# Predict The Fare Amount Of Future Rides Using Regression Analysis

Renu Jindal

## Problem

The fare amount of a ride is influenced by various factors such as distance, duration, traffic conditions, time of day, and demand.

## Goal

In this project I am going  to develop a regression model that can predict the fare amount based on most important features.

# Why this is a Regression problem

- First of all our Target variable is a continuous variable.
- Second there is a Linear Correlation between Independent and Dependent variable.
- The distribution of Target variable is Normally distribution.
- In Regression model we create a straight best fit line to predict the Target variable .

# STEPS OF PROJECT

1 Data reading

2 Exploratory Data Analysis and Data Cleaning

3 Data Visualization

4 Feature Engineering

5 Splitting the Data in to Train and Test set

6 Standardization

7 Building Linear Regression Model

8 Checking VIF

9 Residual Analysis

10 Model Evaluation

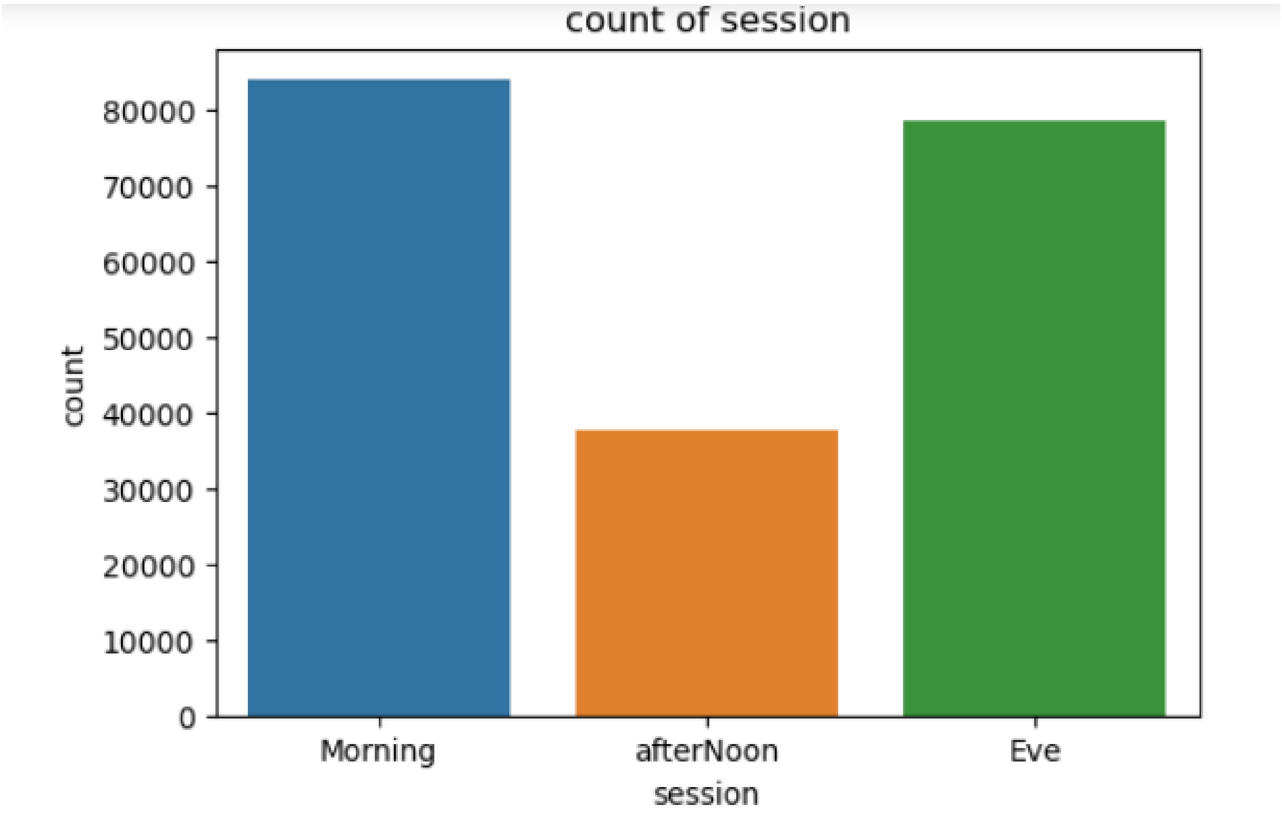
# Data Reading

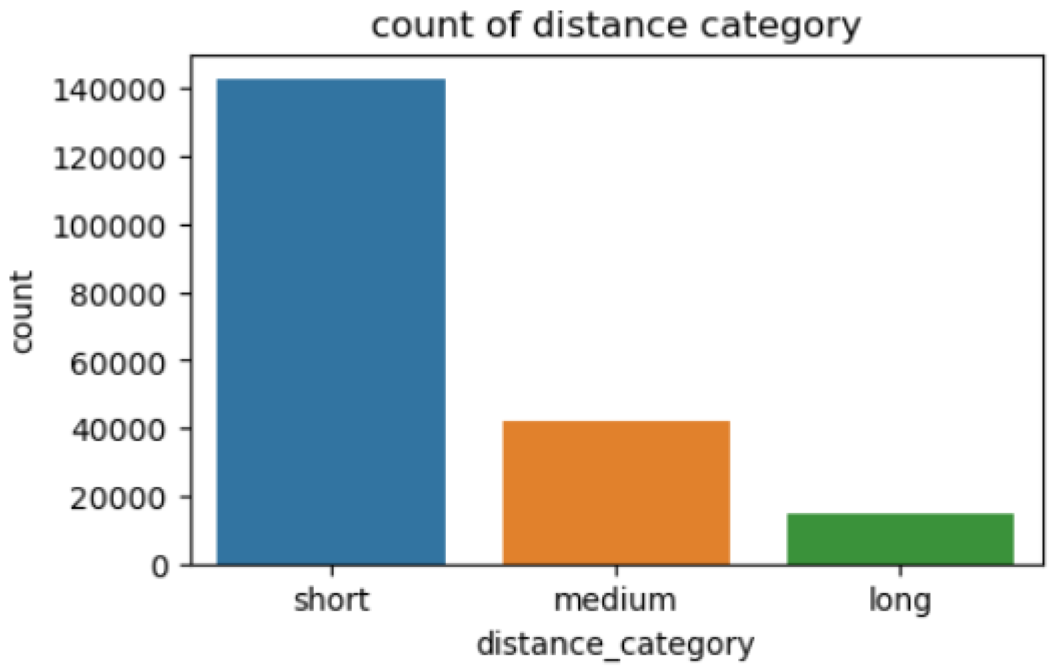
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 9 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   Unnamed: 0         200000 non-null  int64
 1   key                200000 non-null  object
 2   fare_amount        200000 non-null  float64
 3   pickup_datetime    200000 non-null  object
 4   pickup_longitude   200000 non-null  float64
 5   pickup_latitude    200000 non-null  float64
 6   dropoff_longitude  199999 non-null  float64
 7   dropoff_latitude   199999 non-null  float64
 8   passenger_count    200000 non-null  int64
dtypes: float64(5), int64(2), object(2)
memory usage: 13.7+ MB
```

