# Wine Variety Prediction

May 11, 2020

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [79]:
```
wine_train_data = pd.read_csv('./Knight ML Assignment/Data/train.csv')
```

In [18]:
```
wine_train_data.isnull().sum()
```

Out[18]:
```
user_name             19393
country                  35
review_title              0
review_description        0
designation           23647
points                    0
price                  5569
province                 35
region_1              12754
region_2              46708
winery                    0
variety                   0
dtype: int64
```

In [22]:
```
#Let's remove some attributes which might not be much relevant and have many null val
#Although region_1 also have many missing values but let's keep this feature for now.
wine_train_data = wine_train_data.drop(['user_name','region_2','designation'],axis=1)
```

In [26]:
```
wine_train_data.shape
```

Out[26]:
```
(82657, 9)
```

In [34]:
```
# Let's remove null values row in the data.
wine_train_data = wine_train_data.dropna()
wine_train_data.shape
```

Out[34]:
```
(65199, 9)
```

Around 17500 rows had been removed

In [3]:
```
# Ok, so let's try to implement Decision Tree Classifier for the problem
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
```

```
In [33]: features = wine_train_data.columns[:8]

In [67]: X = wine_train_data[features].copy()

In [61]: y = wine_train_data['variety'].copy()

In [116]: # Let's split the data in training and test set
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=15199, random_sta

In [267]: # Here the attributes value is not in numerical. So, we need to encode them first
          # using label encoder.
          from sklearn import preprocessing
          encoder = preprocessing.LabelEncoder()

In [42]: X_train.head()

Out[42]:       country                                          review_title  \
         46479  France   Charles Joguet 2010 Clos du Chêne Vert  (Chinon)
         42016      US  Larner 2009 Estate Grown Syrah (Santa Ynez Val...
         63308      US  Gypsy Dancer 2010 Tribute Pinot Noir (Willamet...
         21212      US  Bargetto 2004 La Vita Regan Vineyards Red (San...
         38707  France  Domaine Chiroulet 2013 Terroir Gascon Red (Côt...

                                          review_description  points  price  \
         46479  Along with the other top cuvées in the Charles...      92   50.0
         42016  Somms will have fun pairing this wine with jus...      92   38.0
         63308  Extra time in bottle has proved rewarding for ...      92   65.0
         21212  Minty, with easy wintergreen, cherry and sanda...      84   60.0
         38707  There is an attractive touch of smokiness to t...      84   14.0

                       province              region_1             winery
         46479      Loire Valley                Chinon     Charles Joguet
         42016        California     Santa Ynez Valley             Larner
         63308            Oregon     Willamette Valley       Gypsy Dancer
         21212        California  Santa Cruz Mountains           Bargetto
         38707  Southwest France     Côtes de Gascogne  Domaine Chiroulet

In [118]: # encoding the training X variable
          for feature in X_train.columns:
              X_train[feature] = encoder.fit_transform(X_train[feature])
          X_train.head(10)

/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  This is separate from the ipykernel package so we can avoid doing imports until
```

Out[118]:

| | country | review_title | review_description | points | price | province | region_1 | winery |
|---|---|---|---|---|---|---|---|---|
| 46479 | 3 | 8397 | 5382 | 12 | 46 | 24 | 179 | 1748 |
| 42016 | 6 | 26547 | 28059 | 12 | 34 | 8 | 740 | 6159 |
| 63308 | 6 | 21323 | 13099 | 12 | 61 | 39 | 939 | 5228 |
| 21212 | 6 | 2926 | 21545 | 4 | 56 | 8 | 736 | 598 |
| 38707 | 3 | 14479 | 32716 | 4 | 10 | 49 | 287 | 3749 |
| 51643 | 6 | 44000 | 43094 | 8 | 16 | 8 | 569 | 9730 |
| 44568 | 6 | 33644 | 41359 | 4 | 11 | 58 | 936 | 7633 |
| 71276 | 3 | 30654 | 32144 | 12 | 91 | 7 | 635 | 6964 |
| 9374 | 6 | 26655 | 40336 | 6 | 16 | 8 | 317 | 6193 |
| 62498 | 6 | 39119 | 45962 | 7 | 14 | 33 | 588 | 8769 |

In [119]:
```
# We also need to transform X_test in the encoder
for feature in X_test.columns:
    X_test[feature] = encoder.fit_transform(X_test[feature])
X_test.head(10)
```

```
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  This is separate from the ipykernel package so we can avoid doing imports until
```

Out[119]:

| | country | review_title | review_description | points | price | province |
|---|---|---|---|---|---|---|
| 14784 | 6 | 5469 | 4819 | 8 | 40 | 55 |
| 32528 | 5 | 14170 | 1973 | 4 | 10 | 9 |
| 27524 | 6 | 11730 | 11239 | 10 | 20 | 8 |
| 45721 | 6 | 6730 | 13389 | 8 | 11 | 8 |
| 7708 | 6 | 13585 | 8194 | 10 | 43 | 8 |
| 33547 | 3 | 3322 | 438 | 8 | 12 | 5 |
| 40267 | 3 | 4735 | 4950 | 12 | 141 | 43 |
| 33754 | 3 | 4926 | 2349 | 10 | 32 | 0 |
| 21798 | 6 | 6402 | 5897 | 7 | 8 | 8 |
| 57275 | 6 | 2017 | 5552 | 8 | 17 | 8 |

```
       region_1  winery
14784       168    2508
32528       111    5859
27524       533    4919
45721        92    2982
7708        242    5607
33547        74    1502
40267       211    2202
33754        12    2291
21798       456    2869
57275       617     790
```

```
In [110]: # Defining the classifier
          variety_predictor = DecisionTreeClassifier(max_leaf_nodes=10, random_state=0)
```

```
In [111]: # Training the model
          variety_predictor.fit(X_train,y_train)
```

```
Out[111]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                      max_features=None, max_leaf_nodes=10,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort=False, random_state=0,
                      splitter='best')
```

```
In [112]: predictions = variety_predictor.predict(X_test)
```

```
In [113]: # Let's check the accuracy score of Decision Tree classifier
          accuracy_score(y_true = y_test, y_pred = predictions)
```

```
Out[113]: 0.299296006316205
```

