# Business Understanding and Problem Discovery

## Business Problem

We have seen variation in prices of a cab ride. What factors affect the prices of cab ride? How much the company can change the price and still not losing customers?

To solve this query, we can run some tests which can provide a solution to the company, that how much variation is good for a company to make more profit and still not losing the valuable customers.

## Data Requirement

Need a data from a well-known cab ride company like Uber. Need features like how much time it will take to reach the destination, pricing, area, weather, time stamp etc.

Data Source :

Link : <https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices?select=Cab-Weather+Data>

## Data Assumption

* Date – Time
* Time taken for a ride
* Price
* Weather
* From station and To station

## Data Limitation

Data should not be more than 10 regions and data should be only for one or two companies.

## Proposed Test

For this problem, We can use Random Forest technique, as it provides tree structure and is best for classification and regression problem.

# Data Description:

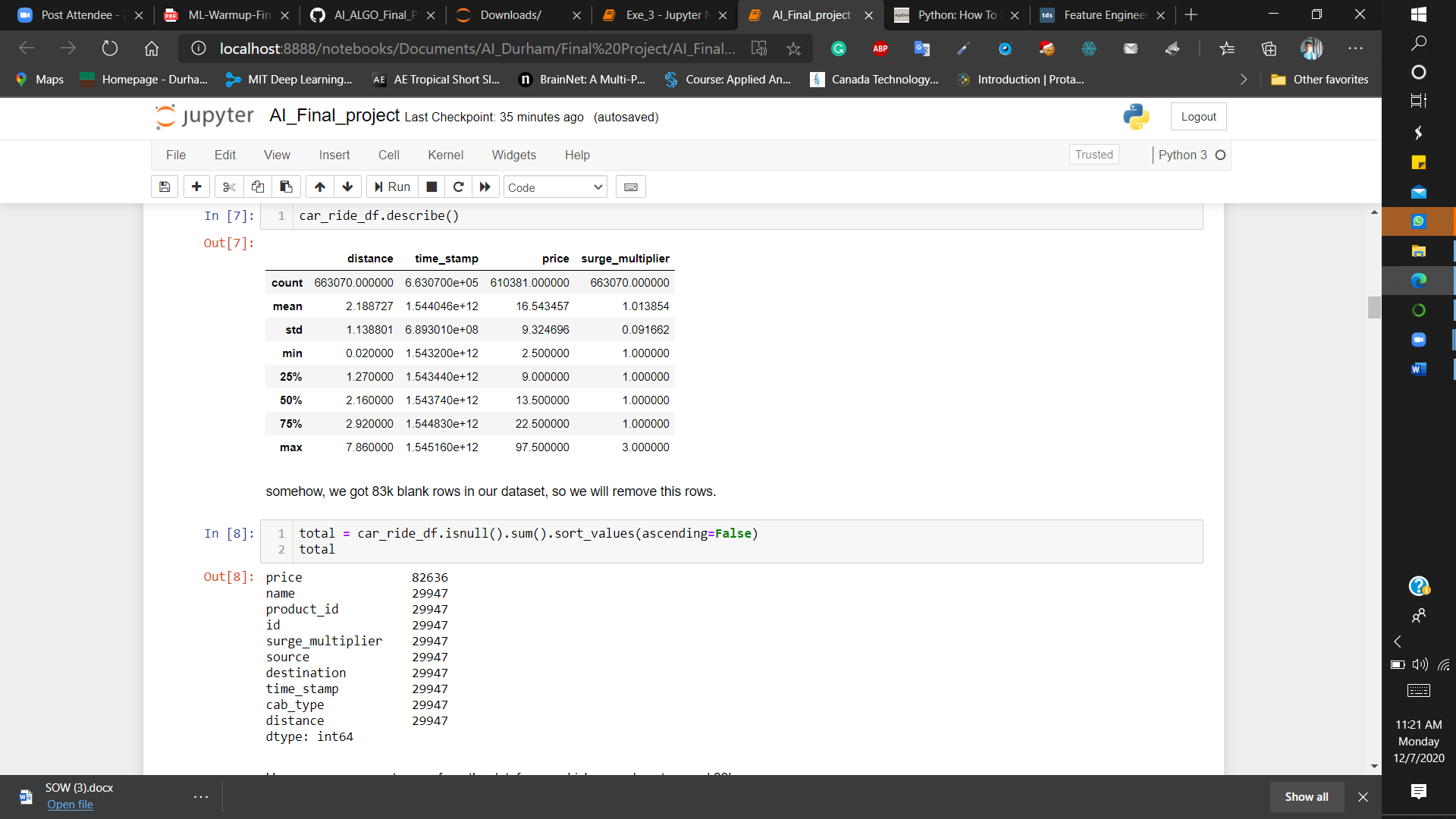
Here we have two data file

1. Cab ride dataset for two companies Uber and Lyft within 12 region.
2. That region weather dataset

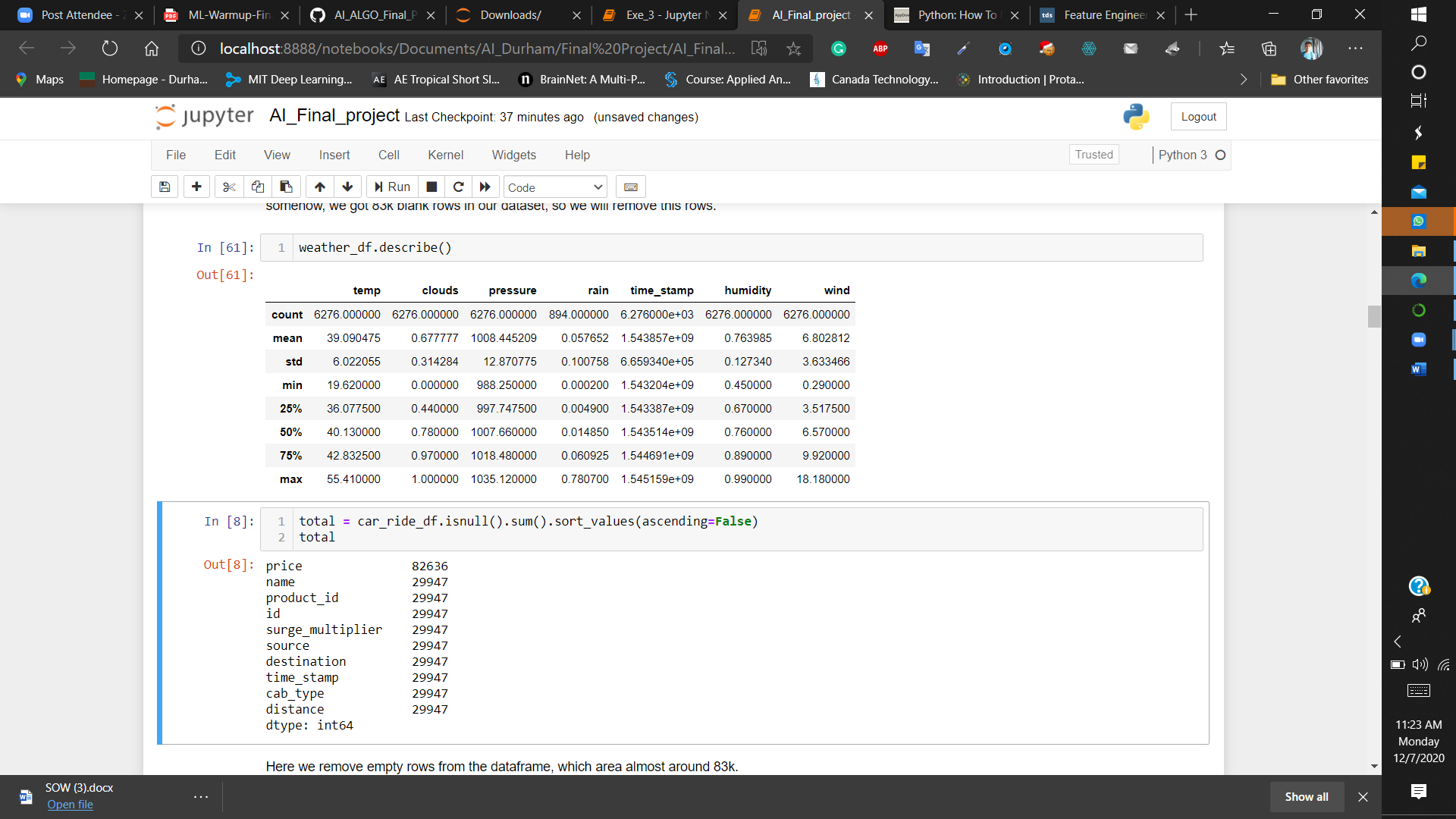
Cab ride Data consist of ‘610381 rows × 11 columns’

