# FOUNDATION OF COMPUTING ASSIGNMENT -02

NUTHAN REDDY VADDI REDDY

UTA ID:1002175133

**DATASET:** This dataset contains information on taxi trips in New York City over a specific period, with features such as pickup time, drop-off time, trip duration, passenger count, and geographical coordinates

**TASK 1 Data Pre-processing:**

1.1 Provide summary statistics for the dataset:

By using the df.describe() function we get the complete summary statistics of the data like including the count, mean, standard deviation, minimum, 25th percentile (first quartile), median (50th percentile or second quartile), 75th percentile (third quartile), and maximum for each column.

| | vendor_id | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | trip_duration |
|---|---|---|---|---|---|---|---|
| count | 100000.000000 | 100000.000000 | 100000.000000 | 100000.000000 | 100000.000000 | 100000.000000 | 100000.000000 |
| mean | 1.537760 | 1.668260 | -73.973466 | 40.750920 | -73.973436 | 40.751833 | 944.415930 |
| std | 0.498575 | 1.320623 | 0.042385 | 0.031886 | 0.046791 | 0.036578 | 3020.231686 |
| min | 1.000000 | 0.000000 | -78.547401 | 36.029301 | -79.817978 | 36.029301 | 1.000000 |
| 25% | 1.000000 | 1.000000 | -73.991783 | 40.737460 | -73.991257 | 40.736018 | 398.000000 |
| 50% | 2.000000 | 1.000000 | -73.981659 | 40.754230 | -73.979721 | 40.754757 | 663.000000 |
| 75% | 2.000000 | 2.000000 | -73.967237 | 40.768375 | -73.963150 | 40.769993 | 1074.000000 |
| max | 2.000000 | 6.000000 | -73.553223 | 41.256882 | -73.363937 | 41.256889 | 86360.000000 |

1.2 Check the data types of the features:

By using df.dtypes function we get datatype of each feature (or) column of the dataset, we can see that there are 4 object data types,4 float data types and 2 int datatypes in the data.

```
id                   object
vendor_id             int64
pickup_datetime      object
dropoff_datetime     object
passenger_count       int64
pickup_longitude     float64
pickup_latitude      float64
dropoff_longitude    float64
dropoff_latitude     float64
store_and_fwd_flag    object
trip_duration          int64
dtype: object
```

1.3 Handle missing or erroneous data:

By using df.isnull().sum() we get the number of null values in each column. In this dataset there are no null values in all the columns.

```
id                   0
vendor_id            0
pickup_datetime      0
dropoff_datetime     0
passenger_count      0
pickup_longitude     0
pickup_latitude      0
dropoff_longitude    0
dropoff_latitude     0
store_and_fwd_flag   0
trip_duration        0
dtype: int64
```

1.4 Convert categorical data to numerical values where necessary:

I am converting the categorical data column store_and_fwd_flag into numerical values by replacing the N and Y into 0 and 1 by using df['store_and_fwd_flag'].replace( {'N': 0, 'Y': 1}) method.

| :time | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | store_and_fwd_flag | trip_duration |
|---|---|---|---|---|---|---|---|
| )5-13 50:11 | 6 | -73.992958 | 40.768280 | -73.981155 | 40.779301 | 0 | 297 |
| )4-21 24:58 | 1 | -73.987854 | 40.747585 | -73.996162 | 40.750702 | 0 | 208 |
| )1-07 28:19 | 6 | -73.985512 | 40.735691 | -74.008110 | 40.739491 | 0 | 652 |
| )5-03 42:26 | 1 | -73.997704 | 40.741165 | -73.985764 | 40.747059 | 0 | 879 |
| )6-04 15:23 | 1 | -73.978264 | 40.752213 | -73.991272 | 40.750263 | 0 | 896 |

1.5 Normalize or standardize numerical features if required:

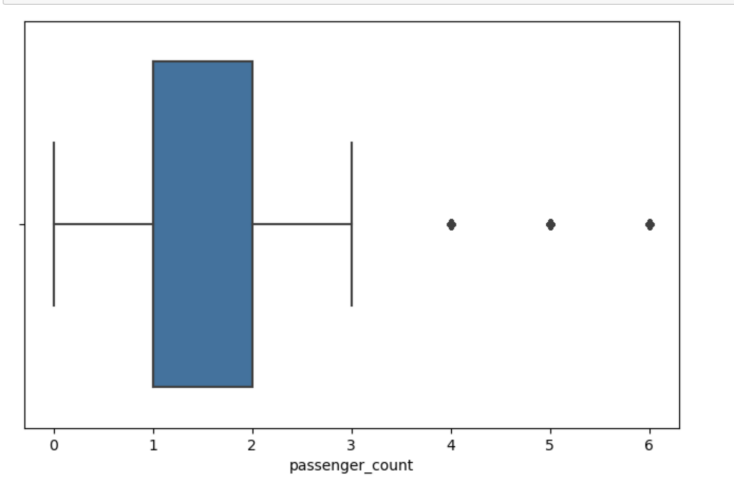
I am performing normalization on the features pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude and trip_duration.