**That clearly doesn't worked out well enough.**

```
In [114]: # Let's now see how Gradient Boosting Classifier works out for this problem
          from sklearn.ensemble import GradientBoostingClassifier
```

```
In [115]: variety_predictor_gbc = GradientBoostingClassifier()
```

```
In [123]: variety_predictor_gbc.fit(X_train,y_train)
```

```
Out[123]: GradientBoostingClassifier(criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='deviance', max_depth=3,
                          max_features=None, max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=100,
                          n_iter_no_change=None, presort='auto', random_state=None,
                          subsample=1.0, tol=0.0001, validation_fraction=0.1,
                          verbose=0, warm_start=False)
```

```
In [124]: predictions = variety_predictor_gbc.predict(X_test)

In [125]: accuracy_score(y_true = y_test, y_pred = predictions)

Out[125]: 0.1623791038884137
```

**This accuracy is even less than Decision Tree**

```
In [126]: # Let's try Random Forest Classifier now
          from sklearn.ensemble import RandomForestClassifier

In [127]: variety_predictor_rfc = RandomForestClassifier()

In [128]: variety_predictor_rfc.fit(X_train,y_train)

/data/anaconda3/lib/python3.6/site-packages/sklearn/ensemble/forest.py:246: FutureWarning: The
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)


Out[128]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
                      oob_score=False, random_state=None, verbose=0,
                      warm_start=False)

In [131]: predictions = variety_predictor_rfc.predict(X_test)

In [132]: accuracy_score(y_true = y_test, y_pred = predictions)

Out[132]: 0.2978485426672807
```

It seems like there is not much impact of 'Region1' feature in the dataset. #### So, let's begin again with raw data and this time we'll see the impact of removing region1 and review description on results.

```
In [395]: wine_train_data = pd.read_csv('./Knight ML Assignment/Data/train.csv')

In [396]: wine_train_data = wine_train_data.drop(['user_name','region_2','designation','region_

In [403]: wine_train_data = wine_train_data.dropna()

In [150]: features = wine_train_data.columns[:6]

In [155]: X = wine_train_data[features].copy()
          y = wine_train_data['variety'].copy()

In [214]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=7057, random_stat
```

```
In [168]: # Encoding training set before modelling
          X_train['country'] = encoder.fit_transform(X_train['country'])
          X_train['review_title'] = encoder.fit_transform(X_train['review_title'])
          X_train['province'] = encoder.fit_transform(X_train['province'])
          X_train['winery'] = encoder.fit_transform(X_train['winery'])

          # Encoding test set
          X_test['country'] = encoder.fit_transform(X_test['country'])
          X_test['review_title'] = encoder.fit_transform(X_test['review_title'])
          X_test['province'] = encoder.fit_transform(X_test['province'])
          X_test['winery'] = encoder.fit_transform(X_test['winery'])
```

```
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html

/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  This is separate from the ipykernel package so we can avoid doing imports until
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  after removing the cwd from sys.path.
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  """
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html

/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:9: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  if __name__ == '__main__':
```

```
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  # Remove the CWD from sys.path while we load stuff.
/data/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:11: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  # This is added back by InteractiveShellApp.init_path()
```

```
In [169]: X_train.describe()
```

Out[169]:

| | country | review_title | points | price | province | winery |
|---|---|---|---|---|---|---|
| count | 70000.000000 | 70000.000000 | 70000.000000 | 70000.000000 | 70000.000000 | 70000.000000 |
| mean | 24.629643 | 32784.777686 | 88.510629 | 36.906886 | 133.318343 | 6628.140957 |
| std | 12.131560 | 18931.309144 | 3.104641 | 42.725820 | 109.957491 | 3662.103382 |
| min | 0.000000 | 0.000000 | 80.000000 | 4.000000 | 0.000000 | 0.000000 |
| 25% | 11.000000 | 16398.750000 | 86.000000 | 18.000000 | 39.000000 | 3786.000000 |
| 50% | 32.000000 | 32730.500000 | 88.000000 | 27.000000 | 62.000000 | 6832.000000 |
| 75% | 35.000000 | 49184.250000 | 91.000000 | 45.000000 | 208.000000 | 9832.000000 |
| max | 37.000000 | 65647.000000 | 100.000000 | 2500.000000 | 345.000000 | 12698.000000 |

```
In [170]: X_test.describe()
```

Out[170]:

| | country | review_title | points | price | province | winery |
|---|---|---|---|---|---|---|
| count | 7057.000000 | 7057.000000 | 7057.000000 | 7057.000000 | 7057.000000 | |
| mean | 20.235227 | 3500.929007 | 88.522460 | 37.127533 | 80.418025 | |
| std | 9.240040 | 2022.485871 | 3.090213 | 52.449262 | 62.222068 | |
| min | 0.000000 | 0.000000 | 80.000000 | 5.000000 | 0.000000 | |
| 25% | 10.000000 | 1746.000000 | 86.000000 | 18.000000 | 28.000000 | |
| 50% | 27.000000 | 3503.000000 | 88.000000 | 28.000000 | 39.000000 | |
| 75% | 28.000000 | 5255.000000 | 91.000000 | 45.000000 | 125.000000 | |
| max | 29.000000 | 6998.000000 | 100.000000 | 3300.000000 | 202.000000 | |

winery

```
count  7057.000000
mean   2096.146663
std    1189.820221
min       0.000000
25%    1098.000000
50%    2128.000000
75%    3145.000000
max    4129.000000
```

```
In [203]: variety_predictor.fit(X_train,y_train)
```

```
Out[203]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                     max_features=None, max_leaf_nodes=10,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=0,
                     splitter='best')
```

```
In [204]: predictions = variety_predictor.predict(X_test)
```

```
In [205]: accuracy_score(y_true = y_test, y_pred = predictions)
```

```
Out[205]: 0.14326200935241604
```

```
In [234]: wine_train_data['winery'].value_counts()
```

