Project Topic

**Spanish High Speed Rail Data Analysis**

# Team Number and Team Members

Team Number 19

Members:

· Manav Malavia (NEU ID 002763769)

· Tianxiang Ren (NEU ID 002104759)

# Introduction

1. **Background**

****

Rail transport plays a big role in Spain’s society, every year there are countless travelers from all over the world taking the rail transportation in Spain to get to their destination. *Renfe*, Spain’s national train network, is usually the best way to get from city to city. Like most countries in Europe, people are basically moving from one big city to another big city, such as Madrid and Barcelona, though in those big cities, rental car is becoming a popular way for people to go on a trip, but still, getting stuck in traffic is always a nightmare and you would need a valid license to drive. Getting the affordable tickets of *Renfe*, and going freely to cities would still be people’s first choice around Spain.

2. **Motivation**

Nowadays, traveling is a huge and popular part of people’s lifestyle in Spain. If we could make some analysis about the data, which could be the pricing of traveling, or customer and train information, we could save money for the passenger or make money for the train station. We may find the best route for traveling or we could even use such data to know more about customer’s preferences and their willingness to pay.

3. **Goal**

We will focus our efforts on assisting consumers in having more comfortable travel experiences in light of what was expressed. Customers find that to be the most helpful and advantageous. With this background, we might perform some analysis on the cost of the tickets, the most popular routes, the types of trains, try to find out the correlation between price and duration and apply machine learning algorithms to compare and uncover key insights that will aid in understanding this rail system.

Dataset Description

The dataset contains main data of all Renfe (Spanish railway system) trips with price description, available seats, time and vehicle information in 2019.

The dataset has 14 columns: id, company, origin, destination, departure, arrival, duration, vehicle\_type, vehicle\_class, price; huge number of trips with individual id for us to track every trip. This data will serve our research well in our scope of the ticket cost, the most popular routes, as well as the most welcomed types of trains.

Data Sources

*https://www.kaggle.com/datasets/thegurusteam/spanish-high-speed-rail-system-ticket-pricing*

# Methodology

With the help of this data set, we will be able to determine the following things:

**1. Exploratory Data Analysis**

Ahead of the upcoming research, we would implement the Exploratory Data Analysis for better understanding of the dataset, detecting any errors and getting more detailed information of the dataset.

With the help of methods such as Univariate Analysis, we could easily find out which is the cheapest route and which is the most preferred vehicle type and most popular vehicle class among customers.

**2. Linear Regression And Gradient Boosting**

In statistics, a scalar response and one or more explanatory variables are modeled using a linear approach called linear regression.

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

We will implement Linear Regression and Gradient Boosting to find the following

1) The change in numeric route values(after Label Encoding route) and origin correlation- 0.9 (refer the heat map)- highest of all features.

2)The change in fare with respect to origin.

3)The change in fare with respect to the month.

**3. K-Means Clustering**

The goal of K-means clustering, a vector quantization technique that originated in signal processing, is to divide n observations into k clusters, each of which has a prototype (cluster centroid or cluster center) that each observation belongs to.

By implementing K-Means Clustering techniques, we will find out the following

1)Clusters -Numerical vehicle type values (after label encoding vehicle type)

and Numerical vehicle class values(after label encoding vehicle class).

2)Clusters -Price and Duration.

3)Clusters- Numeric Route Values (after label encoding route )and Duration

**4. Naive Bayes**

It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class has nothing to do with the presence of any other feature. There are three types of Naive Bayes algorithm under Sci-kit Library: Gaussian, Multinomial and Bernoulli.

We will implement Gaussian and Multinomial naive bayes on features price and fare- get to know the change in price with respect to Fare.

**5. KNN**

The k-nearest neighbors algorithm, sometimes referred to as KNN , is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point.

We will implement KNN algorithm between features -

1. Price, duration, numeric route values(after label encoding route) and Duration.

# 

# Results and Analysis

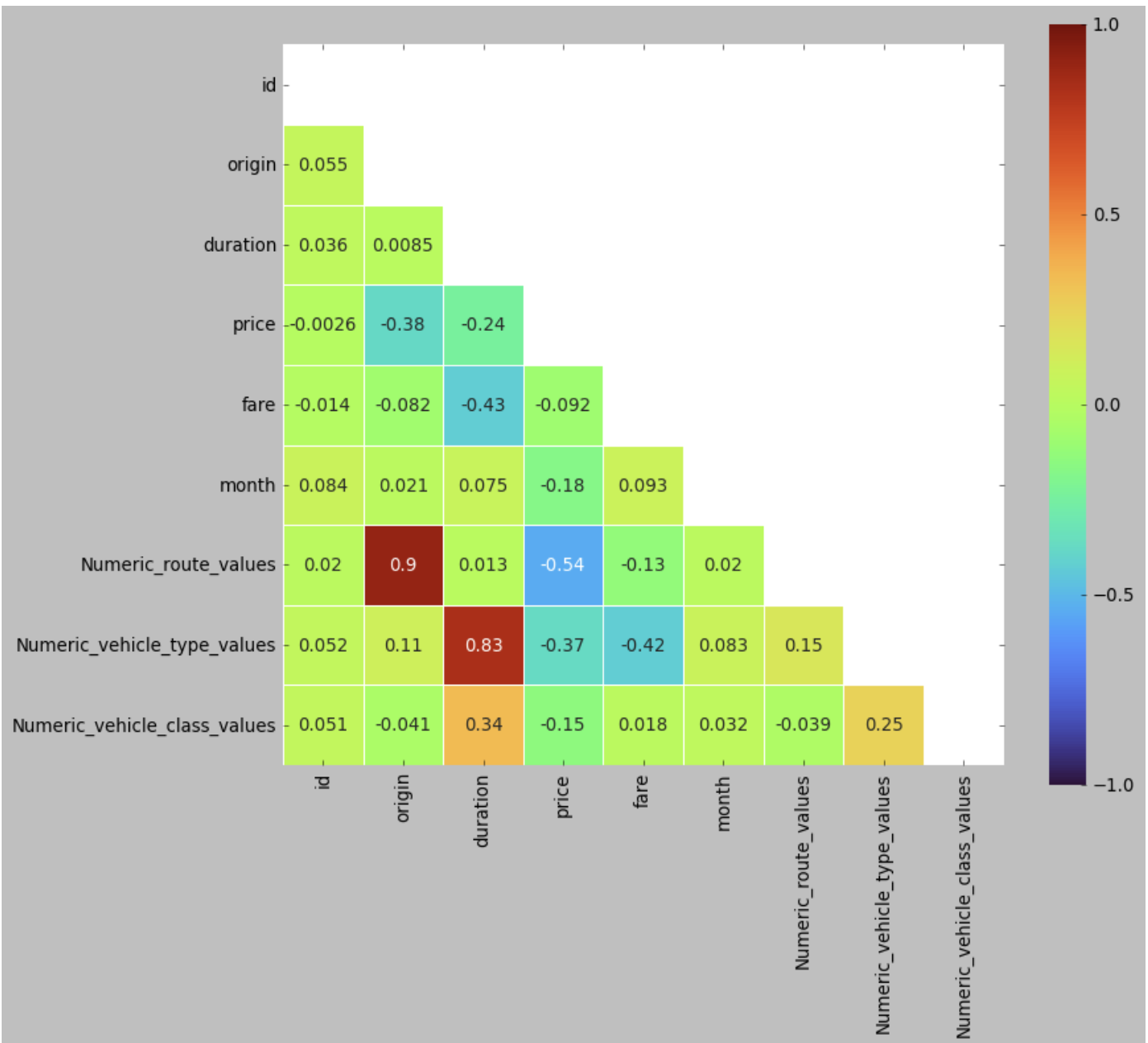
**Exploratory Data Analysis**

The primary goals of EDA are to identify errors, outliers, and various patterns in the data. It enables analysts to comprehend the data more thoroughly before drawing any conclusions. The outcomes of EDA can assist businesses in understanding their customers, growing their business, and making the appropriate decisions.

**Heat Map**

To start with, we would introduce the heat map of the datasets to visualize the data information.

A heat map is a type of data visualization that displays a phenomenon’s size in two dimensions using color. In order to visualize the link between the features, we utilized a heat map.It helps to find features that are best for Machine Learning model building. The heat map transforms the correlation matrix into color coding.



From our heatmap above, we can infer the following:

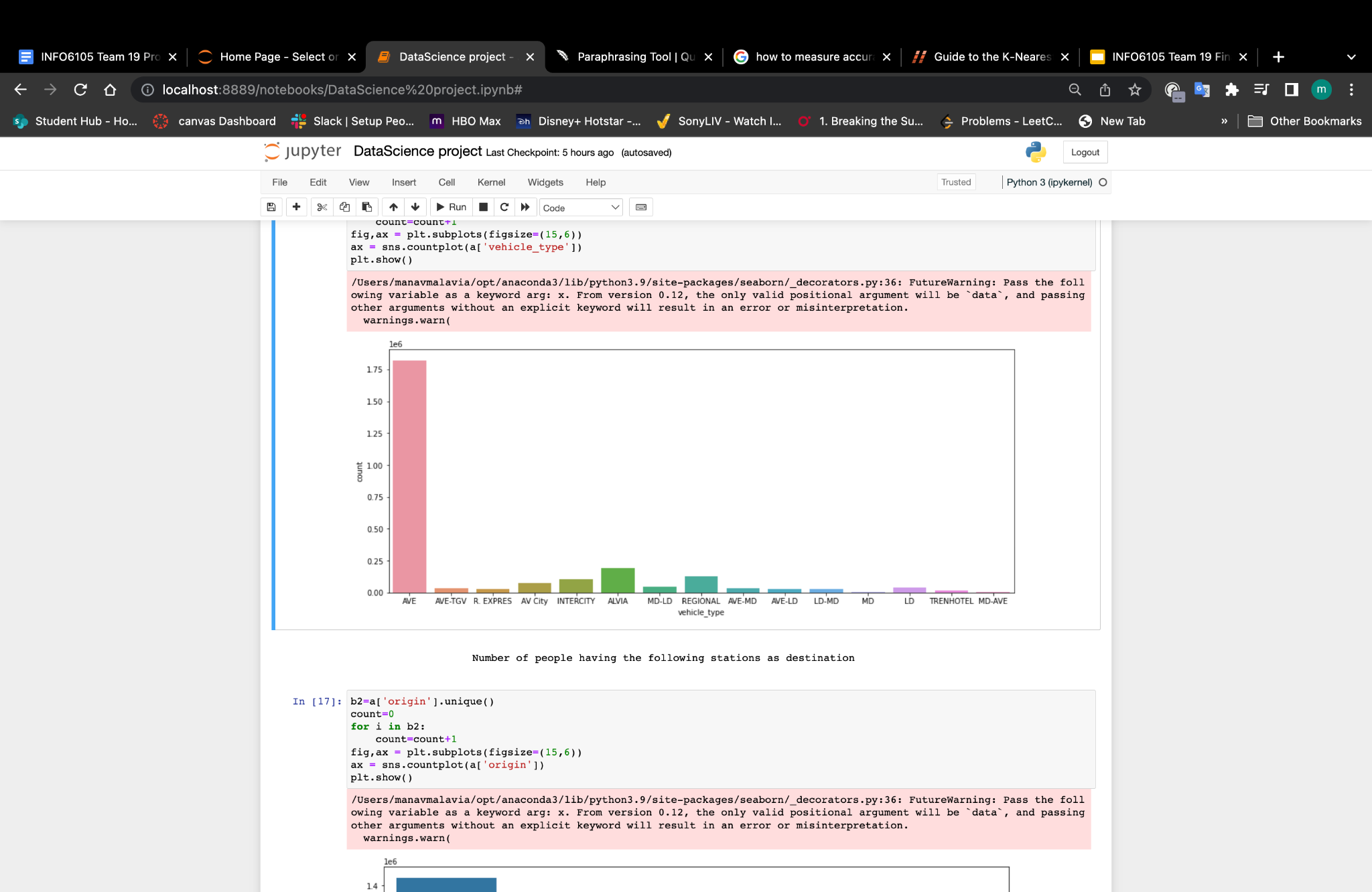
* Features *‘price’* and *‘duration’* have a negative correlation with a value of -0.24
* Features *‘Numeric\_route\_values’* and *‘price’* have a strong negative correlation with a value of -0.24
* Features *‘id’* and *‘duration’* have a positive correlation with a value of 0.036
* Features *‘Numeric\_route\_values’* and *‘origin’* have a strong positive correlation with a value of 0.9.

