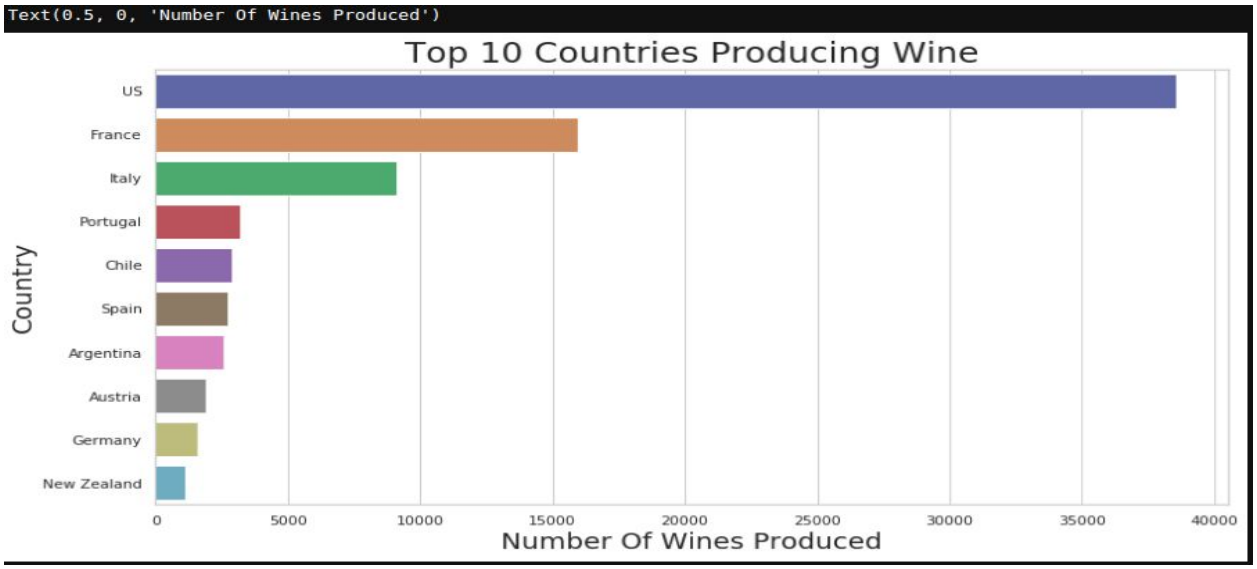
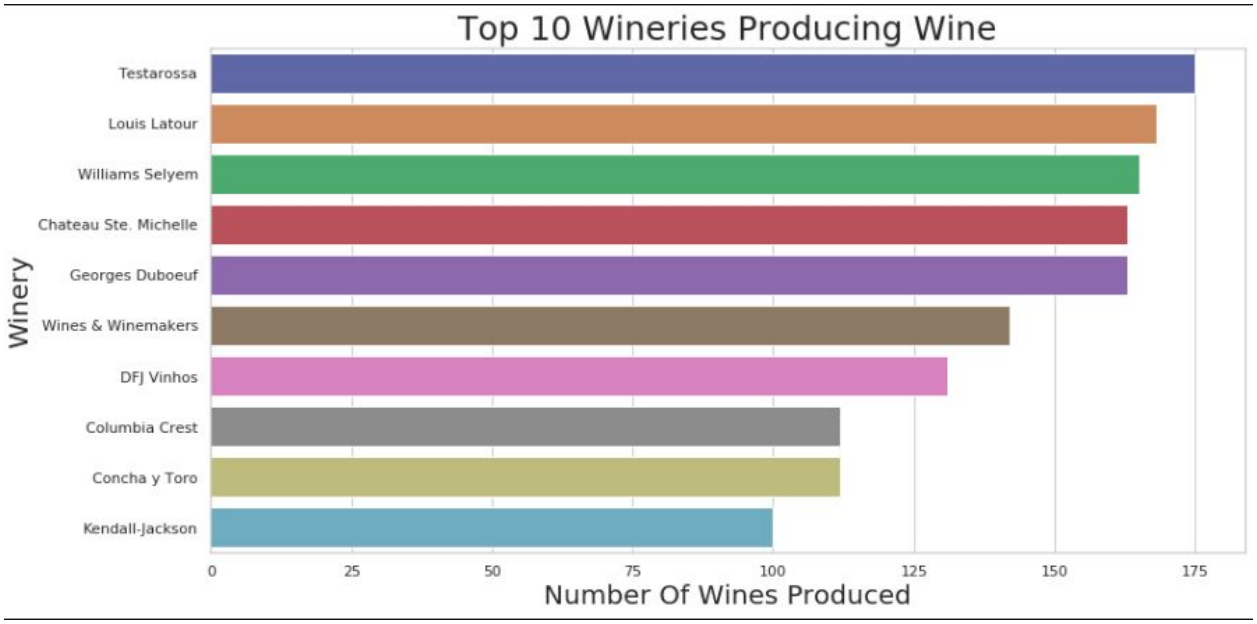
Top 5 Insights from Data

