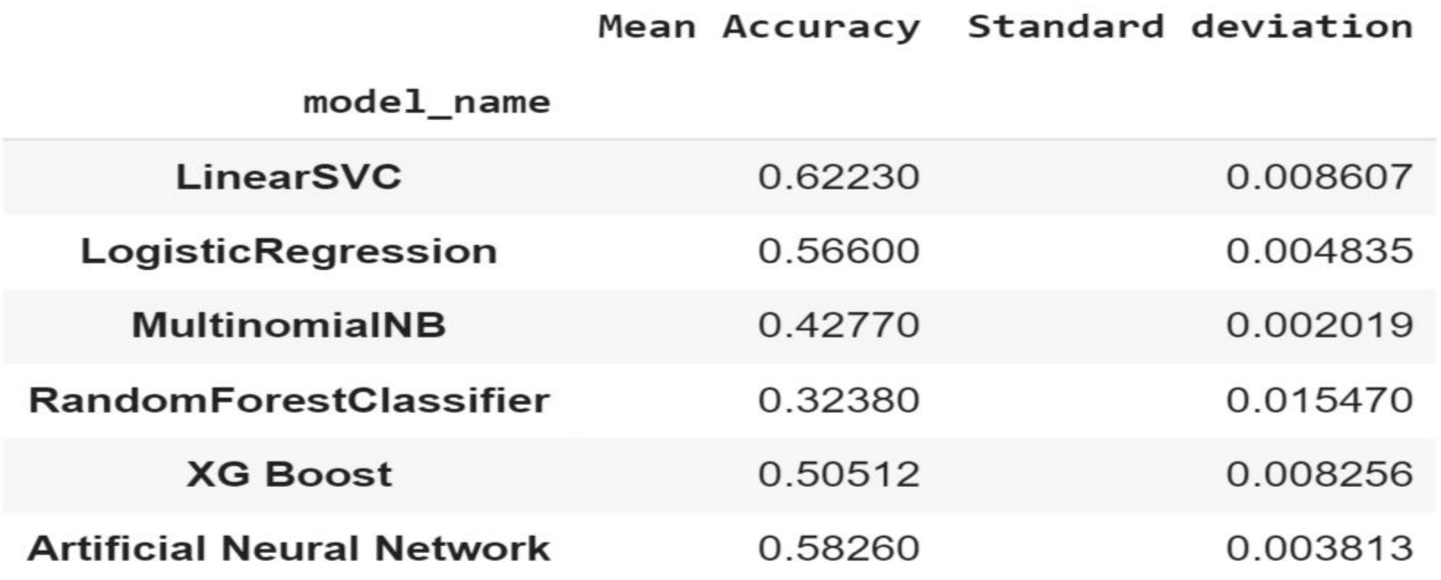
Summary

Cross-Validation scores::

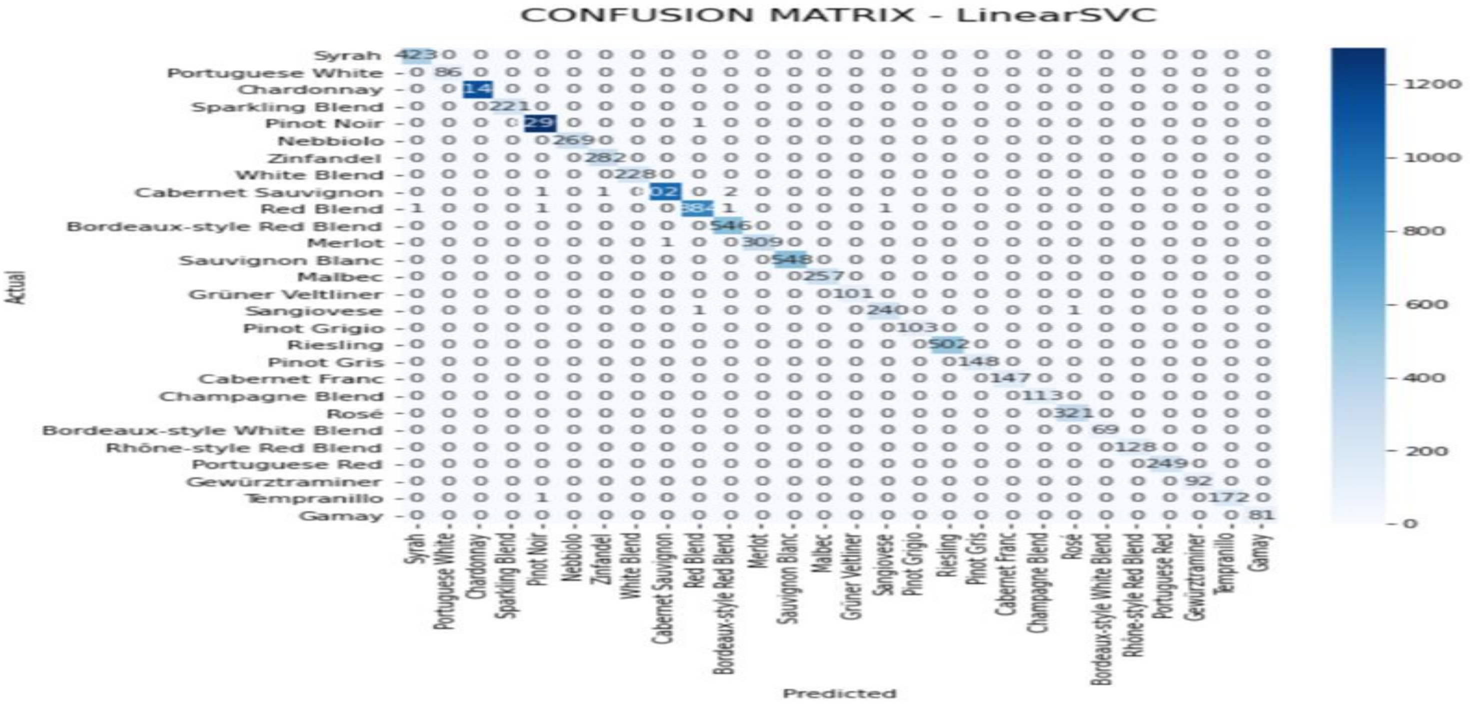


model_name	Mean Accuracy	Standard deviation
LinearSVC	0.62230	0.008607
LogisticRegression	0.56600	0.004835
MultinomialNB	0.42770	0.002019
RandomForestClassifier	0.32380	0.015470
XG Boost	0.50512	0.008256
Artificial Neural Network	0.58260	0.003813