```
df.shape
```

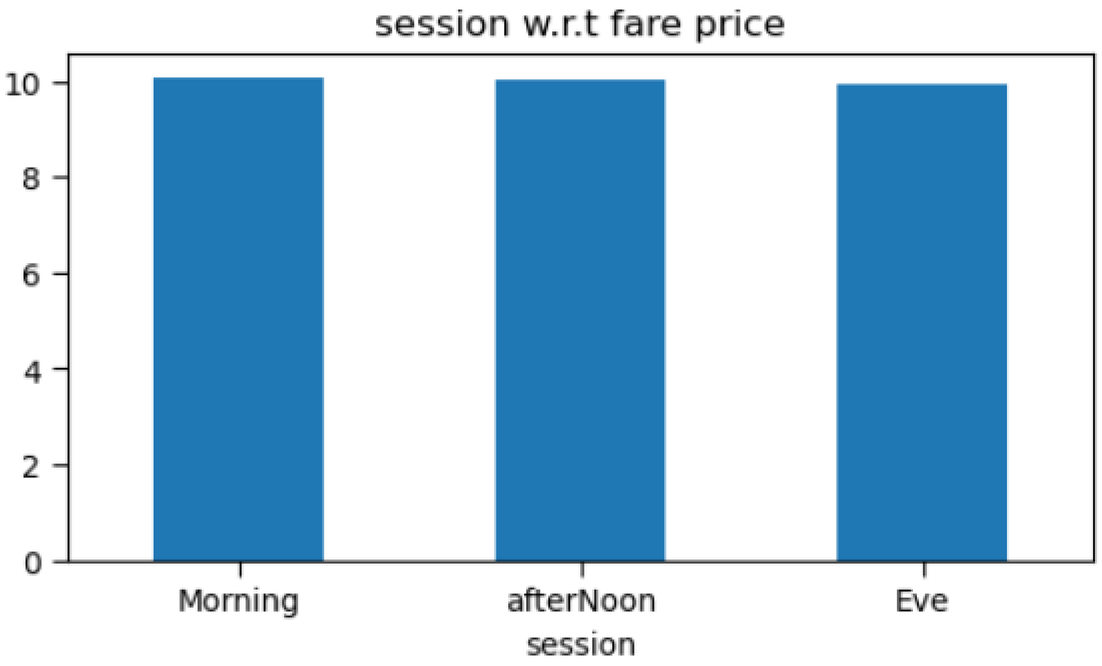
```
(200000, 9)
```

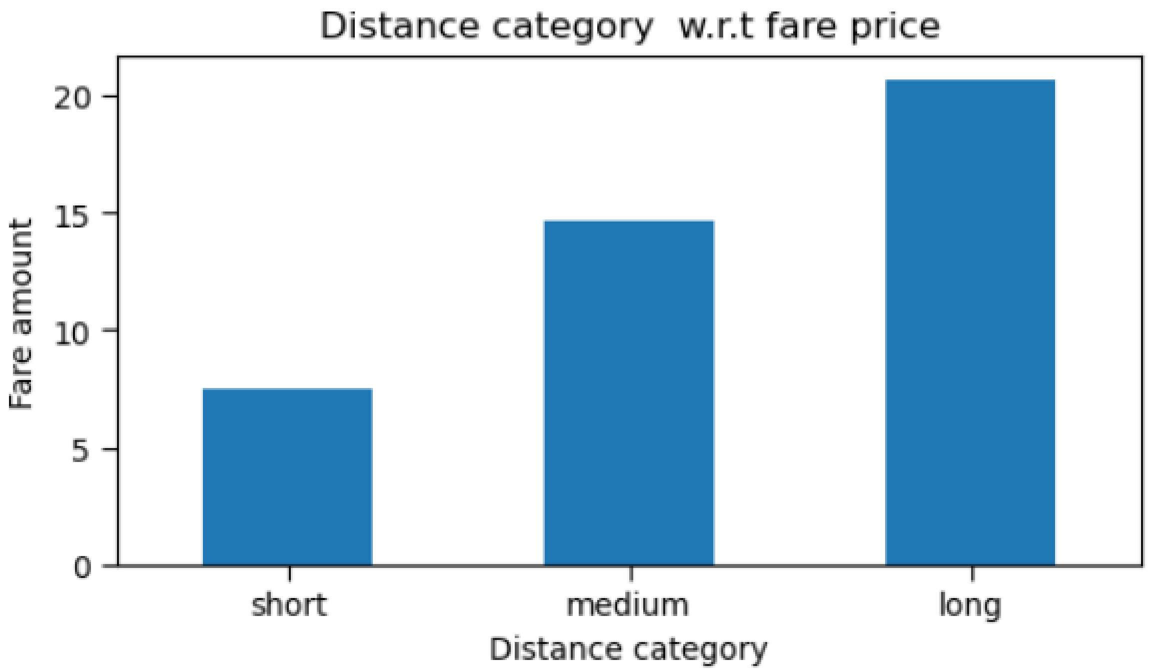
# Visualization



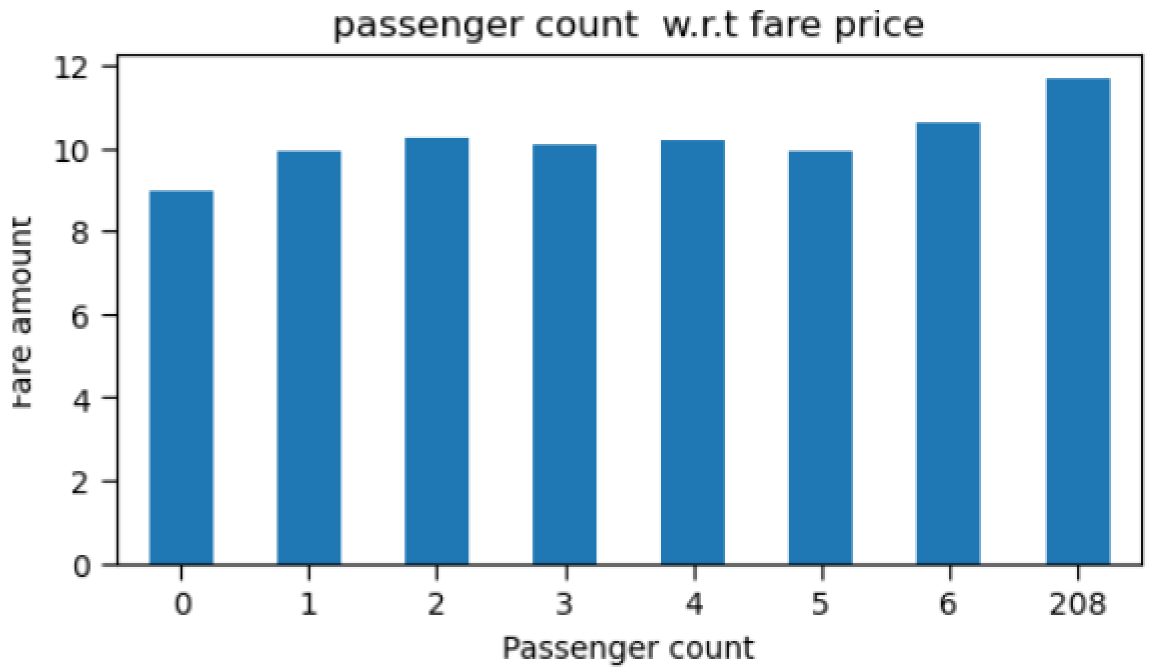



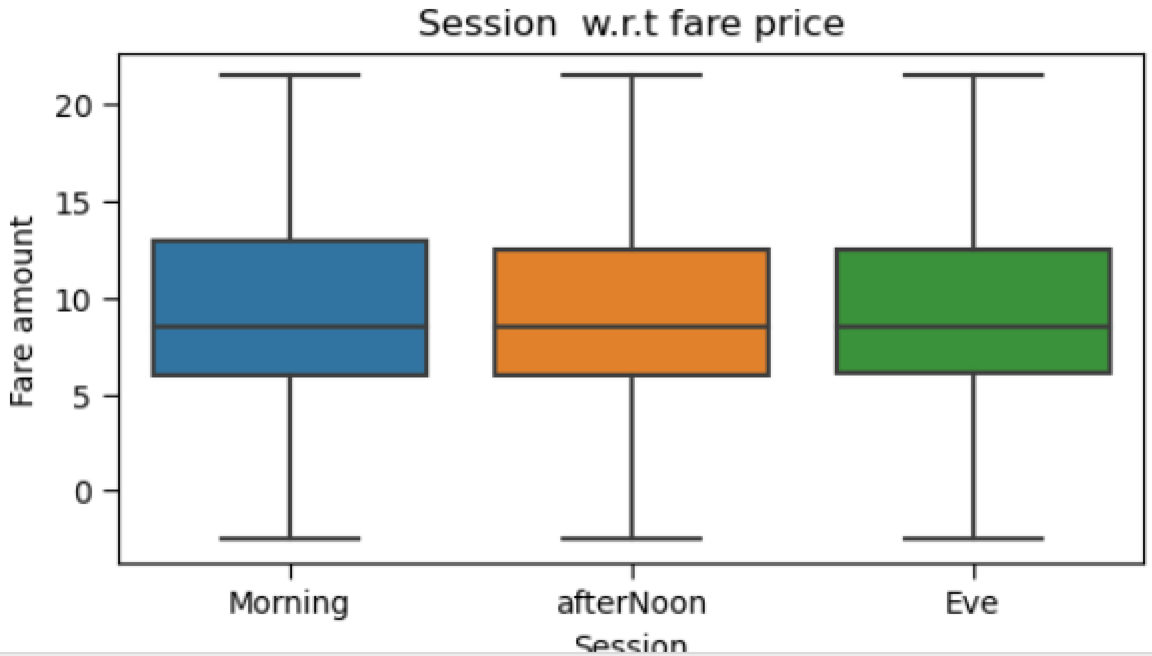
# Bivariate Analysis



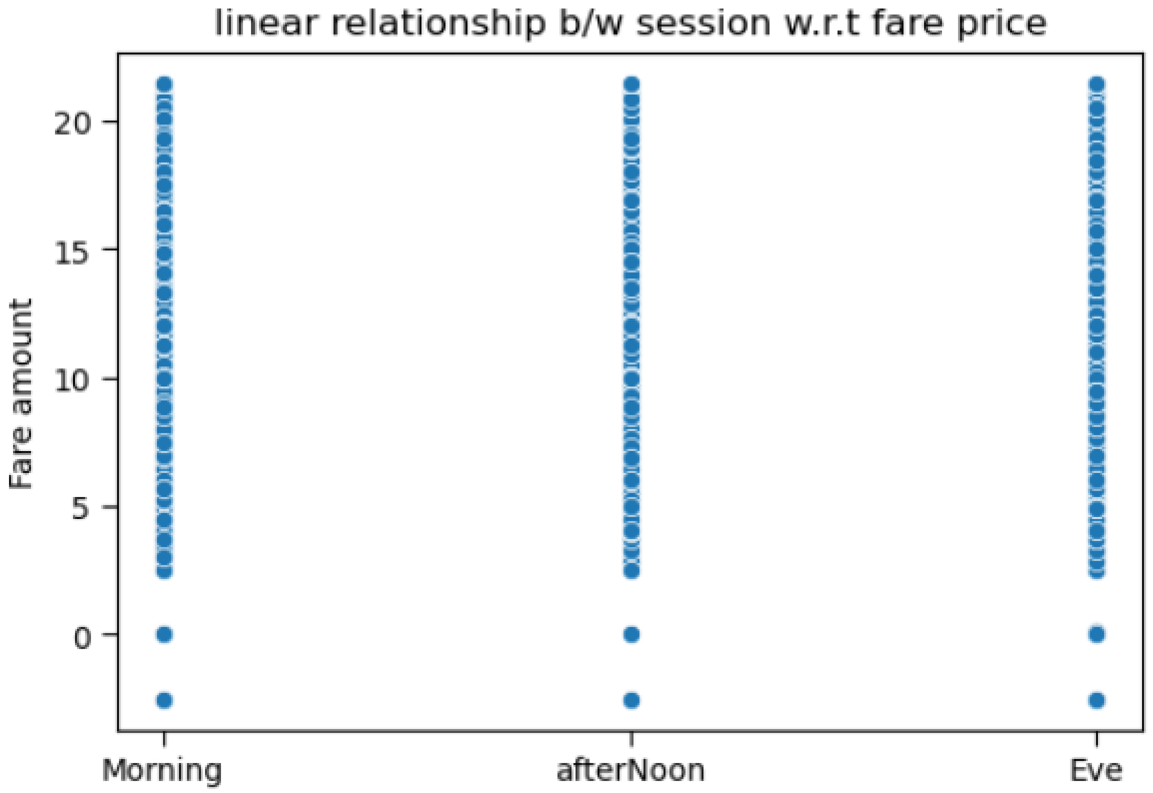



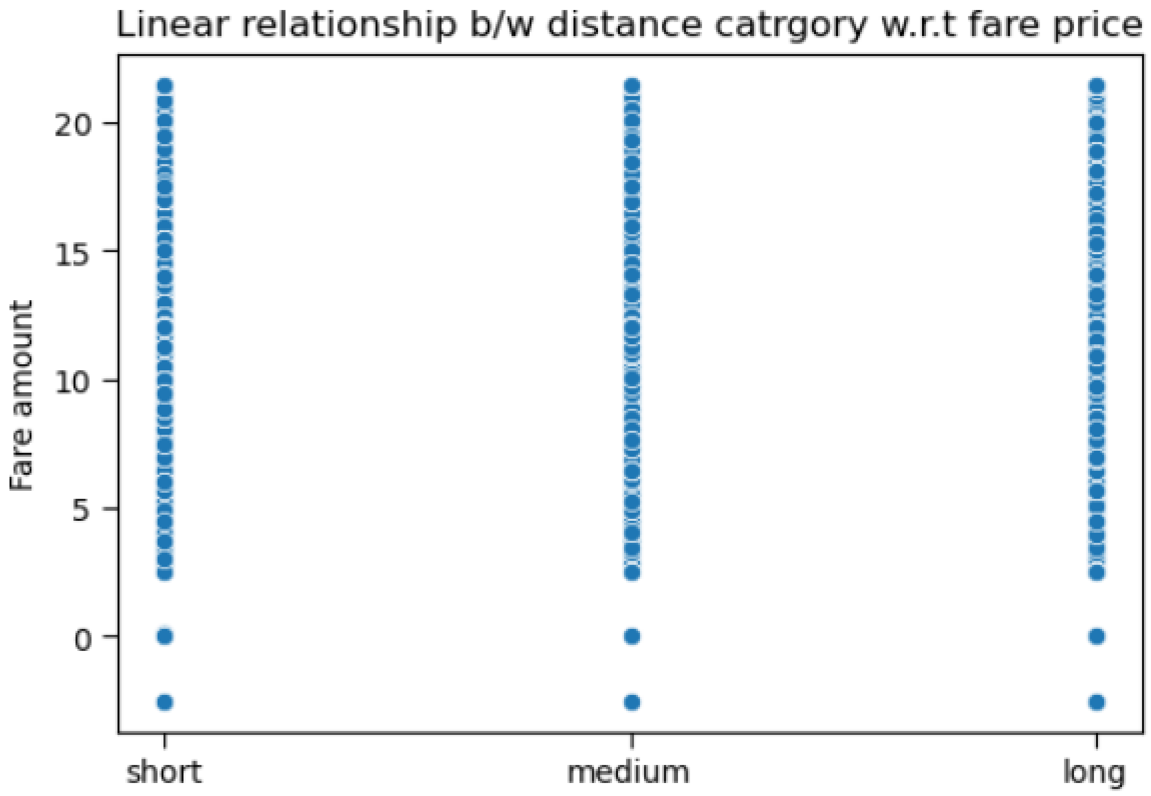
# Bivariate Analysis



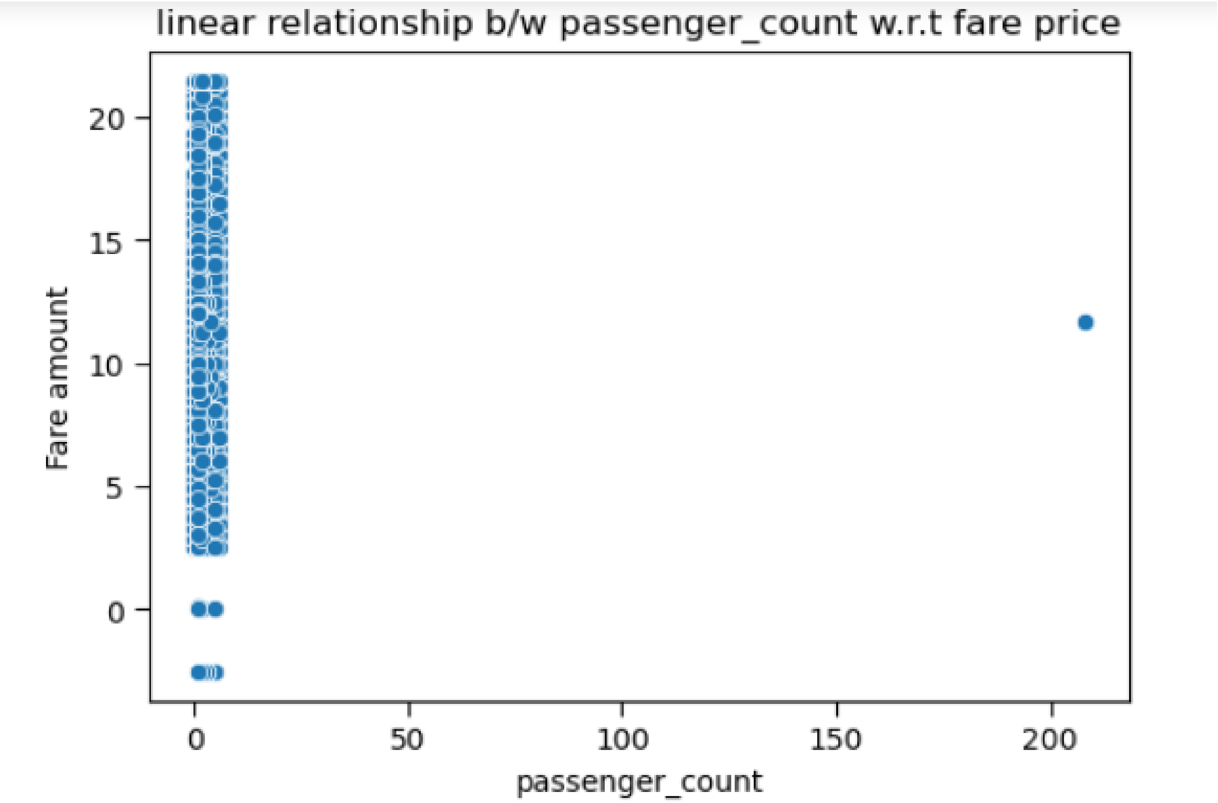



# Linear relationship between Independent and Target Variable

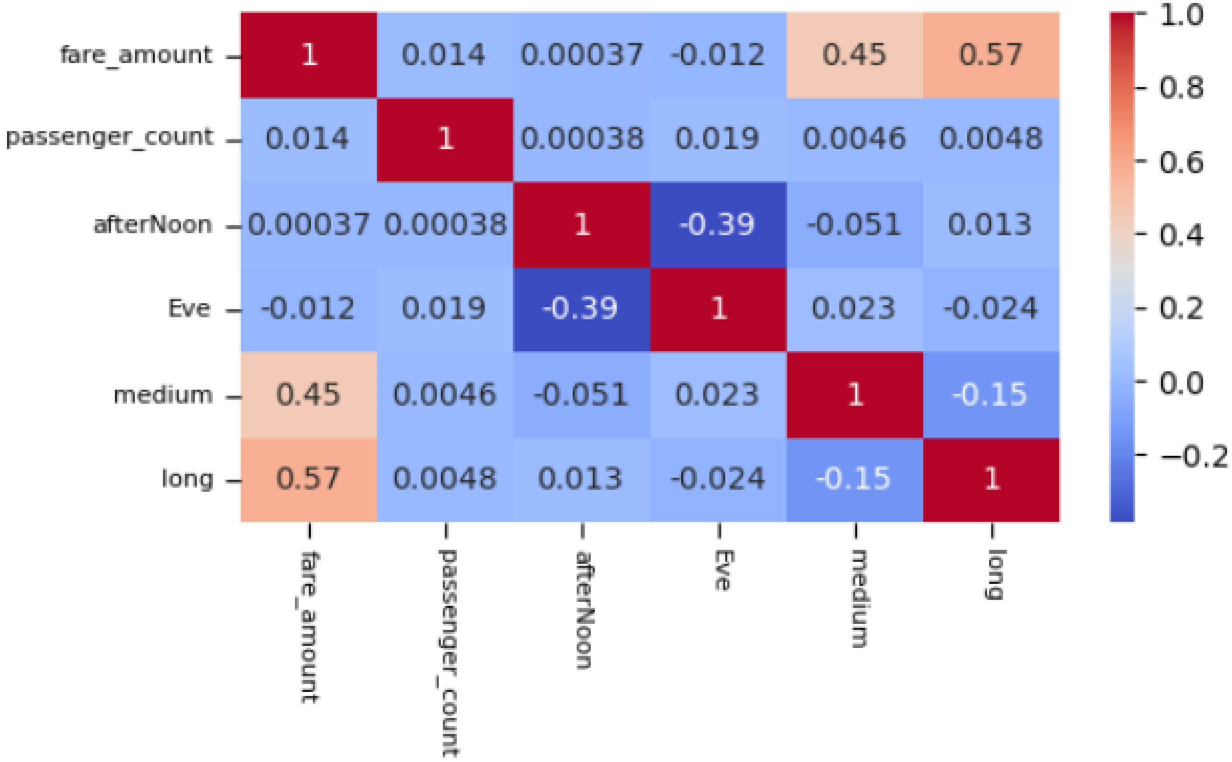