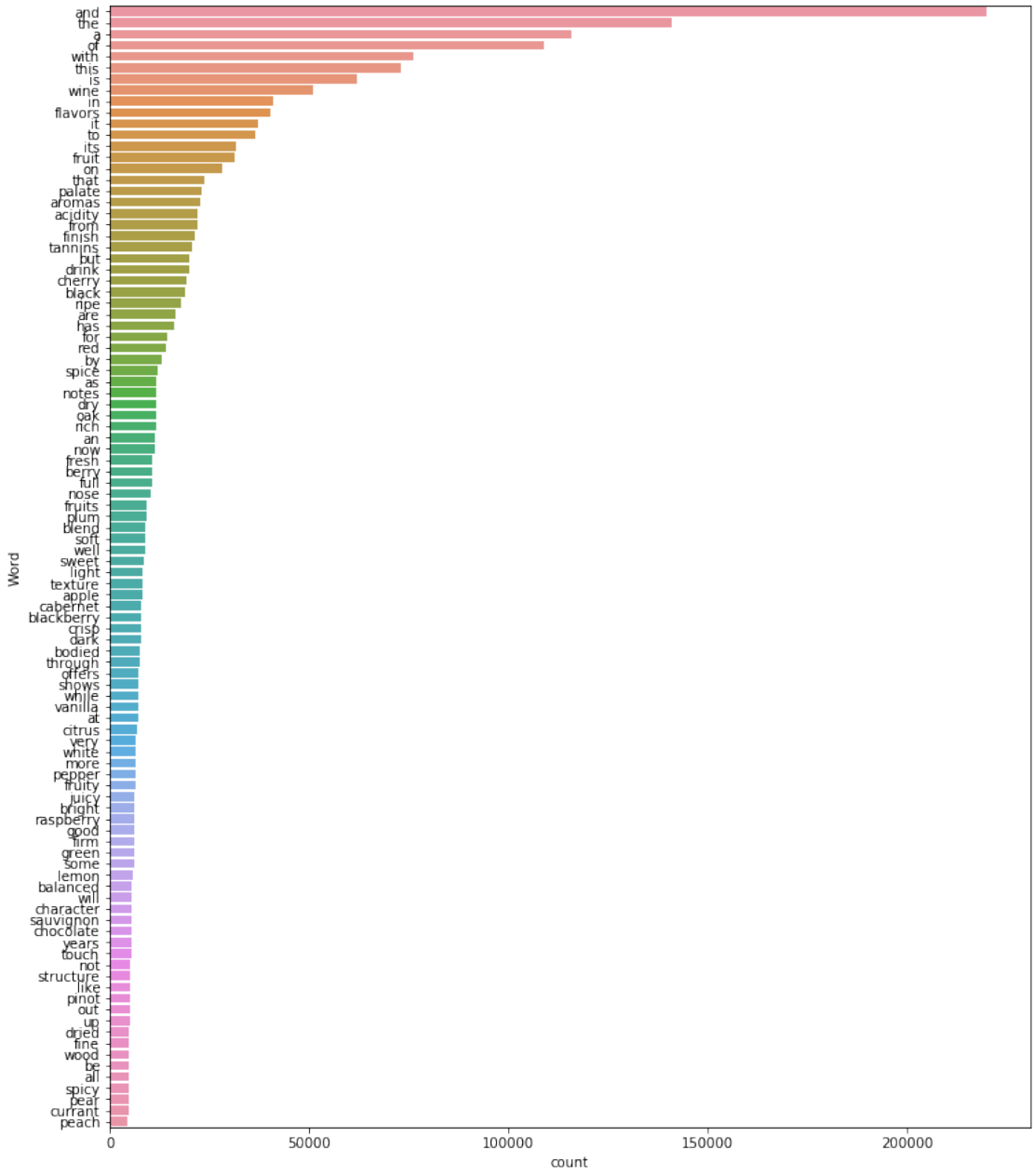
```
Out[234]: Testarossa                          175
          Williams Selyem                     165
          Chateau Ste. Michelle               162
          Louis Latour                        146
          Wines & Winemakers                  132
          DFJ Vinhos                          126
          Columbia Crest                      111
          Concha y Toro                       110
          Georges Duboeuf                     102
          Kendall-Jackson                     100
          Siduri                               99
          Lynmar                               98
          Gary Farrell                         98
          Montes                               90
          Fess Parker                          85
          Martin Ray                           85
          Jean-Luc and Paul Aegerter           82
          Iron Horse                           81
          Foxen                                80
          Chehalem                             80
          Chanson Père et Fils                 80
          Dutton-Goldfield                     79
          Dr. Loosen                           79
```

```
Trapiche                                78
V. Sattui                               77
Louis Jadot                             76
Kenwood                                 75
De Loach                                74
Rodney Strong                           72
Undurraga                               72
                                       ...
Toretti's                                1
Barbed Wire                              1
Terre Nere di Campigli - Vallone         1
Dilao                                    1
Morgan Simpson                           1
Pöckl                                    1
Blustone                                 1
Château Coquillas                        1
Ladera                                   1
Oregon Trails                            1
EdenVale                                 1
Renaissance                              1
Charter Oak                              1
Pieve Santo Stefano                      1
Château Piron                            1
Bryant Family                            1
Château Grand-Portail                    1
Domaine du Touch                         1
Regale Winery and Vineyards              1
Pyrenees                                 1
Áldás                                    1
Château de la Grenière                   1
Waypoint                                 1
Cummins Road                             1
Blue Cove                                1
Josef Brigl                              1
Domaine Saint Gregory                    1
Double Eagle                             1
Binz                                     1
Bodegas Campante                         1
Name: winery, Length: 13073, dtype: int64
```