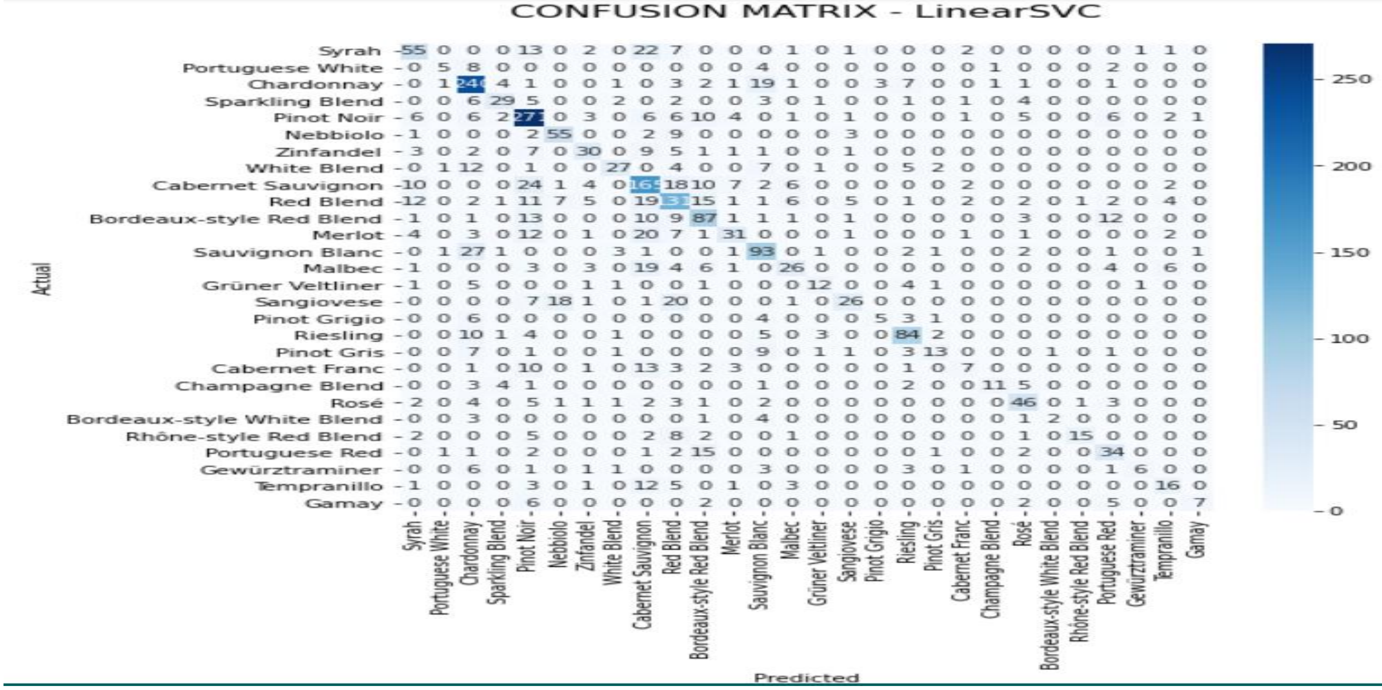
After applying the above models to the batch of 10000 training dataset(reduced) and calculating the cross-validation score for each model for **nfolds=5**, we obtain maximum accuracy with the **"LinearSVC"** model. Thus, we predicted the test data with the "LinearSVC" model.

Data-Visualization::