Weather Data consist of ‘6276 rows × 8 columns’

Cab Ride Dataset:

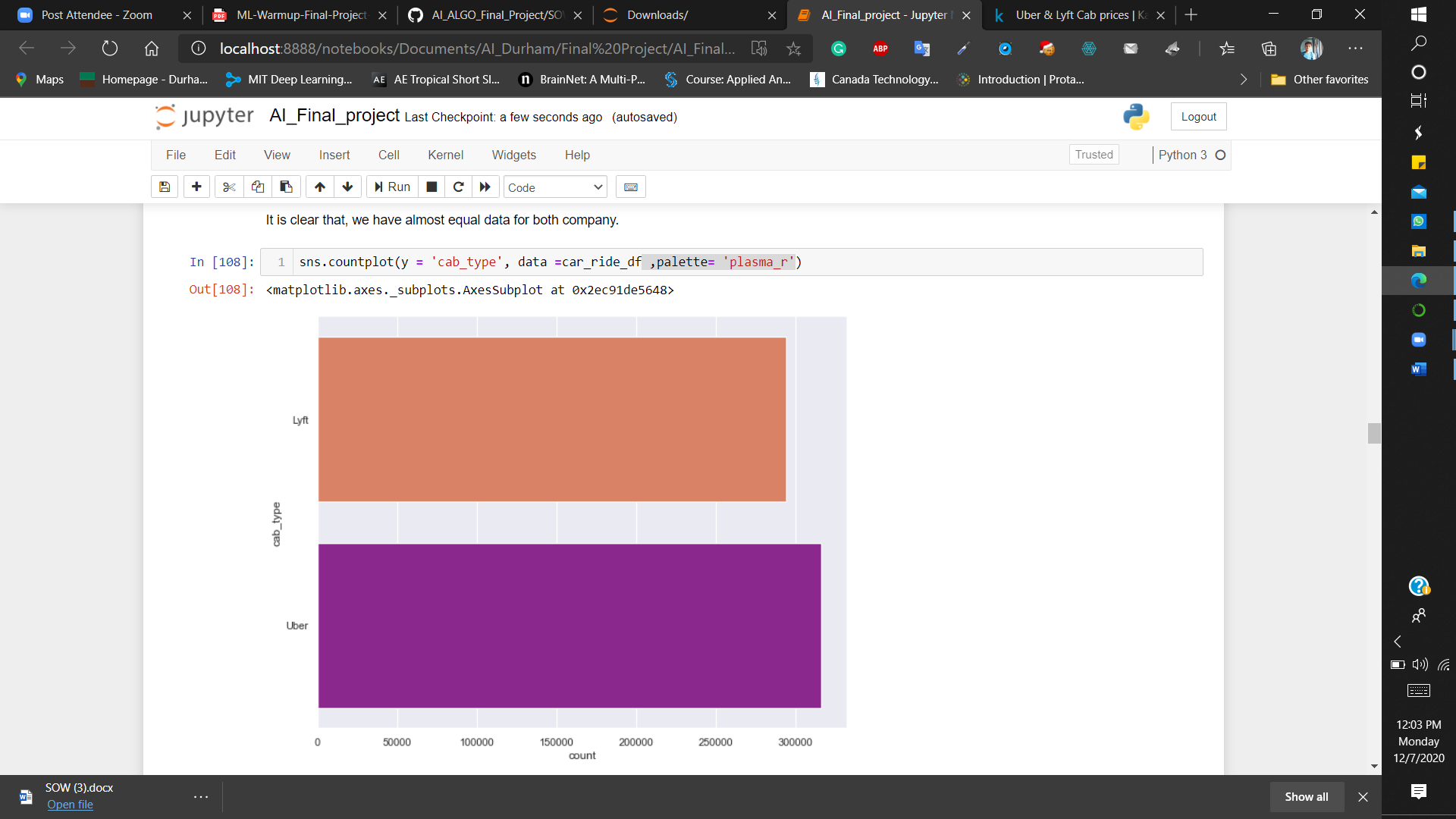


Weather Dataset:

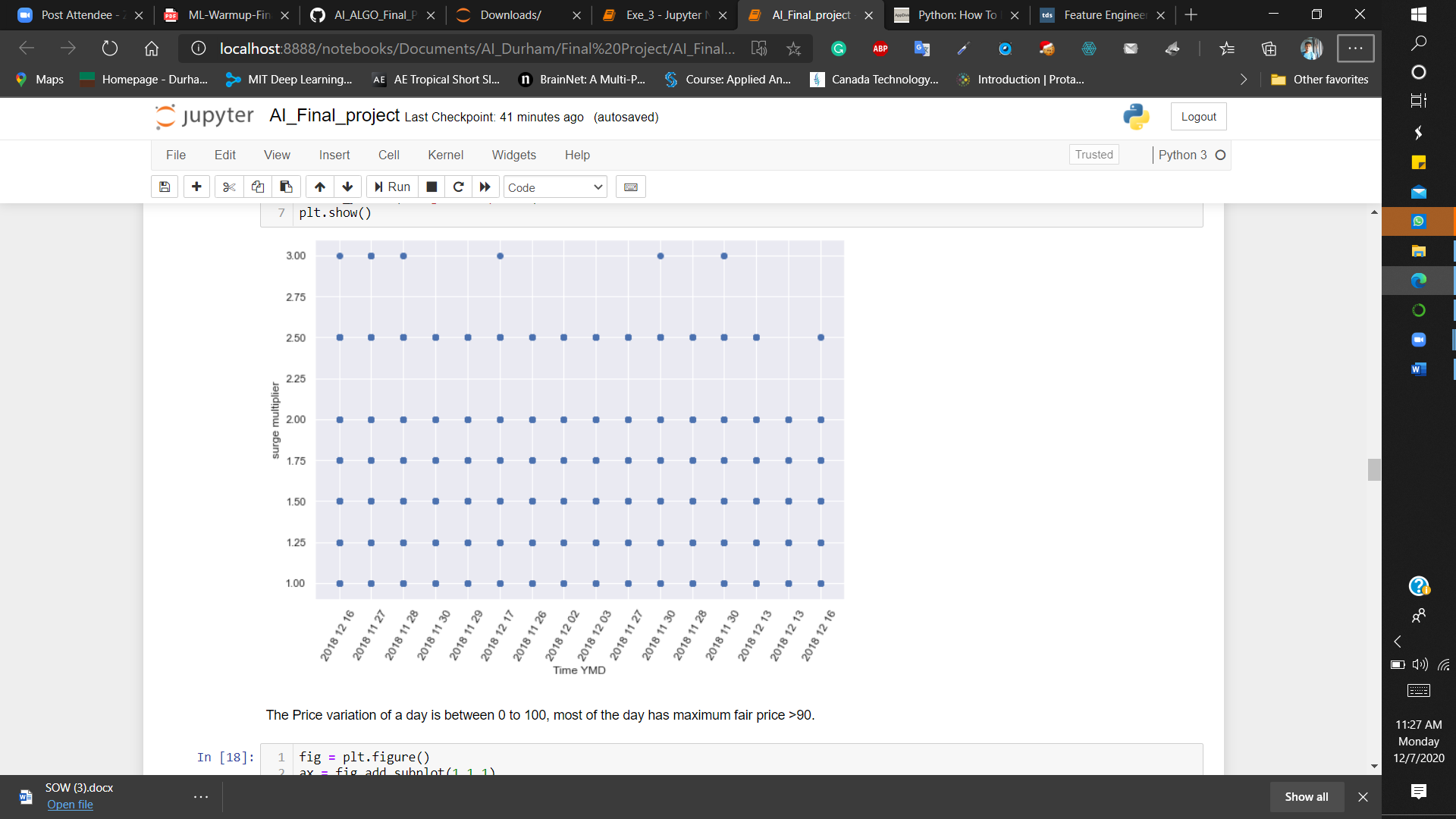


The data is collected from ‘26/11/2018’ to ‘18/12/2018’.