**Univariate Analysis**

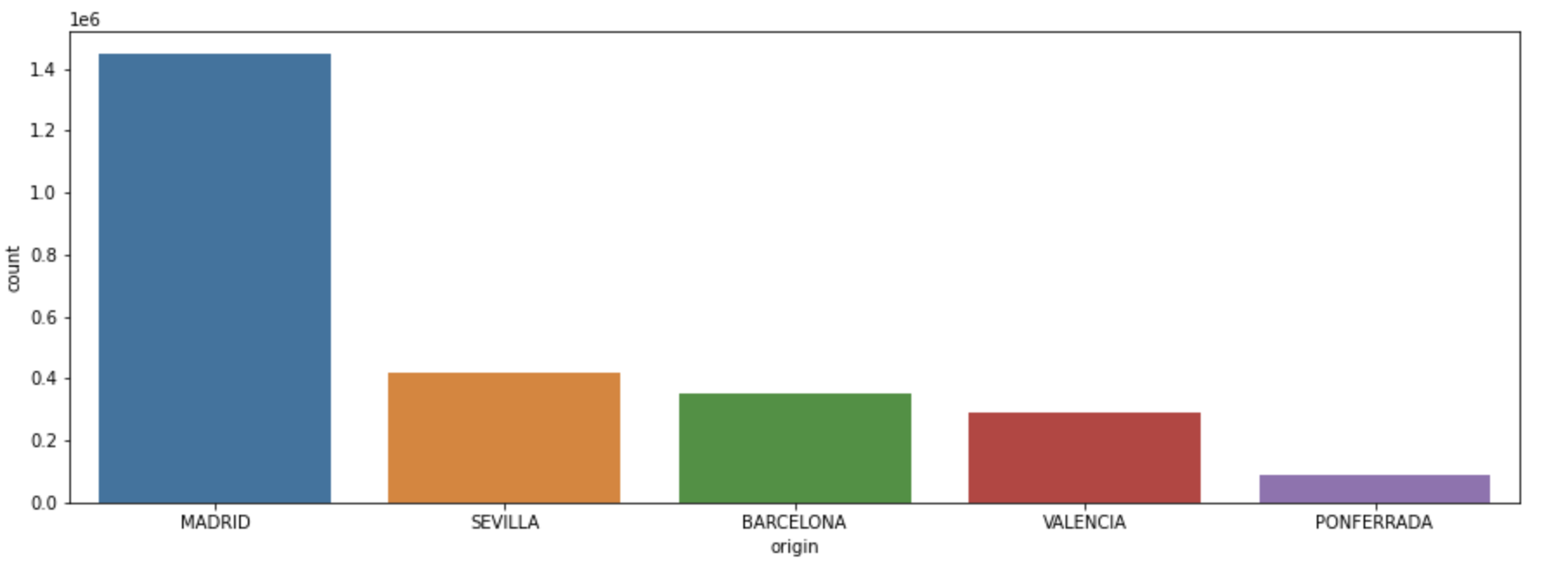
A fundamental method of statistical data analysis is called univariate analysis. Here, there is only one variable in the data, so a cause-and-effect link is not present.

Our next steps is going to do univariate analysis on following aspects:

* Different types of train that runs in Spain



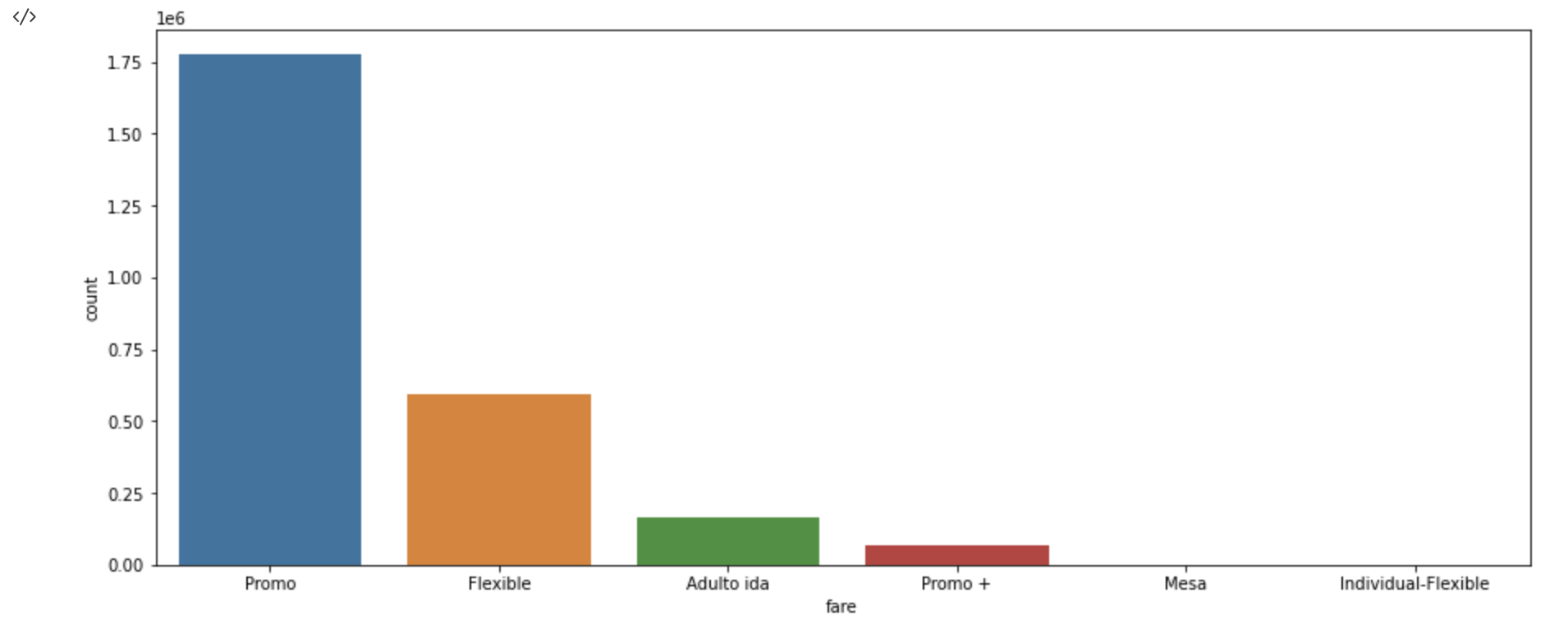
* Number of people having the following stations as destination