**It is pretty much clear now that excluding review_description from data is not gonna lead to a good model. Thus we had to now implement language processing.**

```
In [4]: # So let's start with importing necessary libraries
        import nltk
        import re
        import seaborn as sns
        from sklearn.feature_extraction.text import TfidfVectorizer
```

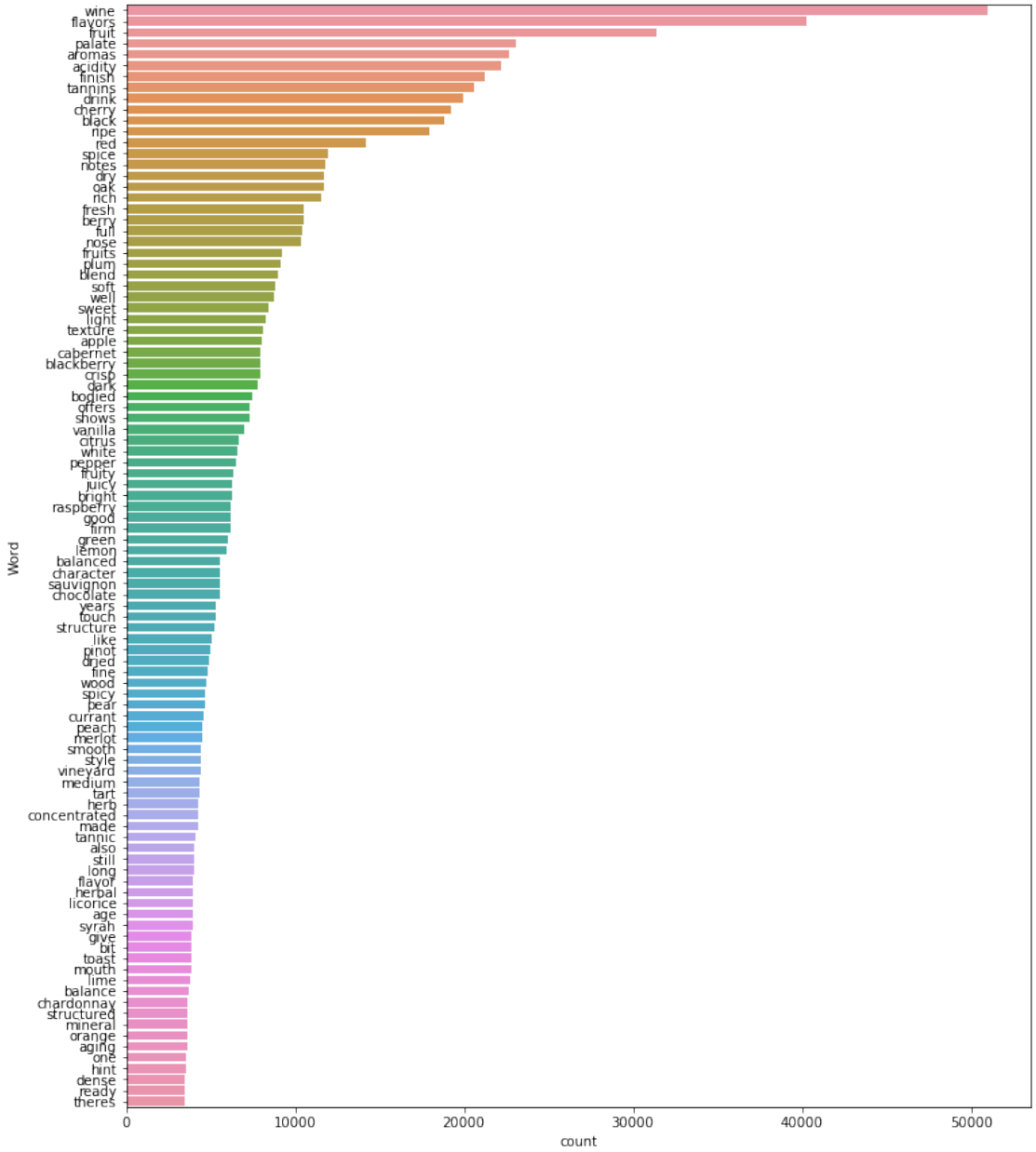
```
In [5]: wine_train_data = wine_train_data.drop(['user_name','designation','price','region_1','

In [6]: wine_train_data = wine_train_data.dropna()
        wine_train_data.shape

Out[6]: (82622, 7)

In [7]: # function for text cleaning like removing \',whitespaces and eveything other than alp
        def clean_text(text):
            # remove backslash-apostrophe
            text = re.sub("\'", "", text)
            # remove everything except alphabets
            text = re.sub("[^a-zA-Z]"," ",text)
            # remove whitespaces
            text = ' '.join(text.split())
            # convert text to lowercase
            text = text.lower()

            return text

In [9]: wine_train_data['review_description'] = wine_train_data['review_description'].apply(la

In [10]: # Now let's define a function to calculate frequency of words in text.
         def freq_words(x, terms = 30):
           all_words = ' '.join([text for text in x])
           all_words = all_words.split()
           fdist = nltk.FreqDist(all_words)
           words_df = pd.DataFrame({'word':list(fdist.keys()), 'count':list(fdist.values())})
           print(words_df.shape)

           # selecting top n most frequent words
           d = words_df.nlargest(columns="count", n = terms)

           # visualize words and frequencies
           plt.figure(figsize=(12,15))
           ax = sns.barplot(data=d, x= "count", y = "word")
           ax.set(ylabel = 'Word')
           plt.show()

In [11]: # Let's explore the most occured words in review_description
         freq_words(wine_train_data['review_description'], 100)

(26280, 2)
```




```
In [12]: # It is clearly visible that there are lots of stopwords in it. Let's work on removing
         from nltk.corpus import stopwords
         stop_words = set(stopwords.words('english'))

         # function to remove stopwords
         def remove_stopwords(text):
             no_stopword_text = [w for w in text.split() if not w in stop_words]
             return ' '.join(no_stopword_text)

In [13]: wine_train_data['review_description'] = wine_train_data['review_description'].apply(la
```