Normalization is a technique used to scale and transform the values of different features on a similar scale, typically between 0 and 1. Min-Max scaling is a specific form of normalization.

| :etime | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | store_and_fwd_flag | trip_duration |
|---|---|---|---|---|---|---|---|
| 05-13 :50:11 | 6 | 0.911950 | 0.906534 | 0.904367 | 0.908641 | 0 | 0.003428 |
| 04-21 :24:58 | 1 | 0.912972 | 0.902575 | 0.902042 | 0.903170 | 0 | 0.002397 |
| 01-07 :28:19 | 6 | 0.913441 | 0.900300 | 0.900191 | 0.901025 | 0 | 0.007538 |
| 05-03 :42:26 | 1 | 0.911000 | 0.901347 | 0.903653 | 0.902473 | 0 | 0.010167 |
| 06-04 :15:23 | 1 | 0.914893 | 0.903460 | 0.902800 | 0.903086 | 0 | 0.010364 |

1.6 Look for errors or outliers in the data:

The outliers in the data seen by plotting the data in the box plot for each key feature.

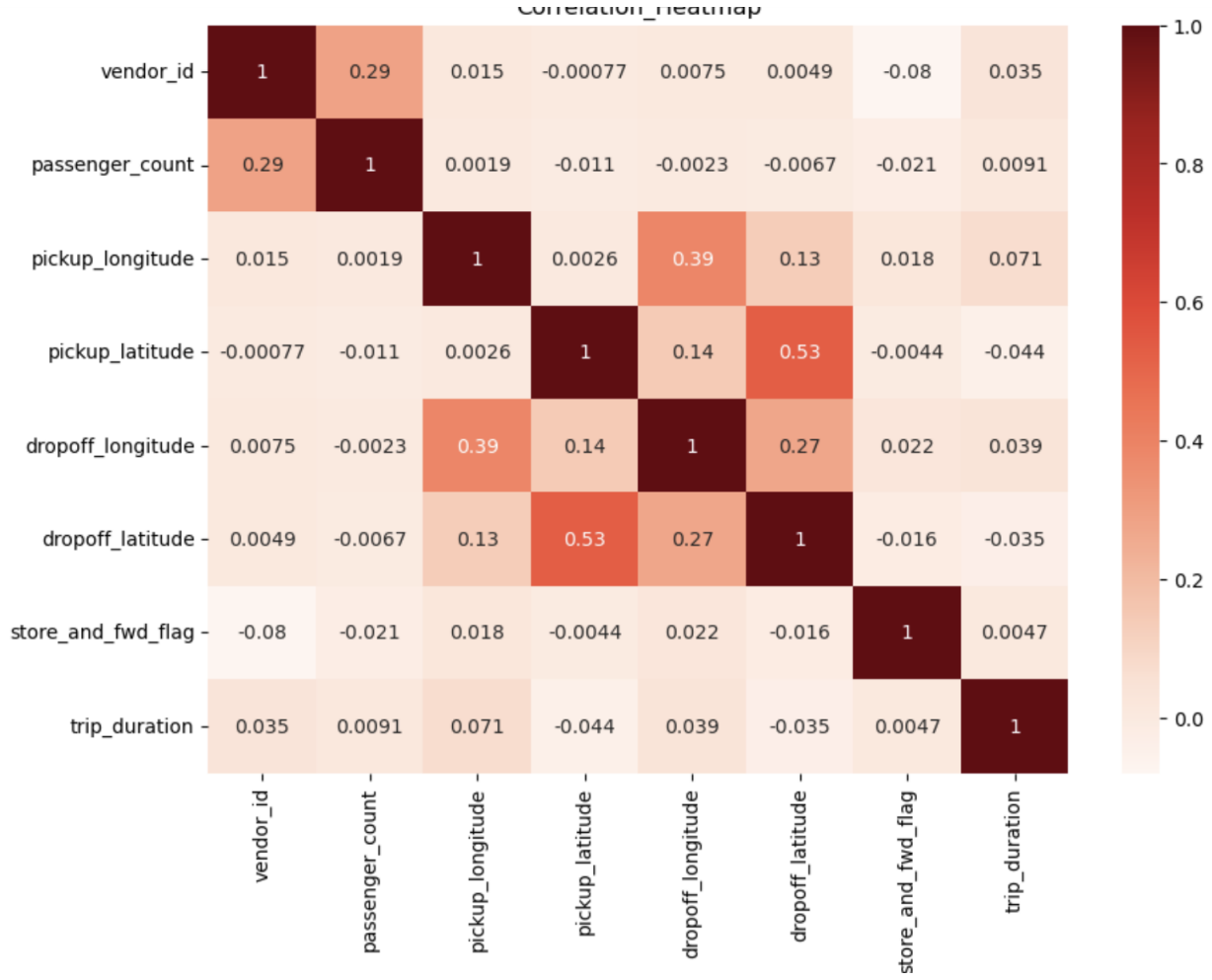



## TASK-2 Discovering Relationships:

2.1 correlation analysis:

