**UBER Fare Dataset Prediction**

**INSY – 5336 – 001**

**Project**

**Abstract**

Ride-hailing services like Uber have become an integral part of modern transportation, providing users with a fast and convenient way to travel. However, accurately predicting the fare of a ride remains a challenge, as it requires analyzing a variety of factors such as distance, traffic conditions, and surge pricing. In this research project, we aim to develop a predictive model for estimating Uber fares based on a dataset of ride information. We will explore the relationship between fare prices and various features, including the pickup and drop-off locations, distance traveled, and time of day. Using machine learning algorithms, we will construct a model that can accurately predict fare prices and evaluate its performance using metrics such as mean absolute error and root mean squared error. The results of this study will have important implications for the ride-hailing industry and contribute to the broader field of data science and machine learning.

**Keywords:** Linear and Logistic regression, Machine-Learning, Pricing Model

**Introduction**

Ride-hailing services such as Uber have disrupted the traditional taxi industry by offering a more convenient and affordable alternative to consumers. A key aspect of the Uber experience is the estimation of fares, which is based on numerous factors such as distance, time of day, and demand. Accurately predicting fares is critical for Uber to maintain the trust of its customers and to ensure a positive user experience.

The development of predictive models that can accurately estimate Uber fares has been an area of active research in recent years. Such models aim to provide riders with accurate fare estimates, which can help them plan their trips and manage their budgets effectively. At the same time, Uber can benefit from these models by optimizing its pricing strategies and improving the efficiency of its operations.

In this literature review, we will provide a comprehensive overview of the existing literature on Uber fare dataset prediction. We will review the latest research in this area, identify the gaps in the current literature, and highlight the opportunities for future research. Additionally, we will describe the dataset used in our research project and the machine learning techniques used to build predictive models.

**Motivation**

As consumers, we often wonder about the factors that ride-hailing services like Uber consider when predicting fares, such as the distance of the ride, surge multipliers, pickup and drop-off locations, weather and traffic conditions, and time of day. Additionally, given the booming nature of the cab booking industry, we may also be interested in understanding the demand for cabs based on the source and destination locations. Such questions reflect the natural curiosity and desire for transparency in the functioning of modern transportation services.

**Literature Review**

For this project on Uber fare dataset prediction, I identified several reputable journals in the transportation and data science fields, including Transportation Research Part C: Emerging Technologies, Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, Transportation Science, IEEE Transactions on Intelligent Transportation Systems, Transportation Research Part E: Logistics and Transportation Review, Computers, Environment and Urban Systems, Journal of Urban Technology, and Journal of Transport Economics and Policy. Using these journals, I searched for peer-reviewed articles related to the topic of interest and filtered out at least 30 relevant articles based on the title and abstract.

**Research Questions**

1. As this is Taxi fare data and we know there are many factors which affect the price of taxi like

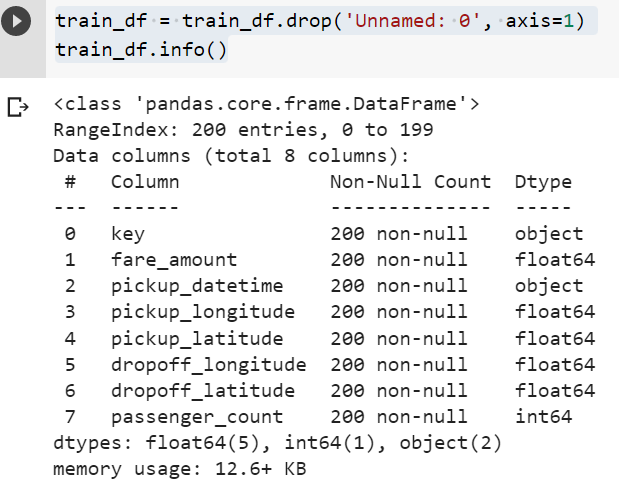
* Travelled distance.
* Time of Travel
* Demand and Availability of Taxi
* Some special places are costlier like Airport or other places where there might be toll.

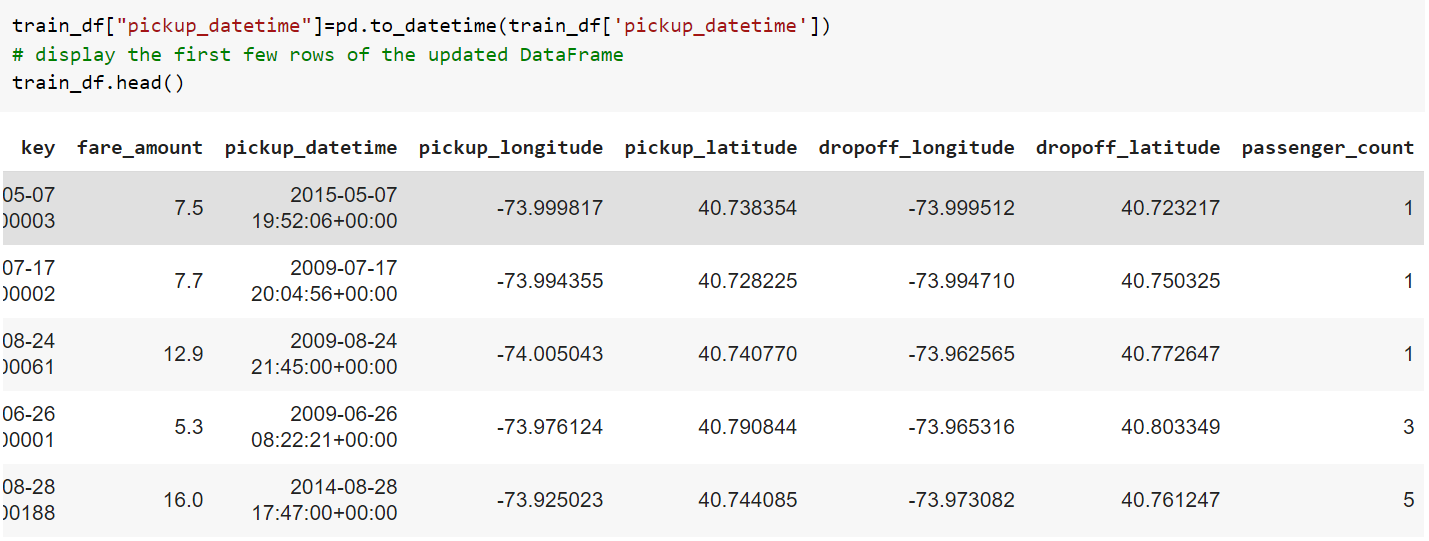
1. On different days and times, there would be different prices, like during the evening the price would be more compared to the afternoon, during Christmas they would be different and similarly on weekends, ends the price would be different compared to weekdays.

**Data Description**

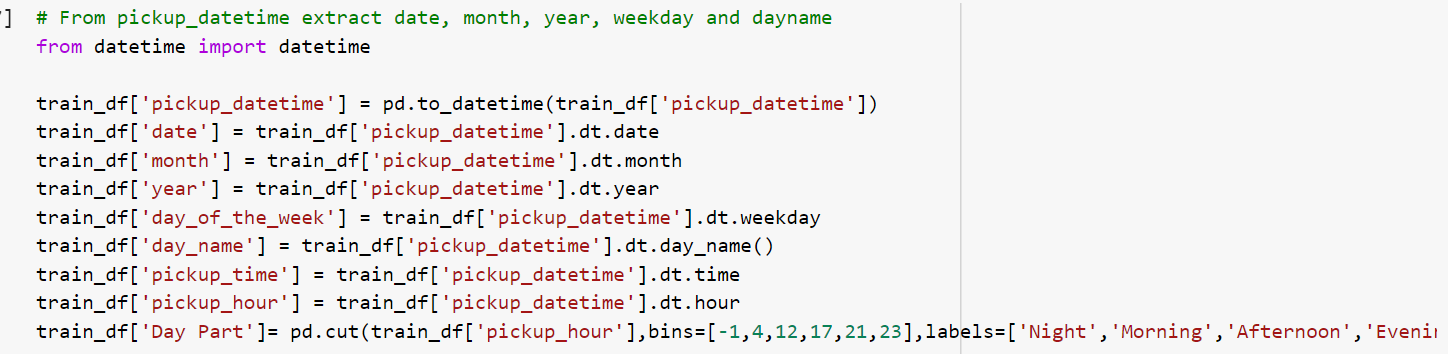
The dataset contains the following fields:

