7/26/2019

Simran Chhabra

for Edwisor.com

Cab Fare Prediction

Contents

1. **Introduction**
   1. Problem Statement
   2. Data
2. **Methodology**
   1. Pre-Processing
      1. Feature Selection
      2. Missing values
      3. Outlier Analysis
   2. Modelling
      1. Dependency Check
      2. Linear Regression
      3. Random Forest Regression
      4. K Neighbours Regression
3. **Conclusion**
   1. Model Evaluation
   2. Model Selection

**Appendix A – Python Code**

**Appendix B – R Code**

**References**

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

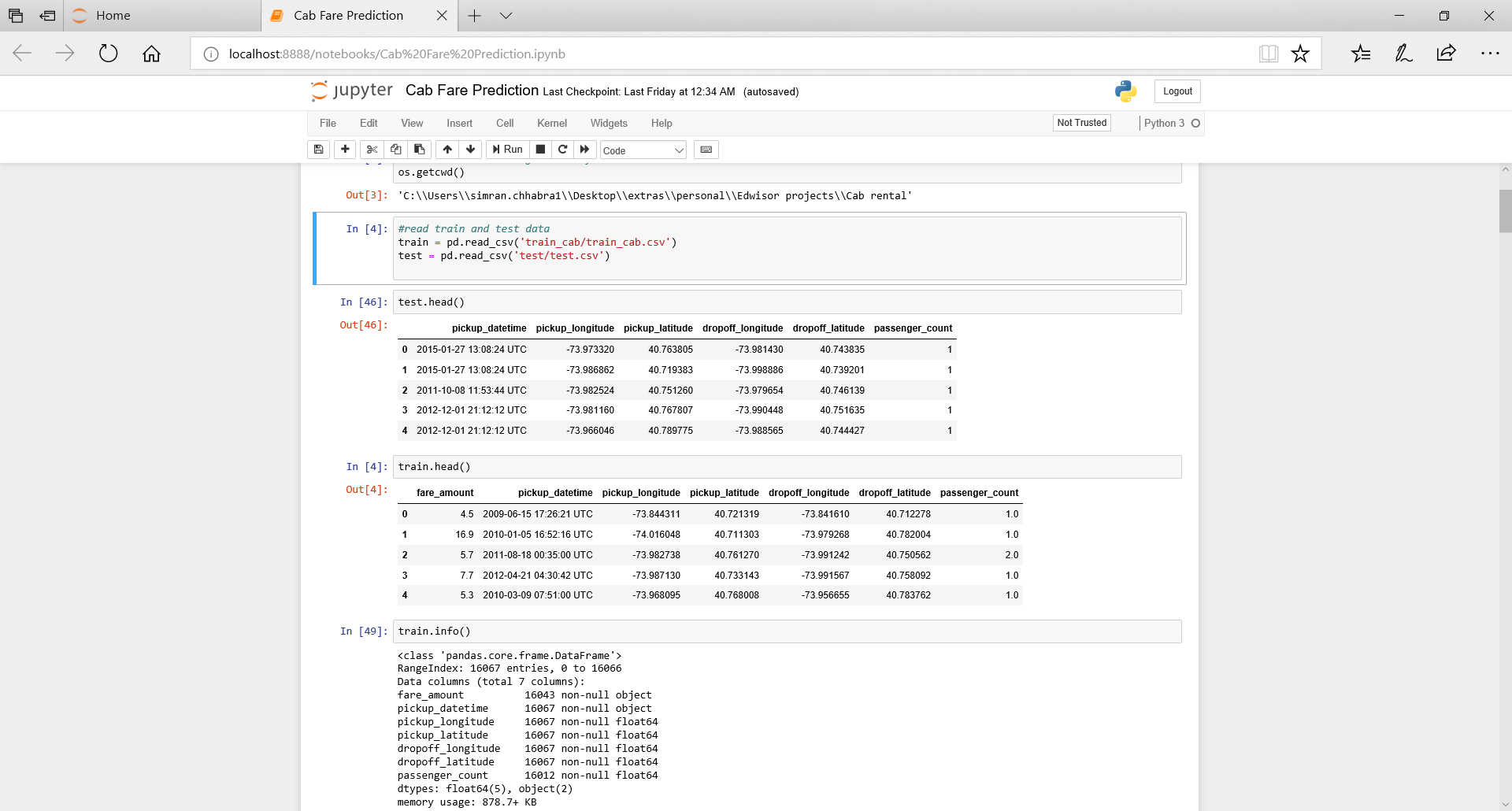
**1.2 Data**

1) train\_cab.zip

2) test.zip

Number of attributes:

* pickup\_datetime - timestamp value indicating when the cab ride started.
* pickup\_longitude - float for longitude coordinate of where the cab ride started.
* pickup\_latitude - float for latitude coordinate of where the cab ride started.
* dropoff\_longitude - float for longitude coordinate of where the cab ride ended.
* dropoff\_latitude - float for latitude coordinate of where the cab ride ended.
* passenger\_count - an integer indicating the number of passengers in the cab ride.



**Chapter 2**

**Methodology**

* 1. **Pre-Processing**

Any predictive modelling requires that we look at the data before we start modelling. However, in Data Mining terms *looking at data* means so much more than looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start with, we first detect anomalies in data like missing values, etc. Dependent variable Fare Amount was observed to be of type String. For prediction purposes, it was changed to Numeric. Anomalies like Fare Amount being negative, count of passengers being less than 1 and greater than 6 were detected. Such values were replaced with null and later filled in using appropriate measures.

* + 1. **Feature Selection**

There could be variables in the data which could not be used directly in the analysis. We try to retrieve helpful information from such data. Then all the non-required fields are removed from the data.

1. Fare does not depend on variables pickup latitude, pickup longitude, drop-off latitude, drop-off longitude, rather it depends on the total distance travelled by the passenger. Thus, we first compute haversine distance using these points and save it in a new variable “distance\_travelled”. We then remove those four variables.
2. We retrieved the values day of the month, month, year and hour of the day from pickup datetime and then found their respective relation to fare. We then removed pickup datetime variable.

**2.1.2 Missing values**

Ideally if a variable has less than 30% missing values then they are filled using appropriate metrics like mean, median, mode or K-nearest neighbours. This will preserve the data and there will be no loss of information. In the data, we had replaced abnormal passenger count and fare values with null. Passenger count is a categorical variable thus we filled the null values in it with mode which will return frequent number of travellers. There were 0.17% of the values in fare which were null. We used mean to fill in these values. Other method could be KNN but since it is a time-consuming process and here not many values were blank, mean was suitable.

