Saurabh Salunkhe

  Northeastern University

Automated price recommendation system

Contents

[Abstract 2](#_Toc505376740)

[Problem Description 2](#_Toc505376741)

[Background Research 2](#_Toc505376742)

[Datasources 3](#_Toc505376743)

[Input Features 3](#_Toc505376744)

[Algorithms Used: 3](#_Toc505376745)

[Result and Analysis: 3](#_Toc505376746)

[References 6](#_Toc505376747)

[Conclusion 6](#_Toc505376748)

# **Abstract**

# **Problem Description**

Big websites or online shopping apps like Mercari often face the problem of predicting how much does an online product cost. Such websites often allow sellers to upload about hundreds of products on their website. Sellers seldom put any price tag on them, which lefts the website owner to decide what price tag to put on that product.

Also, small details can mean a lot of difference in pricing, Poor price labelling can vastly deteriorate the business. Here, our aim as a machine learning engineer is to predict the online product prices by building an algorithm taking into considerations the different features like the product category name, brand name and item condition.

# **Background Research**

The background research will greatly consist of the present models that determine the price of the online products. Firstly, in such websites if the sellers themselves provide the price we have a metric to compare the price our model suggests. However, if there is not price suggested by the sellers, this are the following methods that are used:

There are point of sale (POS) system that do the math for the product and helps set up automated pricing rules for the inventory. Such techniques greatly depend on the immediate profit margin and may not prove as a good metric for labelling the product price.

Another model simply relies on the industry standards. Although this technique can help approximate the price and at the end we may conclude at a safe price but again it not worth trusting as the industry standard is just the estimate and not the exact price.

The current model makes poor assumption about the product simply based on the product brand name or relating the product to previously such encountered similar products. Such techniques do not make a lot of sense since price range of a brand may vary drastically or previously encountered product may be similar looking but from different material or different retailer.

# **Datasources**

Mercari is one of Japans biggest app which is powered by their community. This website has provided us the data in the form of one of the Kaggle Competition.

## **Input Features**

|  |  |
| --- | --- |
| Features | Feature explanation |
| Name | Date from the search event |
| Item Condition | Condition of item provided by seller |
| Category Name | Category of the Listing |
| Brand Name | Name of the brand |
| Price | The price of the item (Target variable) |
| Shipping | 1 if paid by seller, o if paid by Buyer |
| Item Description | The full description of the item |

# **Algorithms Used:**

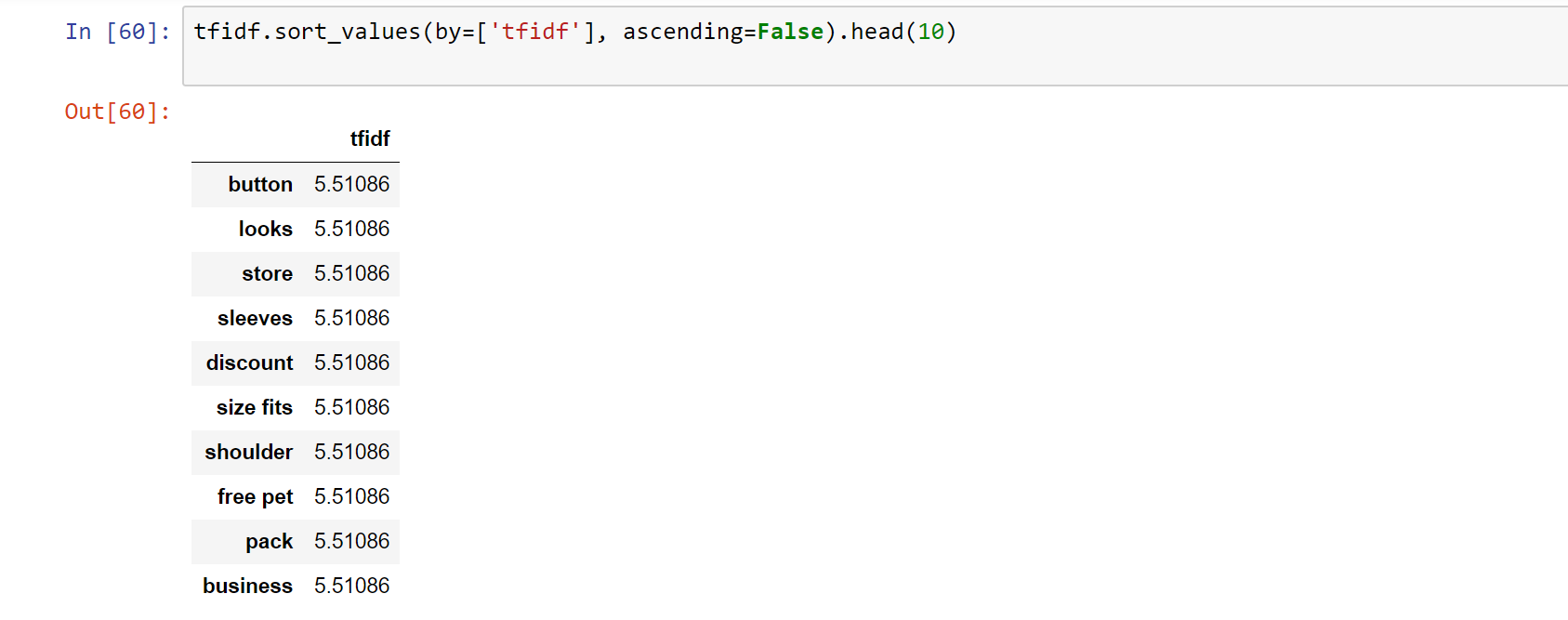
The python libraries used for this project includes pandas, numpy, nltk, matplotlib, scikilt learn, plotly and seaborn.