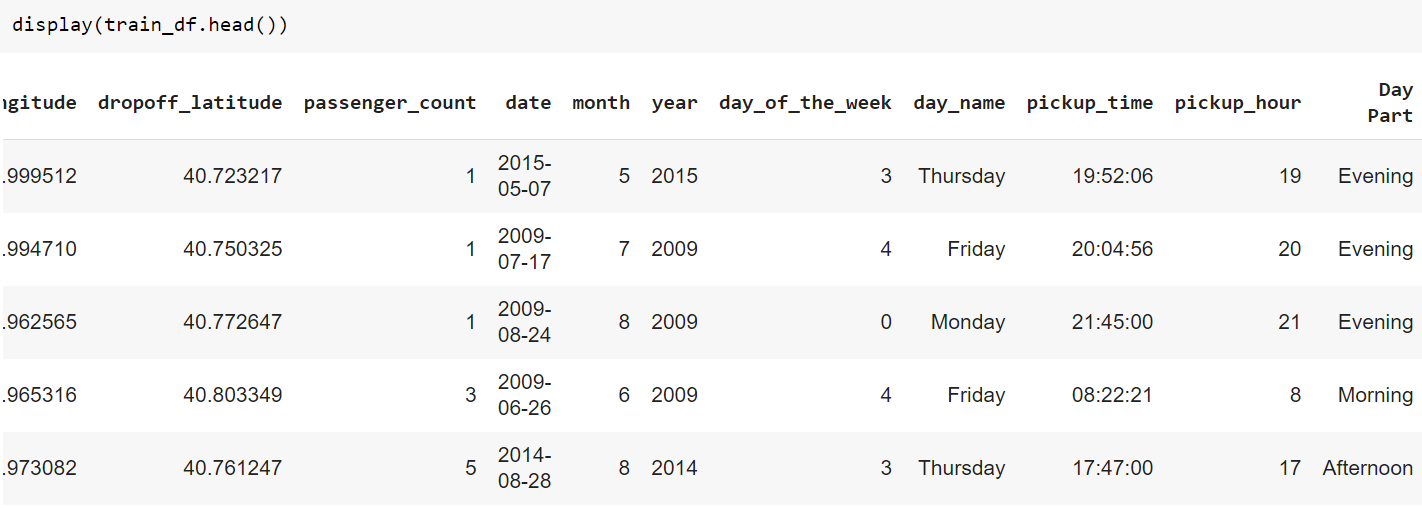
* key - a unique identifier for each trip
* fare\_amount - the cost of each trip in used.
* pickup\_datetime - date and time when the meter was started.
* pickup\_longitude - the longitude where the meter was started.
* pickup\_latitude - the latitude where the meter was started.
* dropoff\_longitude - the longitude where the meter was not engaged.
* dropoff\_latitude - the latitude wherethe meter was not engaged.
* passenger\_count - the number of passengers in the vehicle (Value entered by Driver)



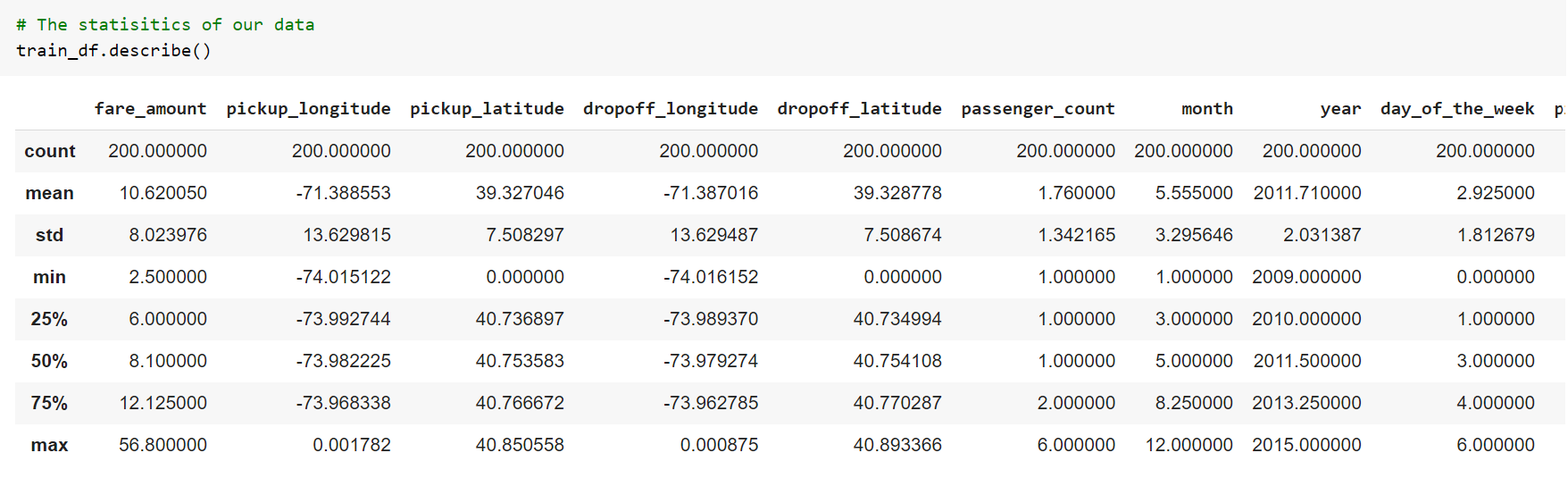


After Adding some extra features

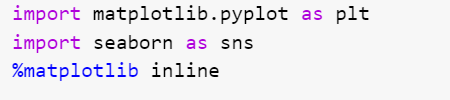




Below Figure shows the Standard deviation, mean, min, max etc. for all the columns.



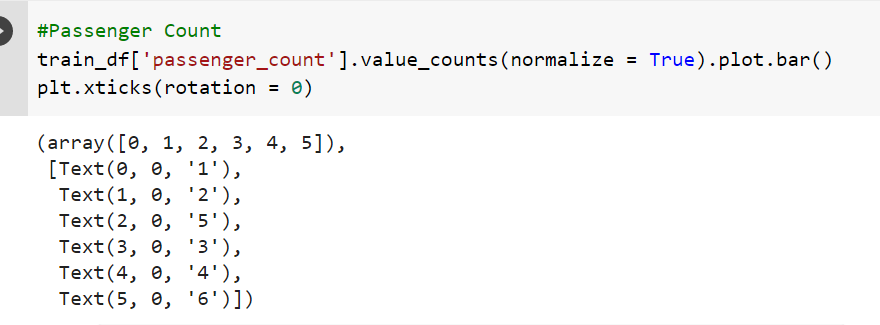
**Expolatery Data Analysis**

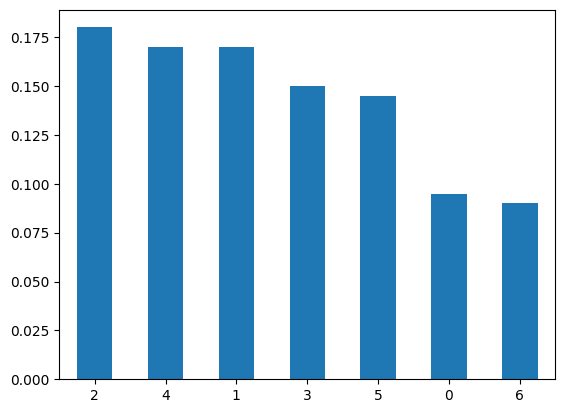


**Categorical Features**

**1) Passenger Count Observation:**

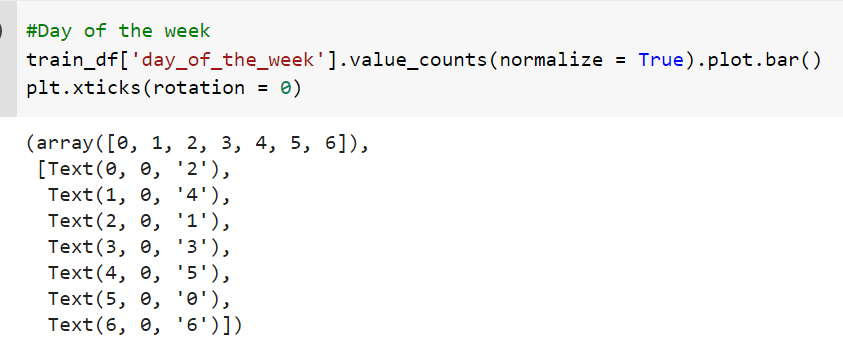
Almost 70% of trips had only 1 passanger