**2.1.3 Outlier Analysis**

We can graph the data on boxplots and check for any outliers (refer figures 1 and 2). These could be corrected by first finding maximum and minimum values for that variable then dropping any values not lying in the range of minimum and maximum.

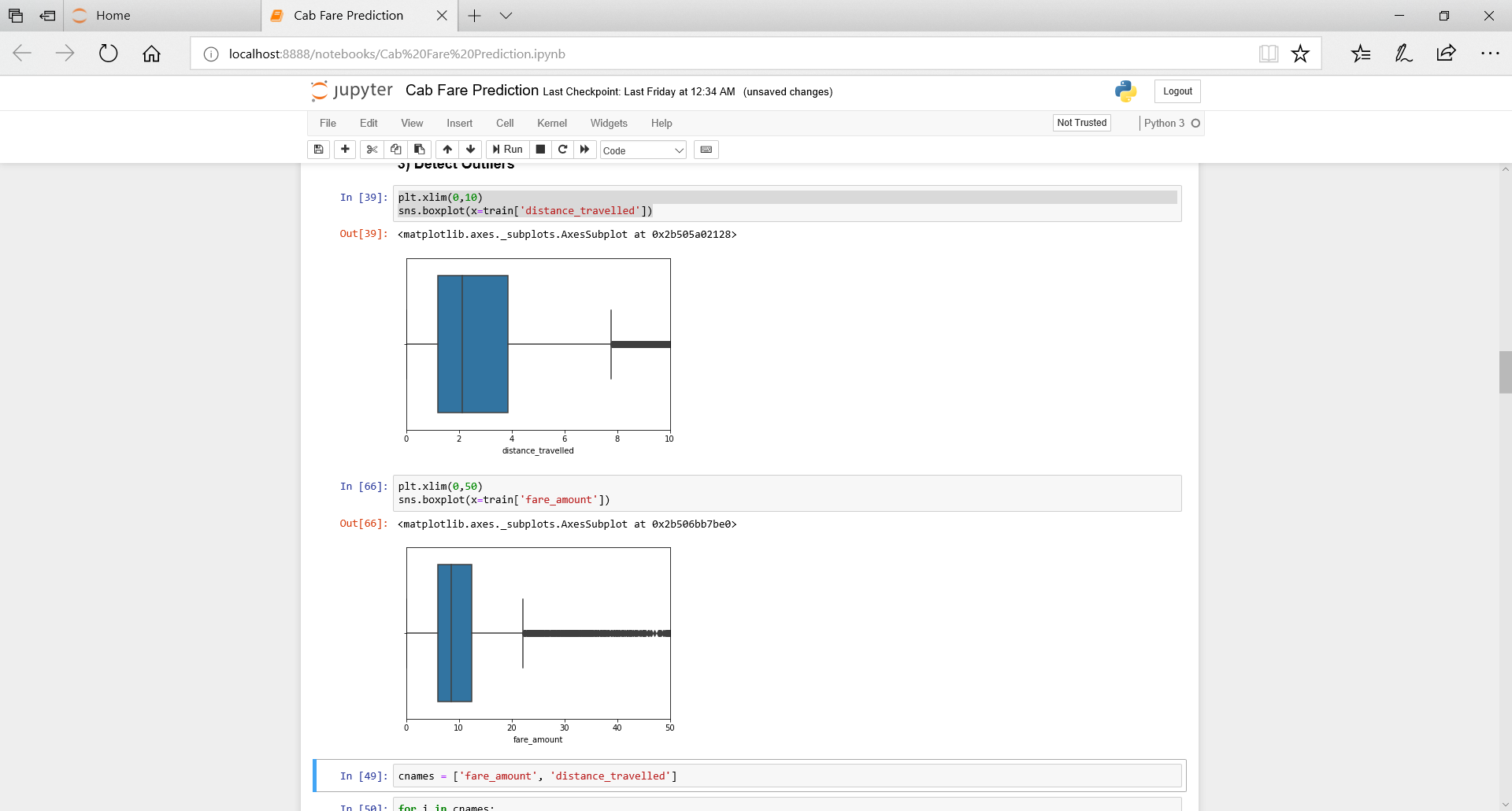
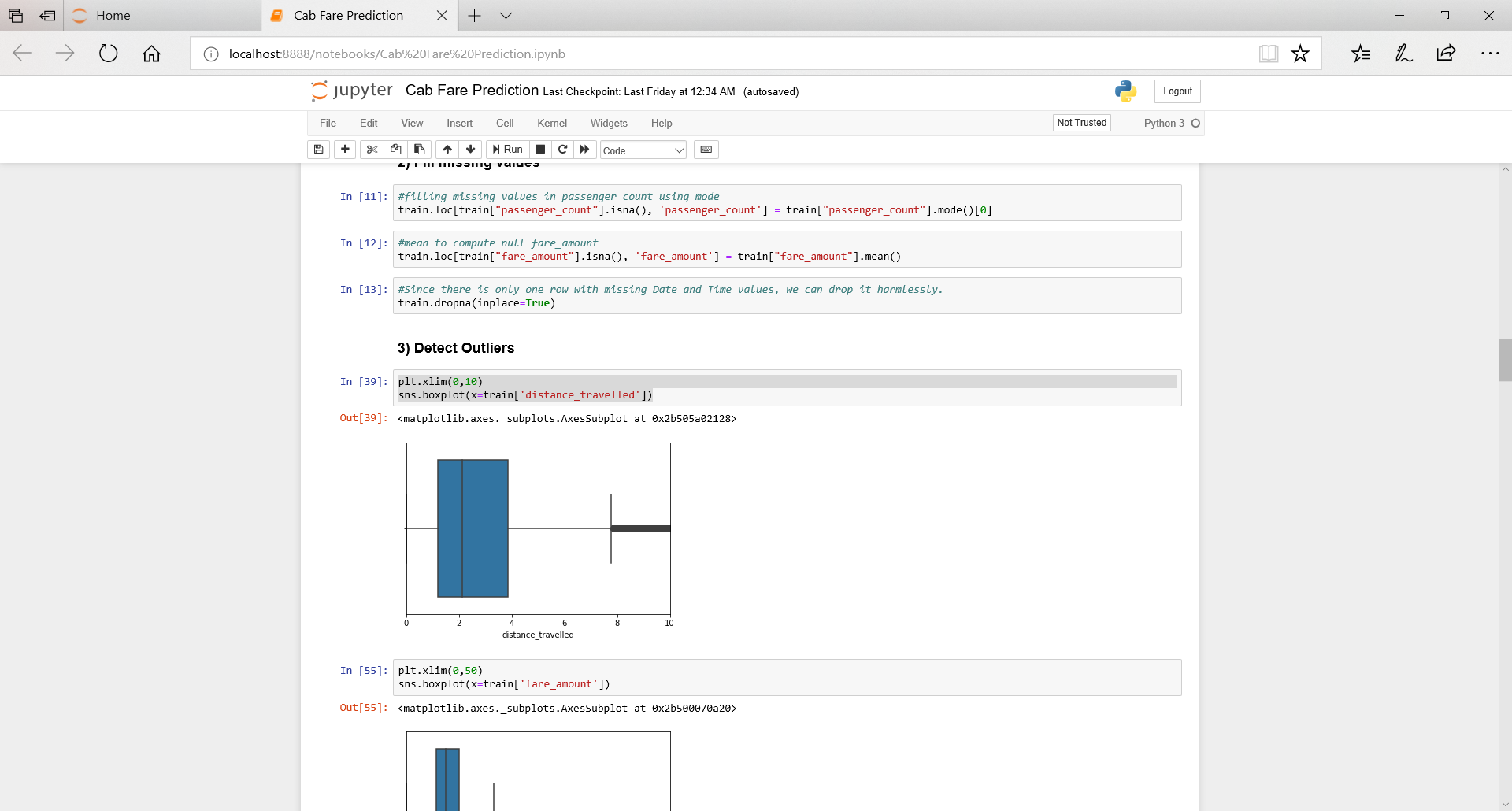
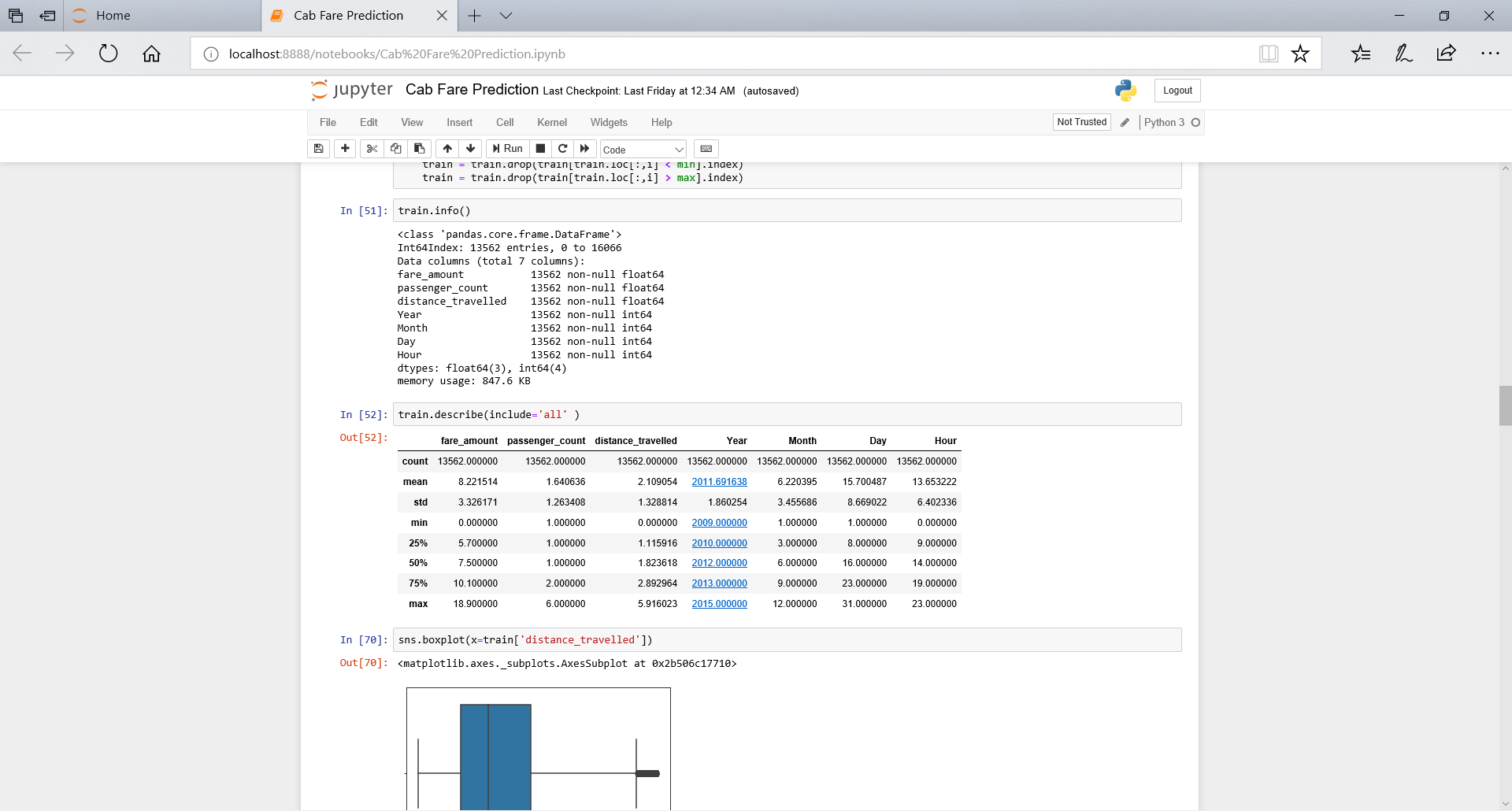