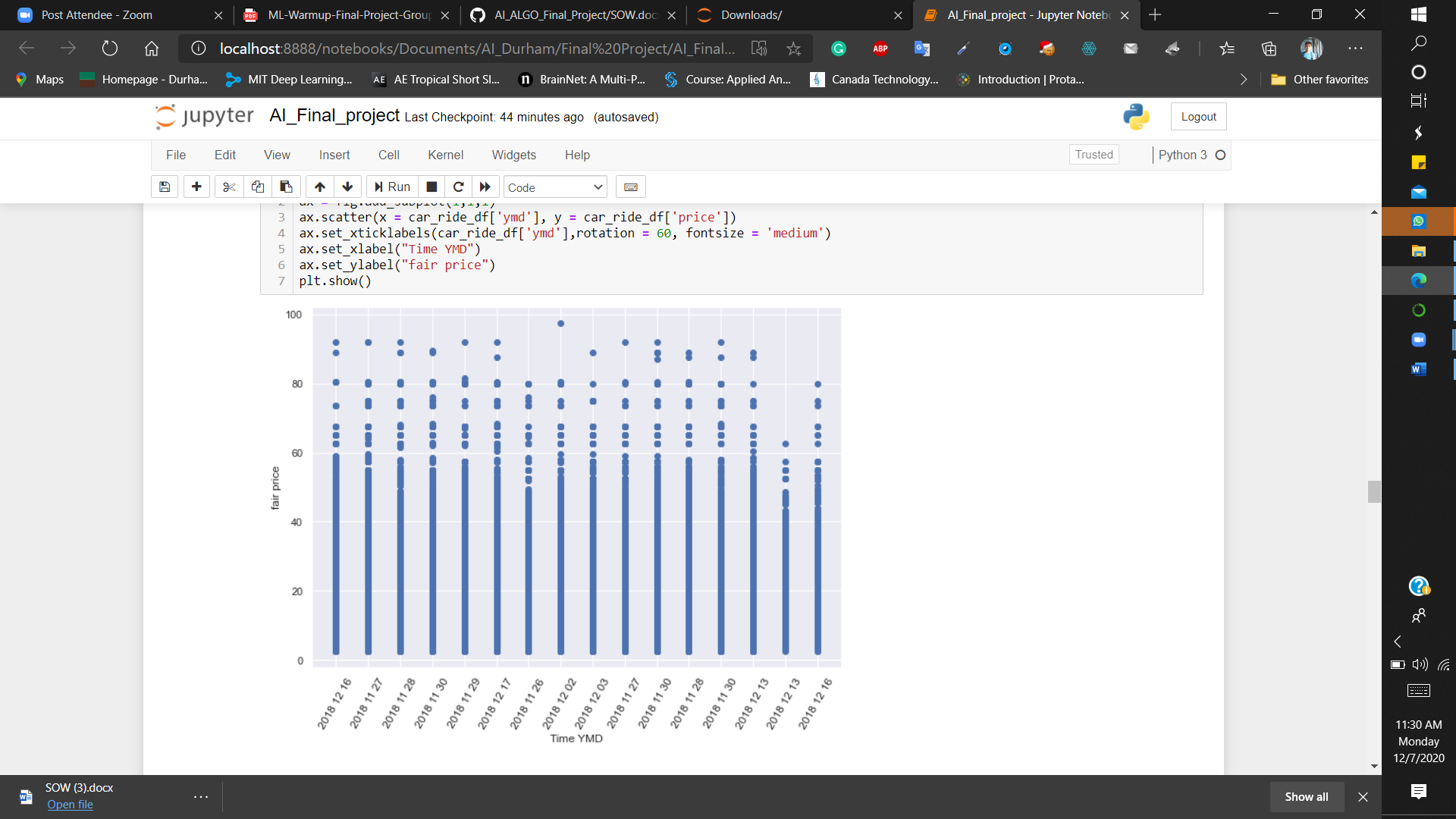
We have almost same amount of data for both companies.