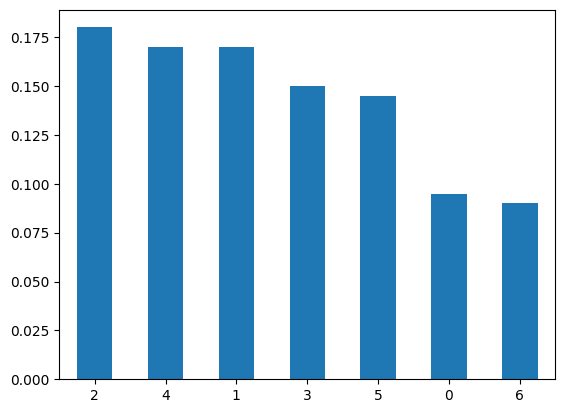




2) Day of the week Observation:

As expected, Friday (weekend eve) and Saturday had the highest number of trips

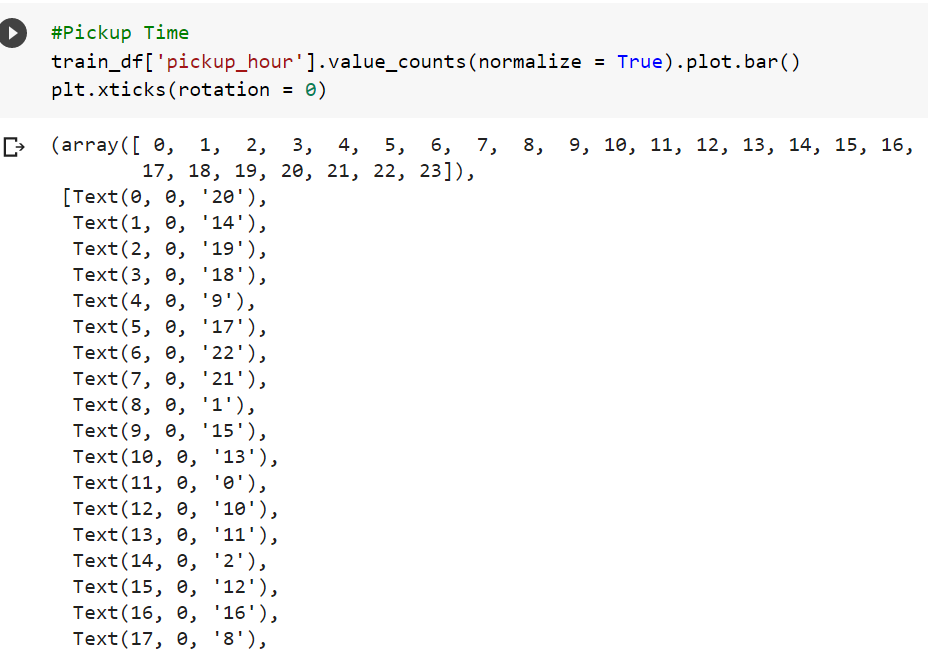


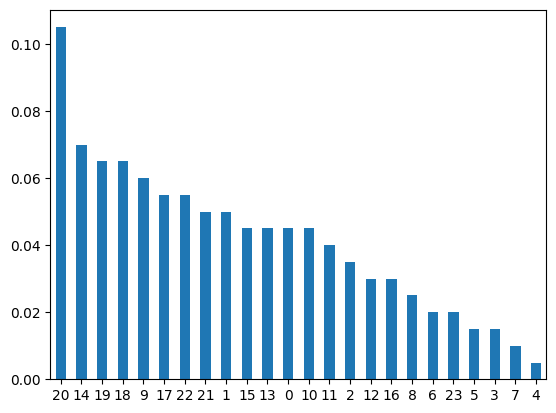


3) Pickup time Observation:

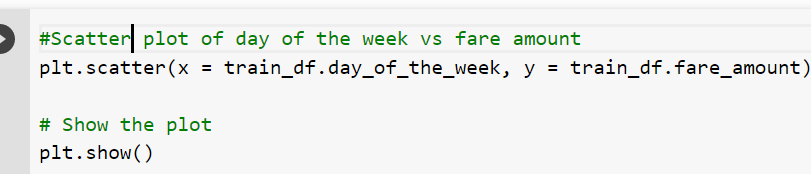
As expected, the number of trips is low during 1am - 5am, and peak between 7pm - 9pm

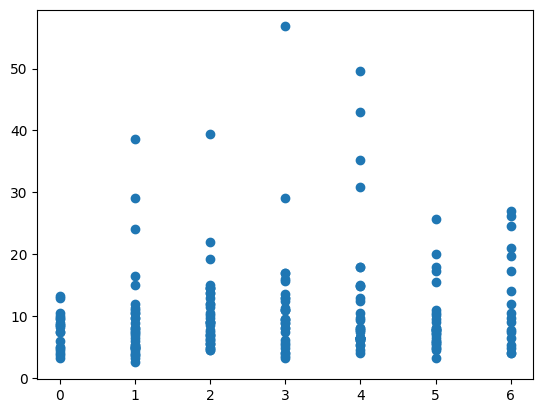
We can clearly see that during weekdays, peak hours are during the day while on weekends, peak hours are late in the night



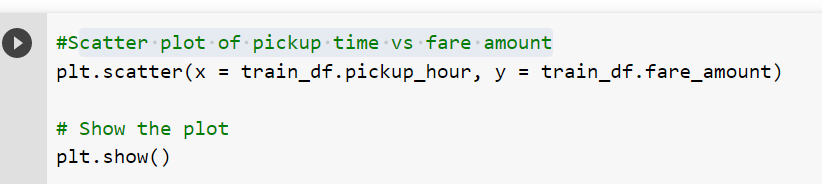


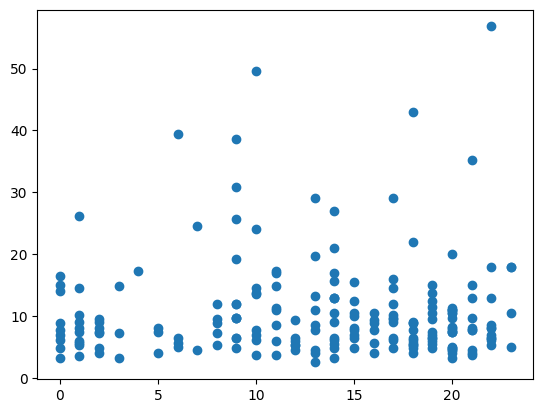
4) Scatter plot of day of the week vs fare amount