| | vendor_id | passenger_count | pickup_longitude | pickup_latitude | dropoff_longitude | dropoff_latitude | store_a |
|---|---|---|---|---|---|---|---|
| **vendor_id** | 1.000000 | 0.286771 | 0.015250 | -0.000773 | 0.007514 | 0.004936 | |
| **passenger_count** | 0.286771 | 1.000000 | 0.001933 | -0.010586 | -0.002347 | -0.006653 | |
| **pickup_longitude** | 0.015250 | 0.001933 | 1.000000 | 0.002637 | 0.388205 | 0.134963 | |
| **pickup_latitude** | -0.000773 | -0.010586 | 0.002637 | 1.000000 | 0.140145 | 0.527573 | |
| **dropoff_longitude** | 0.007514 | -0.002347 | 0.388205 | 0.140145 | 1.000000 | 0.273444 | |
| **dropoff_latitude** | 0.004936 | -0.006653 | 0.134963 | 0.527573 | 0.273444 | 1.000000 | |
| **store_and_fwd_flag** | -0.079918 | -0.021463 | 0.018368 | -0.004397 | 0.021848 | -0.015754 | |
| **trip_duration** | 0.034668 | 0.009061 | 0.070830 | -0.043717 | 0.039345 | -0.035358 | |

Correlation heatmap:




The trip_duration had a stronger positive correlation with pickup_longitude than other features with 0.071 positive correlation, The trip_duration increases when the pickup_longitude will increase.

The trip_duration has positive correlation with passenger_count,vendor_id,drop_longitude and store_and_fwd_flag, The trip_duration increases when the positive correlated features will increase.

The trip_duration had a stronger negative correlation with pickup_latitude than dropoff_latitude with -0.044 negative correlation, The trip_duration decreases when the pickup_longitude will increase.

The pickup_latitude has a high positive correlation with dropoff_latitude with positive correlation among all features, the pickup_latitude will increase when dropoff_latitude is increased.

The trip_duration had a stronger negative correlation with dropoff_longitude than any other feature with -0.044 negative correlation.

2.2 :Regression analysis:

R-squared of 0.008, the model is able to explain for small of the variance in trip duration.

The Omnibus test is 245653.578 it is used to evaluate the overall statistical significance of a regression model.

The F-statistic is 167.9, and the associated p-value is very close to zero . This suggests that at least one of the predictors is significantly related to trip duration .

The increase of pickup_latitude and dropoff_latitude will lead to decrease of the trip duration.

The increase of pickup_longitude  and dropoff_longitude will lead to increase of the trip duration

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | trip_duration | **R-squared:** | 0.008 |
| **Model:** | OLS | **Adj. R-squared:** | 0.008 |
| **Method:** | Least Squares | **F-statistic:** | 167.9 |
| **Date:** | Fri, 17 Nov 2023 | **Prob (F-statistic):** | 1.99e-178 |
| **Time:** | 19:12:56 | **Log-Likelihood:** | 1.9384e+05 |
| **No. Observations:** | 100000 | **AIC:** | -3.877e+05 |
| **Df Residuals:** | 99994 | **BIC:** | -3.876e+05 |
| **Df Model:** | 5 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.0447 | 0.022 | -2.049 | 0.040 | -0.087 | -0.002 |
| **passenger_count** | 0.0002 | 8.34e-05 | 2.685 | 0.007 | 6.05e-05 | 0.000 |
| **pickup_longitude** | 0.2675 | 0.014 | 18.915 | 0.000 | 0.240 | 0.295 |
| **pickup_latitude** | -0.1624 | 0.021 | -7.611 | 0.000 | -0.204 | -0.121 |
| **dropoff_longitude** | 0.1361 | 0.017 | 8.006 | 0.000 | 0.103 | 0.169 |
| **dropoff_latitude** | -0.1841 | 0.019 | -9.628 | 0.000 | -0.222 | -0.147 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 245653.578 | **Durbin-Watson:** | 2.001 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 2308109440.867 |
| **Skew:** | 26.668 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 745.362 | **Cond. No.** | 702. |

## 2.3 limitations of the analysis:

Regression analysis assumes that the relationship between the output and feature variables is linear.

Multicollinearity occurs when independent variables are highly correlated with each other.

Linear regression assumes that the residuals are normally distributed.

Linear regression is sensitive to outliers.

Linear regression models are not suitable for extrapolation beyond the range of observed data.

Outliers can have a significant impact on the results of a regression analysis. It is important to identify and remove outliers before fitting a regression model.

Homoscedasticity assumes constant variance of residuals across all levels of the independent variables

The analysis assumes that the taxi trips are confirmed to New York City. If trips extend beyond the boundaries of the dataset, the model may not perform well in those areas.

The model considers only geographical coordinates as predictors, neglecting other potential factors such as traffic conditions, time of day, road condition and speed limit.

The residuals that should follow distributions that do not follow a normal distribution are violated based on the Jarque-Bera test result.

2.4 creating new features:

Creating new features distance,speed,day_of_week,hour_of_day and time_of_day for understanding the data in a better way.

The distance and speed represents the distance traveled in the trip and speed of the trip.