Figure 1 (distance\_travelled) Figure 2 (fare\_amount)

* 1. **Modelling**

During our pre-processing, we understood that the fare does not depend on the passenger’s location but on the distance travelled. It might also depend on what month of the year it is or what day it is or what time it is.

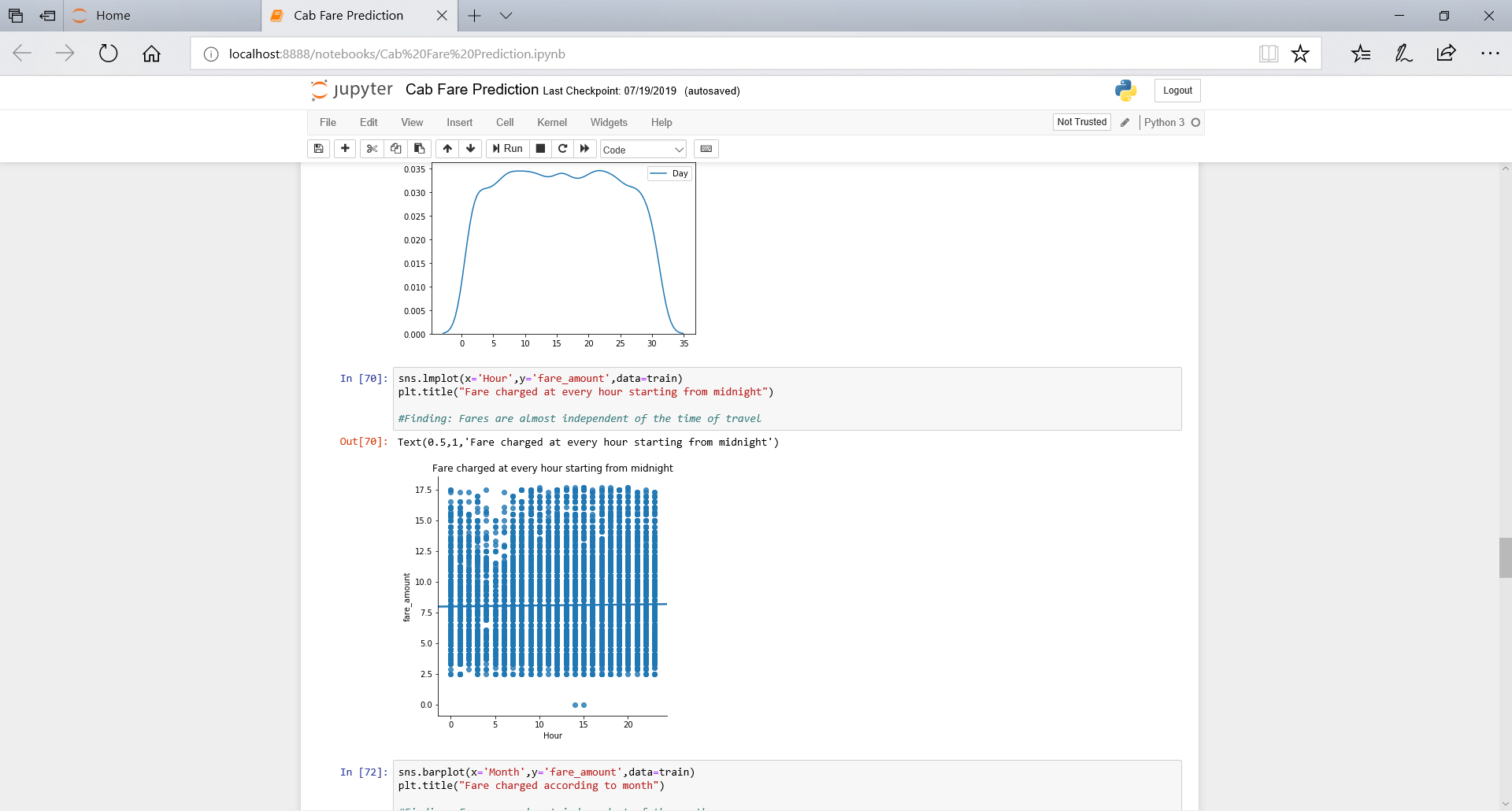
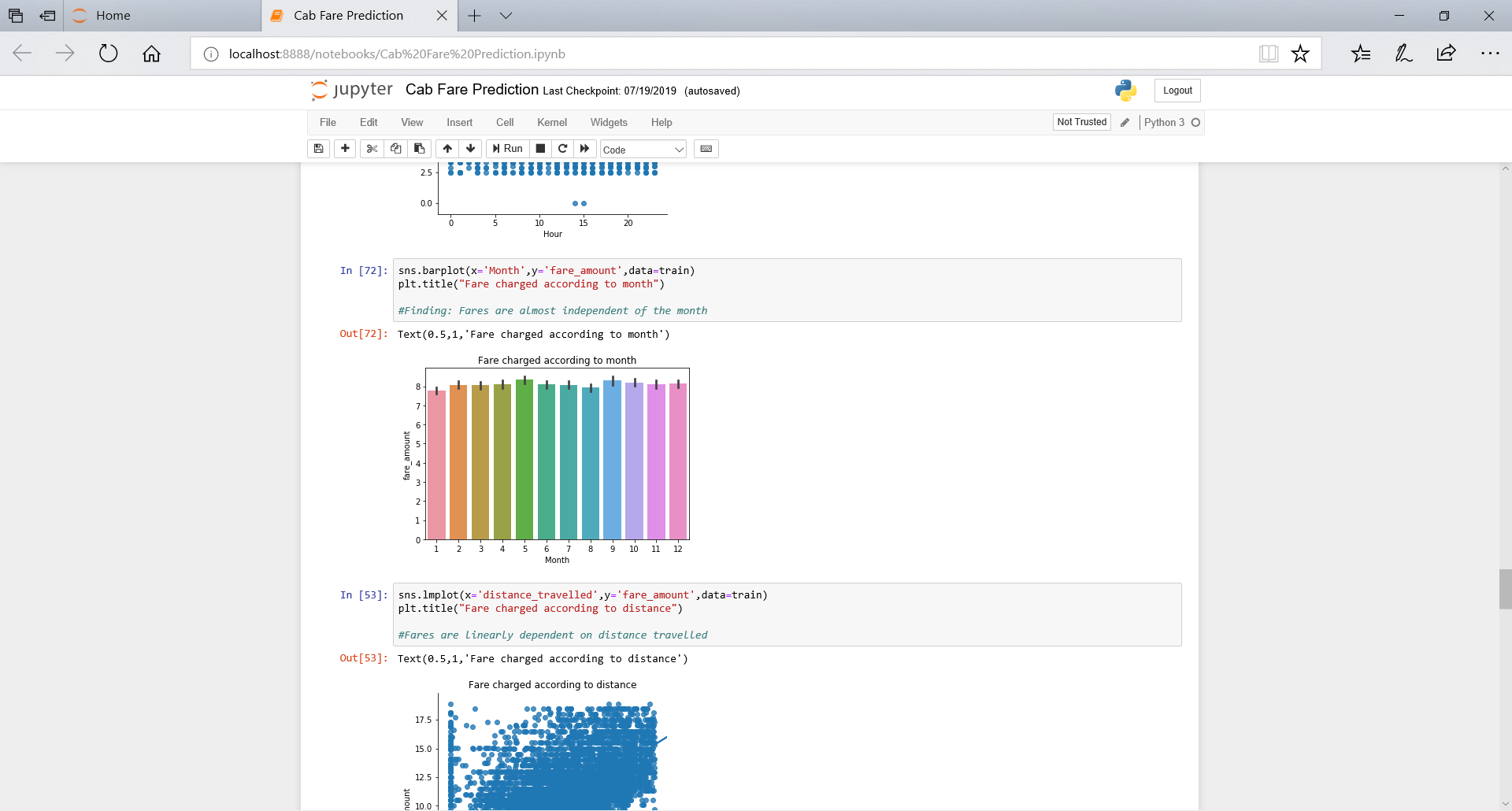
Finally, we have following fields:

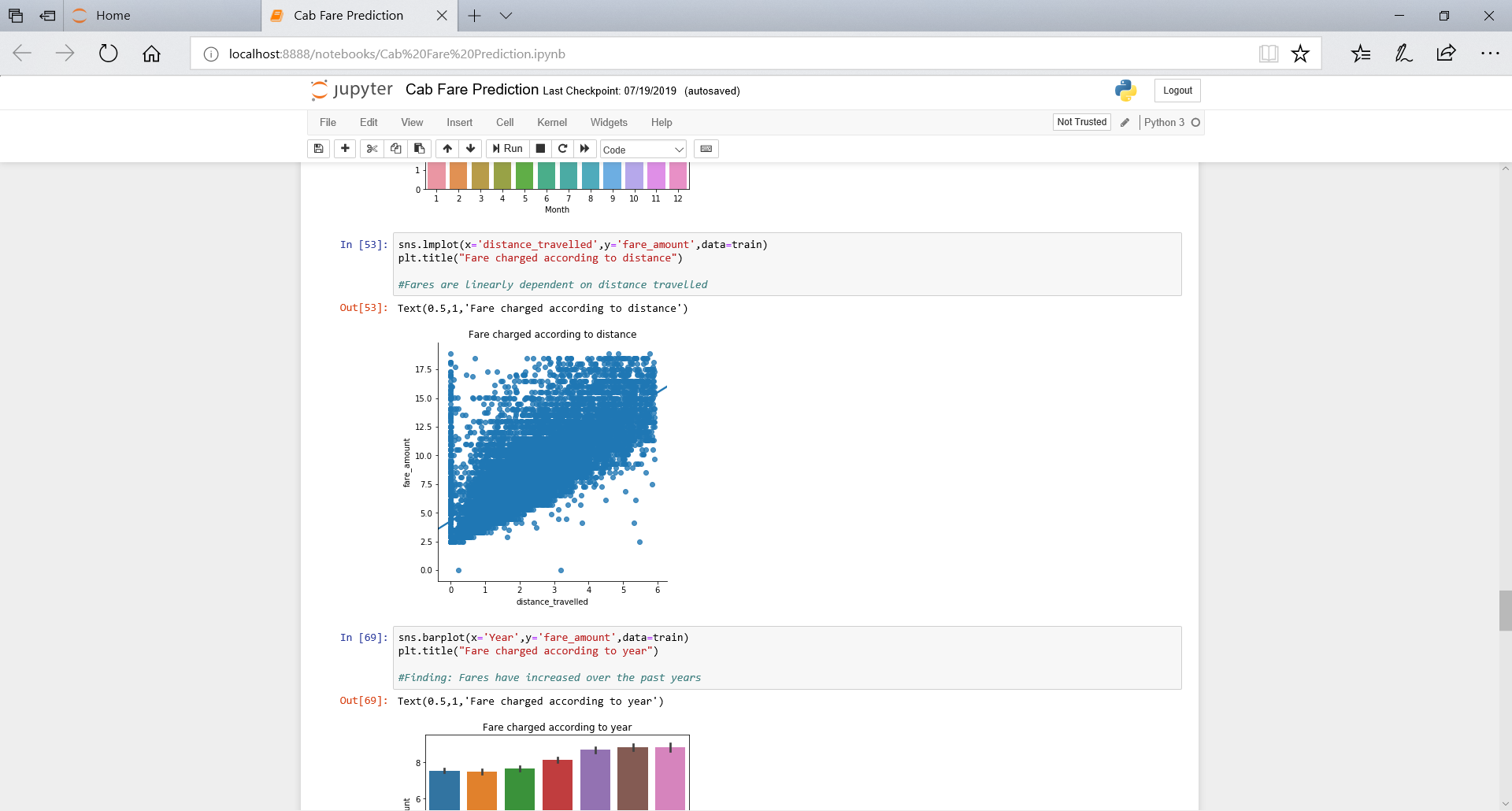
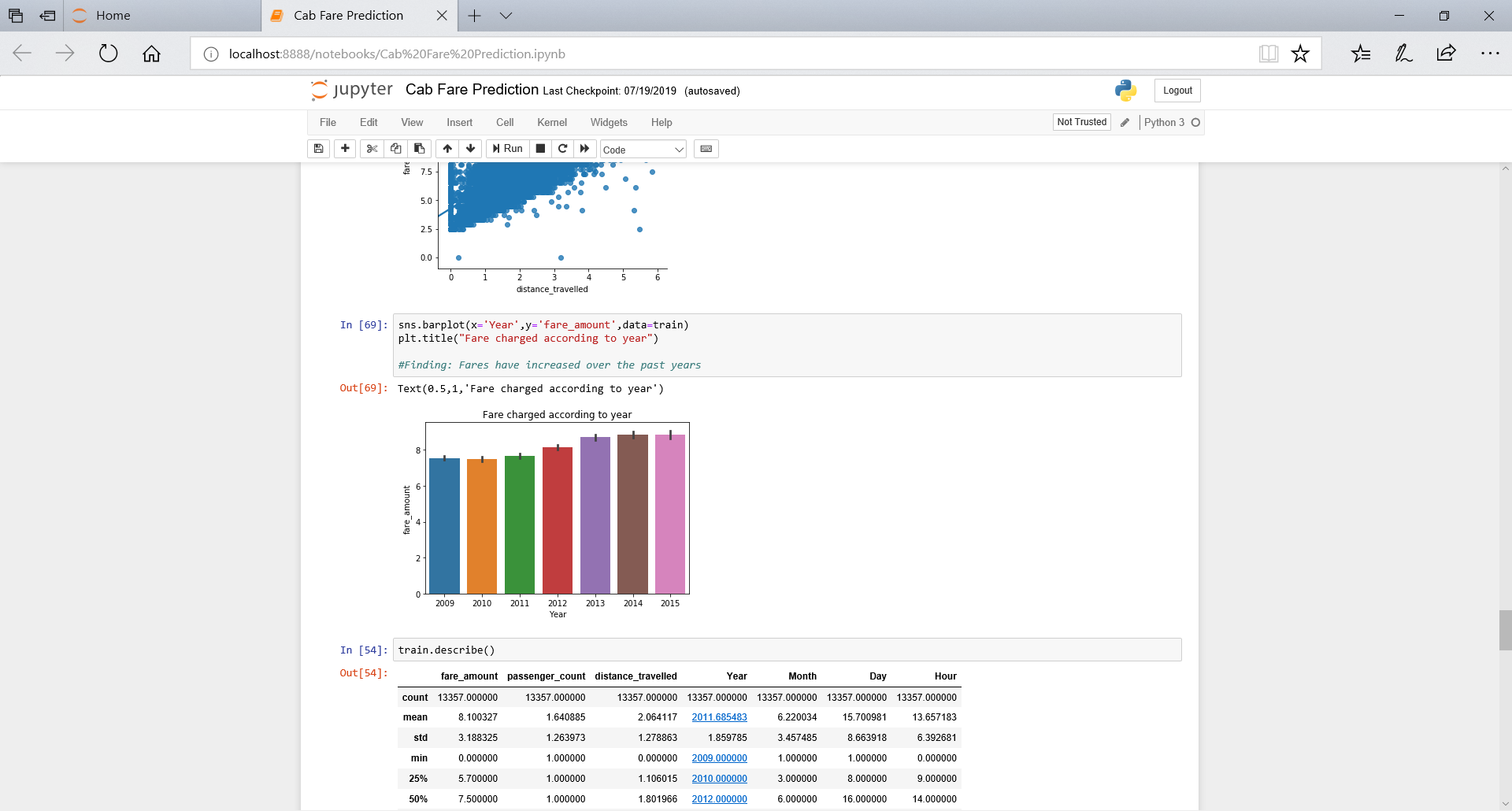