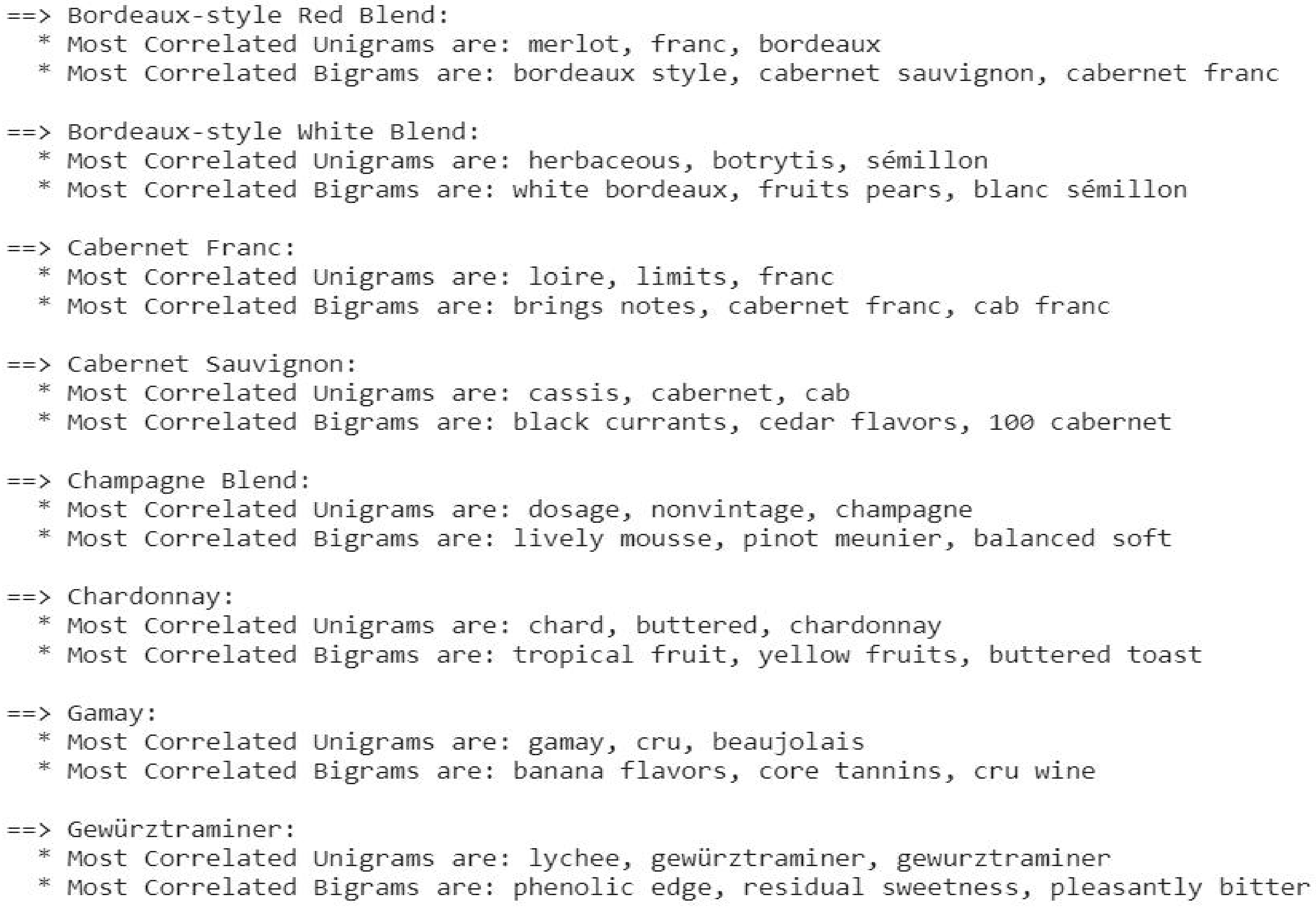
Model performance on training data itself can be visualized by the below confusion matrix.



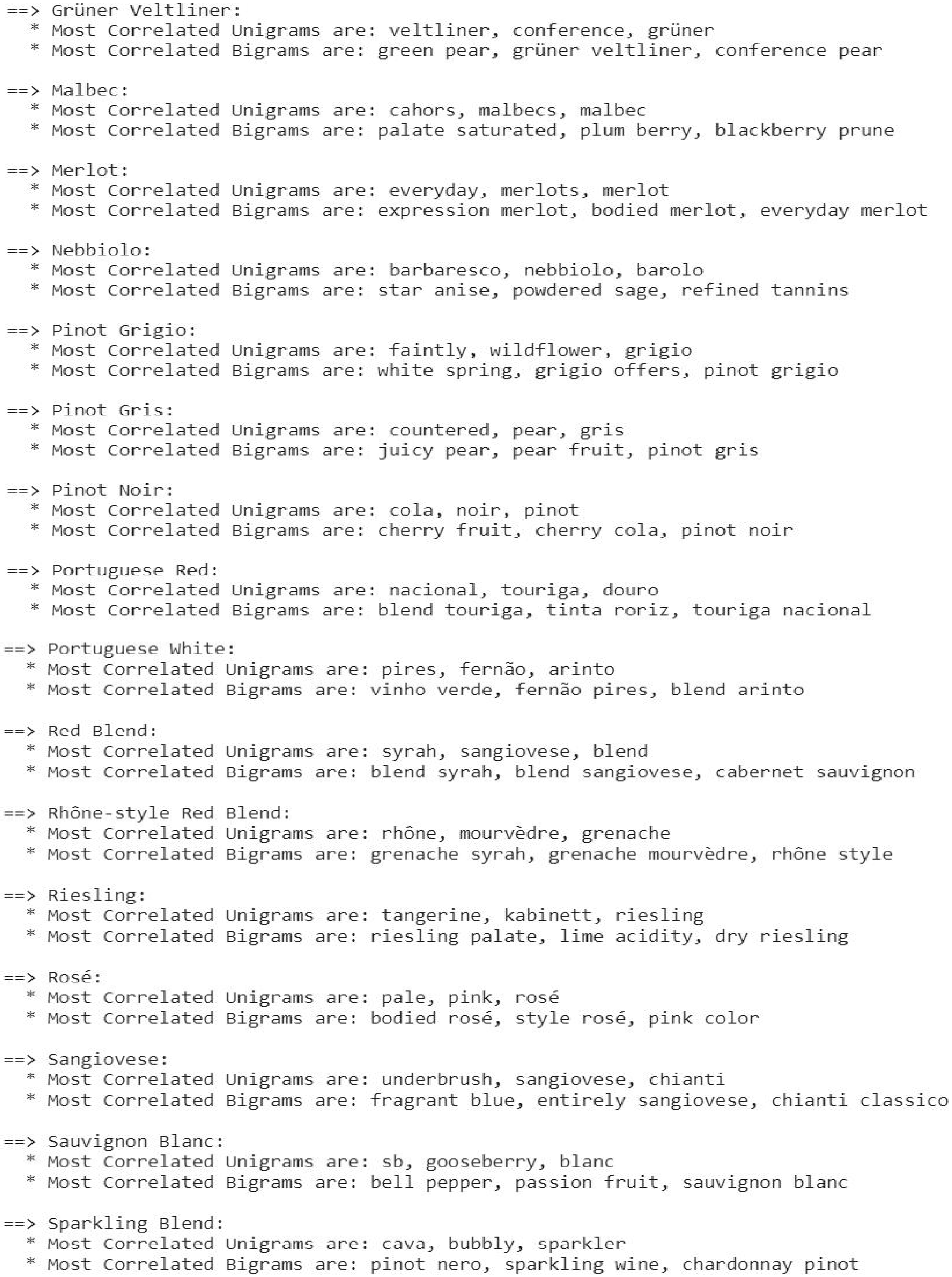
The model performance on validation set is given below.