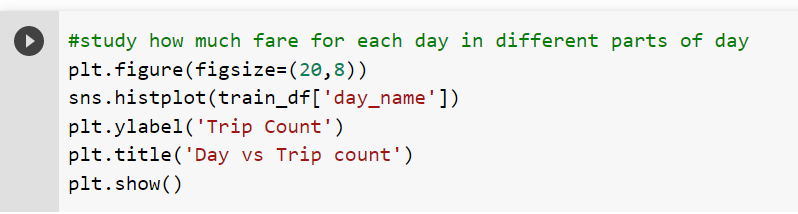


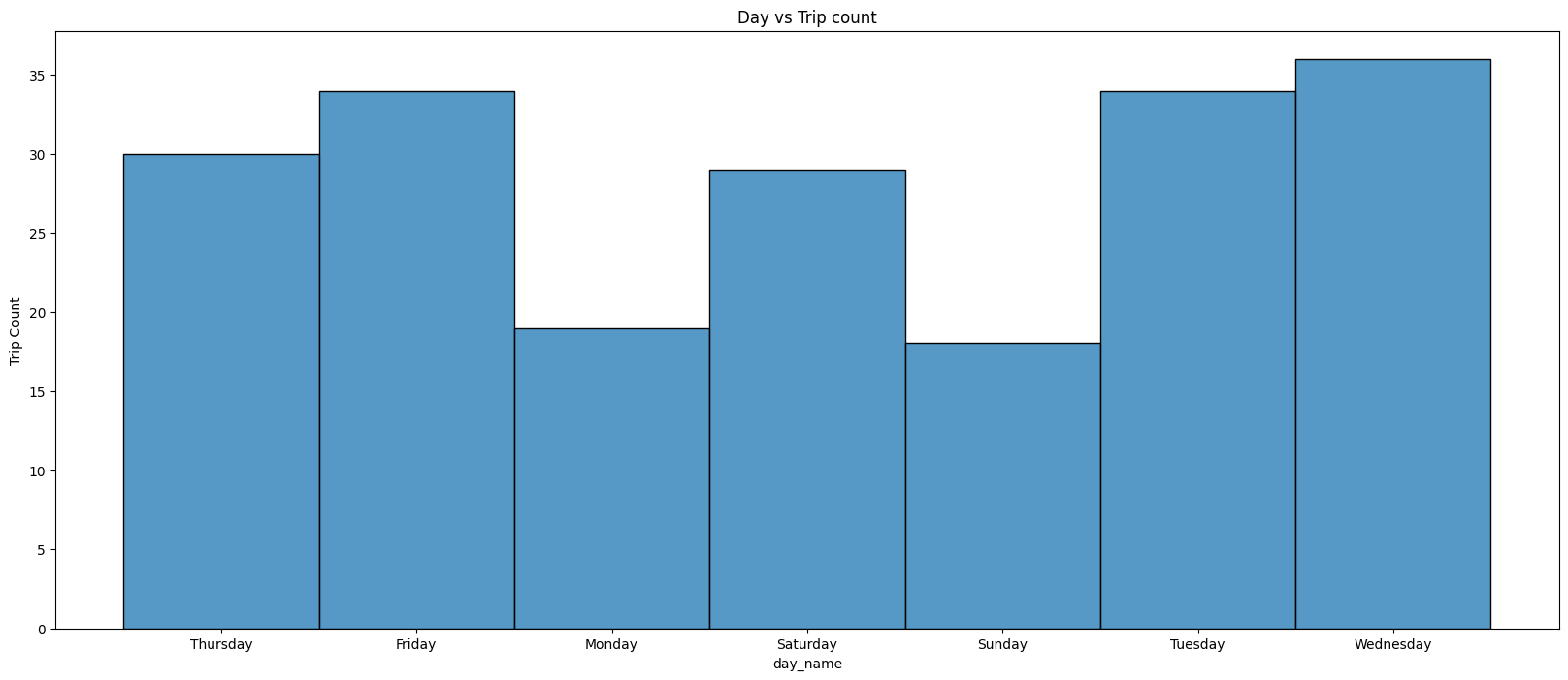
5) Scatter plot of pickup time vs fare amount



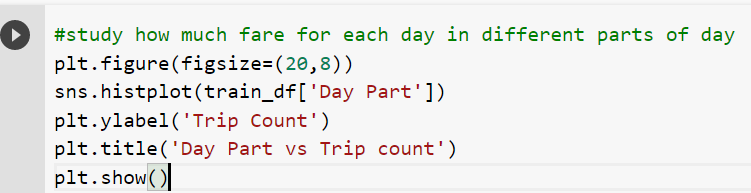


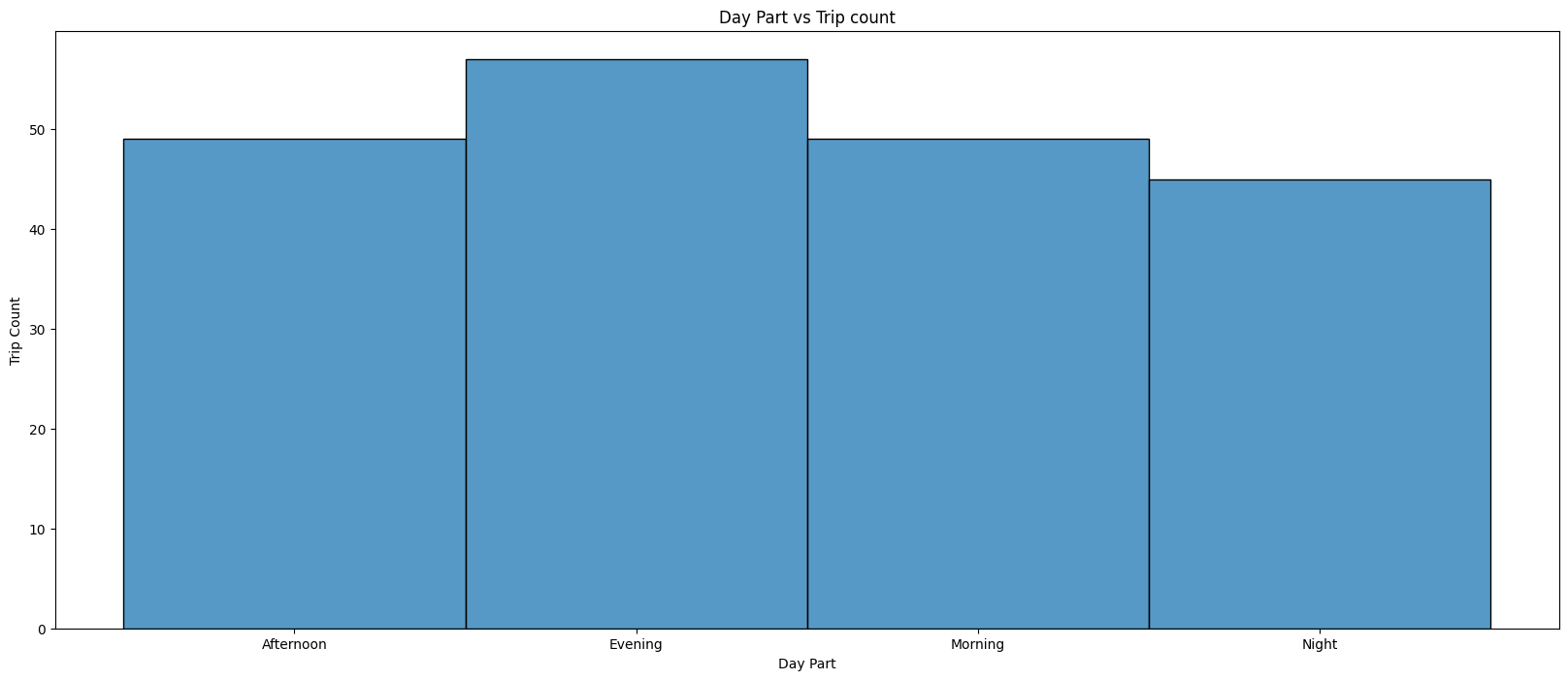
6) Fare for each day in different parts of day