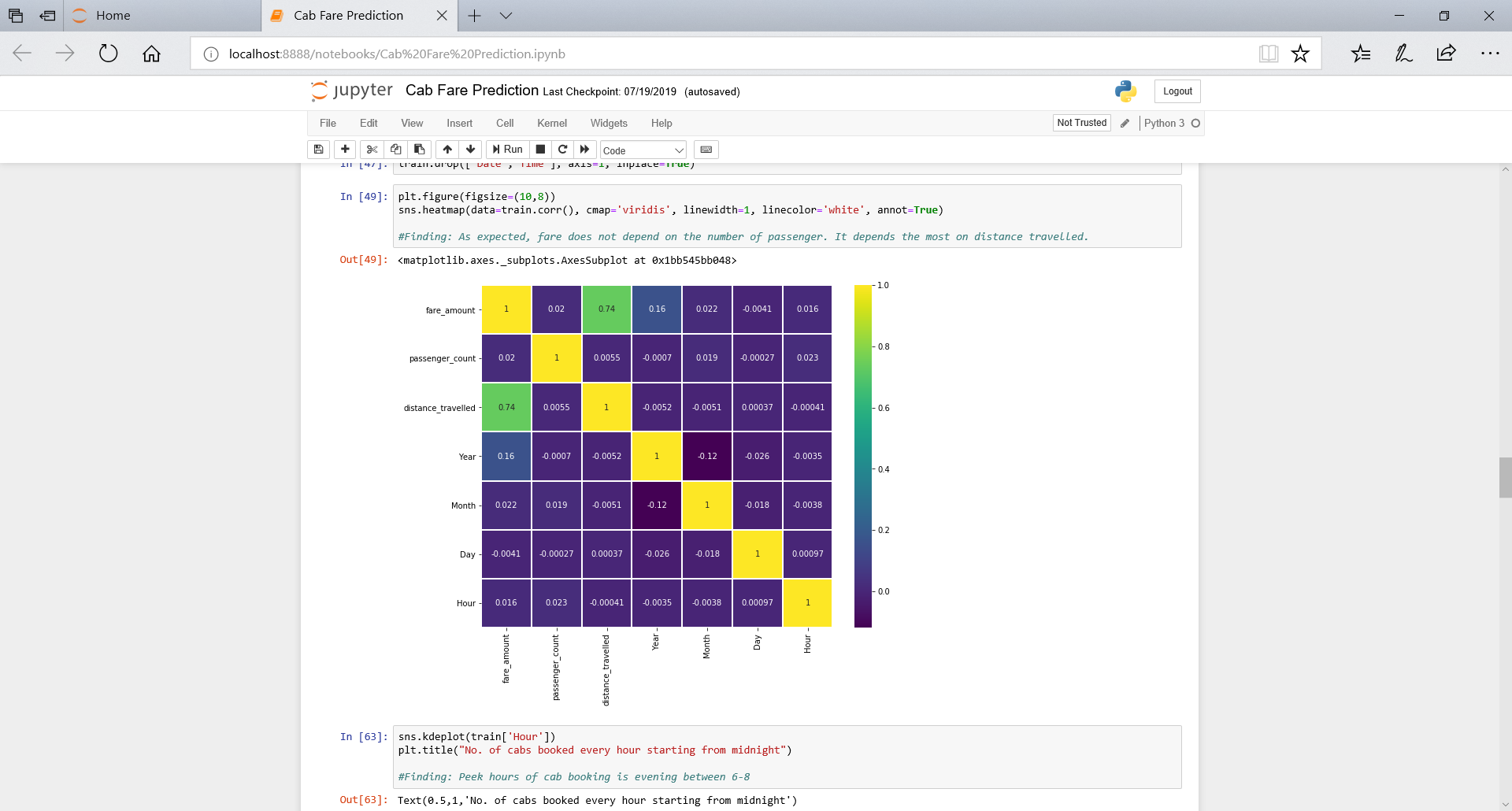
Expected relations:

1. Fare should linearly depend on the distance travelled i.e. if more kms are travelled then fare should be more.
2. It should also depend on the time of the day. Is it morning 5 am when a smaller number of people take cabs or is it 6 pm when most people take cabs. Fare should be higher in the evening 6-9 pm and morning 7-11 am when the demand for the cabs is more.
3. Cabs were not so prominent in initial years of their existence therefore fare should be less in those years.
4. It should not much depend on what day or month it is as the demand is consistent throughout.
   * 1. **Dependency check**

Visualize and check dependency of target variable on independent variables.