In [345]:
```
# Now again let's look at the top 100 words in the revised review_description_cleane
freq_words(wine_train_data['review_description'], 100)
```

```
(26149, 2)
```




**That's very nice. A large set of useless words has been smoothly removed.**

In [59]:
```
# Now, the next step is very important as this is the core of complete model. Now we'
# convert the text in features
tfidf_vectorizer = TfidfVectorizer(max_features=7500)
```

```
In [15]: y = wine_train_data['variety'].copy()

In [16]: # Let's split complete data in train and test data
         X_train, X_test, y_train, y_test = train_test_split(wine_train_data['review_descriptio

In [60]: X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
         X_test_tfidf = tfidf_vectorizer.transform(X_test)

In [18]: # Time to built the model
         from sklearn.linear_model import LogisticRegression

         # Binary Relevance
         from sklearn.multiclass import OneVsRestClassifier

         # Performance metric
         from sklearn.metrics import f1_score

In [40]: lr = LogisticRegression()
         clf = OneVsRestClassifier(lr)

In [61]: # fitting model on train data
         clf.fit(X_train_tfidf, y_train)

/data/anaconda3/lib/python3.6/site-packages/sklearn/linear_model/logistic.py:433: FutureWarning
  FutureWarning)


Out[61]: OneVsRestClassifier(estimator=LogisticRegression(C=1.0, class_weight=None, dual=False
                   intercept_scaling=1, max_iter=100, multi_class='warn',
                   n_jobs=None, penalty='l2', random_state=None, solver='warn',
                   tol=0.0001, verbose=0, warm_start=False),
                   n_jobs=None)

In [62]: # Let's predict the test set
         predictions = clf.predict(X_test_tfidf)

In [63]: predictions

Out[63]: array(['Zinfandel', 'Riesling', 'Portuguese White', ..., 'Zinfandel',
                'Grüner Veltliner', 'Merlot'], dtype='<U26')

In [64]: accuracy_score(y_true = y_test, y_pred = predictions)

Out[64]: 0.6859304084720121
```

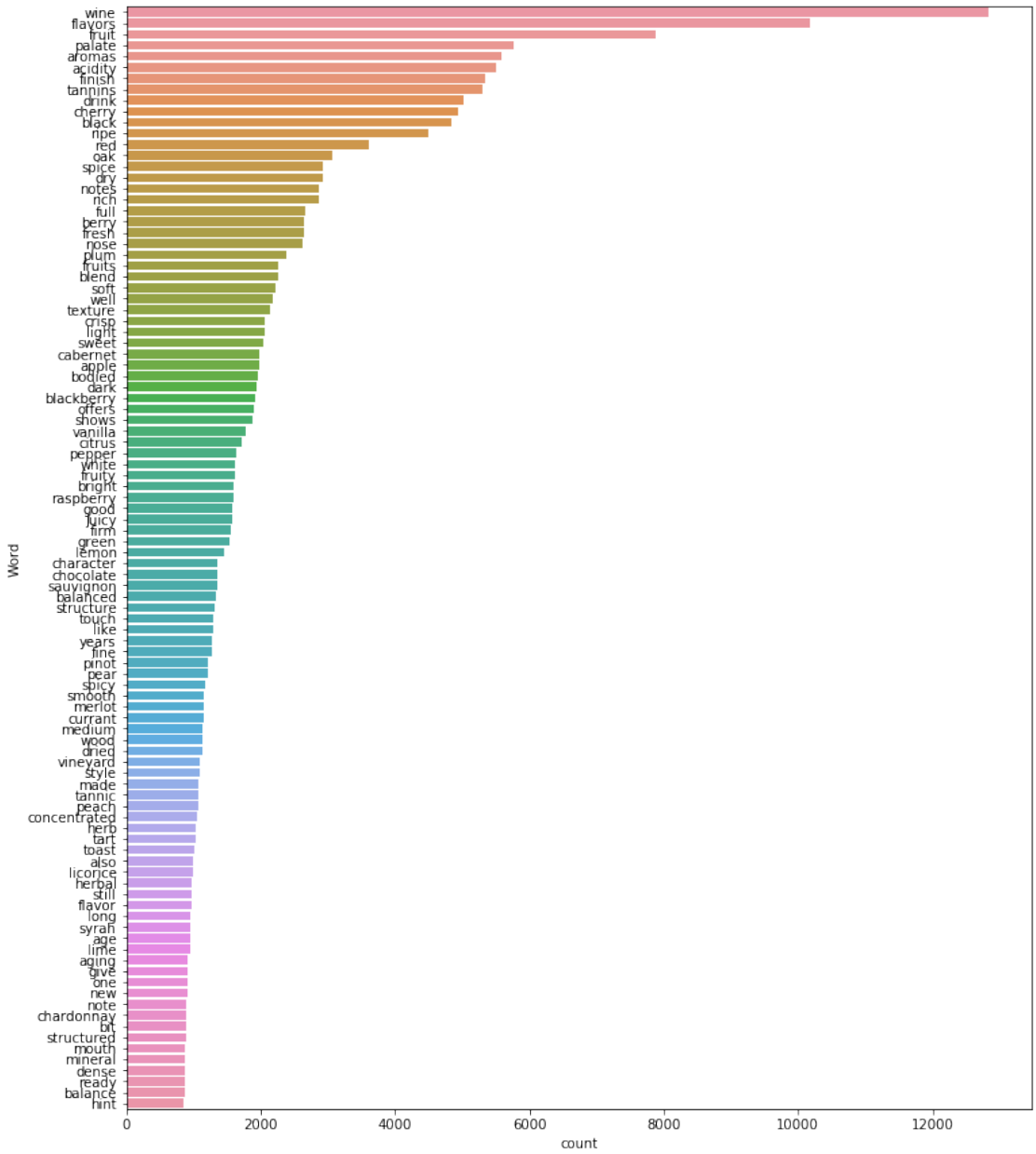
**Well that's a very significant rise in accuracy as compared to previous value, which was around 30%. Through text processing and analysis we've been able to increase accuracy by around 40%**
Now, is the time to predict variety in test data

```
In [65]: wine_test_data = pd.read_csv('./Knight ML Assignment/Data/test.csv')
```