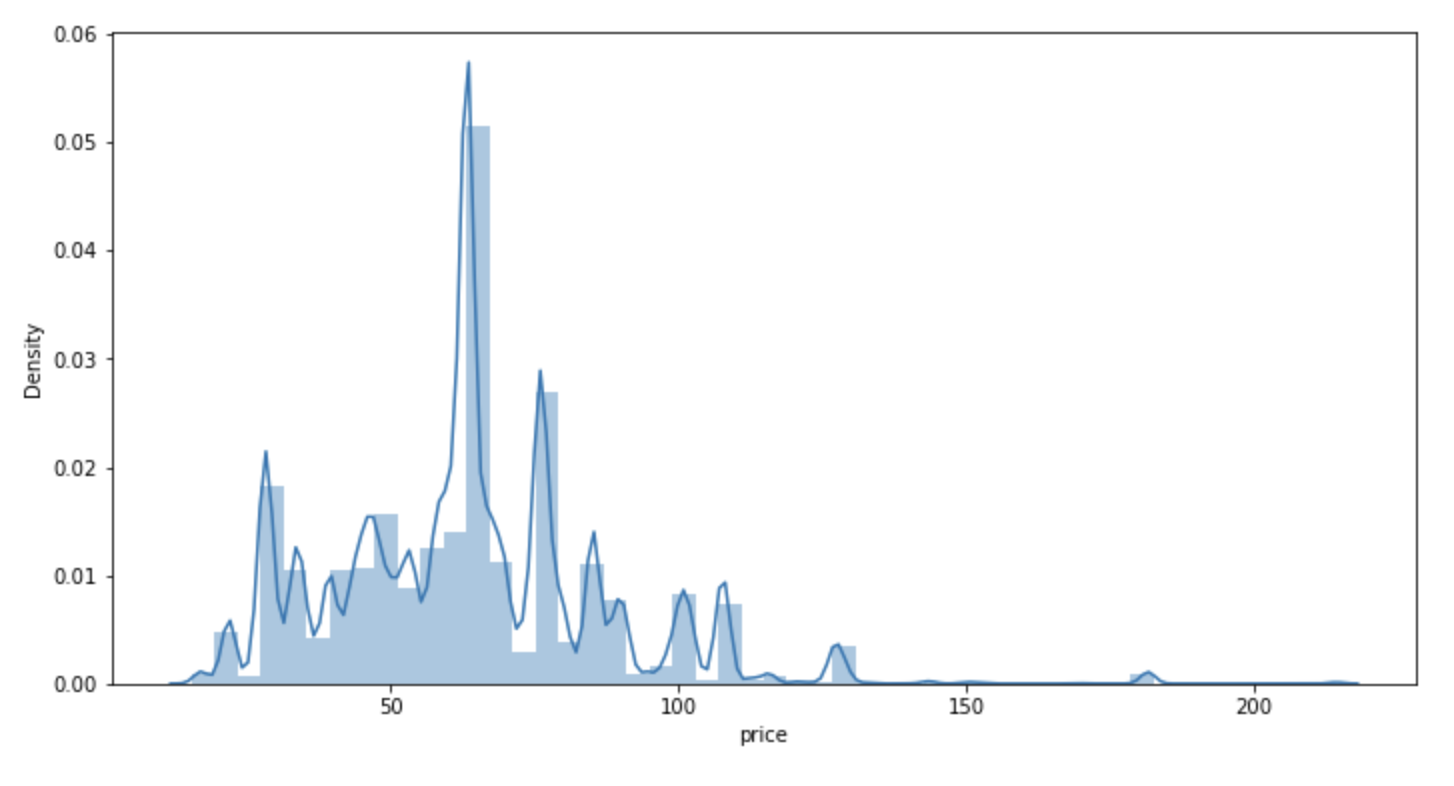
* Number of train of different class



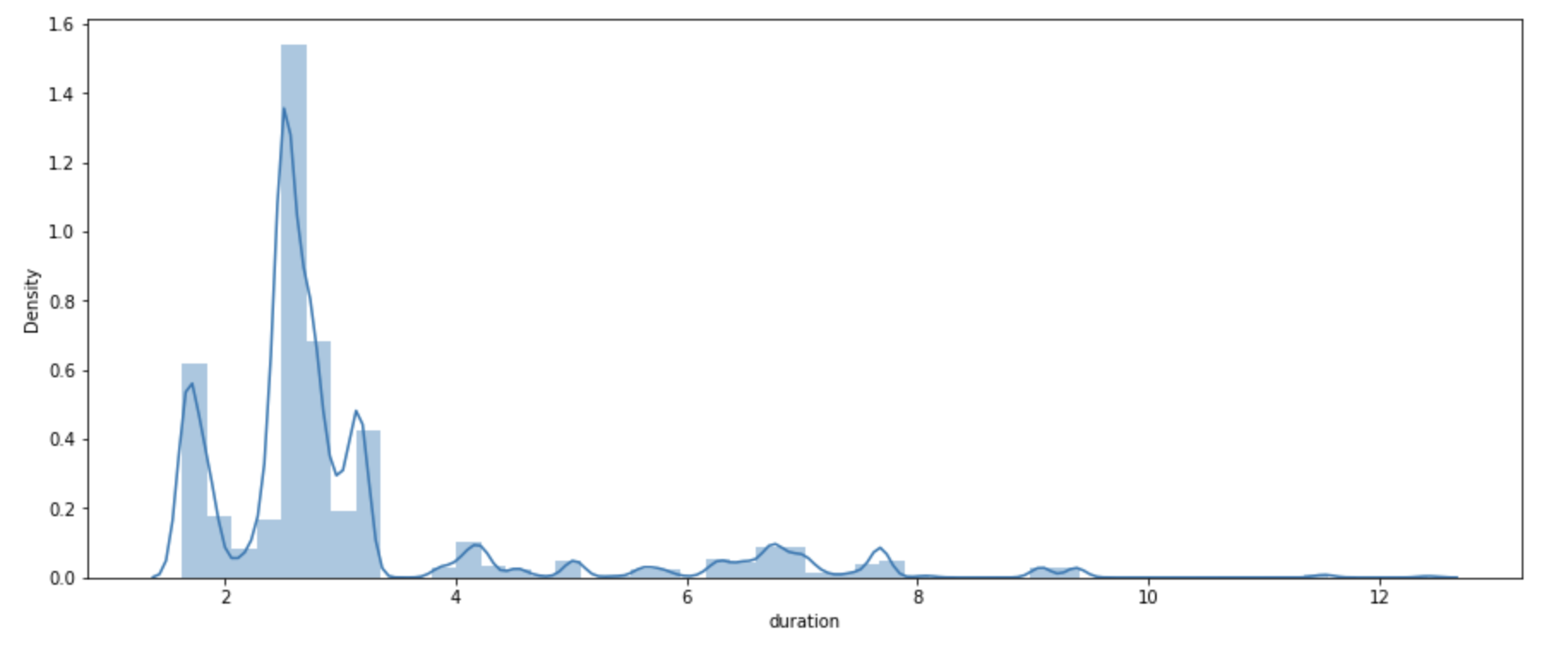
* Number of tickets bought from each category



* Distribution of the ticket prices



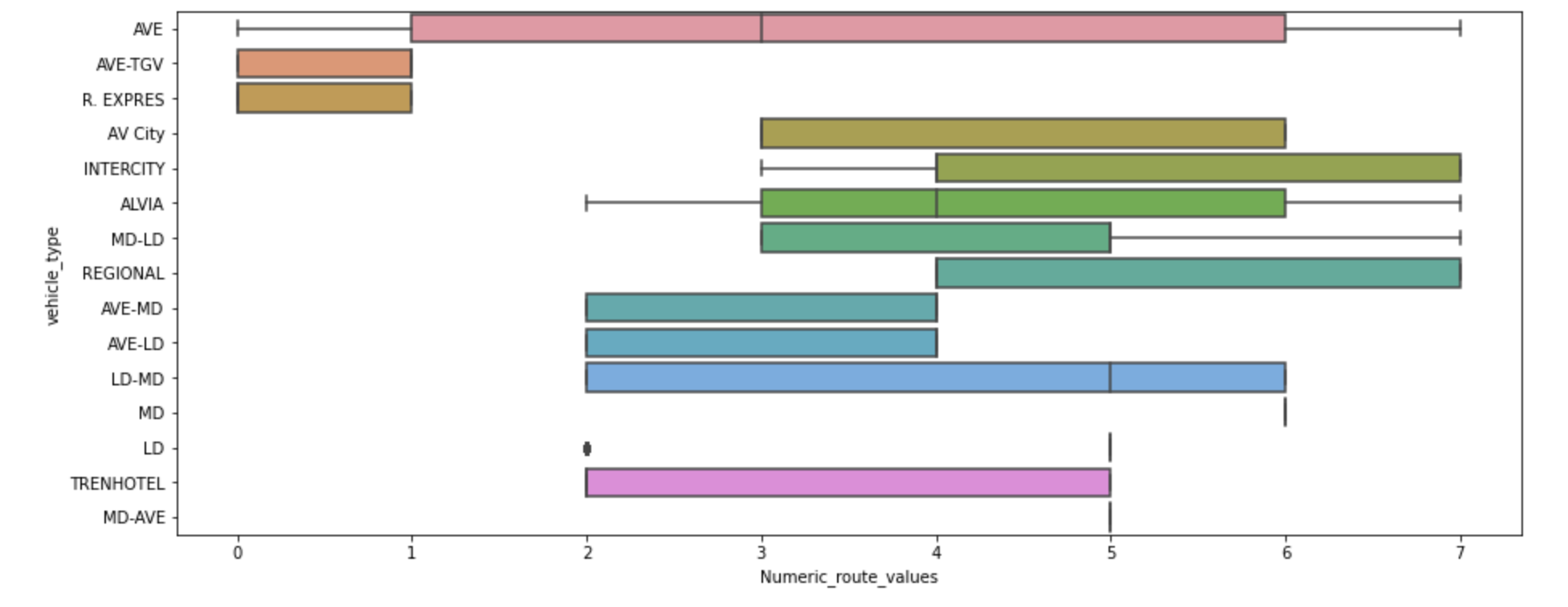
* Distribution of Duration



**Bivariate Analysis**

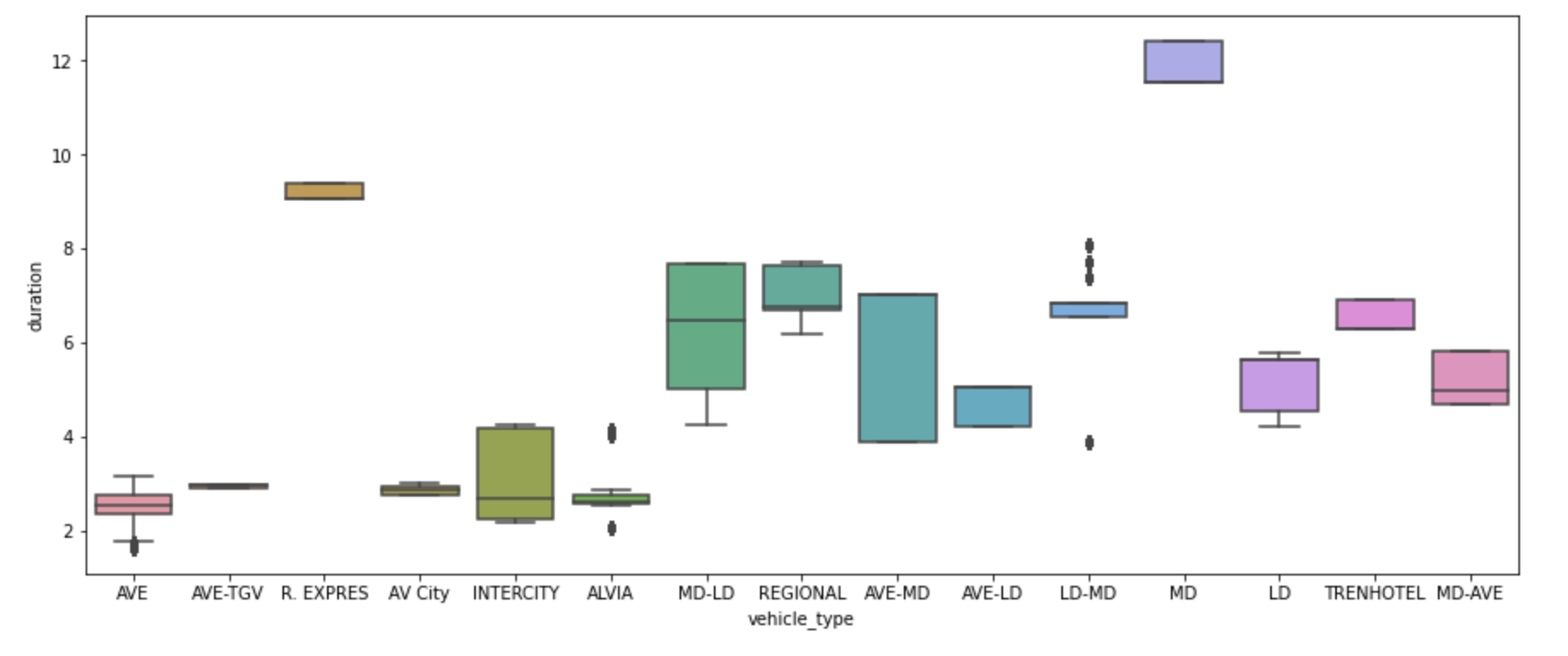
Bivariate analysis is the study of data with two variables. To determine the link between two value sets, a single statistical analysis was utilized. X and Y are the pertinent variables. When using two variables, bivariate analysis is used to determine the relationship between each variable in the dataset and the target variable of interest..

Our next step is to do bivariate analysis as follows:



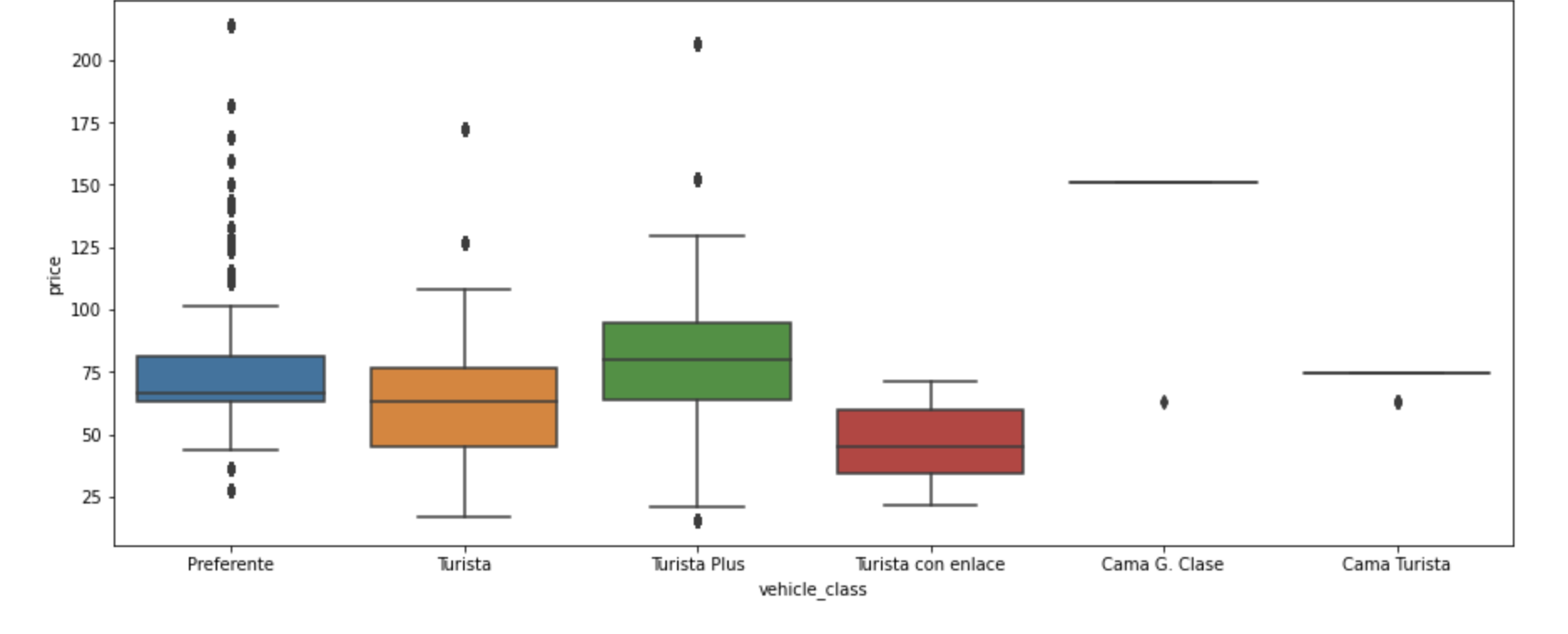
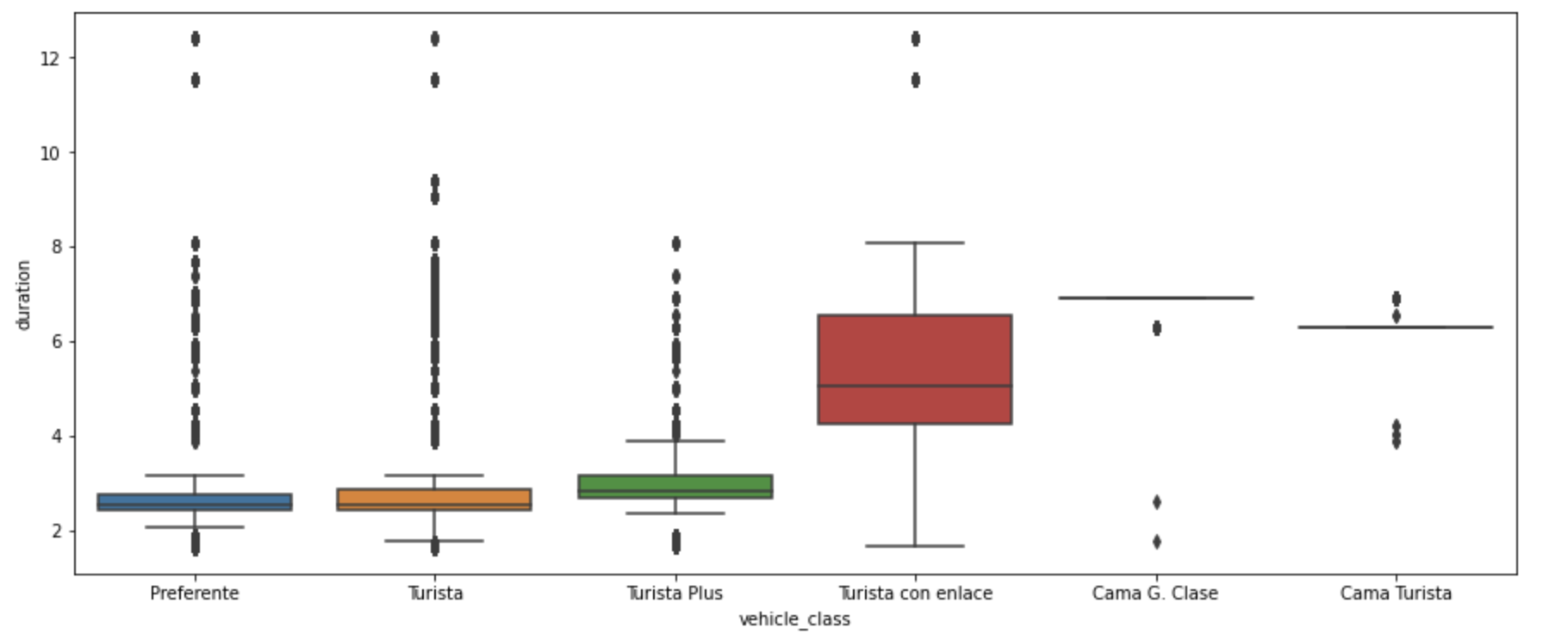
* Using Label Encoder for Bivariate Analysis

With this boxplot, we can get to know the Numeric route values (after label encoding route) and the respective vehicle types preferred by people on that route.



* The one with a high correlation- vehicle\_type\_class and duration

With this boxplot, we can get to know the duration and the respective vehicle types preferred by people .



* Analyzing for Each Routes(There are 8 eight routes overall)

We Have Analyzed the routes and made plots of different routes and the results are as follows:-

1) Madrid to Sevilla and Sevilla to Madrid(combining rows from both routes and using a single Dataframe):-

* Cheapest returm fare from madrid to sevilla

(by sorting the routes by price)- 50.4

* finding the relationship between duration and vehicle type
* Finding and printing respective routes and vehicle types by using groupby Vehicle\_Types
* Finding and printing respective routes and vehicle classes by using groupby Vehicle\_Types

2) Madrid to Barcelona and Barcelona to Madrid(combining rows from both routes and using a single Dataframe):-

* Cheapest returm fare from madrid to sevilla

(by sorting the routes by price)- 70.6

* finding the relationship between duration and vehicle type
* Finding and printing respective routes and vehicle types by using groupby Vehicle\_Types
* Finding and printing respective routes and vehicle classes by using groupby Vehicle\_Types

3) Madrid to Valencia and Valencia to Madrid(combining rows from both routes and using a single Dataframe):-

* Cheapest returm fare from madrid to sevilla

(by sorting the routes by price)- 30.9

* finding the relationship between duration and vehicle type
* Finding and printing respective routes and vehicle types by using groupby Vehicle\_Types
* Finding and printing respective routes and vehicle classes by using groupby Vehicle\_Types

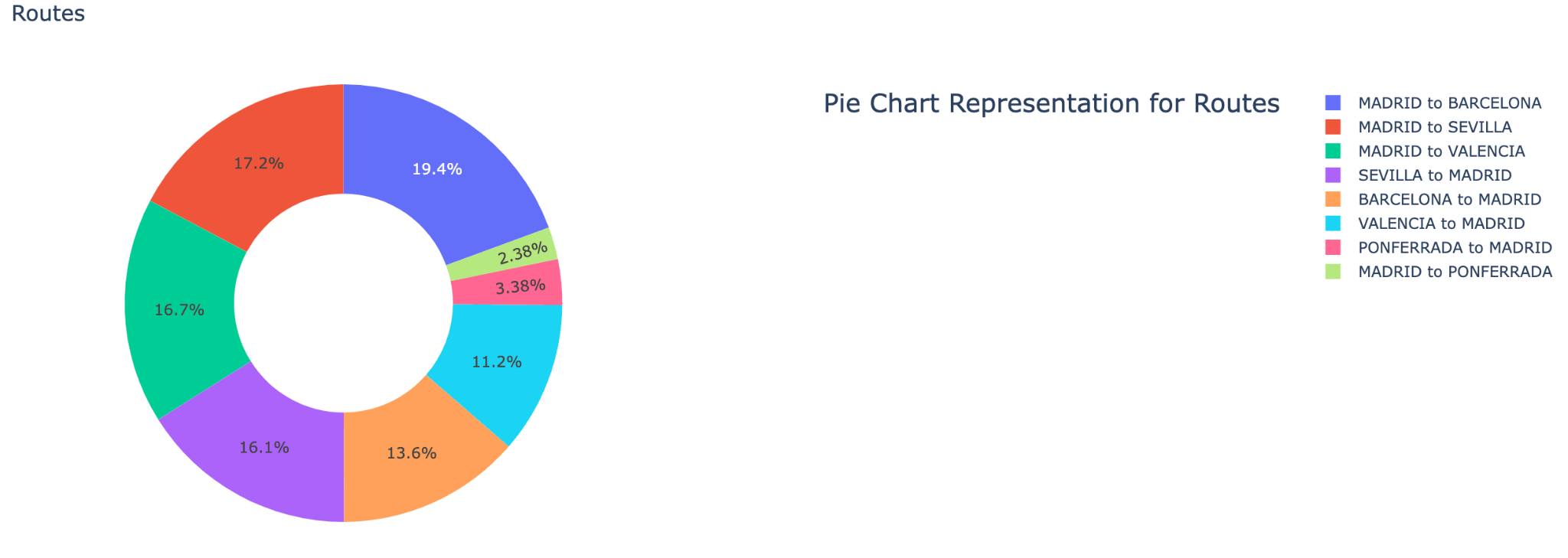
4) Madrid to Ponferrada and Ponferrada to Madrid(combining rows from both routes and using a single Dataframe):-

* Cheapest returm fare from madrid to sevilla

(by sorting the routes by price)- 46.599999

* finding the relationship between duration and vehicle type
* Finding and printing respective routes and vehicle types by using groupby Vehicle\_Types
* Finding and printing respective routes and vehicle classes by using groupby Vehicle\_Types

Next we are making a pie chart for Routes in a DataFrame.



The pie chart showed that the train system's principal routes connect large cities like Madrid, Barcelona, Seville, and Valencia, whereas Ponferrada, a minor city, receives less consideration.

# 

We can also infer from the pie chart above that Cama Turista and Cama G.Clase are less

* Finding the cheapest route and vehicle\_type

Our next step is going to find the cheapest route and vehicle type, let’s sort the data by the price, and we get the result of