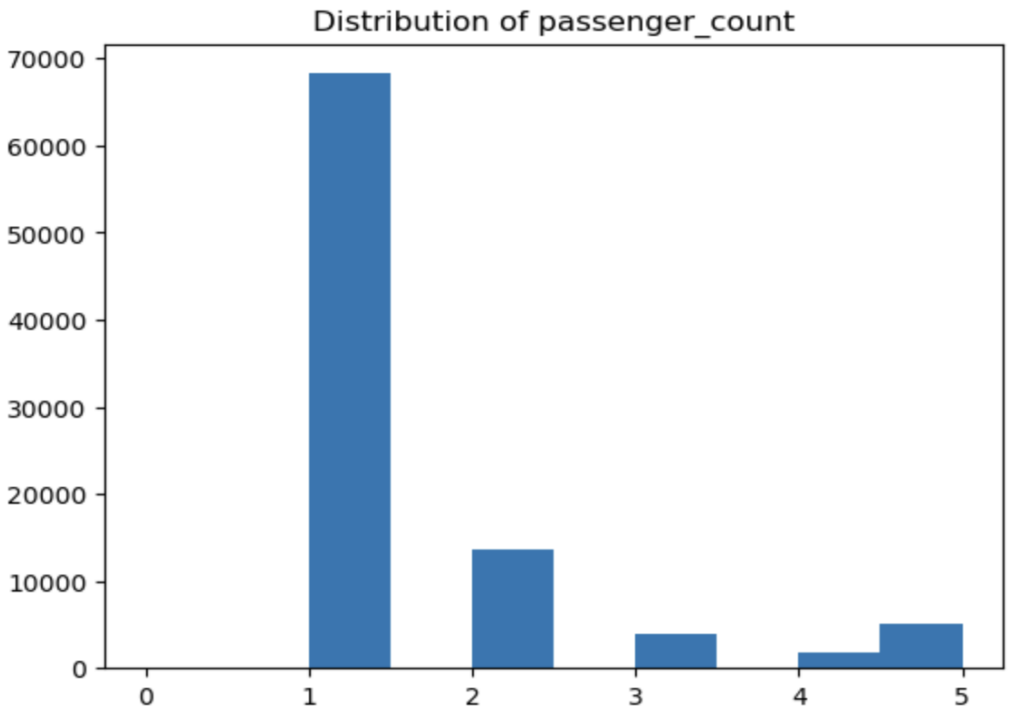
The distance is calculated by pickup features and dropoff features.

The time_of_day represents the time in which the trip occurred.

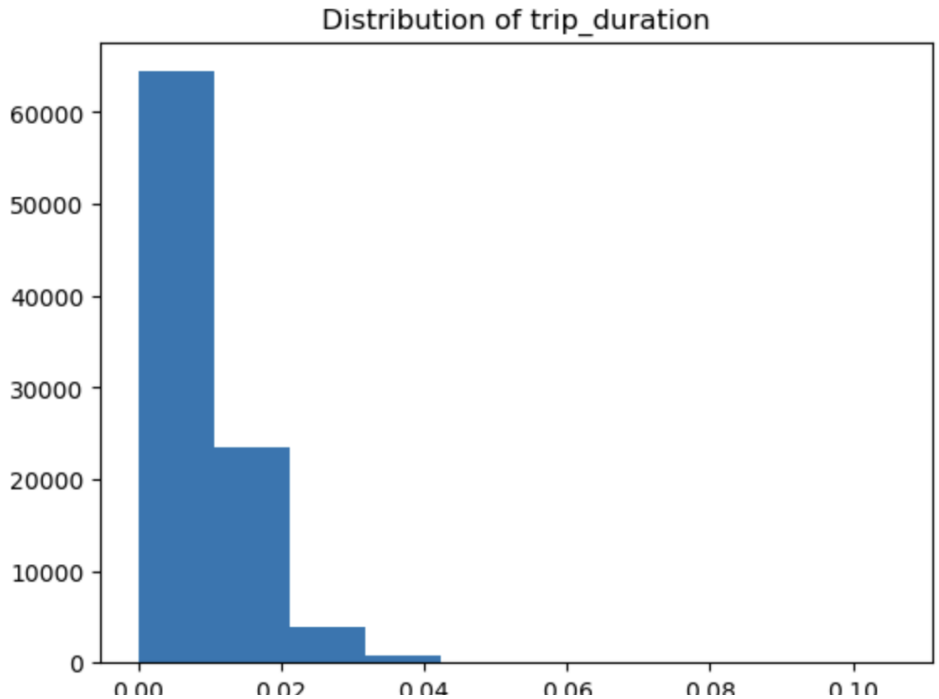
| ff_longitude | dropoff_latitude | store_and_fwd_flag | trip_duration | distance | speed | day_of_week | hour_of_day | time_of_day |
|---|---|---|---|---|---|---|---|---|
| 0.904367 | 0.908641 | 0 | 0.003428 | 0.875614 | 9.196681e+05 | 4 | 8 | Morning |
| 0.902042 | 0.903170 | 0 | 0.002397 | 1.218397 | 1.829905e+06 | 3 | 9 | Morning |
| 0.900191 | 0.901025 | 0 | 0.007538 | 1.477050 | 7.053818e+05 | 3 | 22 | Evening |
| 0.903653 | 0.902473 | 0 | 0.010167 | 0.827190 | 2.929010e+05 | 1 | 19 | Evening |
| 0.902800 | 0.903086 | 0 | 0.010364 | 1.346652 | 4.677809e+05 | 5 | 15 | Afternoon |