In [69]:
```
# Before predicting, the review_description column has to be cleaned and prepared.
wine_test_data['review_description'] = wine_test_data['review_description'].apply(lamb
wine_test_data['review_description'] = wine_test_data['review_description'].apply(lamb
```

In [70]:
```
freq_words(wine_test_data['review_description'], 100)
```

```
(15418, 2)
```



**The most frequent words in test data are more or less similar to what was observed in train data.**

```
In [71]: # Let's prepare the variable that will be passed to the predictor.
         X_test_tfidf = tfidf_vectorizer.transform(wine_test_data['review_description'])

In [72]: predictions = clf.predict(X_test_tfidf)

In [76]: predictions

Out[76]: array(['Pinot Noir', 'Malbec', 'Chardonnay', ..., 'Cabernet Sauvignon',
                'Malbec', 'Cabernet Sauvignon'], dtype='<U26')

In [77]: # Let's assign these prediction as a 'variety' feature in test dataset.
         wine_test_data['variety'] = predictions
         wine_test_data

Out[77]:
```

| | user_name | country | \ |
|---|---|---|---|
| 0 | @paulgwineã | US | |
| 1 | @wineschach | Argentina | |
| 2 | @vboone | US | |
| 3 | @wineschach | Argentina | |
| 4 | @kerinokeefe | Italy | |
| 5 | NaN | US | |
| 6 | @vossroger | France | |
| 7 | @vboone | US | |
| 8 | @wineschach | Chile | |
| 9 | @kerinokeefe | Italy | |
| 10 | @paulgwineã | US | |
| 11 | @vossroger | France | |
| 12 | @vossroger | France | |
| 13 | @wawinereport | US | |
| 14 | @paulgwineã | US | |
| 15 | @vossroger | Austria | |
| 16 | @vboone | US | |
| 17 | @mattkettmann | US | |
| 18 | @paulgwineã | US | |
| 19 | @vossroger | France | |
| 20 | NaN | US | |
| 21 | @paulgwineã | US | |
| 22 | @wineschach | Spain | |
| 23 | @vossroger | Portugal | |
| 24 | NaN | US | |
| 25 | @laurbuzz | France | |
| 26 | @vboone | US | |
| 27 | @vossroger | France | |
| 28 | @vboone | US | |
| 29 | NaN | Italy | |
| ... | ... | ... | |

```
20635     @mattkettmann           US
20636               NaN           US
20637               NaN           US
20638        @vossroger       France
20639        @vossroger       France
20640       @AnneInVino      Austria
20641       @paulgwineă           US
20642     @mattkettmann           US
20643     @wawinereport           US
20644        @vossroger       France
20645        @vossroger       France
20646       @paulgwineă           US
20647     @mattkettmann           US
20648       @wineschach        Chile
20649        @vossroger     Portugal
20650      @kerinokeefe        Italy
20651            @JoeCz  New Zealand
20652       @wineschach        Chile
20653     @wawinereport           US
20654     @wawinereport           US
20655               NaN           US
20656      @kerinokeefe        Italy
20657               NaN           US
20658        @vossroger     Portugal
20659     @mattkettmann           US
20660  @gordone_cellars           US
20661         @laurbuzz       France
20662     @mattkettmann           US
20663       @wineschach        Spain
20664               NaN           US

                                          review_title  \
0      Boedecker Cellars 2011 Athena Pinot Noir (Will...
1      Mendoza Vineyards 2012 Gran Reserva by Richard...
2                  Prime 2013 Chardonnay (Coombsville)
3      Bodega Cuarto Dominio 2012 Chento Vineyard Sel...
4              SassodiSole 2012  Brunello di Montalcino
5      Mount Pleasant Winery NV Villagio Off Dry Whit...
6      Château Trolliet Lafite 2010 Red (Côtes de Ber...
7      Dry Creek Vineyard 2013 Cabernet Sauvignon (Dr...
8      Viña Casas Patronales 2010 Mixtura Red (Maule ...
9      Contadi Castaldi NV Rosé Sparkling (Franciacorta)
10     Spindrift Cellars 2011 Pinot Noir (Willamette ...
11             Eric Taillet NV Egali't Brut  (Champagne)
12                          Bougrier 2013  Rosé d'Anjou
13         Almquist 2013 Pinot Noir (Horse Heaven Hills)
14     Brittan Vineyards 2014 Gestalt Block Pinot Noi...
15     Bründlmayer 2009 Steinmassel Trockenbeerenausl...
```

```
16     Mounts 2014 Estate Grown Cabernet Sauvignon (D...
17                      HammerSky 2013 Rosé (Paso Robles)
18     Novelty Hill 2009 Stillwater Creek Vineyard Ca...
19              Olivier Leflaive 2006  Bâtard-Montrachet
20                 Schug 1999 Pinot Noir (Sonoma Valley)
21                    Lenné Estate 2014 LeNez Pinot Noir
22                   Cruz 2011 Sagra  (Ribera del Duero)
23     Wines & Winemakers 2014 Casa Ermelinda Freitas...
24           Butterfly Kiss 2010 Chardonnay (California)
25     Les Collines du Bourdic 2016 Éclat de Gris Ros...
26     Medlock Ames 2013 Heritage Merlot (Alexander V...
27                     Pierre Gruber 2015  Mâcon-Villages
28            Decoy 2015 Sauvignon Blanc (Sonoma County)
29                      Montevetrano 2008 Red (Campania)
...                                                   ...
20635              Opaque 2014 Darkness Red (Paso Robles)
20636  Domaine Chandon 2006 Unoaked Chardonnay (Sonom...
20637  Iron Horse 2002 Russian Cuvée Sparkling (Green...