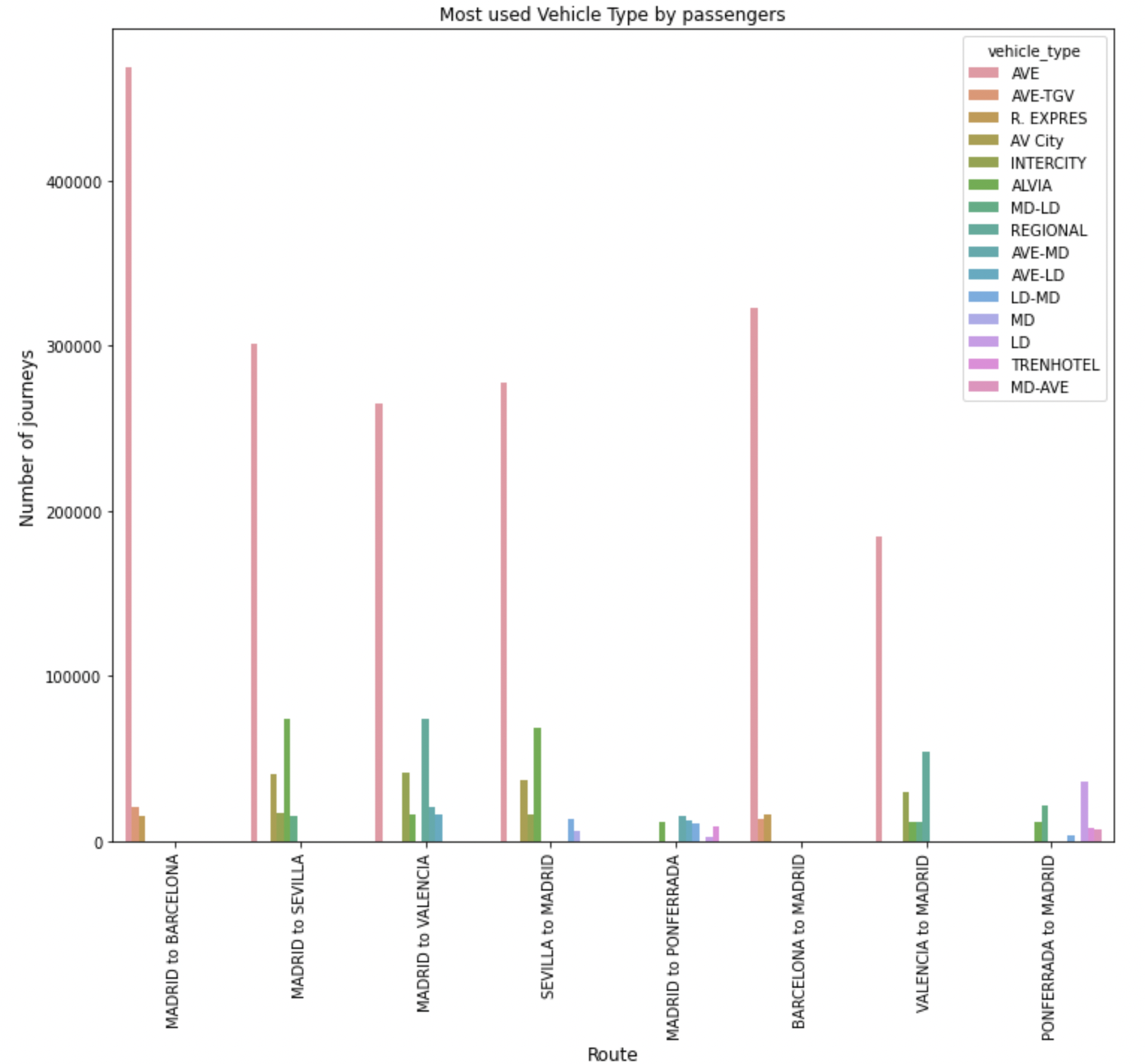
* *Results for the*  cheapest route and vehicle\_type:-

*Madrid to Valencia using the vehicle\_type -INTERCITY- VEHICLE\_CLASS- Turista plus with the least duration-4.17*

* We get 15.45 which is the cheapest price.
* Finding the Shortest Route in terms of Duration

With the same strategy, we would sort the data by duration and get the result:

Then we are going to plot the most used vehicle type based on the routes:

* *Results for* the Shortest Route in terms of Duration:-*Valencia to Madrid is has the least duration with 1.63 hours using the vehicle\_class-turista and type -AVE*
* 

We can clearly infer from the above image that the MID-AVE is the most widely used

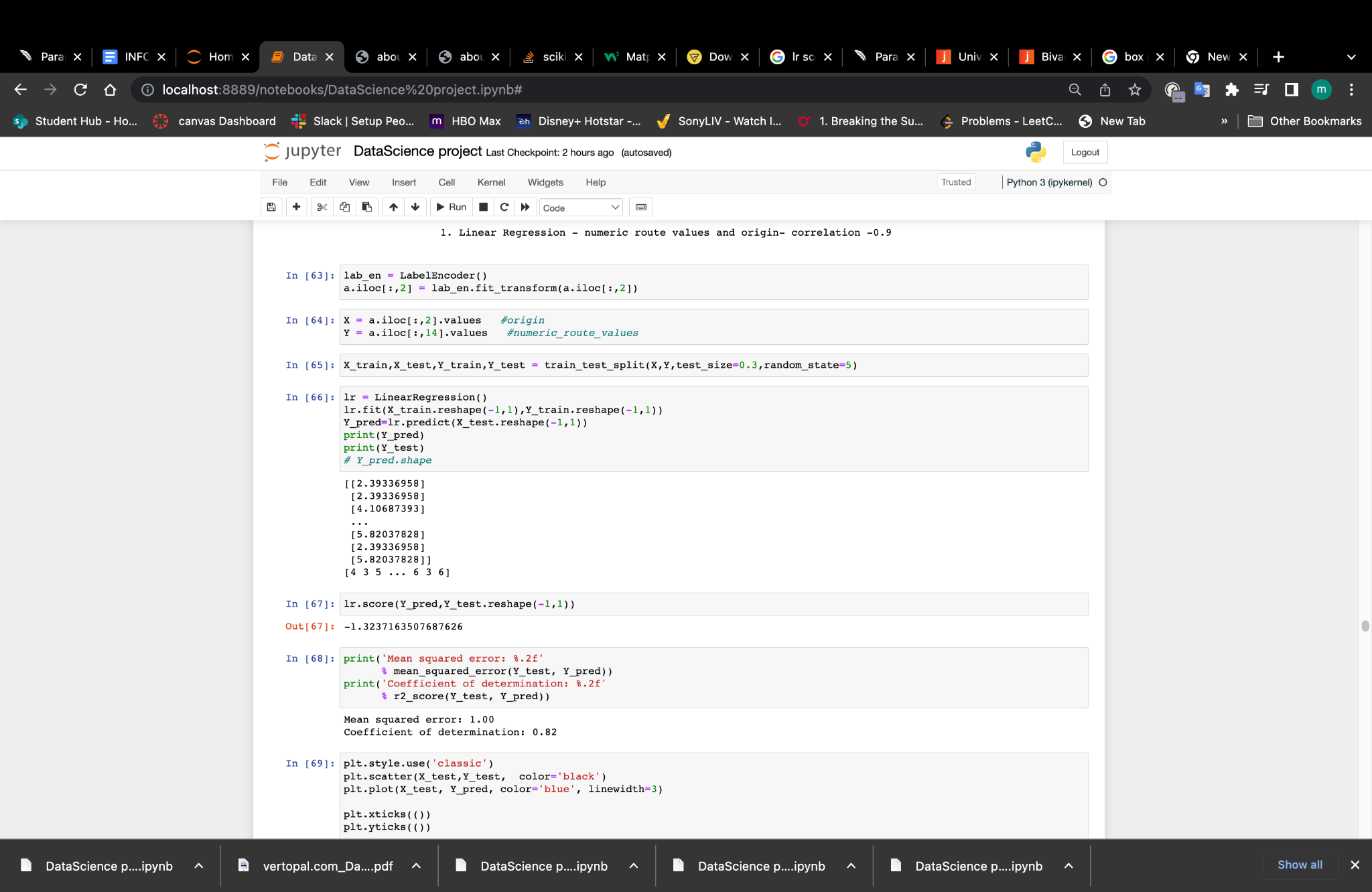
With all the information and analysis above, we are going to apply machine learning algorithms in the next steps.

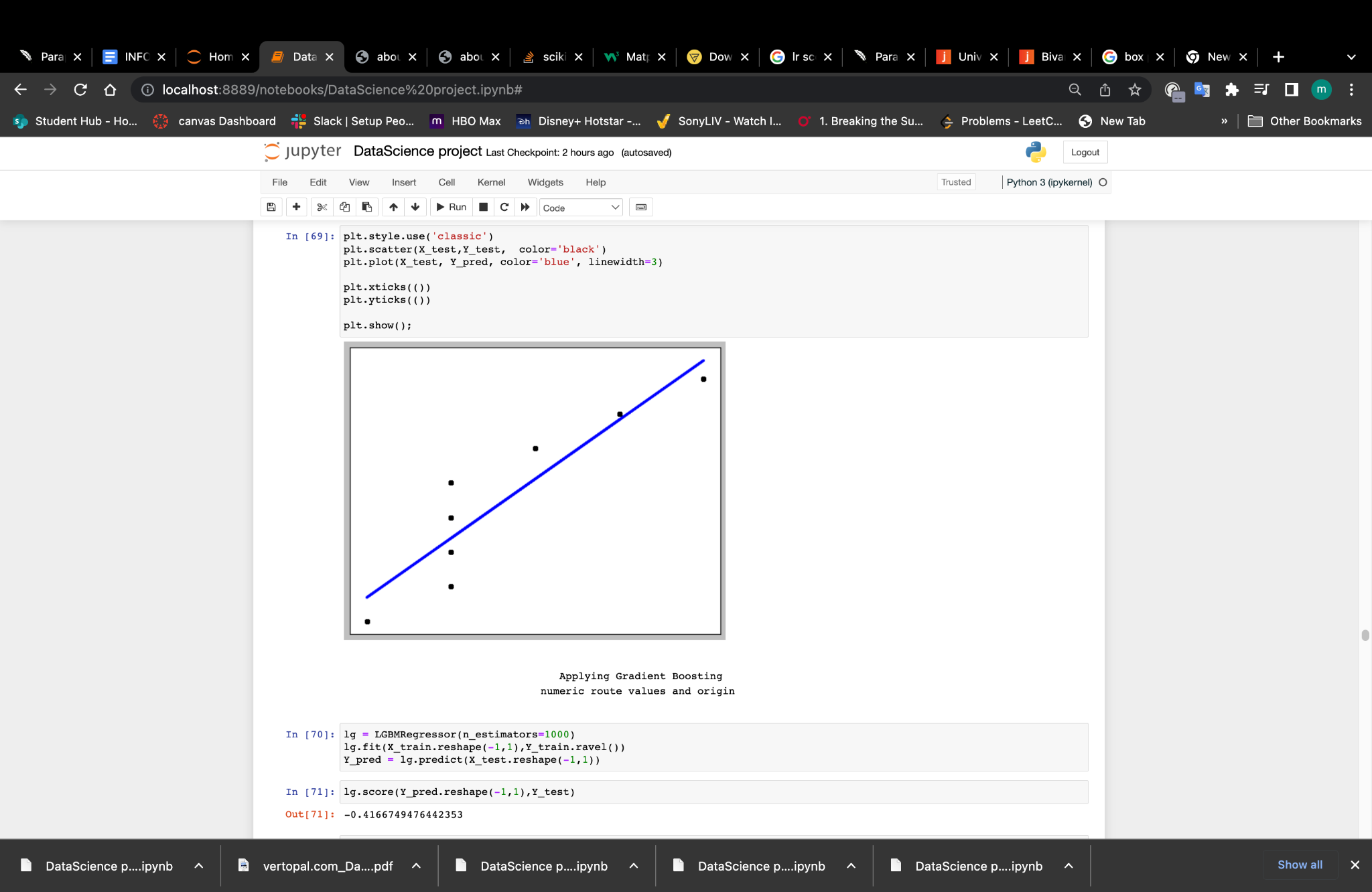
**Applying Machine Learning Algorithms**

**1. Linear Regression**

Implement Linear Regression and Gradient Boosting to find:-

1) The change in numeric route values(after Label Encoding route) and origin- correlation- 0.9 (refer the heat map)- highest of all features.