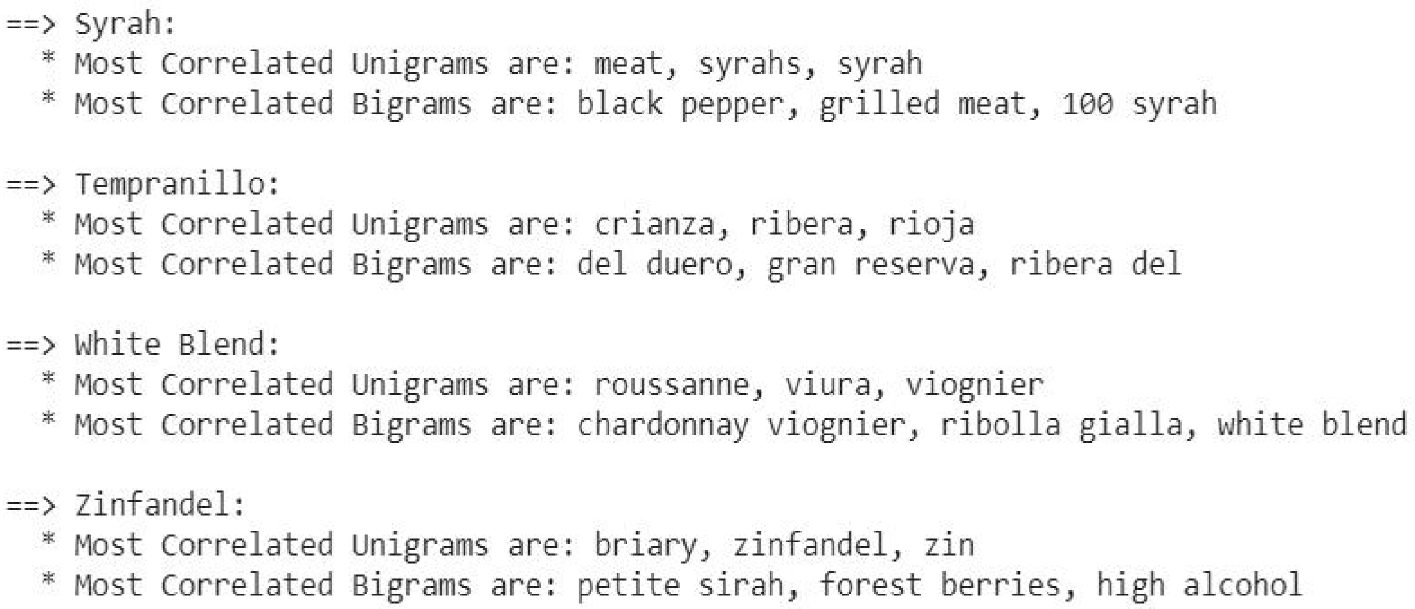
Most frequent unigrams(Single word), bigrams(Double word) in the “review_description” for all the varieties of wines. These can be used to describe a variety of wine.



```
==> Bordeaux-style Red Blend:
  * Most Correlated Unigrams are: merlot, franc, bordeaux
  * Most Correlated Bigrams are: bordeaux style, cabernet sauvignon, cabernet franc

==> Bordeaux-style White Blend:
  * Most Correlated Unigrams are: herbaceous, botrytis, sémillon
  * Most Correlated Bigrams are: white bordeaux, fruits pears, blanc sémillon

==> Cabernet Franc:
  * Most Correlated Unigrams are: loire, limits, franc
  * Most Correlated Bigrams are: brings notes, cabernet franc, cab franc

==> Cabernet Sauvignon:
  * Most Correlated Unigrams are: cassis, cabernet, cab
  * Most Correlated Bigrams are: black currants, cedar flavors, 100 cabernet

==> Champagne Blend:
  * Most Correlated Unigrams are: dosage, nonvintage, champagne
  * Most Correlated Bigrams are: lively mousse, pinot meunier, balanced soft

==> Chardonnay:
  * Most Correlated Unigrams are: chard, buttered, chardonnay
  * Most Correlated Bigrams are: tropical fruit, yellow fruits, buttered toast

==> Gamay:
  * Most Correlated Unigrams are: gamay, cru, beaujolais
  * Most Correlated Bigrams are: banana flavors, core tannins, cru wine

==> Gewürztraminer:
  * Most Correlated Unigrams are: lychee, gewürztraminer, gewurztraminer
  * Most Correlated Bigrams are: phenolic edge, residual sweetness, pleasantly bitter
```