The main algorithms used in the analysis includes scikilt learn tfidf and eucledian distance, nltk’s word tokenization, scikit learn k-means algorithms.

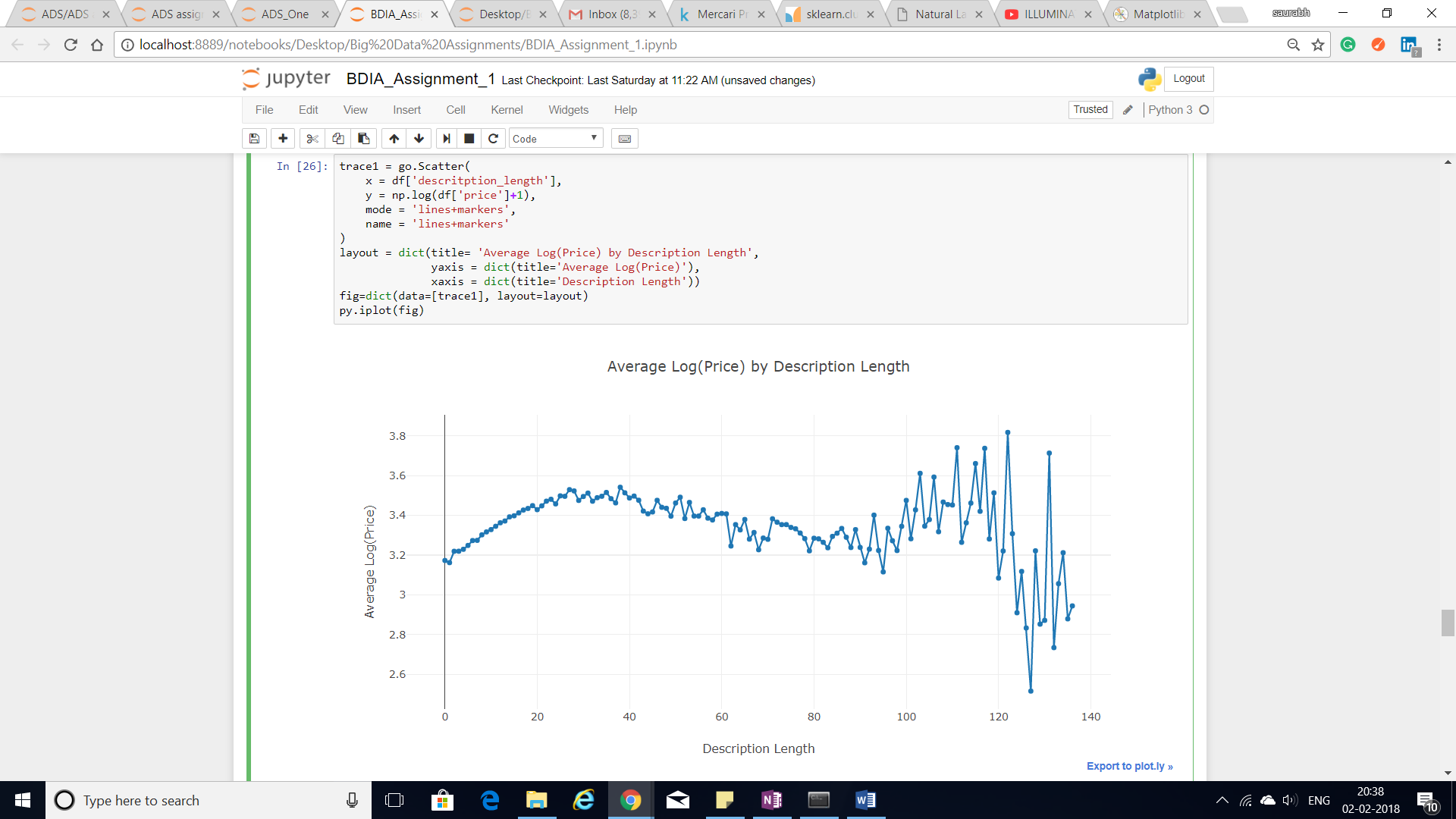
# **Result and Analysis:**

The report consists of three analyses. Since, we are dealing with text data and are supposed to find out price based on this text data, it makes a lot of sense for us to find which words carry how much weightage. Major part of our analysis will include text pre-processing and finding the term frequency and inverse term frequency of our data-sets.