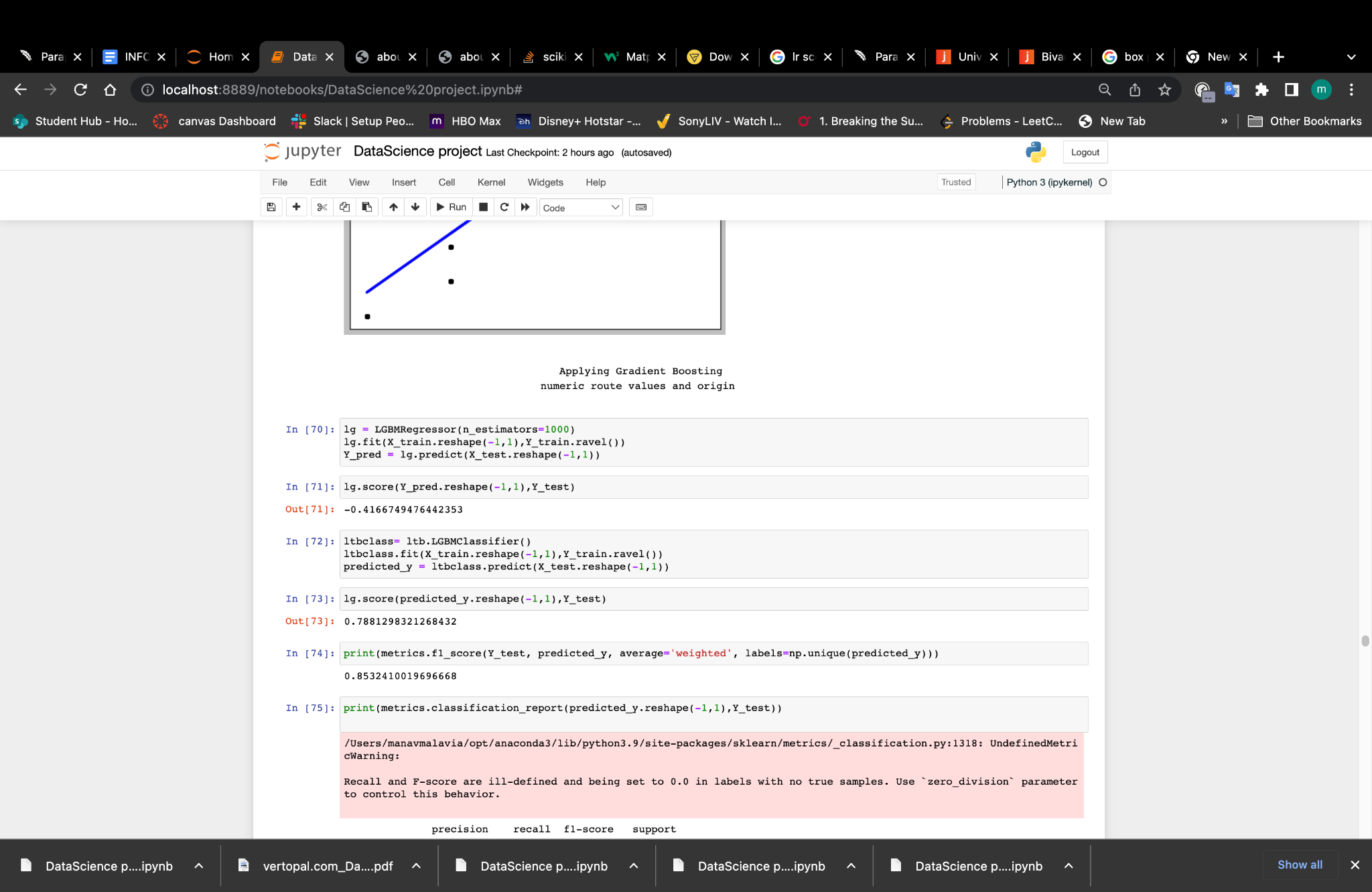


Calculated the Mean squared error and Coefficient of determination, the Mean squared error would measure regression accuracy, Coefficient of determination would help us perform model selection, e.g., finding the appropriate penalty coefficient for regularization.

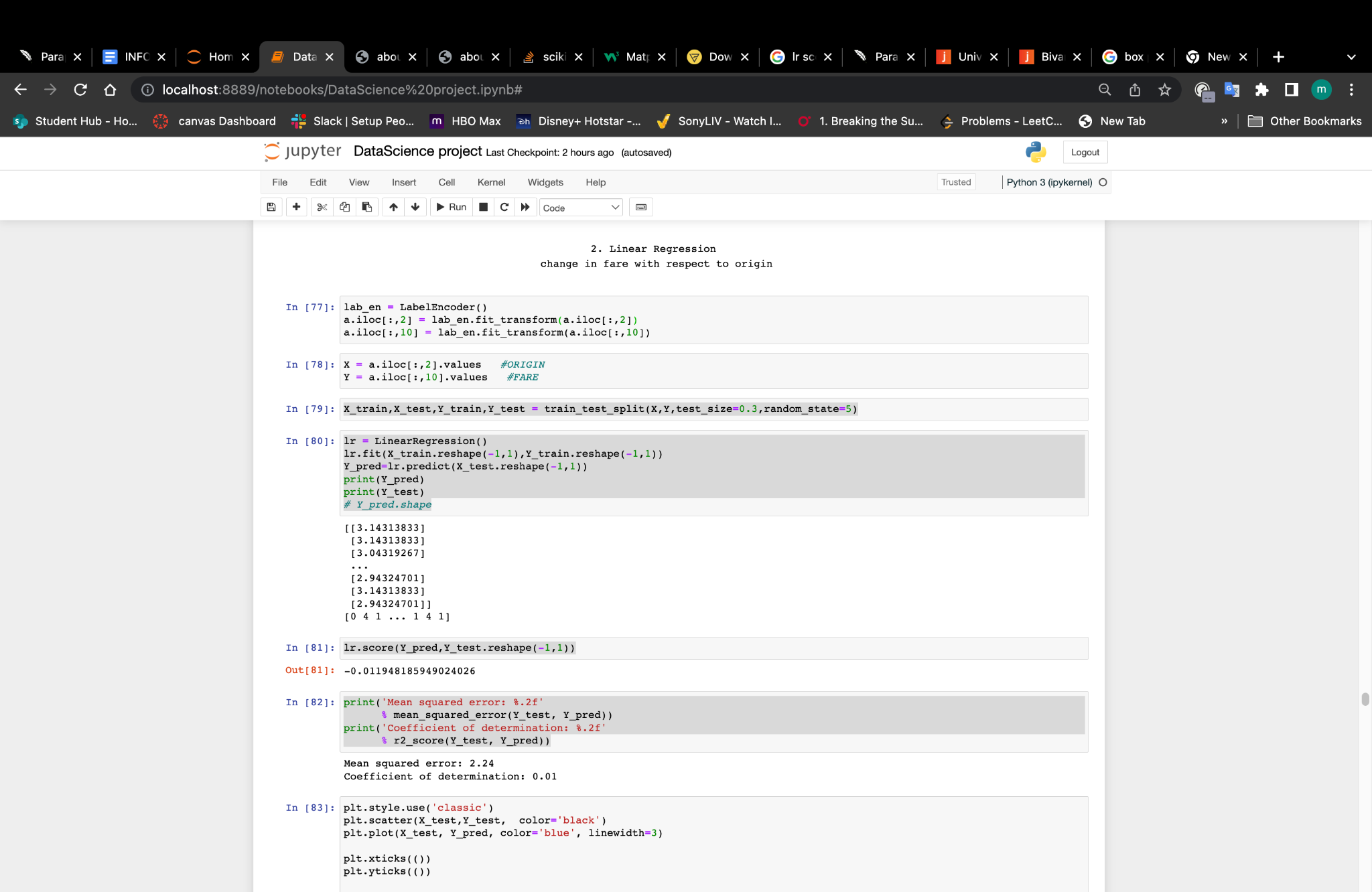
Result: Mean squared error: 1.00 , Coefficient of determination(R2 score): 0.82

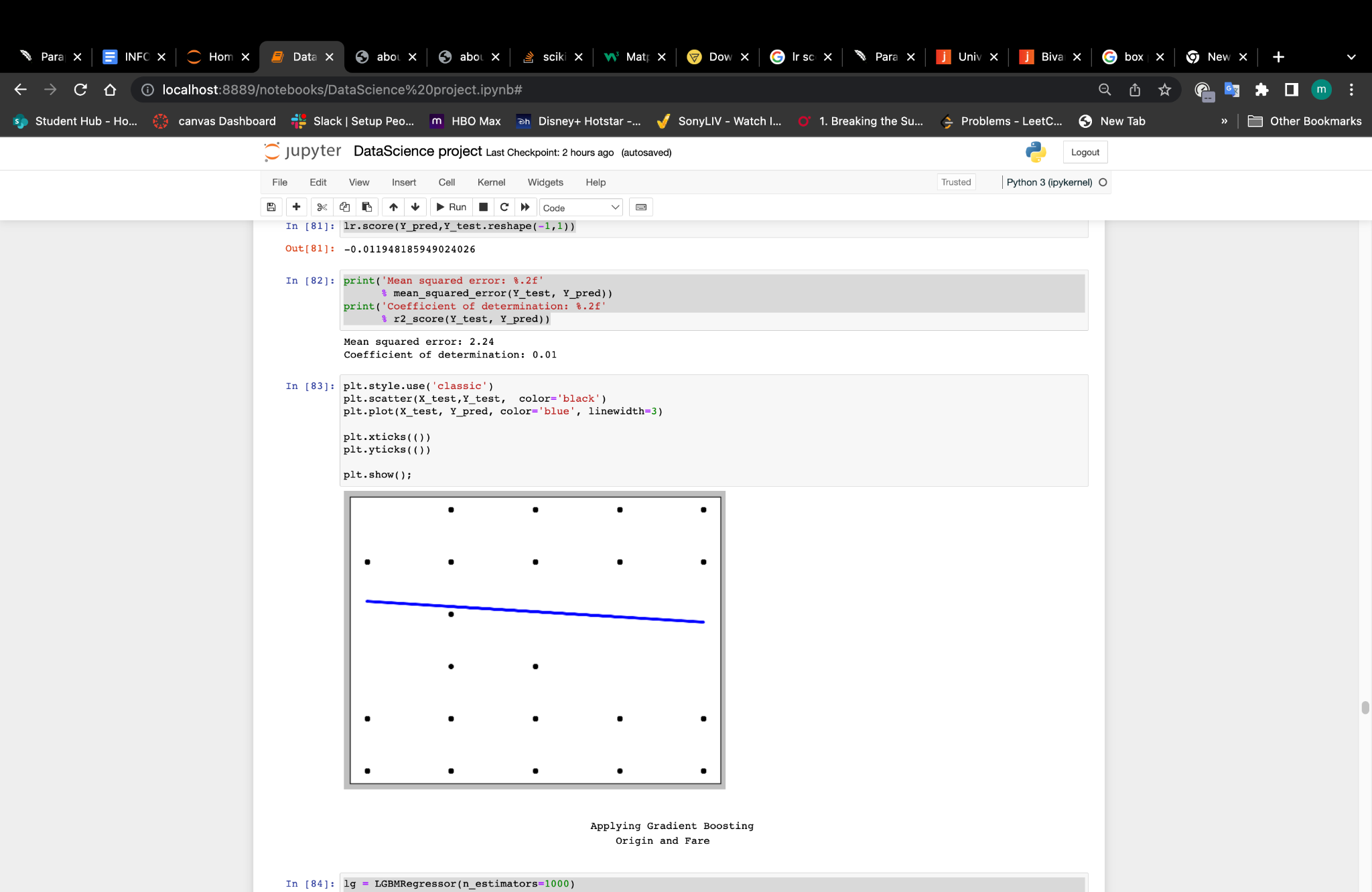
The result is a good fit, the origin and fare is highly correlated.

* **Applying Gradient Boosting**(Regressor and Classifier)-(printing the confusion matrix and f1 score)
* lg score of Regressor :- -0.4166 and lg score of Classifier- 0.7881



2)The change in fare with respect to origin.