20638  Château Ponzac 2014 Eternellement Malbec (Cahors)
20639        Joseph Mellot 2008 La Montarlet  (Sancerre)
20640  Rudi Pichler 2012 Wösendorfer Kollmütz Smaragd...
20641    Castle Rock 2010 Pinot Noir (Willamette Valley)
20642  Dierberg 2012 Drum Canyon Vineyard Estate Grow...
20643  Fielding Hills 2013 Tribute Estate Riverbend V...
20644  Duménil NV Rosé Vieilles Vignes Premier Cru Br...
20645          Château Jeanguillon 2015  Entre-Deux-Mers
20646  Ash Hollow 2007 Headless Red Red (Walla Walla ...
20647  La Montagne 2012 John Sebastiano Vineyard Pino...
20648  Errazuriz 2009 The Blend Limited Edition Red (...
20649  Herdade do Esporão 2014 Quatro Castas Red (Ale...
20650            Andriano 2015 Pinot Grigio (Alto Adige)
20651  Yealands 2016 Single Vineyard Sauvignon Blanc ...
20652  MontGras 2012 Intriga Maxima Cabernet Sauvigno...
20653     Almquist 2010 Garagiste Rosso Red (Washington)
20654  J. Bookwalter 2015 Suspense Conner Lee Vineyar...
20655              Huntington 2006 Zinfandel (California)
20656  Tenuta di Morzano 2010 Emilio  (Chianti Superi...
20657  Sheldrake Point 2014 Estate Bottled Chardonnay...
20658        Quinta do Cruzeiro 2012 White (Vinho Verde)
20659  Blackjack 2012 Black Cap Reserve Chardonnay (S...
20660  Yorkville Cellars 2013 Rennie Vineyard Caberne...
20661  Château Ribaute 2015 Senhal d'Aric Rosé (Corbi...
20662  Daou 2014 Reserve Cabernet Sauvignon (Paso Rob...
20663  Peñascal 2011 Tempranillo Rosé (Vino de la Tie...
20664  Langtry 2005 Tephra Ridge Vineyard Cabernet Sa...

                                      review_description  \
0      nicely differentiated companion stewart bottli...
```

```
1      charred smoky herbal aromas blackberry trend d...
2      slightly sour funky earth restrained white age...
3      concentrated midnight black malbec delivers ar...
4      earthy aromas suggesting grilled porcini leath...
5      timid nose squeaks notes chinese five spice wa...
6      structured wine berry fruits black currants fi...
7      remarkable value delivering terms varietal cha...
8      staunch herbal aromas tobacco olive rubber van...
9      blend chardonnay pinot nero copper colored win...
10     offers light berry melon flavors tart slightly...
11     equal proportions three main champagne grapes ...
12     cut sulfur wine whose bitterness hides potenti...
13     varietally designated pinot noirs extreme rari...
14     black fruited offering winerys signature stamp...
15     always star sweet wines tba riesling preserves...
16     seductive wine concentrated full bodied entire...
17     theres respectably complex nose pink wine lemo...
18     novelty hills estate grown cabernet sauvignon ...
19     bright star olivier leflaives firmament wondro...
20     workhorse wine pinot master pretty darned good...
21     fruit forward entry level release nonetheless ...
22     dominant aromas char rubber give way chocolaty...
23     oak aged wine intense rich packed black fruits...
24     simple chardonnay tastes sweet white sugared p...
25     soft white cherry strawberry hit pink floral a...
26     made cabernet sauvignon malbec petit verdot es...
27     feels fittingly warm southern burgundy wine ri...
28     baked pineapple mango rise fore nutty slightly...
29     cabernet sauvignon merlot aglianico blend sout...
...                                                  ...
20635  kitchen sink blend zinfandel syrah cabernet sa...
20636  oak influences chard acidic side bright clean ...
20637  like sparkling wines little sweet pinot noir c...
20638  top wine trilogy made ch teau ponzac designed ...
20639  strong currant aromas soft texture hints caram...
20640  subdued pear fruit tightly coiled inside rich ...
20641  ubiquitous brand fine job willamette valley pi...
20642  savory nose calls mind mushrooms saut ed itali...
20643  aromas vanilla coconut barrel spice dark fruit...
20644  soft wine shows relatively high dosage giving ...
20645  lively crisp wine creamy texture perfumed char...
20646  cabernet sauvignon estate vineyards quite sour...
20647  slightly sour orange cranberry aromas meet tou...
20648  unusual blend mourv dre syrah petit verdot gre...
20649  concentrated blend local grapes syrah cabernet...
20650  opens aromas chamomile flower ripe bartlett pe...
20651  considerable step quality basic yealands sauvi...
20652  currant aromas blend notes tomato herbs brick ...
```

```
20653  wine blend sangiovese barbera cranberry raspbe...
20654  aromas herb freshly sliced green pepper cherry...
20655  grapes statewide blend came lodi paso robles c...
20656  aromas menthol pressed violet espresso whiff c...
20657  exuberant aromas fresh pineapple mango guava i...
20658  typical touch fizz wine fruity light bright dr...
20659  inviting scents melting butter hot dinner roll...
20660  clearly focused fruit driven wine bonanza dark...
20661  herbal tones bay rosemary upfront nose copper ...
20662  mocha cream pencil shaving dried herb aromas s...
20663  loud citrus berry aromas precede overloaded pa...
20664  ripe fruit firm tannins mountain cab shows att...
```