**TASK-3 DATA VISUALIZATION:**

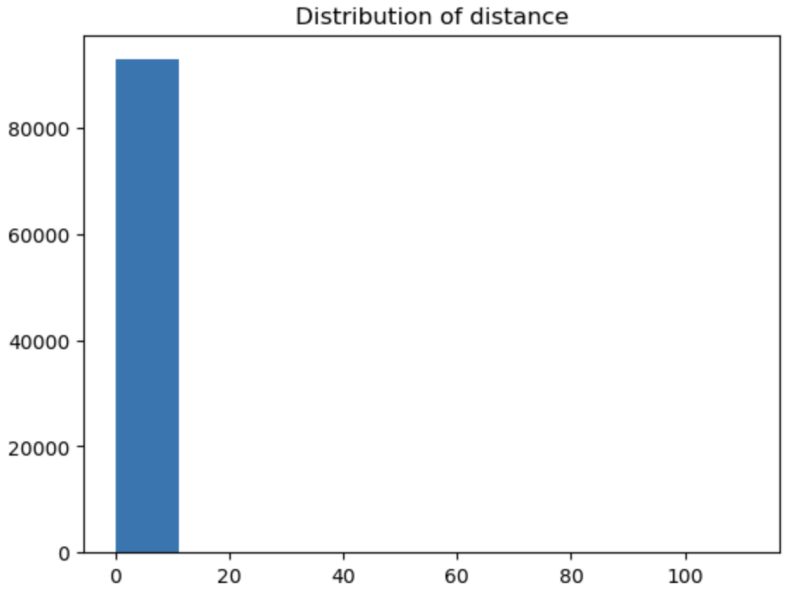
visualizations can help to understand the shape and spread of the data, such as whether the data is normally distributed or skewed.



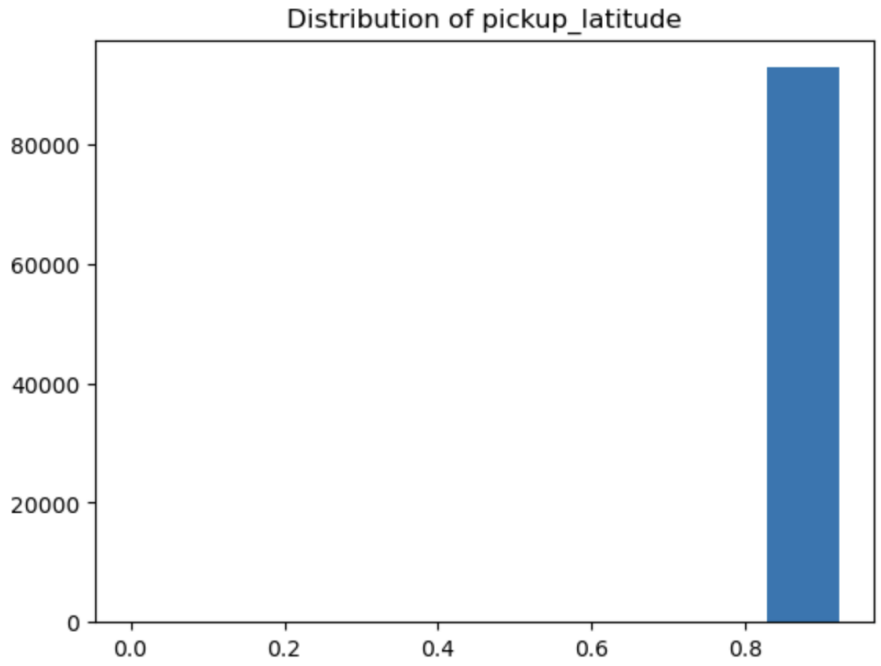

In the total number of trips with the passenger_count 1 has a maximum number of trips and the trips with passenger_count 4 is less.