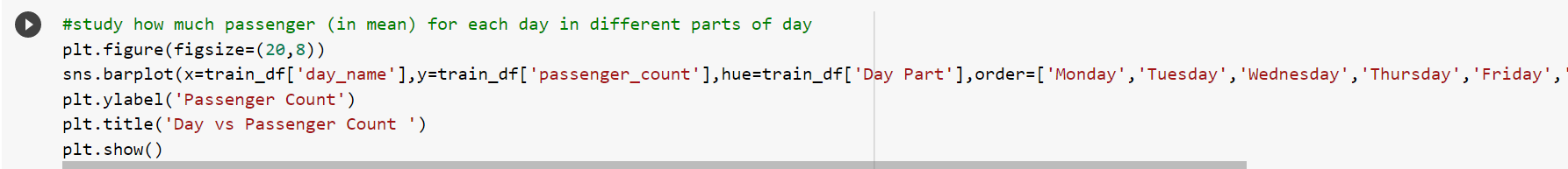


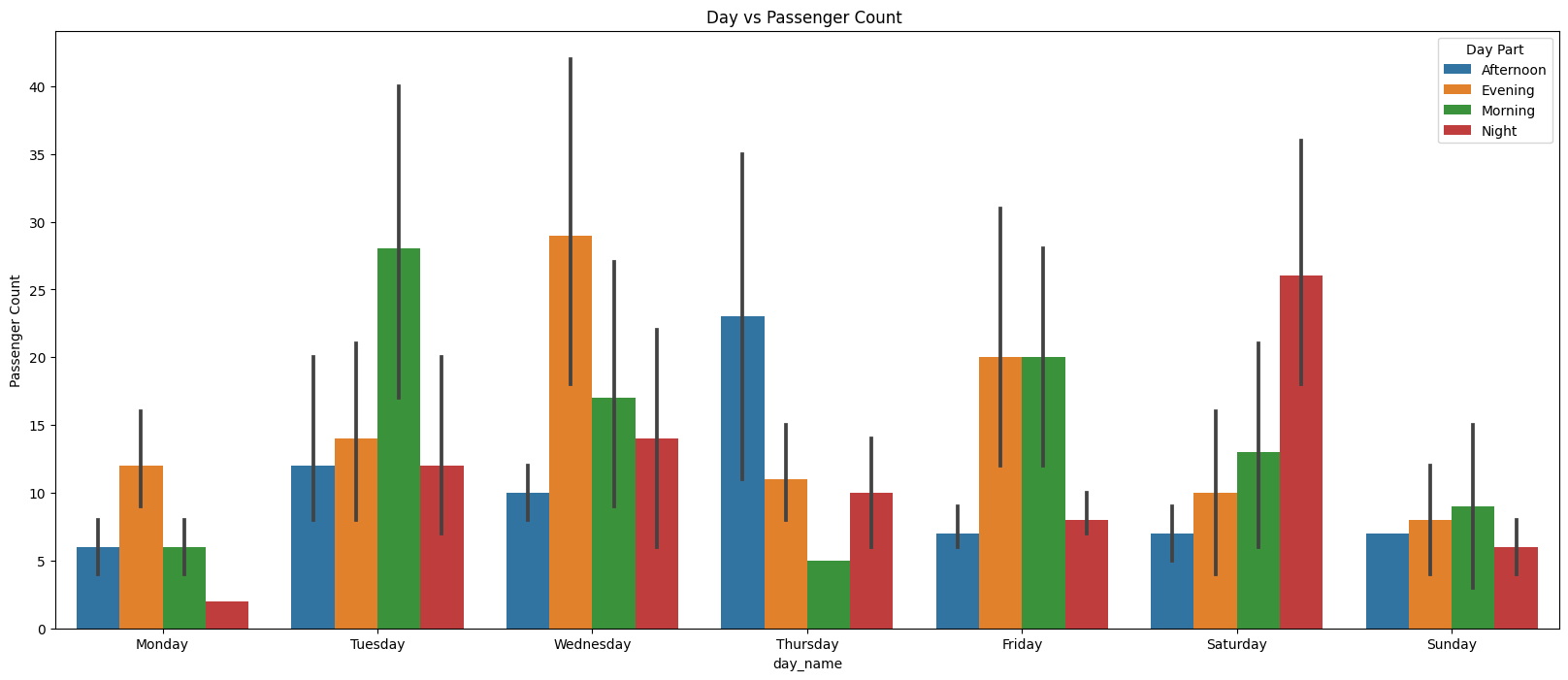
7) Fare for each day in different parts of day



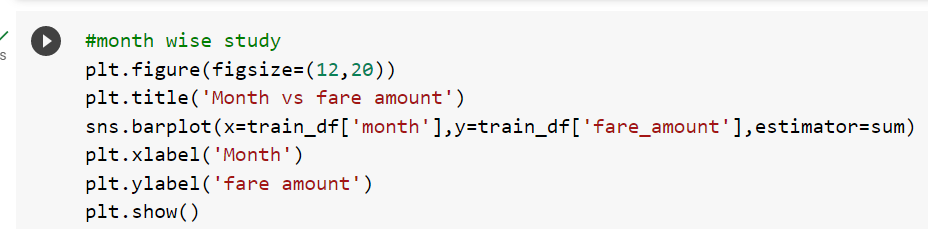


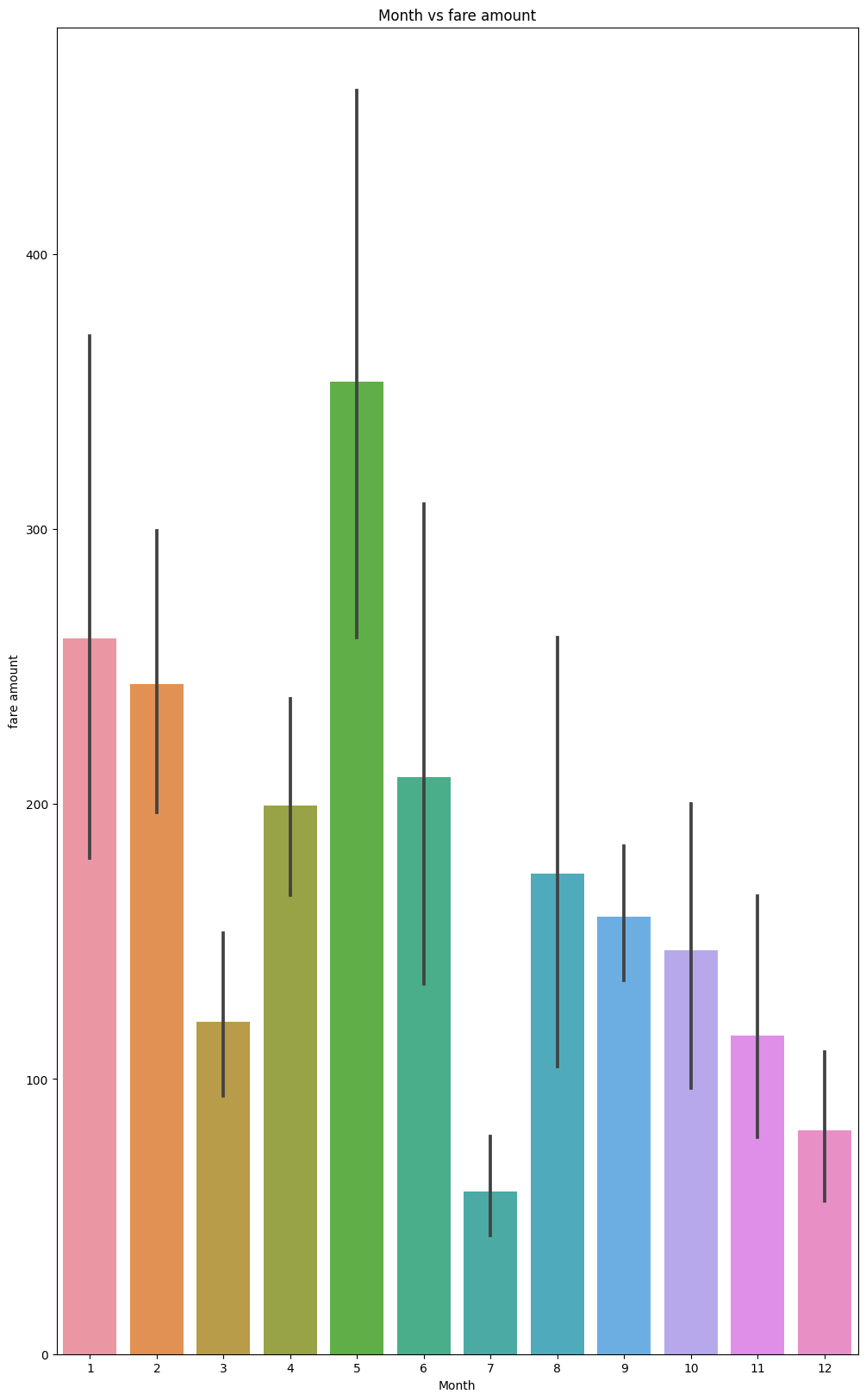
8) Study on how much passenger (in mean) for each day in different parts of day