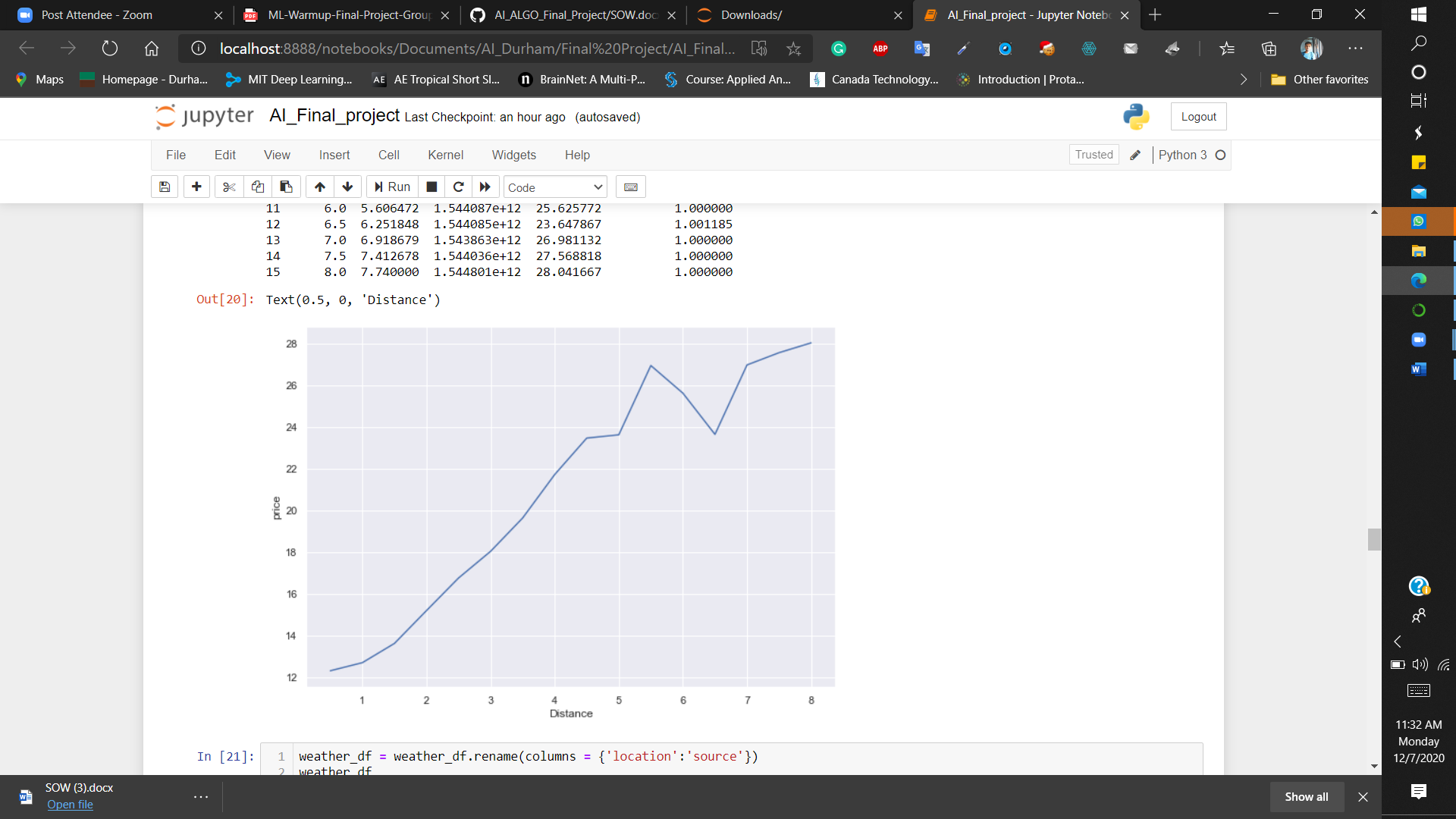
Every day variation in surge multiplier can be seen this graph, and it is interesting that there are 5 days where surge multiplier goes to 3. Which is price is tripled by that time.