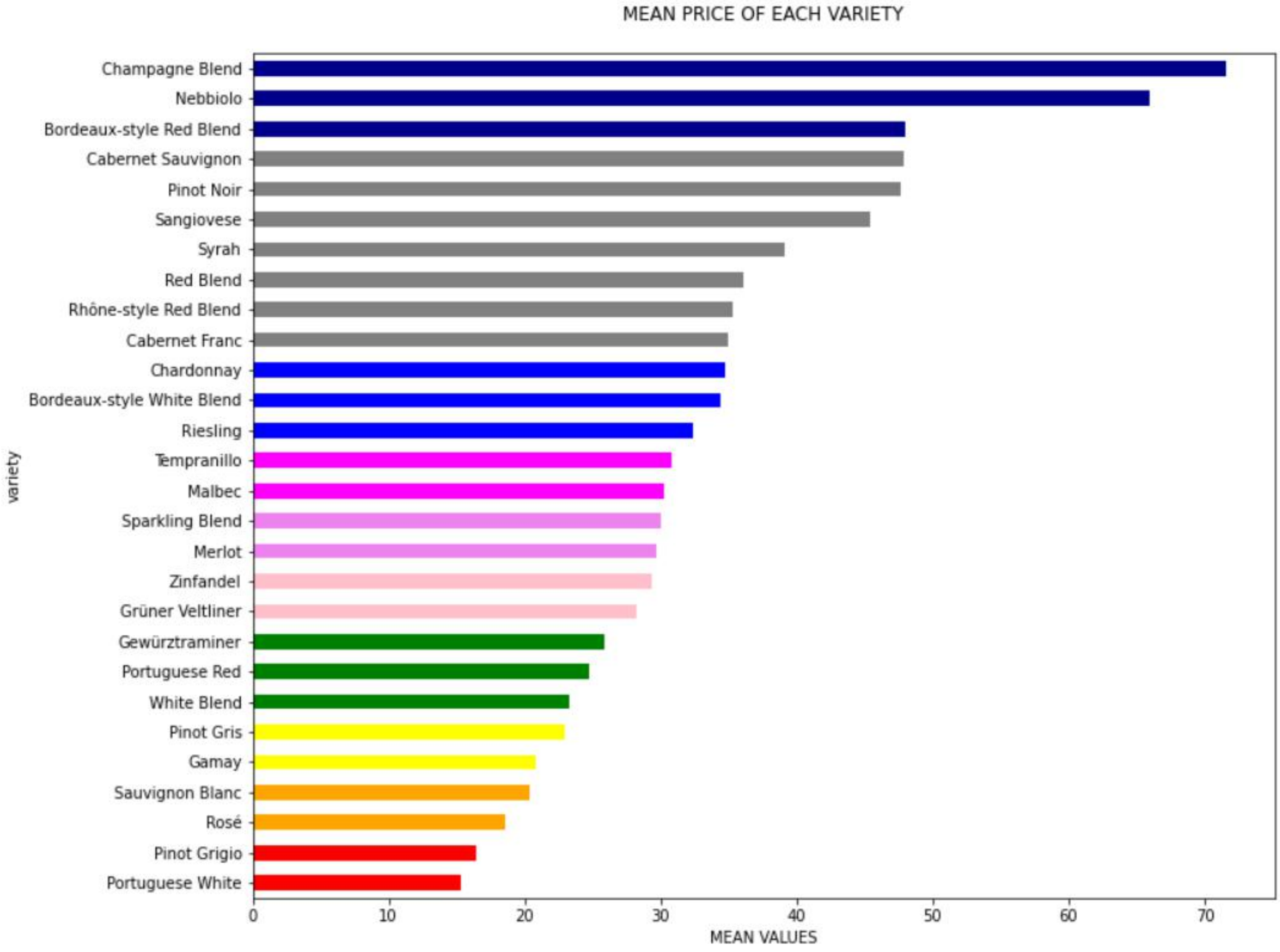


```
==> Grüner Veltliner:
  * Most Correlated Unigrams are: veltliner, conference, grüner
  * Most Correlated Bigrams are: green pear, grüner veltliner, conference pear

==> Malbec:
  * Most Correlated Unigrams are: cahors, malbecs, malbec
  * Most Correlated Bigrams are: palate saturated, plum berry, blackberry prune

==> Merlot:
  * Most Correlated Unigrams are: everyday, merlots, merlot
  * Most Correlated Bigrams are: expression merlot, bodied merlot, everyday merlot

==> Nebbiolo:
  * Most Correlated Unigrams are: barbaresco, nebbiolo, barolo
  * Most Correlated Bigrams are: star anise, powdered sage, refined tannins

==> Pinot Grigio:
  * Most Correlated Unigrams are: faintly, wildflower, grigio
  * Most Correlated Bigrams are: white spring, grigio offers, pinot grigio

==> Pinot Gris:
  * Most Correlated Unigrams are: countered, pear, gris
  * Most Correlated Bigrams are: juicy pear, pear fruit, pinot gris

==> Pinot Noir:
  * Most Correlated Unigrams are: cola, noir, pinot
  * Most Correlated Bigrams are: cherry fruit, cherry cola, pinot noir

==> Portuguese Red:
  * Most Correlated Unigrams are: nacional, touriga, douro
  * Most Correlated Bigrams are: blend touriga, tinta roriz, touriga nacional

==> Portuguese White:
  * Most Correlated Unigrams are: pires, fernão, arinto
  * Most Correlated Bigrams are: vinho verde, fernão pires, blend arinto

==> Red Blend:
  * Most Correlated Unigrams are: syrah, sangiovese, blend
  * Most Correlated Bigrams are: blend syrah, blend sangiovese, cabernet sauvignon

==> Rhône-style Red Blend:
  * Most Correlated Unigrams are: rhône, mourvèdre, grenache
  * Most Correlated Bigrams are: grenache syrah, grenache mourvèdre, rhône style

==> Riesling:
  * Most Correlated Unigrams are: tangerine, kabinett, riesling
  * Most Correlated Bigrams are: riesling palate, lime acidity, dry riesling

==> Rosé:
  * Most Correlated Unigrams are: pale, pink, rosé
  * Most Correlated Bigrams are: bodied rosé, style rosé, pink color

==> Sangiovese:
  * Most Correlated Unigrams are: underbrush, sangiovese, chianti
  * Most Correlated Bigrams are: fragrant blue, entirely sangiovese, chianti classico

==> Sauvignon Blanc:
  * Most Correlated Unigrams are: sb, gooseberry, blanc
  * Most Correlated Bigrams are: bell pepper, passion fruit, sauvignon blanc

==> Sparkling Blend:
  * Most Correlated Unigrams are: cava, bubbly, sparkler
  * Most Correlated Bigrams are: pinot nero, sparkling wine, chardonnay pinot
```