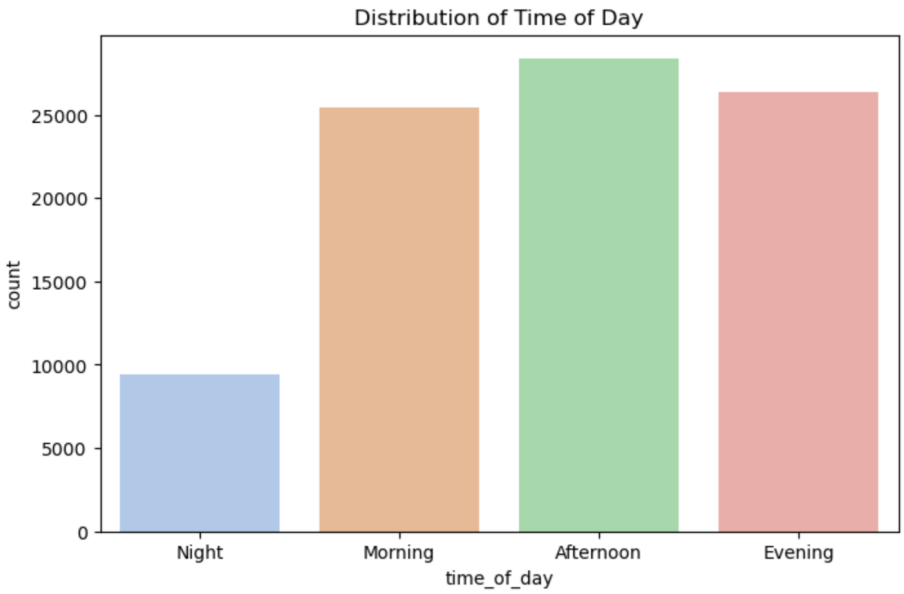

In most of the trips the trip_duration lies between 0 to 0.2.




In all the trips the distance lies between 0 to 20.



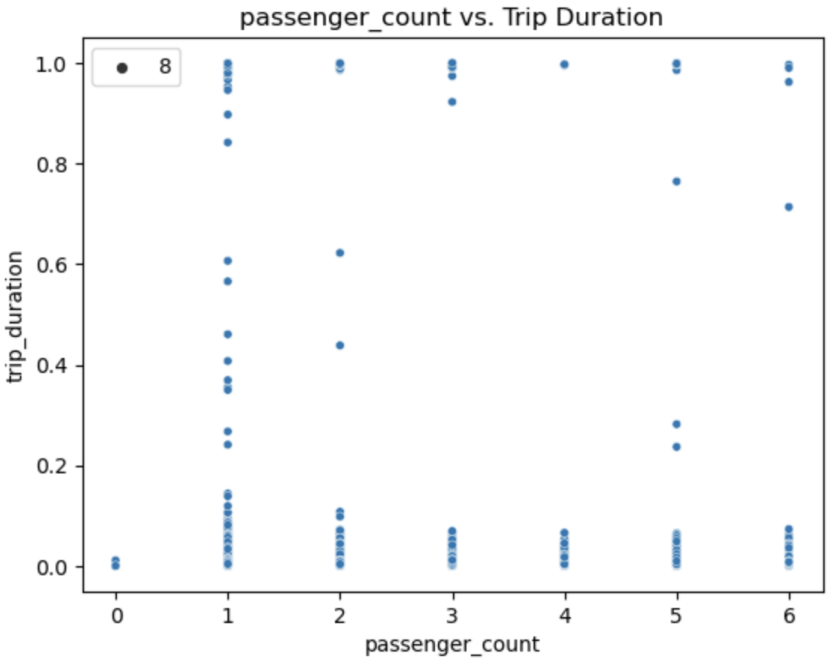

In all the trips the pickup_latitude lies above 0.8.

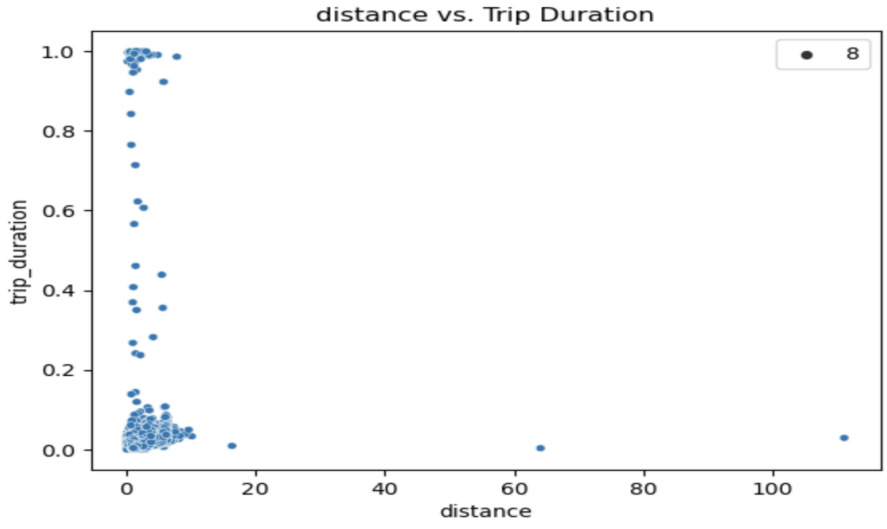



Most of the trips are in the afternoon and less trips are at night.