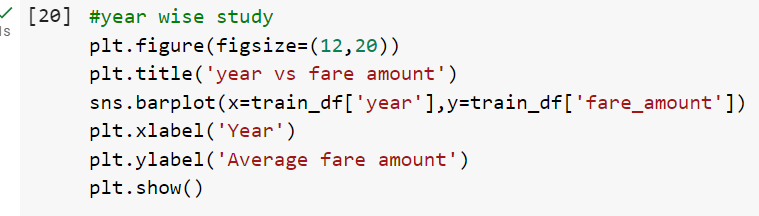


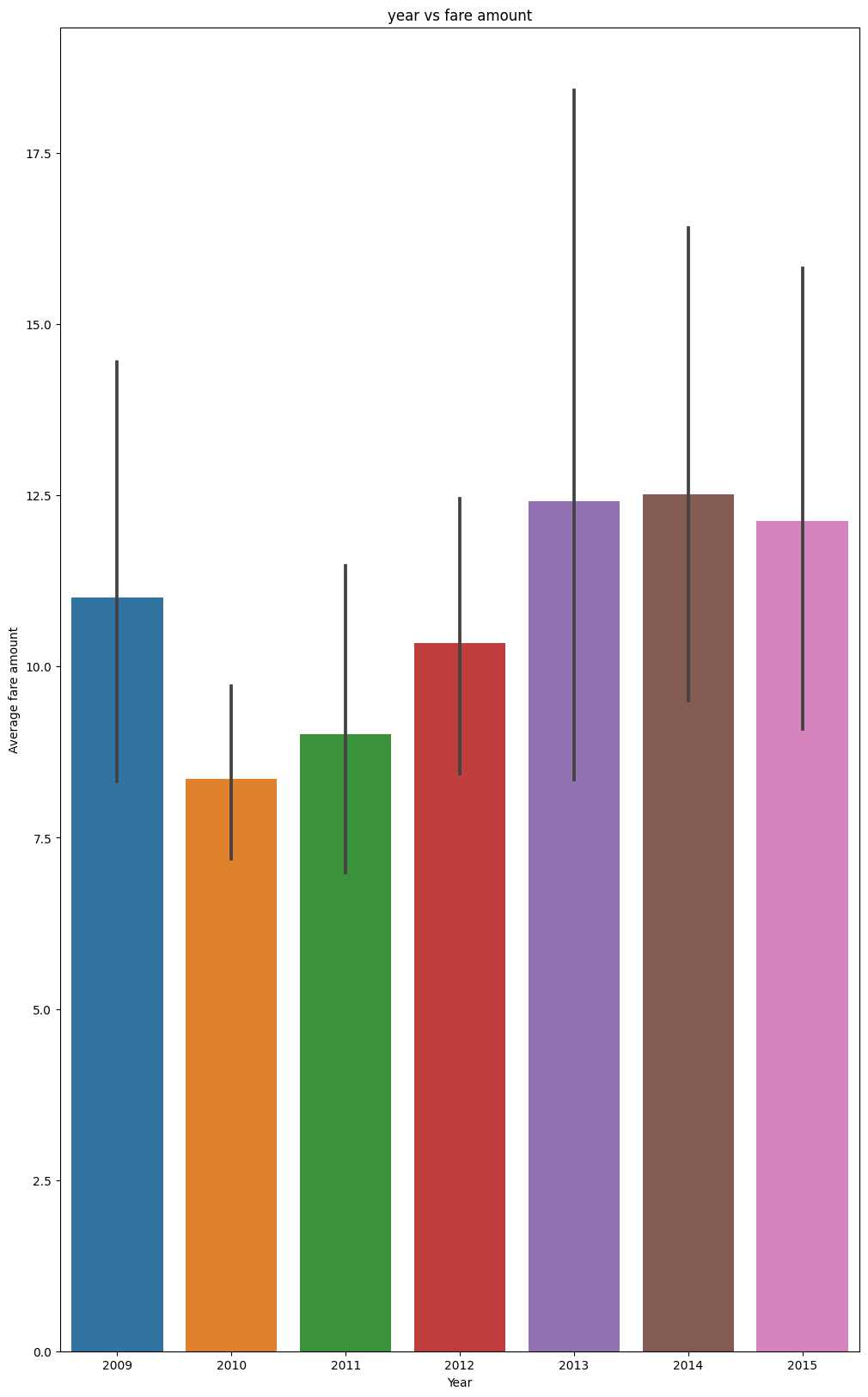
9) Month Wise Study



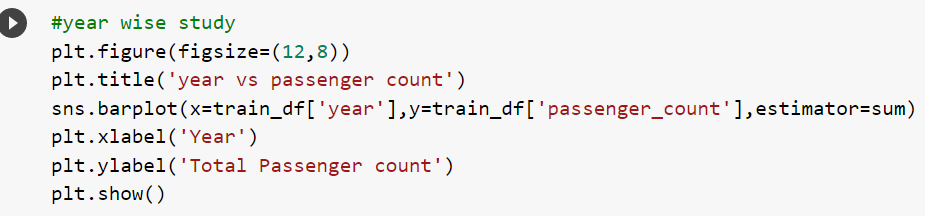


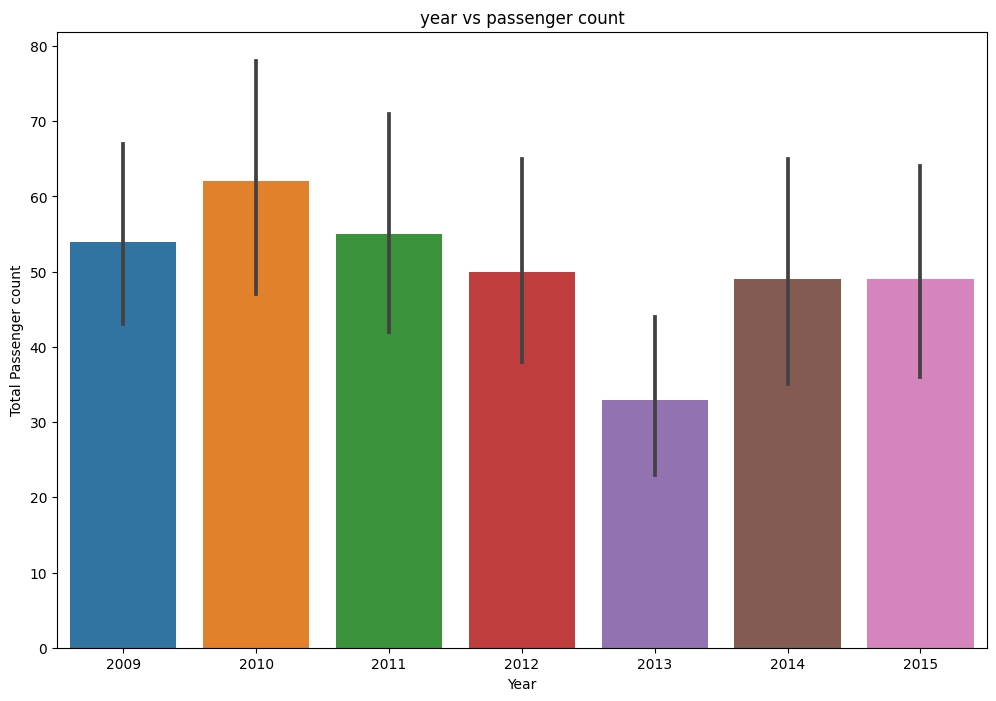
10) Year Wise Study Vs Fare Amount



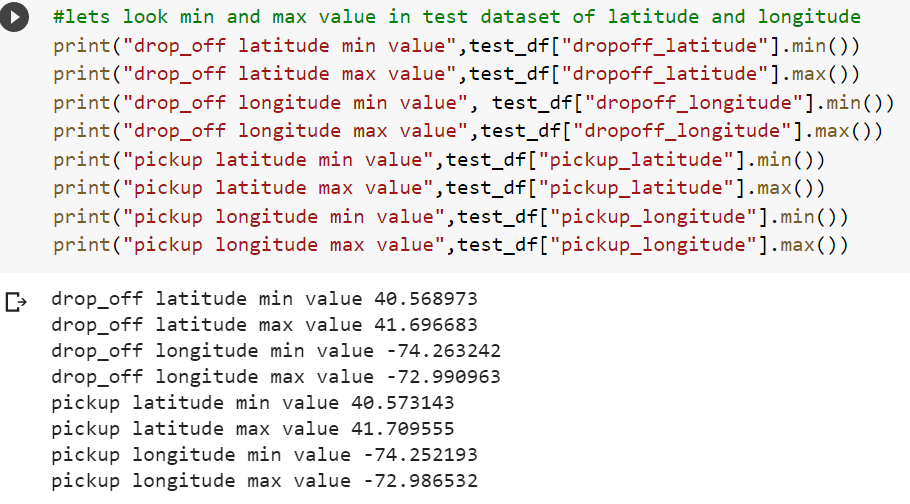


11) Year Wise Study Vs Passenger Count

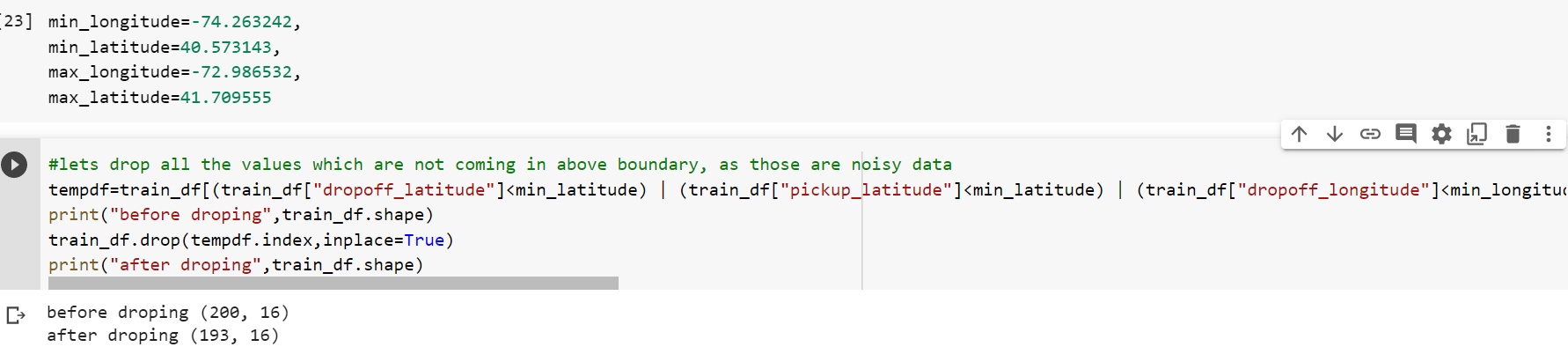