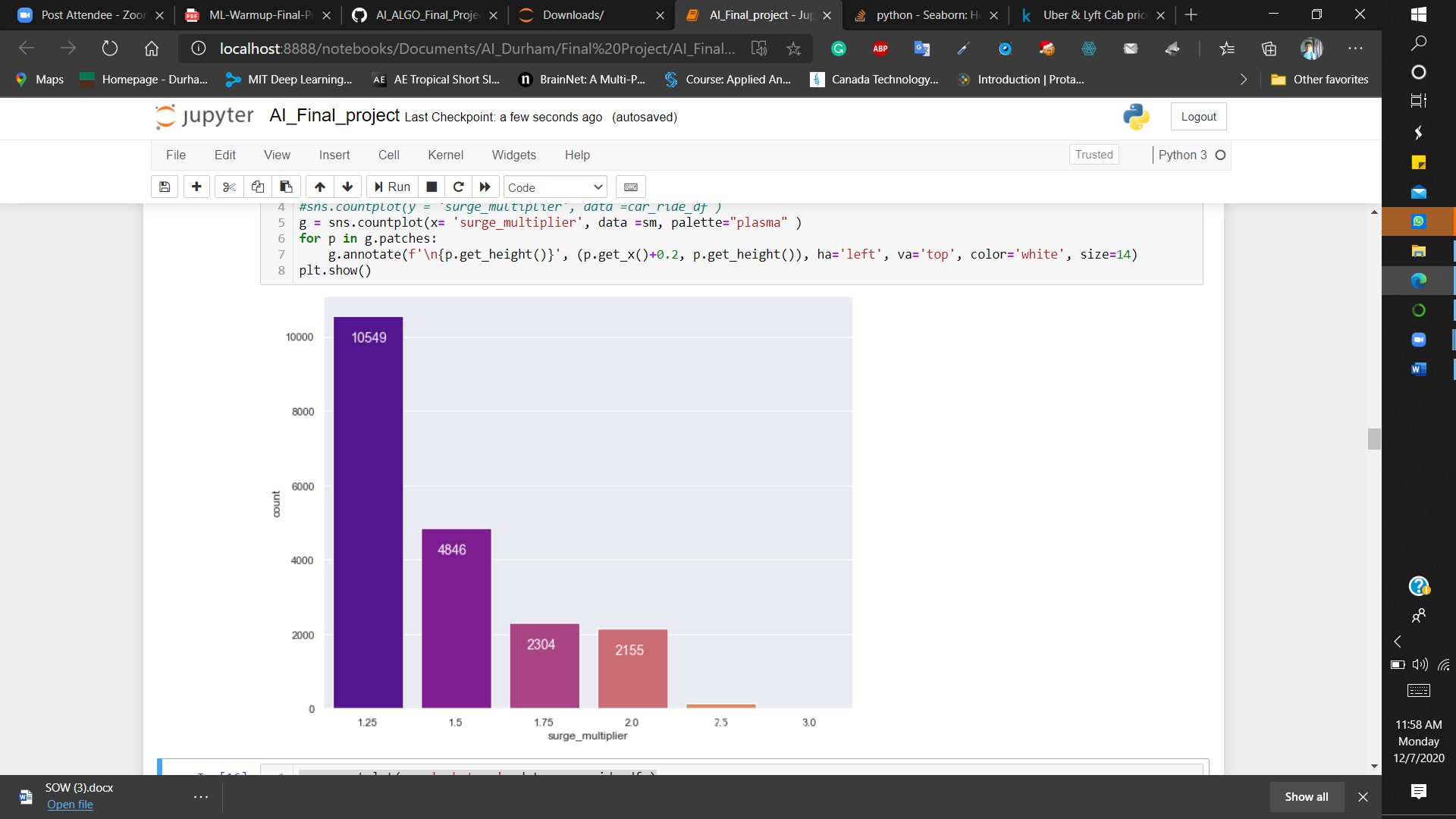
The fair price for most of the trip is lies between 0-60 but there are some exceptional cases where it is goes higher up to 90.



Distance over price : Distance increases the price of ride is also increasing; but there is an interesting depth at distance 6.5 to 7 where the price goes below 24.



Count of surge\_multiplier for greather than 1 value.



## Model Usage

For this example, I have purposed to use random forest for classification problem. Here, we try to train and test our model based on the given dataset car ride share and weather dataset. Both data are very well formatted before use it in model.

### Why Random Forest?

It is the best algorithm in classification and regression problem. Also, it provides high accuracy through cross validation. Random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data. If there are more trees, it won’t allow over-fitting trees in the model.

### Feature Extraction:

This is a crucial part, where data needs to be changed in binary or numerical format from the character or string format, changing float precision for several field required.

### Train – Test Dataset:

Model training part I split data, 80% data for training and rest 20% is for testing purpose.

### Model Outcome