300000	0.0000	3.790000
5	1	1	7	193	2	0.883333	48.2802	5.462703
6	1	1	193	193	1	0.066667	0.0000	6.035000
7	1	5	193	193	1	1.166667	0.0000	11.764167
8	4	1	193	193	1	1.000000	0.0000	4.544667
9	2	1	246	48	1	11.450000	1126.5380	15.425833
10	2	1	246	79	1	16.866667	3862.4160	21.483750
11	1	1	163	161	2	14.433333	1287.4720	15.223077
12	1	1	161	144	1	25.283333	5310.8220	27.638125
13	1	1	43	239	1	5.616667	1721.9938	14.399630
14	1	1	143	25	1	37.333333	12488.4784	33.088889

Challenges

- We face computational burden because of huge volume of data

- We didn’t do the scaling because of duplicates in the input data while splitting into train and test

- The reason for this issue is majority of data containing discrete values as inputs. The probability of repeated values in input features are high.

- Limited number of parameters used for hyper tuning of most of the models because of large volume of data and time consumption.

Future Scopes

- The data is analyzed through CRISP-DM methodology with the aim to build a dynamic price prediction model that predicts optimal prices that can balance both margins as well as conversion rates.

- As there are millions of taxi trips recorded every year, implementing this project in a big data platform and integrating with customer records is the scope of future study.

- Also, some customers cancel the bookings, therefore the study can further be extended in predicting the taxi booking confirmation or rejection from a customer after generating the dynamic price for the taxi trips.

- Creating responsive websites for all devices.

Queries ?

Thanks