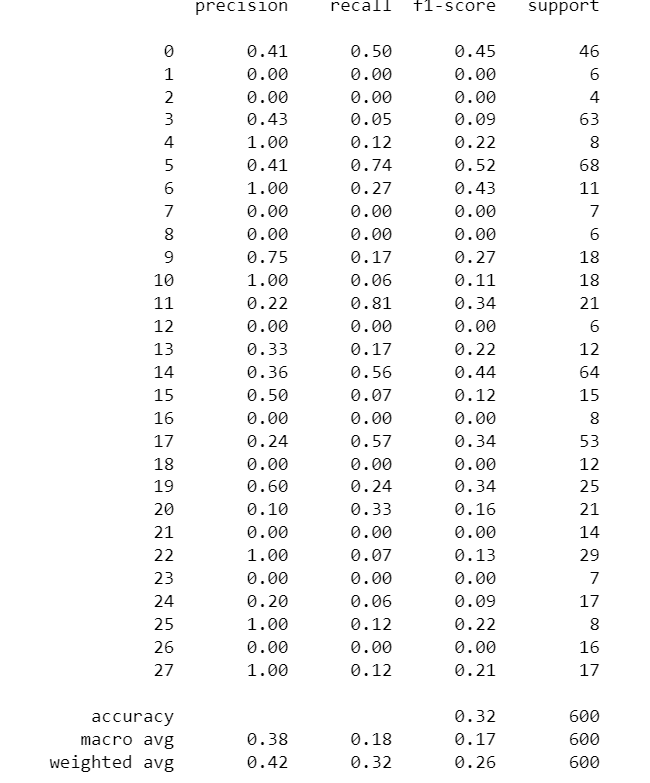
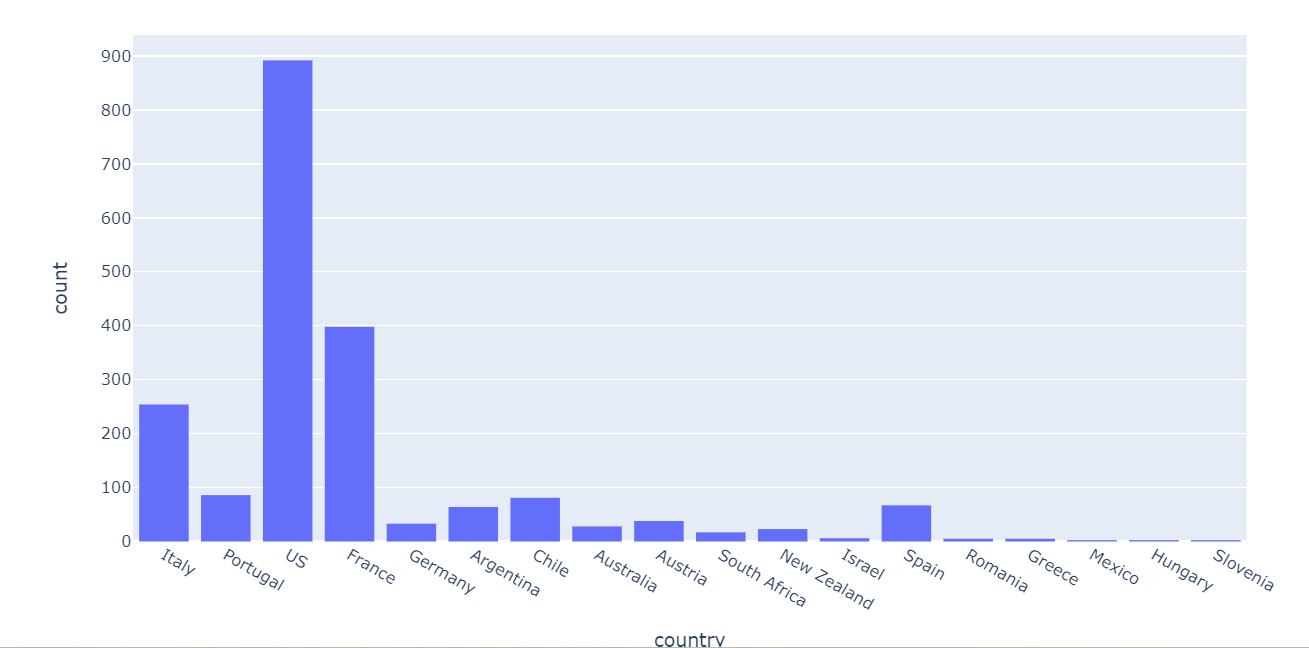
**TOP 5 actionable insights from the data –**