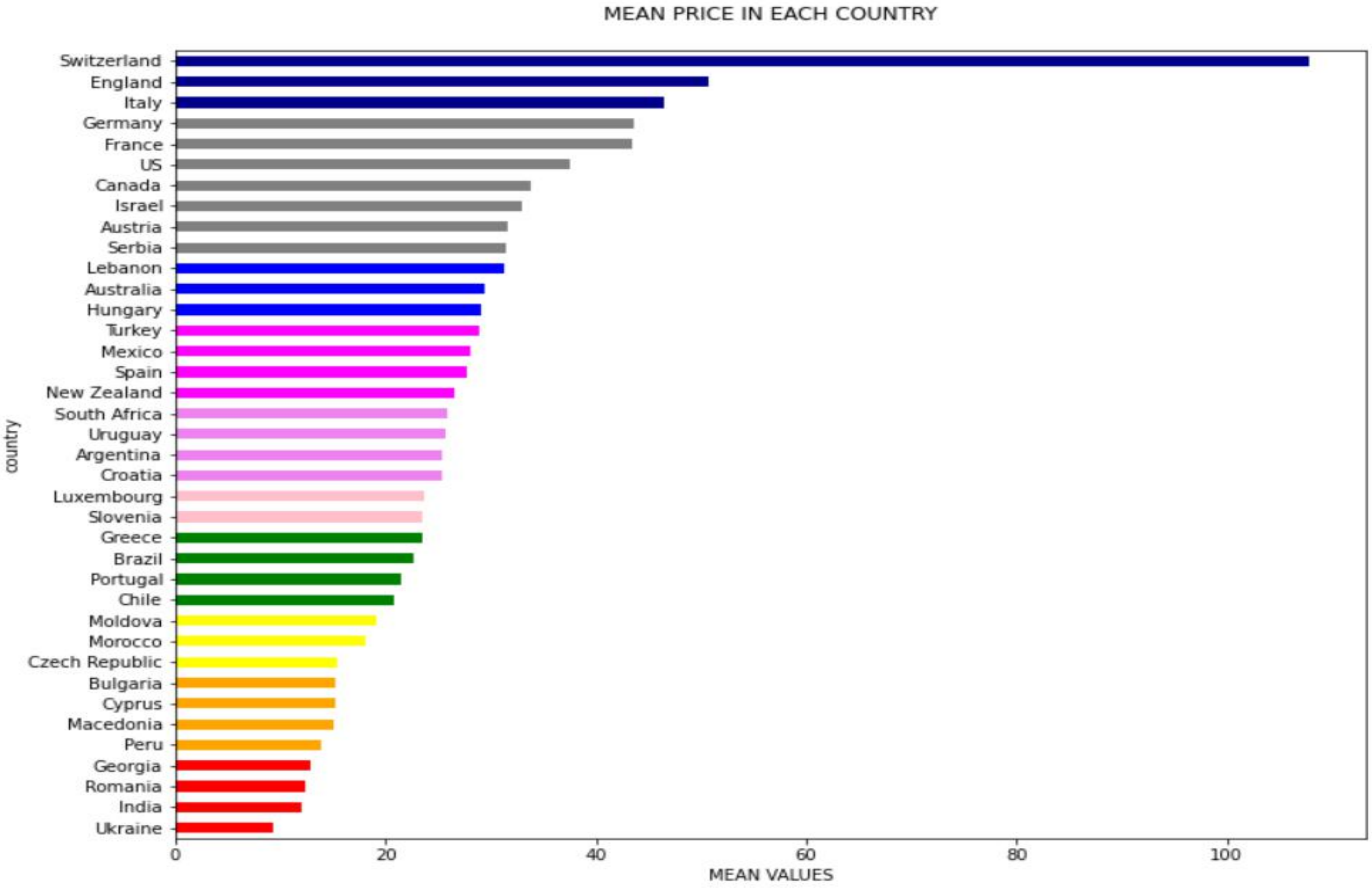


```
==> Syrah:
  * Most Correlated Unigrams are: meat, syrahs, syrah
  * Most Correlated Bigrams are: black pepper, grilled meat, 100 syrah

==> Tempranillo:
  * Most Correlated Unigrams are: crianza, ribera, rioja
  * Most Correlated Bigrams are: del duero, gran reserva, ribera del

==> White Blend:
  * Most Correlated Unigrams are: roussanne, viura, viognier
  * Most Correlated Bigrams are: chardonnay viognier, ribolla gialla, white blend

==> Zinfandel:
  * Most Correlated Unigrams are: briary, zinfandel, zin
  * Most Correlated Bigrams are: petite sirah, forest berries, high alcohol
```