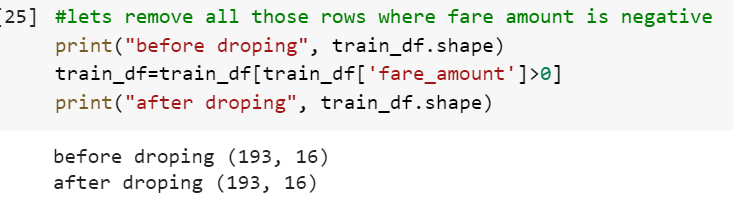


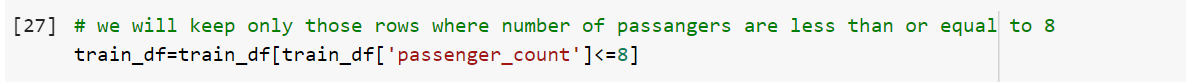
**Data Cleaning**

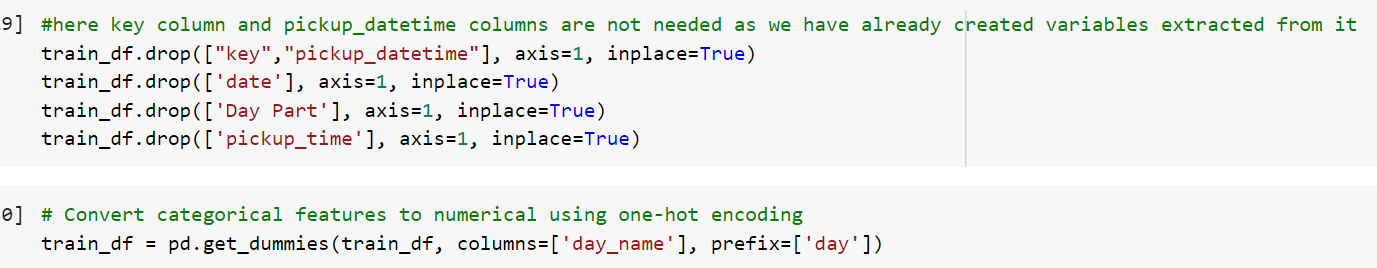


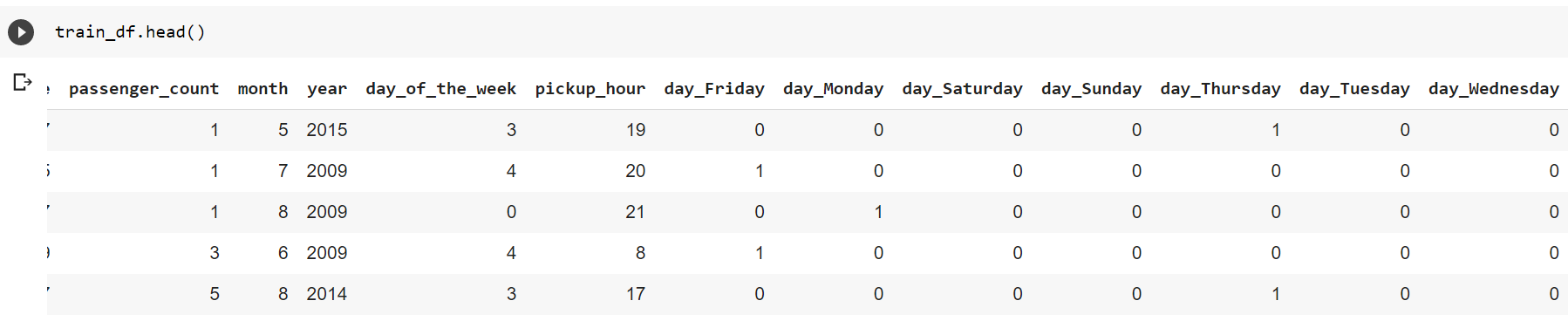
# we can see what range of latitude and longitude of our test dataset is, let us keep the range same in our train set so that even noisy data is remove and we have only the values