1. Top 10 wine producing countries



Outcome: US is the largest wine producing country

2.Top 10 Wineries Producing Wine

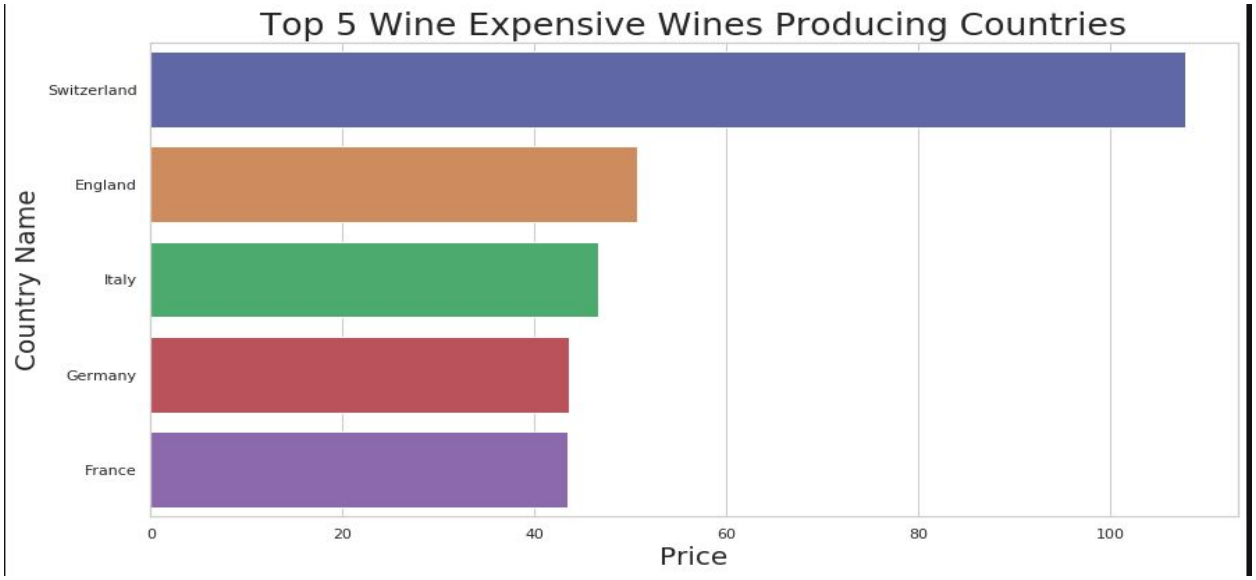


Outcome:Testarossa is the top winery in terms of production.

3.Distribution Of Prices Of Wines

Outcome:Almost 50% of the wines have price less than 30
Only 1.2% of wines cost more than 300

4.Top 5 Wine Expensive Wines Producing Countries



Outcome:Switzerland produces most expensive wines

5.Top 5 Wine Cheapest Wines Producing Countries

Outcome:Ukraine is the cheapest wine producer.

****These plots are made using libraries matplotlib and seaborn.For code of the above visualisations please refer to attached Jupyter notebook.**

Task:2

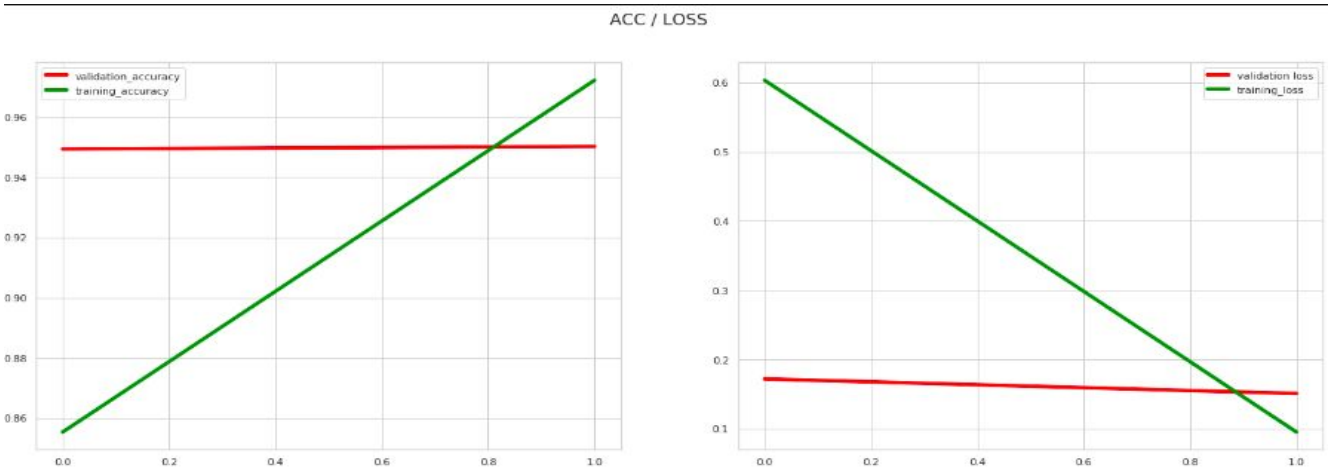
Building a predictive model for predicting the wine "variety"

Initial Approach:

1. Initially just took the "review description" column as input feature and "variety" as output feature.
2. Performed Pre Processing steps like:Punctuation Removal,conversion of letters to lowercase,Tokenization and stop word removal.
3. Also encoded the output variable "variety" using Sklearn library Label Encoder.
4. Then is splitted my train Dataset and used the model SVM but the accuracy was not good,about 60%.This may be due to the non linear separation between different varieties.
5. So i tried a neural network.In this on train set I am getting about 85% but validation accuracy still not upto the mark which is 65%

Final And Better Approach:

1. I again explored the Dataset and found that review title is an important feature and can't be ignored.The reason is some of the review titles directly contain the variety name and many reviews contain the review year.
2. So I made a new feature combining the columns “review title” and “review description”.
3. Again I performed pre processing steps which include Punctuation Removal,conversion of letters to lowercase,Tokenization and stop word removal.
4. Also encoded the output variable “variety” using Sklearn library Label Encoder
5. Then is splitted my train Dataset and used a neural network with a single layer having 200 neurons and “sigmoid” activation in the output layer.
6. I used 2 epochs and results are amazing.I got approximately 97% accuracy on the train and 95% on validation set.Here is the visualisation of the training process.



Report Made By:Saksham Jain
E-mail: sakshamj74@gmail.com