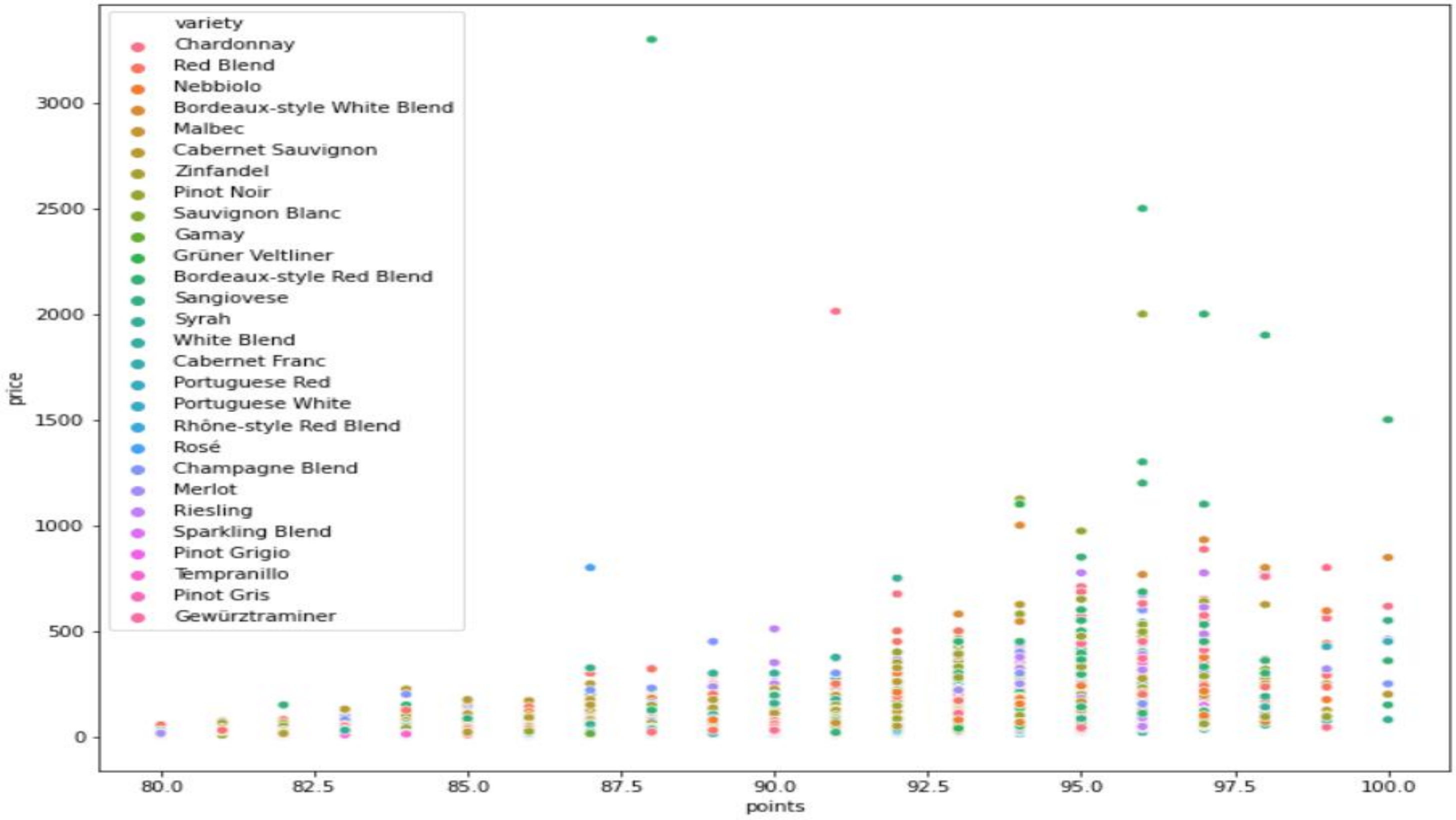
Plot of Mean price of each variety



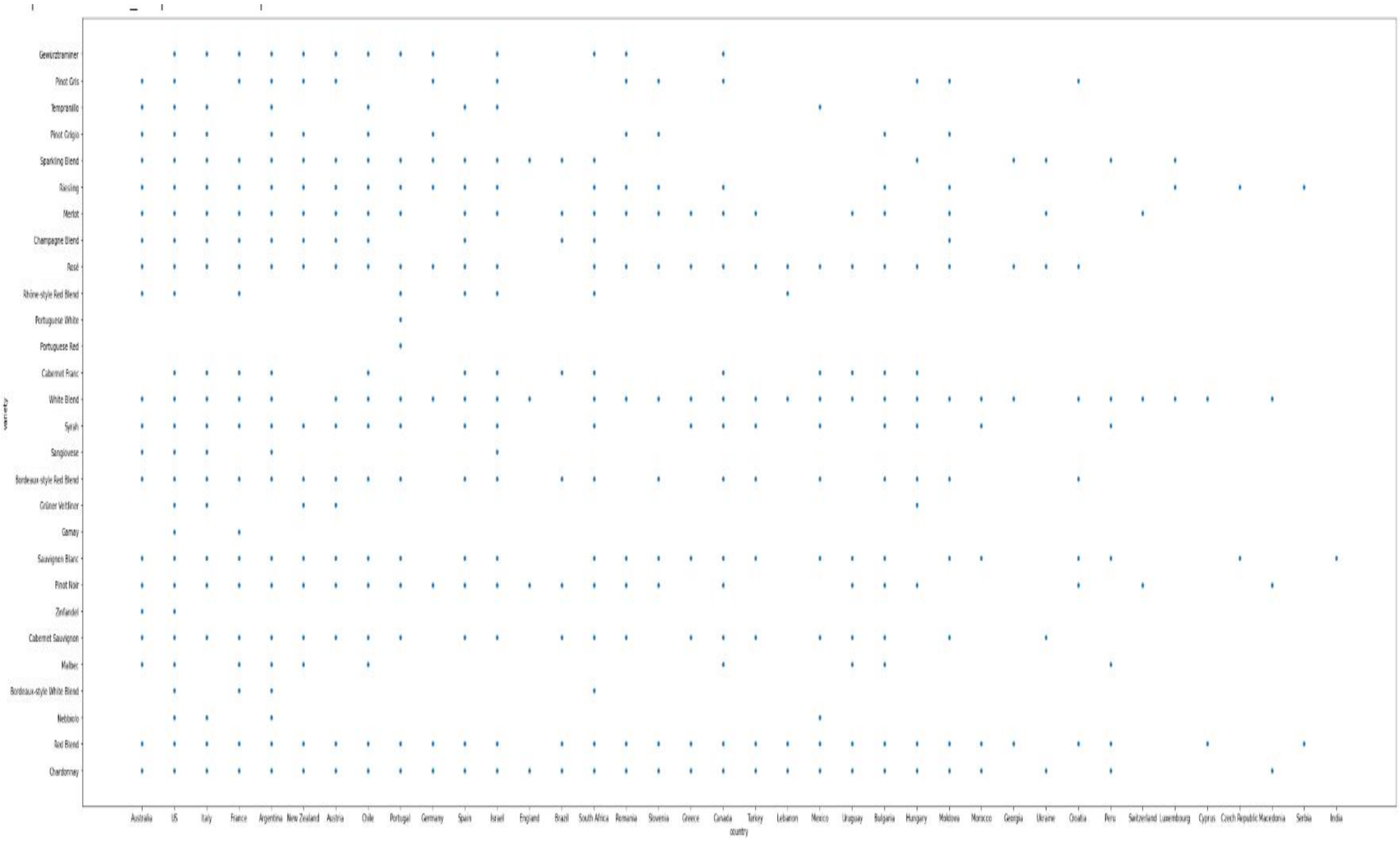
Plot of Mean Price of wine in each country