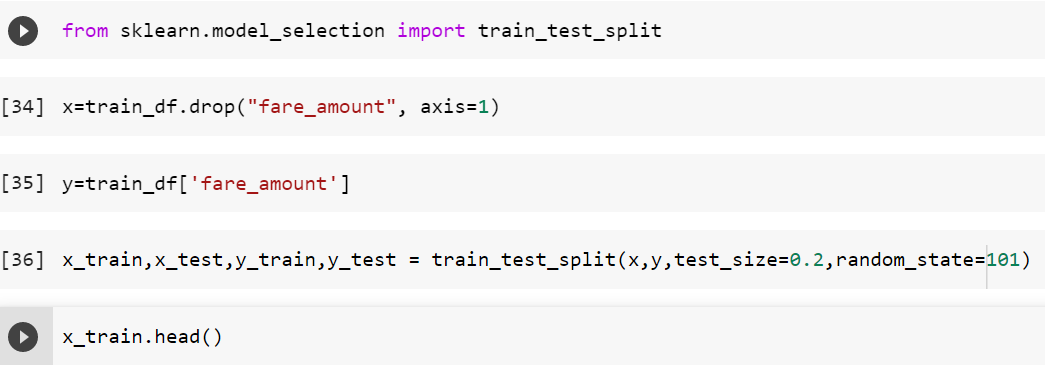


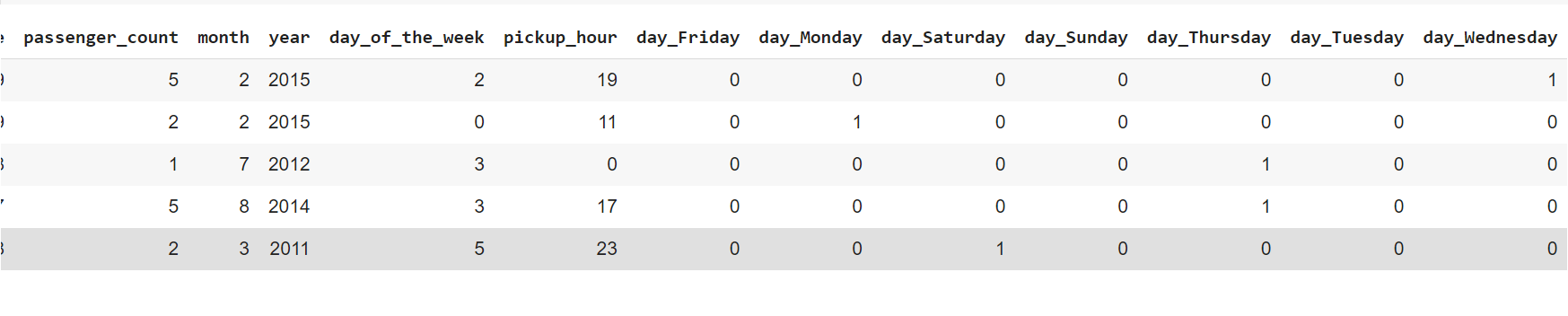


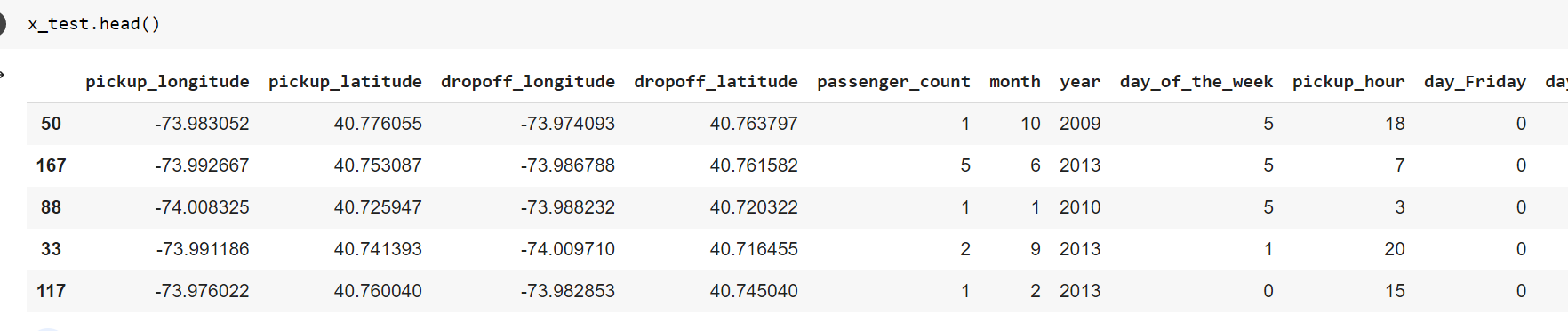


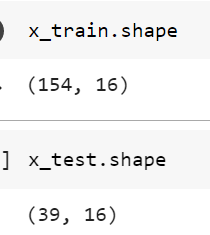


# let us divide the data set into train and validation test set









**Linear Regression Model:**

Linear regression is a type of supervised learning algorithm used in machine learning and statistical analysis. It is used to predict a continuous output variable (also called a response or dependent variable) based on one or more input variables (also called predictors or independent variables).

The linear regression model assumes that there is a linear relationship between the input variables and the output variable. In other words, the output variable can be expressed as a linear combination of the input variables, where each input variable is multiplied by a corresponding weight or coefficient, and the sum of these products is added to a constant term (also called the intercept or bias term).

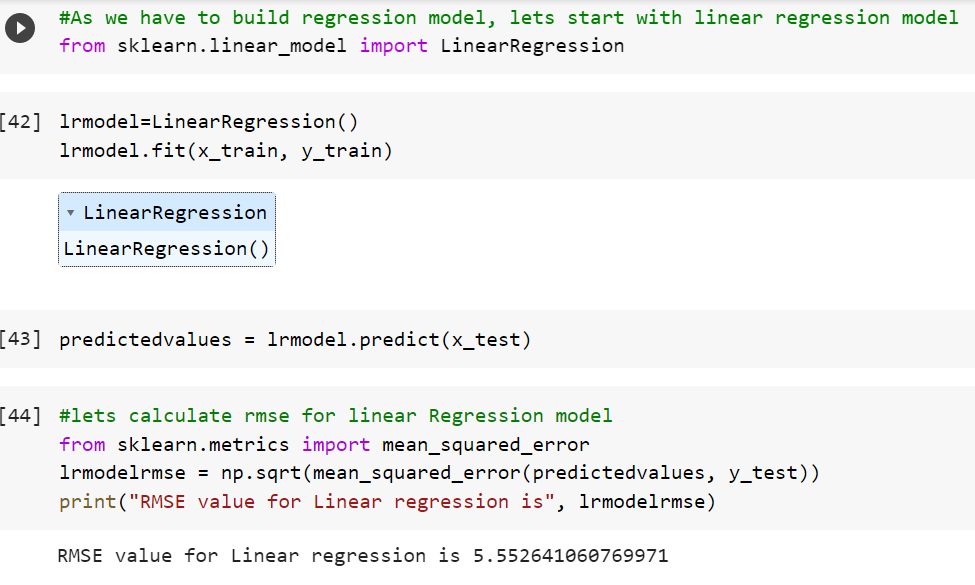
The general equation for a linear regression model with one input variable is:

y = β0 + β1x + ε

where y is the output variable, x is the input variable, β0 is the intercept term, β1 is the coefficient of x, and ε is the error term, which represents the variability of the output variable that is not explained by the input variable.

The goal of linear regression is to estimate the values of the coefficients β0 and β1 that best fit the observed data, to make accurate predictions for new input values. This is typically done by minimizing the sum of squared errors (SSE) between the predicted values and the actual values of the output variable.

Linear regression can be used for both simple linear regression (with one input variable) and multiple linear regression (with two or more input variables). It is a simple and widely used technique for predicting continuous outcomes and is often used as a baseline model for more complex algorithms.



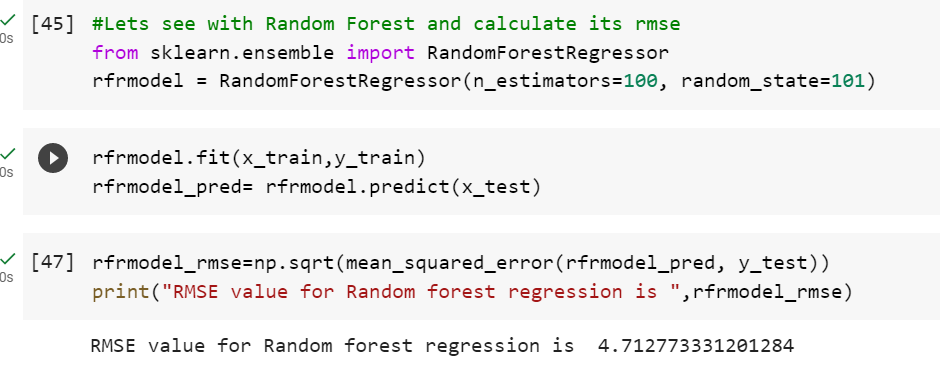
**Random Forest Model**

Random forest is a supervised learning algorithm used in machine learning for both classification and regression problems. It is an ensemble method that combines multiple decision trees to make more accurate predictions.

A decision tree is a simple model that uses a series of binary splits to classify or predict the output variable based on the input variables. Each split divides the input space into two subsets based on a single input variable and a threshold value. The decision tree is built recursively by choosing the input variable and threshold that best separates the data into the two subsets.

Random forest combines multiple decision trees to reduce the variance and improve the accuracy of the predictions. It works by generating a set of random subsets of the training data, and then building a decision tree for each subset using a random subset of the input variables. This process is repeated to generate a forest of decision trees.

To make a prediction for a new input, the random forest algorithm aggregates the predictions of all the individual decision trees. For classification problems, the final prediction is typically the class that receives the most votes from the individual trees. For regression problems, the final prediction is typically the average of the individual tree predictions.



# RandomForest Regressor is giving good value, so we can use it as final model.