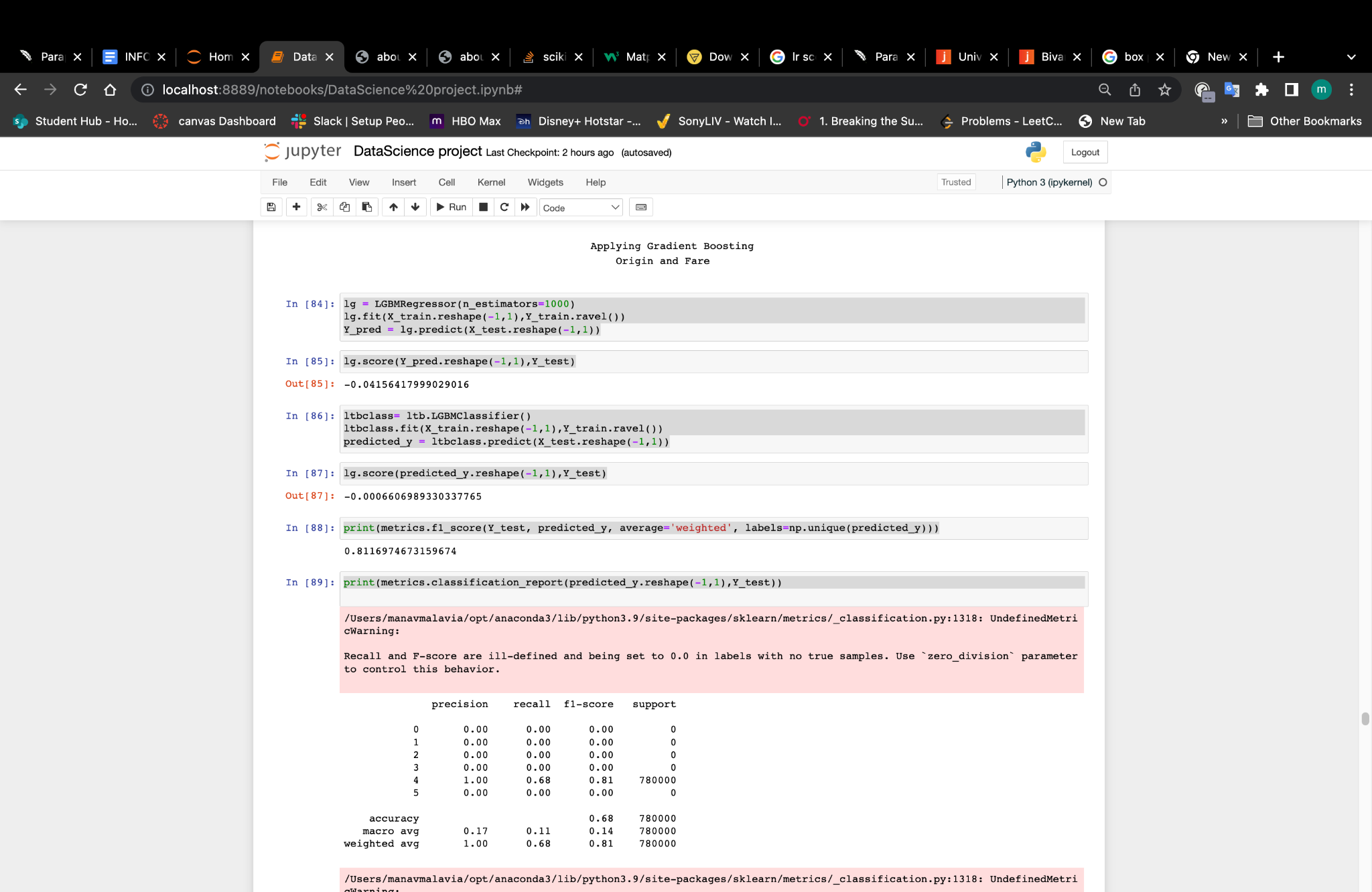


Calculated the Mean squared error and Coefficient of determination, the Mean squared error would measure regression accuracy, Coefficient of determination would help us perform model selection, e.g., finding the appropriate penalty coefficient for regularization.

Result: Mean squared error: 2.24, Coefficient of determination(R2 score): 0.01

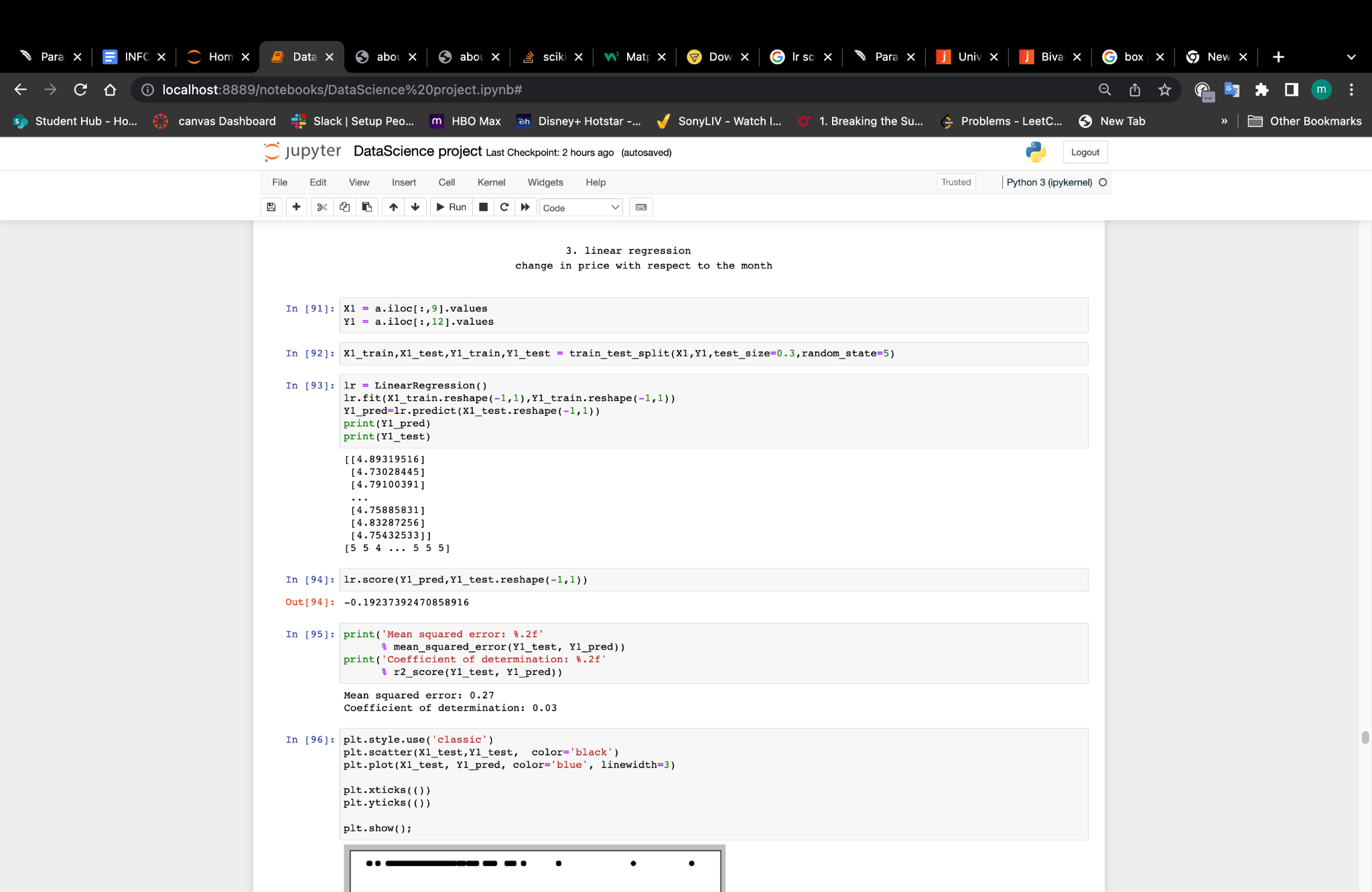
The result is less fit, the origin and fare is less correlated, so we move on.

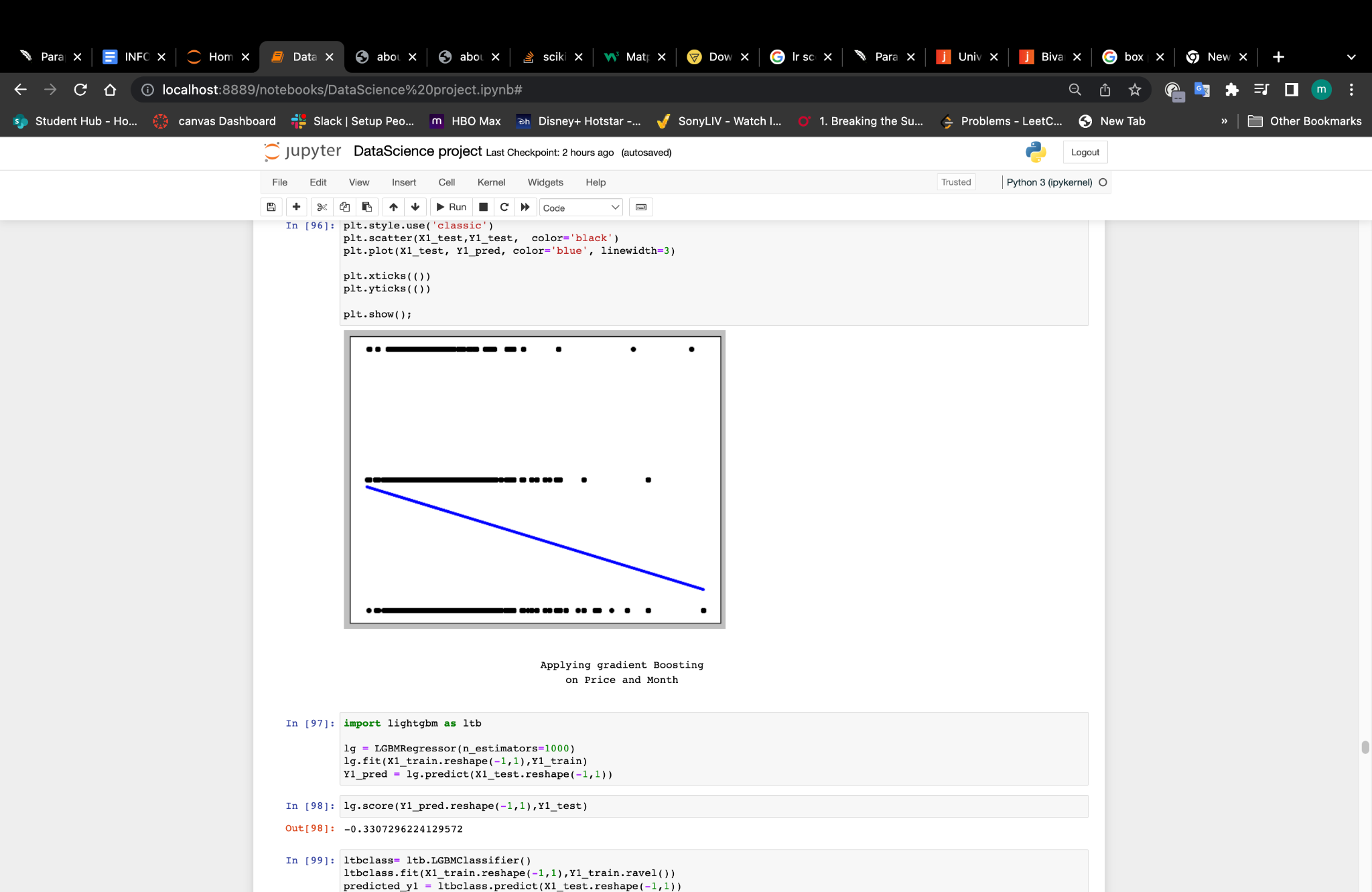
* **Applying Gradient Boosting**(Regressor and Classifier)-(printing the confusion matrix and f1 score)
* lg score of Regressor :- -0.4156 and lg score of Classifier- -0.0006





3)The change in fare with respect to the month.