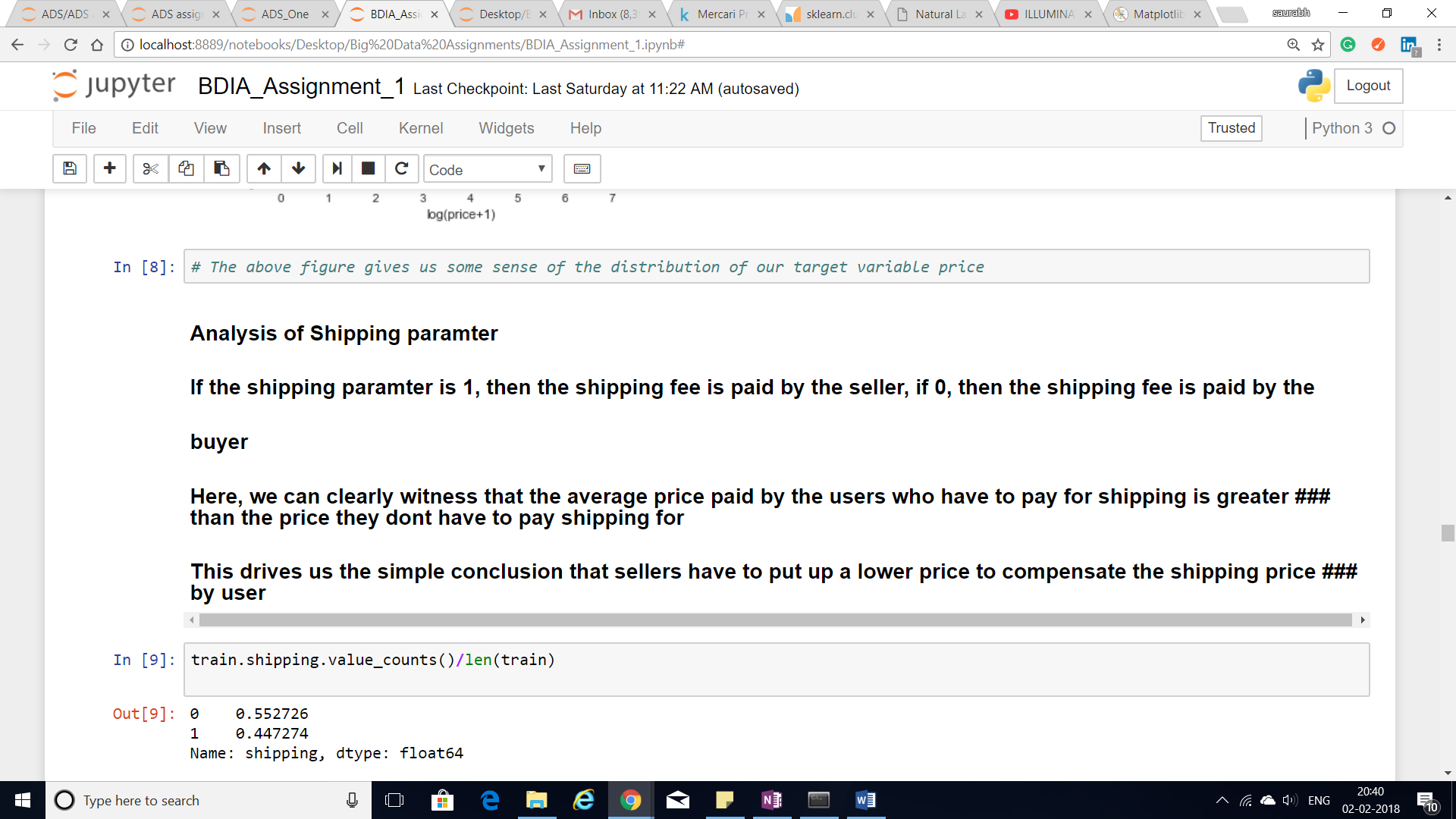




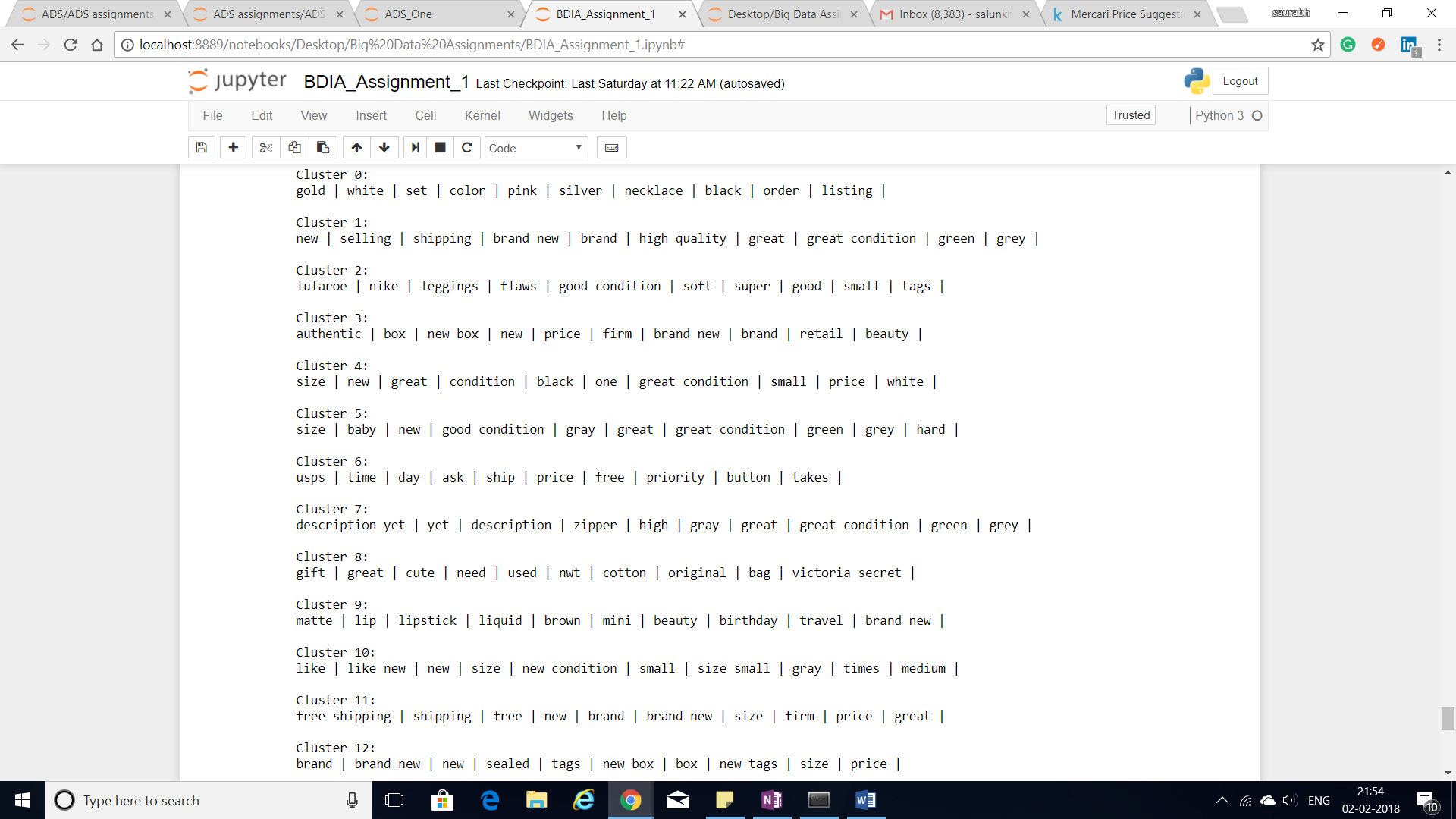
Second part of our analysis will include how the description of the product influences the price of the product. Here, we will take help of natural language processing to process text data and visualization library to show our interesting findings.



Third analysis will be a simple analysis to find how the shipping factor decides the cost of the product for both the entities- seller and buyer.



Our final analysis consists of K-means clustering, where we will utilize the k-means algorithm for clustering the similar words together which can give us insight about a lot of things.



# **References**

<https://www.kaggle.com/c/mercari-price-suggestion-challenge/data>

<http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

<http://www.nltk.org/>

<https://matplotlib.org/>

# **Conclusion**

We first started by stating our problem statement of predicting the online prices and presenting the background results. After performing exploratory data analysis, we got a sense of the data and we were in a position to anticipate the algorithms. Since, our data consisted of text data, we performed natural language processing to remove stopwords, stemwords and punctuation etc. We further stated useful analysis like how shipping factor can affect the average price. We also found out the most significant and the least significant words from the corpus of our data by using the TF-IDF metrics. Furthermore, we applied K-means algorithm to cluster the words that belong to a particular category and this was instrumental in getting an insight about the price of the product.