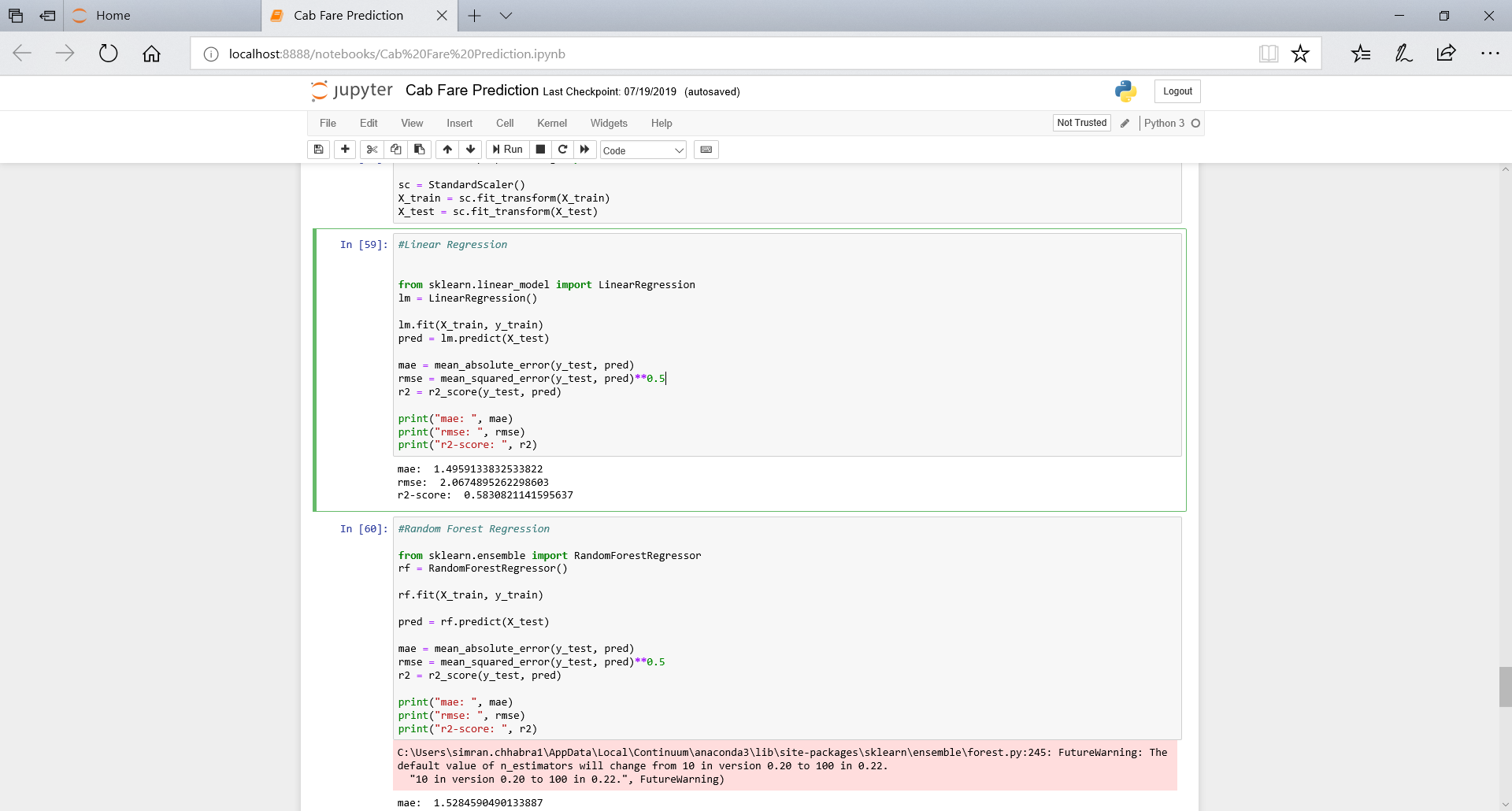
Dependencies match mostly with our expectations except for the point 2 i.e. fare doesn’t really depend on the time of the day the cab is booked. Other than that, there is a strong linear dependency on distance travelled and some dependency on the year.

The data is first split into train and test data to predict and then check the error percentage. We move further to apply regression models on the train data to predict test values. The best model is picked up and then the values are predicted for the provided test data and results are shared.

* + 1. **Linear Regression**

We first check for assumptions needed for Linear regression to be performed that are: linear relationship, normal distribution, no multicollinearity and no auto-correlation. Data is scaled if not done already.

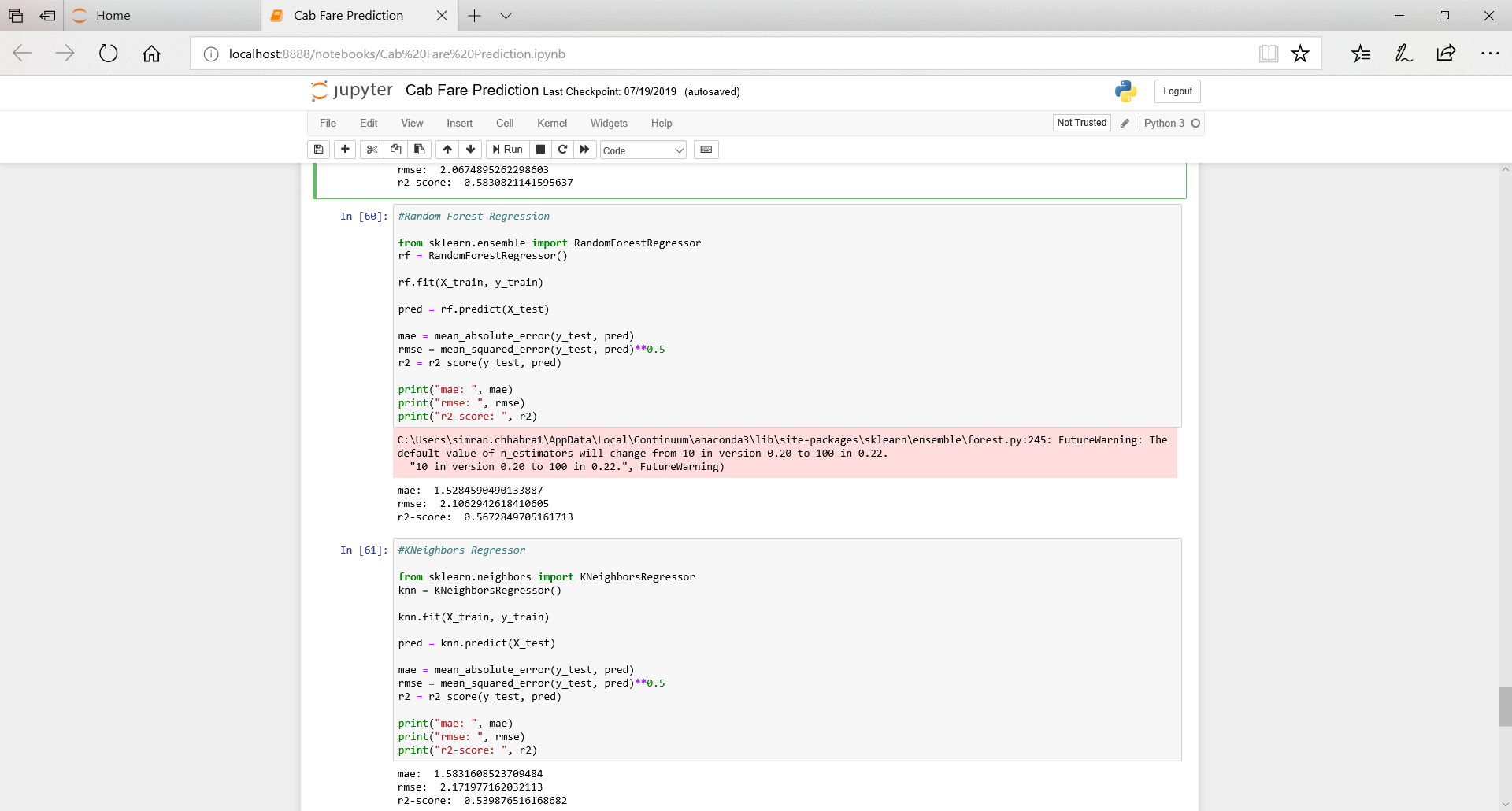
We get following error metrics on performing linear regression:



* + 1. **Random Forest Regression**

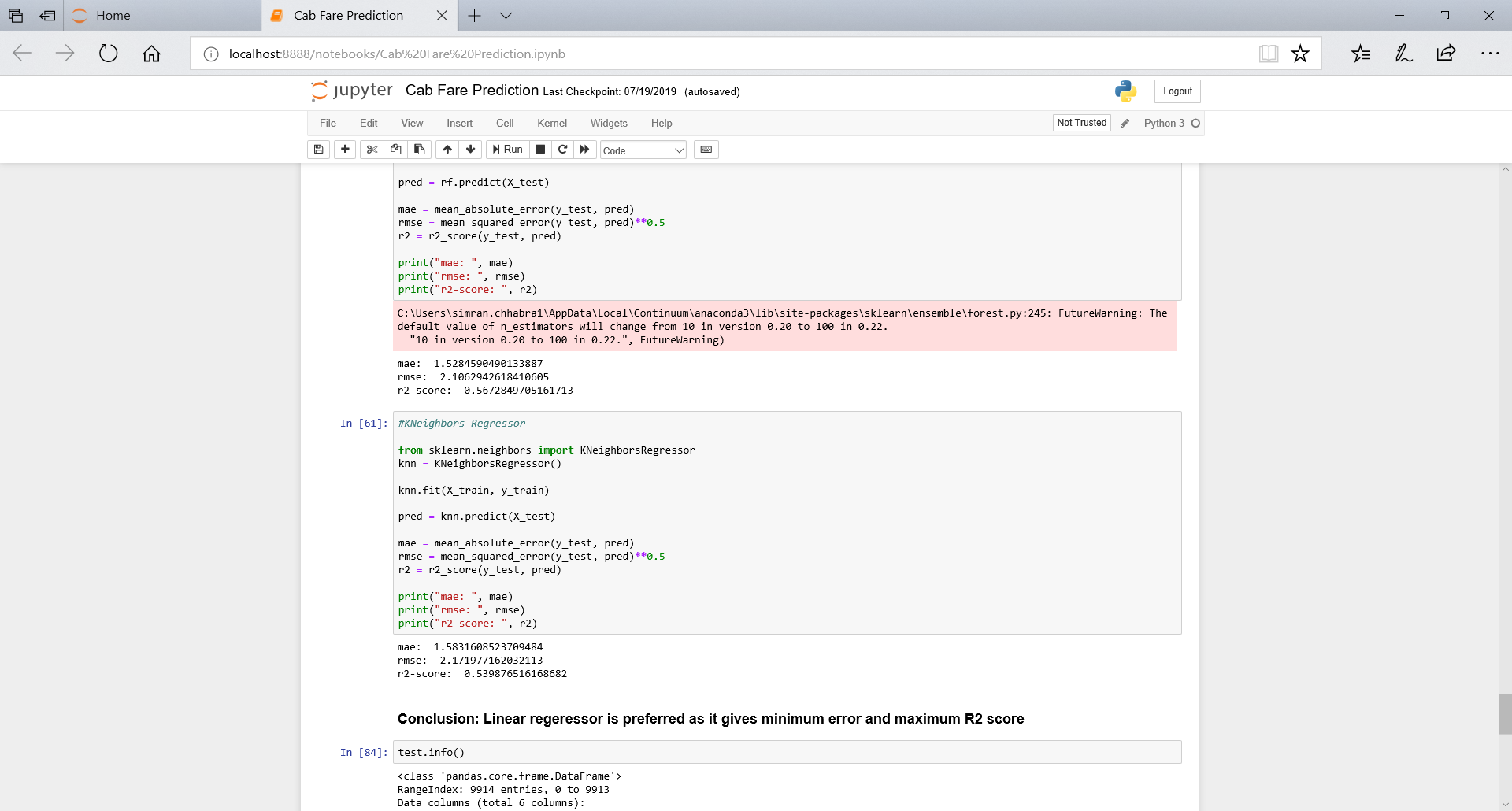
Now we try and use a different regression model to predict our target variable which Decision Tree regression.

We get following error metrics on performing it:



* + 1. **K Neighbours Regression**

Next, test values are predicted using K Neighbours regression and the following results are observed:



**Chapter 3**

**Conclusion**

* 1. **Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. We have used error metrics like mean absolute error (MAE), root mean squared error (RMSE) and R-squared (r2-score) to compare the performance of different models.

MAE and RMSE give us the error in the prediction values as compared to the actual values and r2-score tells us what proportion of variance in the target variable was successfully explained by the independent variables.

**3.2 Model Selection**

Less the value in error metrics and more the value in R-squared parameter, better is the model. Comparing these values in the three models, Linear Regression gives the least MAE and RMSE score and highest r2-score thus it would be the best suited model for our prediction.

**Appendix A**

Please find attached python notebook “Python\_cab\_rental.ipynb” file.

**Appendix B**

Please find attached “R\_cab\_rental.R” file.

**References**

<https://learning.edwisor.com/>