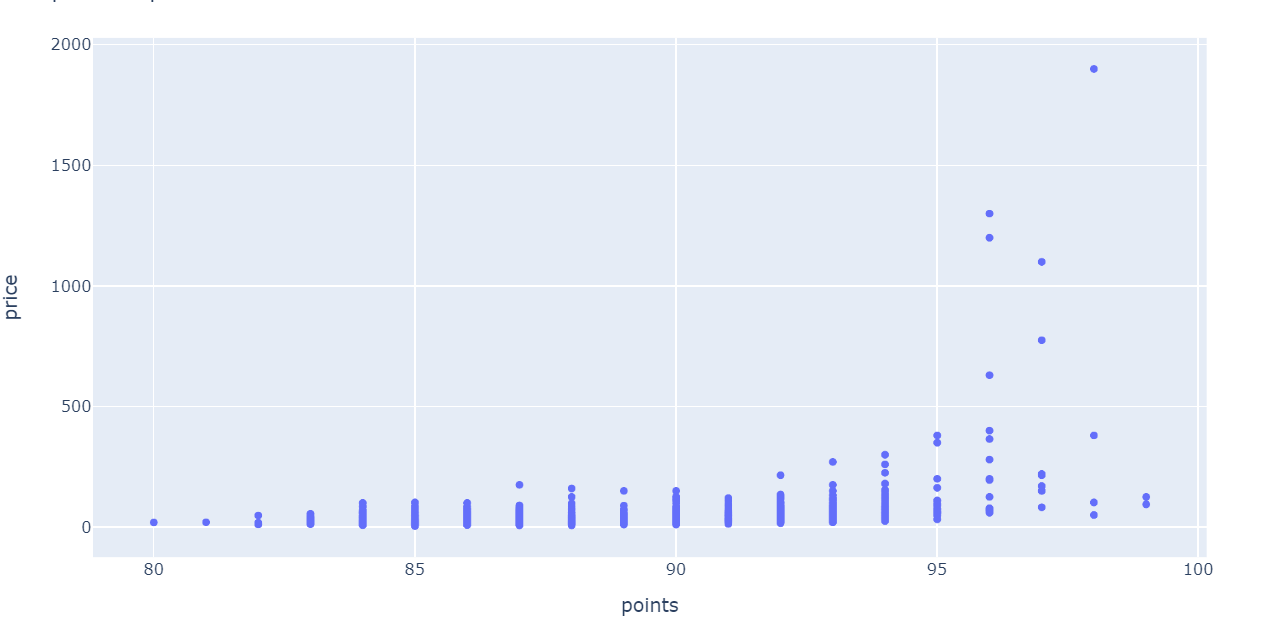
1) I found that the grapes quality column i.e the **variety** column has 28 different types, (the target variable) had a lot of features. This made our precision and F1-score very less as expected. If we could reduce these features, it would help us in better precision and accuracy.



2) Some Countries like US, Italy, Portugal etc. had a lot of rows compared to other countries. This made data a bit imbalanced. So, we can find a way to balance this dataset, it would be helpful thus increasing our accuracy.



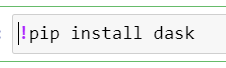
3) As we can conclude that as the price of the wine gradually increases, the points of that wine is increasing.



4) The **number of components** we used were 7. In short, we divided the reviews in a scale of 1 to 7 depending on work frequency of words appearing in each reviews. We can play around this value and increase and decrease number of components depending on our requirement.



5) As the data was too big, I used **Dask** library for parallel computing. Pandas library was not able to handle this big data this over 100000 rows and 10000 columns, so I used Dask.



import dask.dataframe as dd



I have made assumptions and stated them inside notebook. Please go through the notebook for in-depth analysis.

At last, I want to thank Sensegrass for providing me with the opportunity to work on this assignment. I found the assignment challenging but also an opportunity to learn a lot and work on my skills.

Yours Sincerely,

Saurabh Bannagare