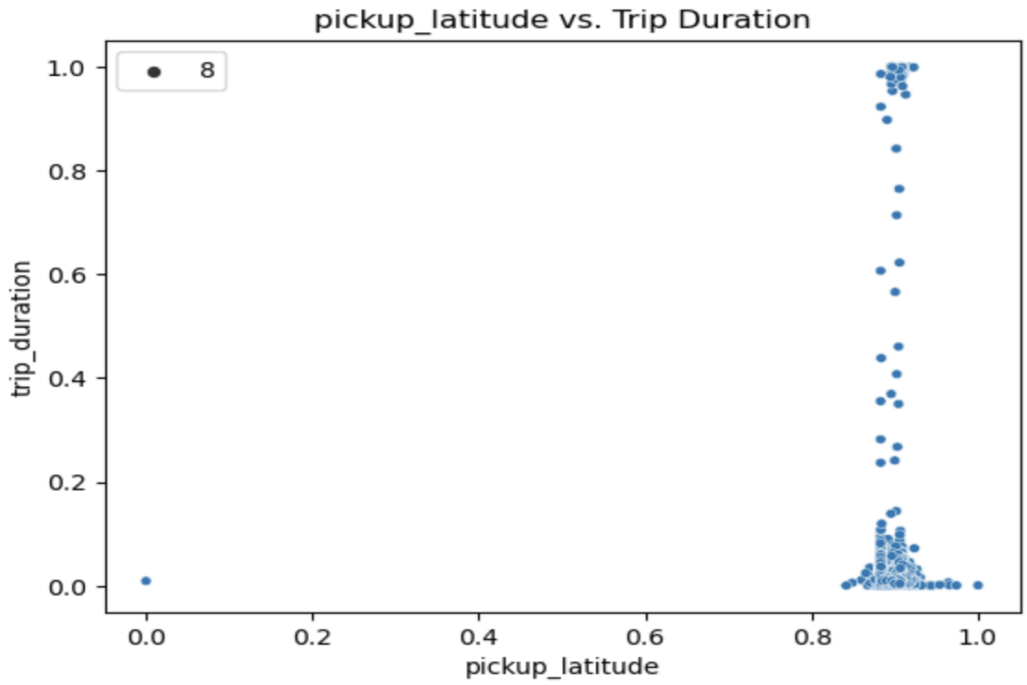
3.2Visualize the relationship between different features and the trip duration:

The passenger_count with 1 will have many high trip_duration and with 4 passenger_count will have many trip_duration is.

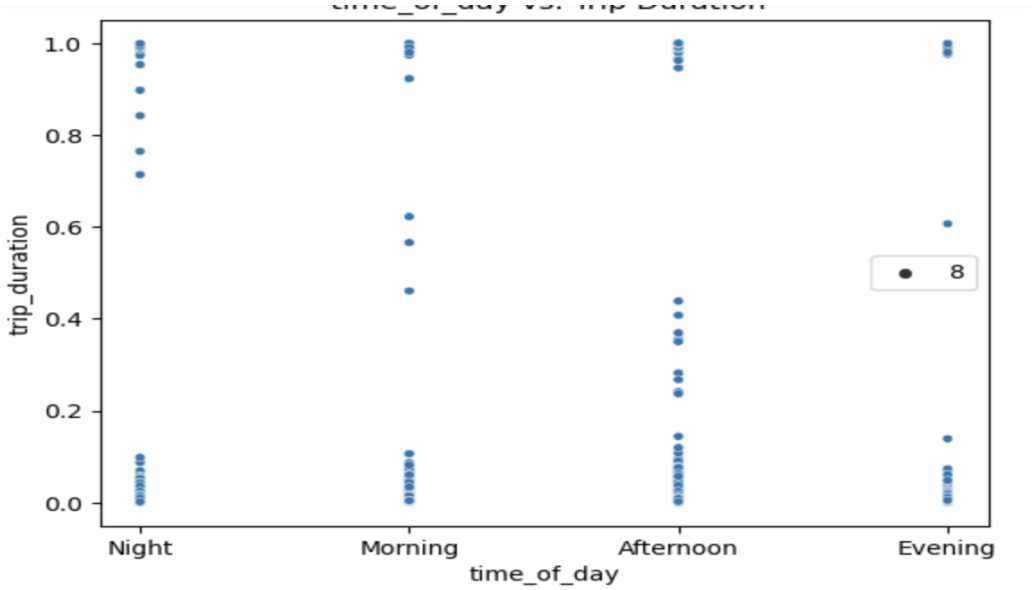


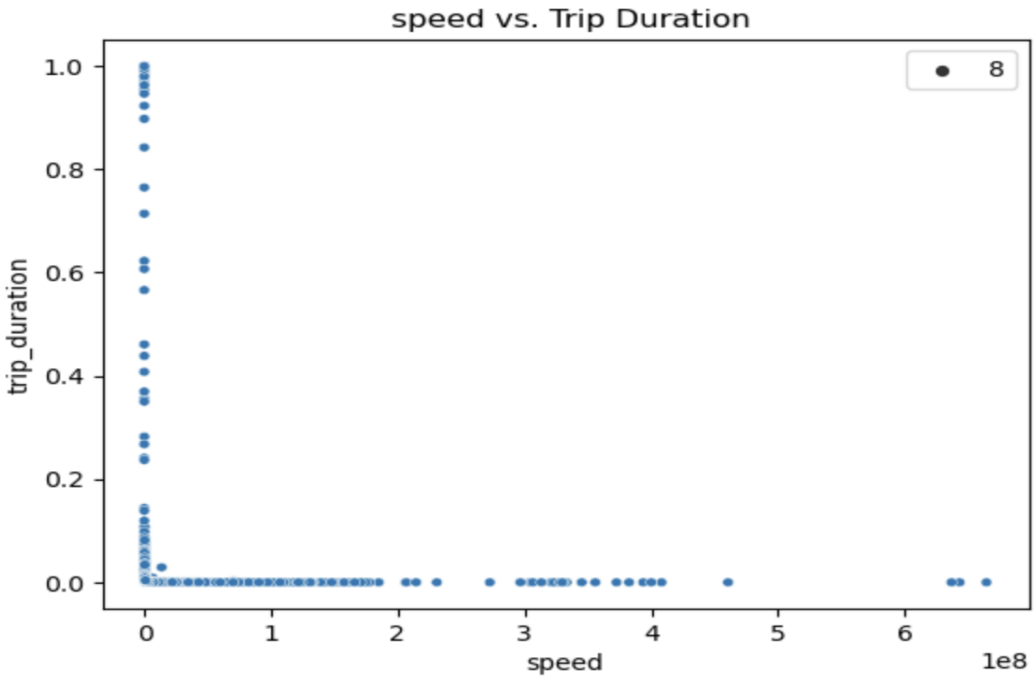
The maximum and less trip_duration lies between the distance 0 to 20 kms, There some of the trip_duration is less with high distance.