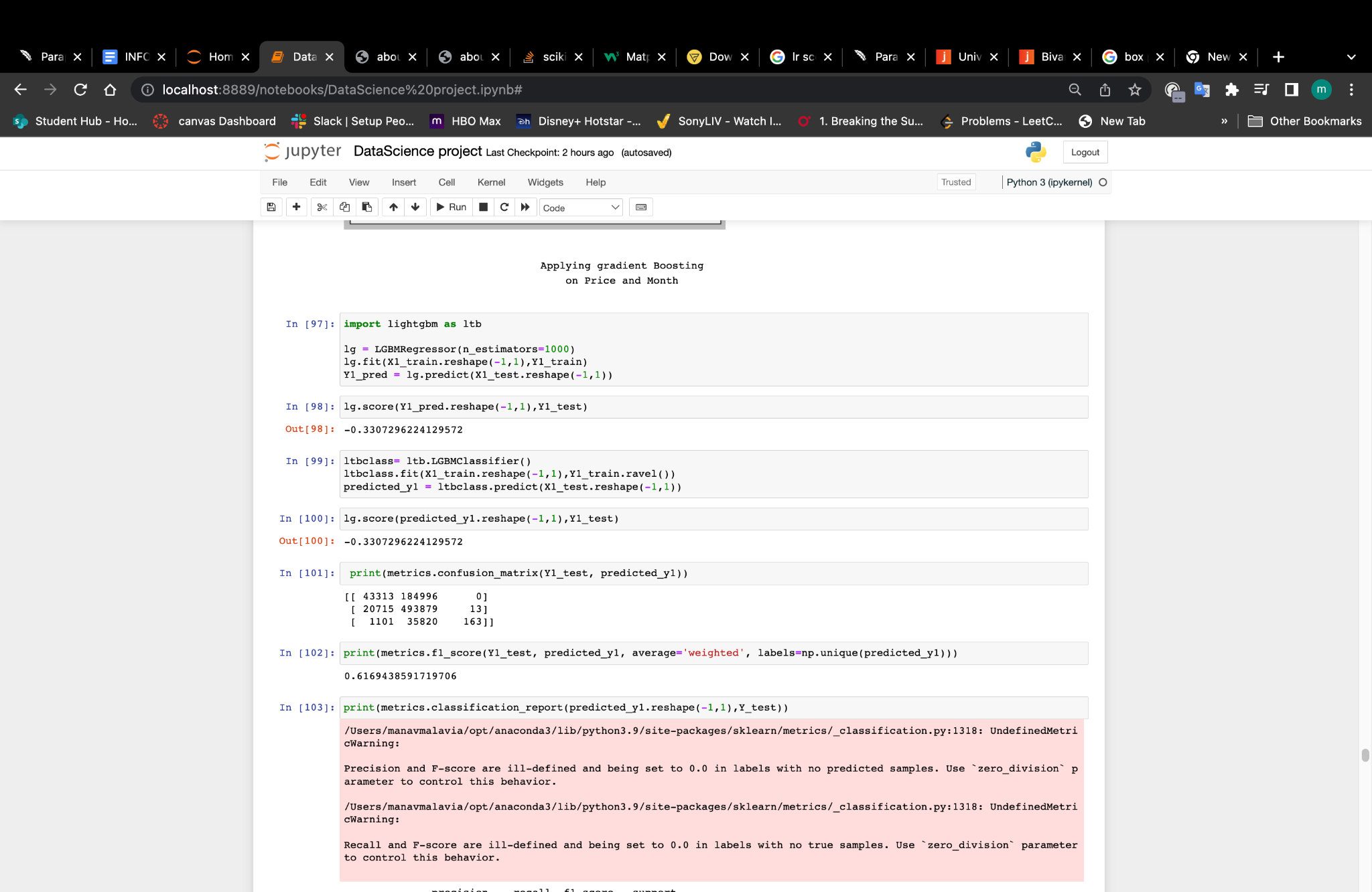


Calculating the Mean squared error and Coefficient of determination, the Mean squared error would measure regression accuracy, Coefficient of determination would help us perform model selection, e.g., finding the appropriate penalty coefficient for regularization.

Result: Mean squared error: 0.27 , Coefficient of determination(R2 score): 0.03

The result is less fit, the origin and fare is less correlated, so we move on.

* Applying Gradient Boosting(Regressor and Classifier)-(printing the confusion matrix and f1 score)
* lg score of Regressor :- -0.3307 and lg score of Classifier- -0.3307



**1.1 About Gradient Boosting:-**

One kind of machine learning boosting is gradient boosting. It is based on the assumption that when prior models are coupled with the best possible upcoming model, the overall prediction error is minimized.

Implement Gradient boosting to find out the correlation between month and fare, as well as price and month.

We are implementing LGBMRegressor and LGBMClassifier to train and predict the data regarding month and fare,

**2.K-Means Clustering**

By implementing K-Means Clustering techniques, we will find out -

1)Clusters -Numerical vehicle type values (after label encoding vehicle type)

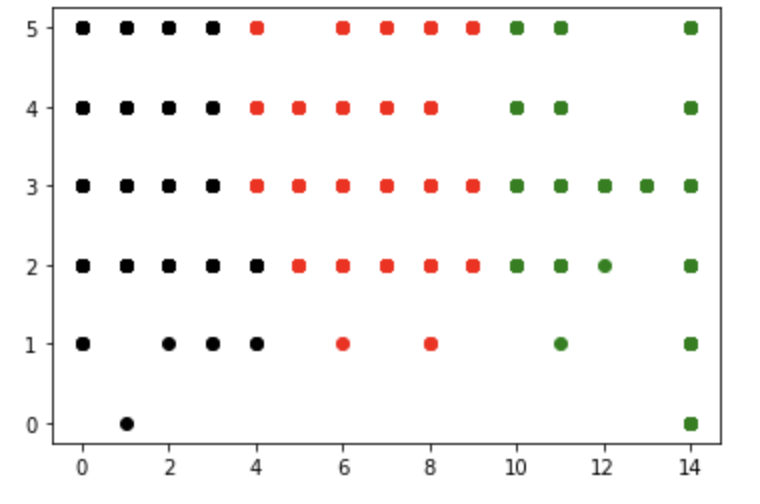
and Numerical vehicle class values(after label encoding vehicle class).

2)Clusters -Price and Duration.

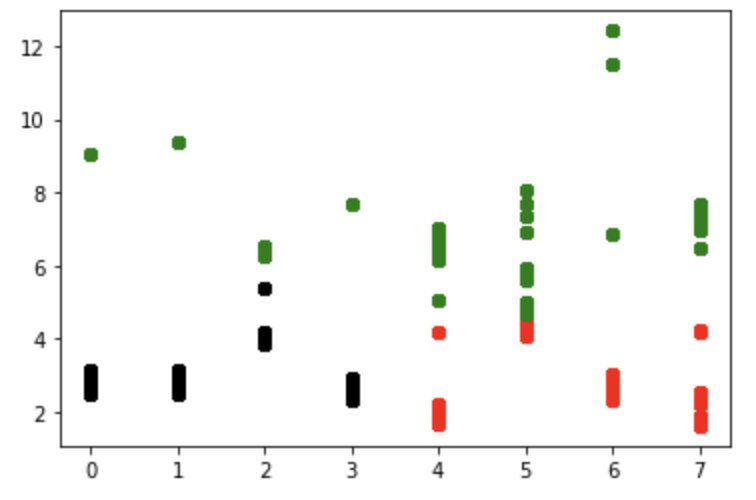
3)Clusters- Numeric Route Values (after label encoding route )and Duration

To start with, we will do the plot of the ‘duration’ and ‘price’, we assume there are 3 clusters, and we have the image:

And if we do the plot of vehicle type and vehicle class we will have such clusters:-



Moreover, if it is the routes and duration then the clusters are:-



**3.Naive Bayes**

Multinomial Naive Bayes assumes a feature vector in which each member denotes the frequency of appearance. This method is highly effective when the samples are created using a common dictionary or when NLP is being used. Instead, the Gaussian Naive Bayes is more appropriate for general classification applications because it is based on a continuous distribution.

We implemented Gaussian and Multinomial naive bayes on features -price and fare.

3.1. Gaussian Naive Bayes:-

The accuracy score obtained after training and testing the model is 0.68833333333334. We would use Gaussian on the features of price and fare. The fact that it has a decent score indicates that our model is good.

3.2. Multinomial bayes theorem:-

When modeling feature vectors with each value denoting, for instance, the frequency or number of occurrences of a phrase, a multinomial distribution is helpful. The model's accuracy score is 0.683925 as a consequence of training and testing. It has a respectable score but is little below the Gaussian.

We can therefore infer from the above that the Gaussian Naive Bayes algorithm is more accurate than the Multinomial Naive Bayes algorithm. This indicates that our model matches the Gaussian Naive Bayes well.

**4.KNN**

KNN is a method that is supervised and nonlinear, as opposed to K-Means Clustering. We start with the inputs "price" and "duration," then train and test the data using the KNeighborsRegressor (n neighbors=5).

And the outcomes are as follows:

mean absolute error: 1.1179487179898289e-06

mean square error: 6.497435897435924e-08

root mean square error: 0.00025490068453097424

The KNN algorithm is our ideal approach because it has such a high score, thus we implemented regressor.score(X test, y test) to obtain the score. The result is 0.9999999743603023.

# Conclusion

In this project, we take the “Spanish High Speed Rail Data Analysis” dataset from kaggle to implement the EDA analysis and several machine learning algorithms to get detailed information and try to figure out the relationships between the variables(values).

To create an insightful view of the data, we used univariate and bivariate analysis. Through these analyses, we learned about the characteristics of a single variable (like the most popular vehicle type) and the relationship between two variables. We also checked the dataset to make sure there were no errors and removed any null values to create a smooth path for our later machine learning implementations.

We attempted to model the dataset in the machine learning portion using the linear regression, k-means clustering, naive bayes, and KNN methods. We received R2 scores of 0.01 with "fare" and "origin," 0.03 with "price" and "month," and an R2 score of 0.82 from the linear regression, which was the best match for the features "numeric route values" and "origin." The K-Means Clustering also creates an image of continuous clusters, which is not appropriate for our dataset. The Naive Bayes algorithm (both Gaussian and Multinomial) gave us a respectable score of about 0.68, however the KNN algorithm generated a score of 0.99, suggesting that it matches our data exceedingly well. As a result, our model fits well with Linear Regression, Naive Bayes, and KNN but based on accuracy KNN fits the best with a score of 0.999.

# References

https://www.kaggle.com/code/scsaurabh/spanish-train-ticket-price-prediction-renfe

https://www.kaggle.com/datasets/thegurusteam/spanish-high-speed-rail-system-ticket-pricing/discussion/163835

https://realpython.com/knn-python/

https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/