# Linear Relationship b/w Passenger Count w.r.t Target variable



# Correlation matrix of all Independent feature and Target variable



# Model accuracy by taking all Independent features

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.4745 | 0.003 | -171.496 | 0.000 | -0.480 | -0.469 |
| passenger_count | 0.0090 | 0.002 | 5.462 | 0.000 | 0.006 | 0.012 |
| afterNoon | 0.0451 | 0.005 | 9.865 | 0.000 | 0.036 | 0.054 |
| Eve | -0.0018 | 0.004 | -0.483 | 0.629 | -0.009 | 0.005 |
| medium | 1.3435 | 0.004 | 328.574 | 0.000 | 1.335 | 1.351 |
| long | 2.4915 | 0.006 | 391.516 | 0.000 | 2.479 | 2.504 |

| | | | |
|---|---|---|---|
| Omnibus: | 21689.605 | Durbin-Watson: | 1.994 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 66964.924 |
| Skew: | 0.809 | Prob(JB): | 0.00 |
| Kurtosis: | 5.977 | Cond. No. | 4.44 |

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | fare_amount | R-squared: | 0.620 |
| Model: | OLS | Adj. R-squared: | 0.620 |
| Method: | Least Squares | F-statistic: | 4.576e+04 |
| Date: | Fri, 26 Apr 2024 | Prob (F-statistic): | 0.00 |
| Time: | 10:20:24 | Log-Likelihood: | -1.3084e+05 |
| No. Observations: | 140000 | AIC: | 2.617e+05 |
| Df Residuals: | 139994 | BIC: | 2.618e+05 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

# Model accuracy improve by Dropping “Eve” Independent Variable

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | fare_amount | **R-squared:** | 0.620 |
| **Model:** | OLS | **Adj. R-squared:** | 0.620 |
| **Method:** | Least Squares | **F-statistic:** | 5.720e+04 |
| **Date:** | Fri, 26 Apr 2024 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 10:20:25 | **Log-Likelihood:** | -1.3084e+05 |
| **No. Observations:** | 140000 | **AIC:** | 2.617e+05 |
| **Df Residuals:** | 139995 | **BIC:** | 2.617e+05 |
| **Df Model:** | 4 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -0.4754 | 0.002 | -224.505 | 0.000 | -0.480 | -0.471 |
| **passenger_count** | 0.0090 | 0.002 | 5.453 | 0.000 | 0.006 | 0.012 |
| **afterNoon** | 0.0459 | 0.004 | 10.905 | 0.000 | 0.038 | 0.054 |
| **medium** | 1.3435 | 0.004 | 328.575 | 0.000 | 1.335 | 1.351 |
| **long** | 2.4915 | 0.006 | 391.598 | 0.000 | 2.479 | 2.504 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 21694.127 | **Durbin-Watson:** | 1.994 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 66984.822 |
| **Skew:** | 0.809 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 5.977 | **Cond. No.** | 4.10 |

# THANK YOU