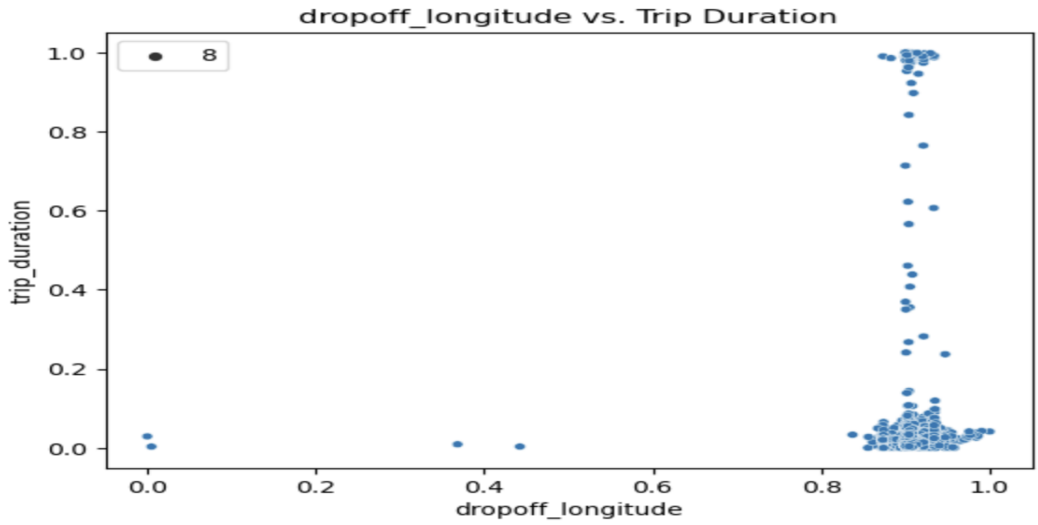



The maximum pickup_latitude lies between the trip_duration 0.0 to 0.2



The trip_duration is high  in the afternoon and low in the night.



**TASK -4 Algorithm and Data Structure Efficiency:**

**The dataset should be sorted and viewed in ascending order of trip duration. New data is added to the dataset frequently, where these new trips should show up at the end of the sorted list.**

Linked List data structure is the best data structure for insertions of new elements at the end of sorted list without the need to shift existing elements and more efficient in sorting data.

When new trips are added frequently, appending to the end of a linked list has a constant time complexity O(1).

The nodes in a linked list contain references to the elements that come after it, it is easy to maintain in  the sorted order.

linkedlist are dynamic in size where there is no limit on the number elements are inserted in the list.

Linked lists can be easily sorted by adjusting the links between nodes without the need for data movement.

where the size of the dataset is not known before the time, linked lists may be more memory-efficient than arrays.

Each new node in a linked list is placed separately, and the entire data does not need to be shifted.

**A new field is added to the data representing the passenger's phone number. It will be used to quickly filter out the trips made by a specific passenger.**

When new data (a phone number) is added and a goal is to fast filter trips by specific passenger, hash tables perform well.

Hash tables are good for obtaining data based on a given key because they offer constant-time average-case complexity for search operations without looking at the entire data.

Hash table provides uniqueness in keys as there are no same phone numbers in the data and provides the data of the specific passenger faster than other data structures.

The Hash Table allows for direct access to the data associated with a specific phone number without scanning the data.

When data is updated frequently or new trips are added for various passengers, hash tables work well.

**TASK -5 Final Analysis:**

The distance between pickup and dropoff will have more impact on the trip duration.

The passenger count has a significant impact on trip duration. Decreasing passenger count can reduce the trip duration.

The time_of_day will have an impact on the trip_duration, The night of the day will have less trip_duration than afternoon due less traffic.

Recommendations:

Choosing the route with short distance between pickup and dropoff locations will decrease the trip_duration.

Choosing the way with less traffic can influence the decrease in trip_duration.

By giving pricing incentives for passengers who choose to travel during off-peak hours.

Encourage shared rides for passengers with similar routes to reduce the number of vehicles on the road and optimize travel time.