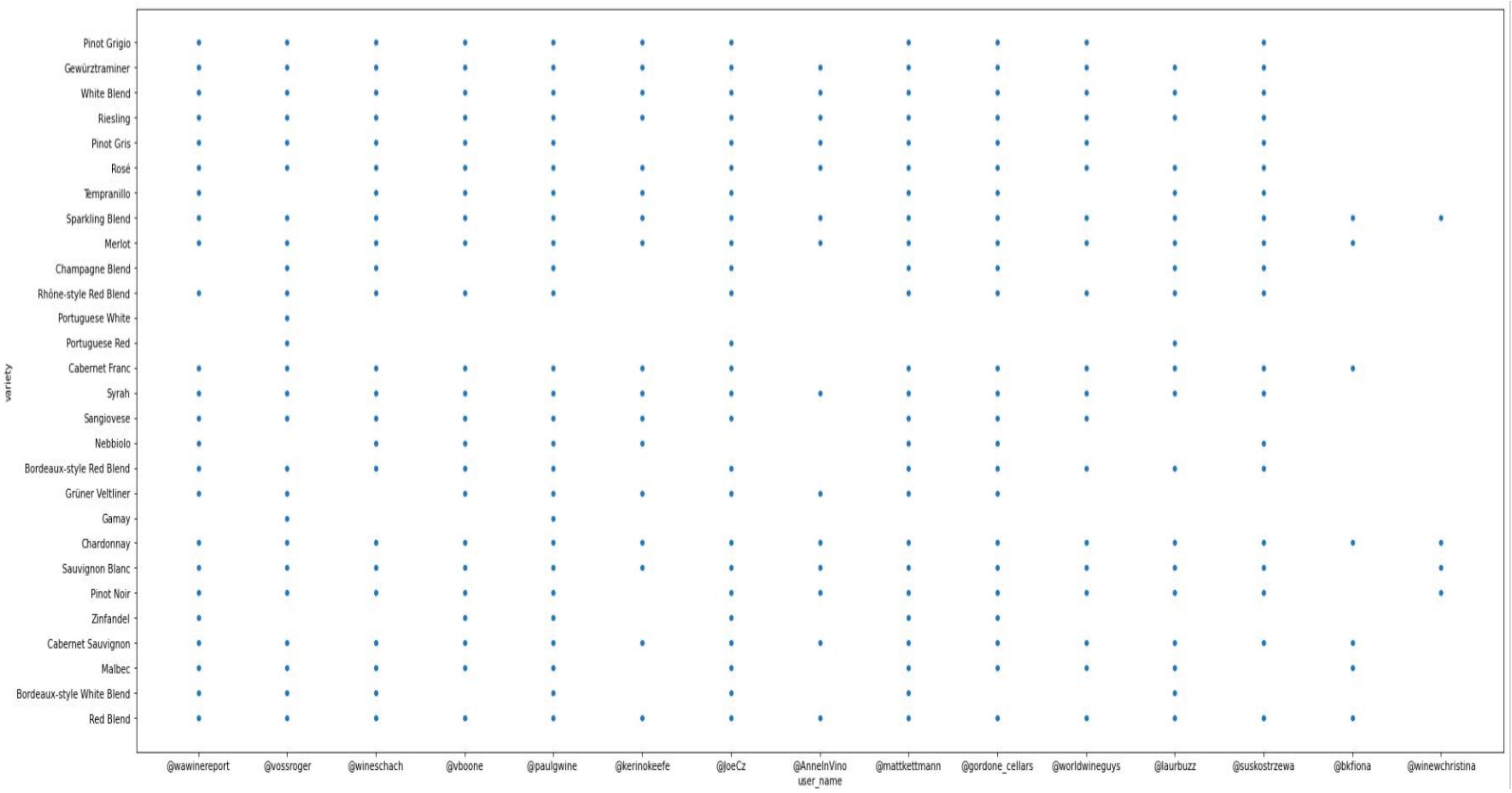
Below is a scatterplot for Cost Vs Price



Below is the plot for Countries vs Variety available



Below is the variety vs user_name



Actionable Insights from data are::

1. **Classification provides the probability of review being related to a variety of wine. We can also show which is the next most probable wine.**
2. **We can also recommend wines to the reviewer based upon the similarity between his taste and matching wines with similar qualities.**
3. **We can use the unigrams and Bigrams to describe the variety of wine.**
4. **The mean price of wine in Switzerland is more than double the mean price in any country.**
5. **Points received for a vine is almost independent of the Price.**