| | designation | points | price | \ |
|---|---|---|---|---|
| 0 | Athena | 88 | 35.0 | |
| 1 | Gran Reserva by Richard Bonvin | 90 | 60.0 | |
| 2 | NaN | 87 | 38.0 | |
| 3 | Chento Vineyard Selection | 91 | 20.0 | |
| 4 | NaN | 90 | 49.0 | |
| 5 | Villagio Off Dry | 82 | 9.0 | |
| 6 | NaN | 89 | NaN | |
| 7 | NaN | 88 | 26.0 | |
| 8 | Mixtura | 88 | 25.0 | |
| 9 | Rosé | 88 | 24.0 | |
| 10 | NaN | 87 | 25.0 | |
| 11 | Egali't Brut | 88 | 35.0 | |
| 12 | NaN | 82 | 13.0 | |
| 13 | NaN | 84 | 35.0 | |
| 14 | Gestalt Block | 93 | 60.0 | |
| 15 | Steinmassel Trockenbeerenauslese | 95 | NaN | |
| 16 | Estate Grown | 88 | 45.0 | |
| 17 | NaN | 88 | 28.0 | |
| 18 | Stillwater Creek Vineyard | 89 | 30.0 | |
| 19 | NaN | 95 | 310.0 | |
| 20 | NaN | 90 | 15.0 | |
| 21 | LeNez | 91 | 30.0 | |
| 22 | Sagra | 82 | 11.0 | |
| 23 | Casa Ermelinda Freitas Monte da Baía Reserva | 91 | 18.0 | |
| 24 | NaN | 82 | 12.0 | |
| 25 | Éclat de Gris | 87 | 11.0 | |
| 26 | Heritage | 87 | 50.0 | |
| 27 | NaN | 87 | 19.0 | |
| 28 | NaN | 86 | 20.0 | |
| 29 | NaN | 92 | NaN | |
| ... | ... | ... | ... | |
| 20635 | Darkness | 90 | 30.0 | |
| 20636 | Unoaked | 85 | 20.0 | |
| 20637 | Russian Cuvée | 88 | 31.0 | |

| | | | |
|---|---|---|---|
| 20638 | Eternellement | 93 | 44.0 |
| 20639 | La Montarlet | 85 | 20.0 |
| 20640 | Wösendorfer Kollmütz Smaragd | 93 | 75.0 |
| 20641 | NaN | 87 | 14.0 |
| 20642 | Drum Canyon Vineyard Estate Grown | 91 | 44.0 |
| 20643 | Tribute Estate Riverbend Vineyard | 90 | 34.0 |
| 20644 | Rosé Vieilles Vignes Premier Cru Brut | 86 | 40.0 |
| 20645 | NaN | 86 | 11.0 |
| 20646 | Headless Red | 86 | 18.0 |
| 20647 | John Sebastiano Vineyard | 92 | 50.0 |
| 20648 | The Blend Limited Edition | 90 | 40.0 |
| 20649 | Quatro Castas | 89 | 22.0 |
| 20650 | NaN | 89 | 20.0 |
| 20651 | Single Vineyard | 90 | 25.0 |
| 20652 | Intriga Maxima | 90 | 90.0 |
| 20653 | Garagiste Rosso | 90 | 20.0 |
| 20654 | Suspense Conner Lee Vineyard | 90 | 60.0 |
| 20655 | NaN | 87 | 15.0 |
| 20656 | Emilio | 87 | 22.0 |
| 20657 | Estate Bottled | 88 | 14.0 |
| 20658 | NaN | 85 | 17.0 |
| 20659 | Black Cap Reserve | 89 | 48.0 |
| 20660 | Rennie Vineyard | 91 | 34.0 |
| 20661 | Senhal d'Aric | 84 | 20.0 |
| 20662 | Reserve | 94 | 85.0 |
| 20663 | Tempranillo | 80 | 9.0 |
| 20664 | Tephra Ridge Vineyard | 87 | 40.0 |

| | province | region_1 | \ |
|---|---|---|---|
| 0 | Oregon | Willamette Valley | |
| 1 | Mendoza Province | Mendoza | |
| 2 | California | Coombsville | |
| 3 | Mendoza Province | Mendoza | |
| 4 | Tuscany | Brunello di Montalcino | |
| 5 | Missouri | Missouri | |
| 6 | Southwest France | Côtes de Bergerac | |
| 7 | California | Dry Creek Valley | |
| 8 | Maule Valley | NaN | |
| 9 | Lombardy | Franciacorta | |
| 10 | Oregon | Willamette Valley | |
| 11 | Champagne | Champagne | |
| 12 | Loire Valley | Rosé d'Anjou | |
| 13 | Washington | Horse Heaven Hills | |
| 14 | Oregon | McMinnville | |
| 15 | Niederösterreich | NaN | |
| 16 | California | Dry Creek Valley | |
| 17 | California | Paso Robles | |
| 18 | Washington | Columbia Valley (WA) | |

| | | |
|---|---|---|
| 19 | Burgundy | Bâtard-Montrachet |
| 20 | California | Sonoma Valley |
| 21 | Oregon | Willamette Valley |
| 22 | Northern Spain | Ribera del Duero |
| 23 | Península de Setúbal | NaN |
| 24 | California | California |
| 25 | Languedoc-Roussillon | Pays d'Oc |
| 26 | California | Alexander Valley |
| 27 | Burgundy | Mâcon-Villages |
| 28 | California | Sonoma County |
| 29 | Southern Italy | Campania |
| ... | ... | ... |
| 20635 | California | Paso Robles |
| 20636 | California | Sonoma County |
| 20637 | California | Green Valley |
| 20638 | Southwest France | Cahors |
| 20639 | Loire Valley | Sancerre |
| 20640 | Wachau | NaN |
| 20641 | Oregon | Willamette Valley |
| 20642 | California | Sta. Rita Hills |
| 20643 | Washington | Wahluke Slope |
| 20644 | Champagne | Champagne |
| 20645 | Bordeaux | Entre-Deux-Mers |
| 20646 | Washington | Walla Walla Valley (WA) |
| 20647 | California | Sta. Rita Hills |
| 20648 | Aconcagua Valley | NaN |
| 20649 | Alentejano | NaN |
| 20650 | Northeastern Italy | Alto Adige |
| 20651 | Awatere Valley | NaN |
| 20652 | Maipo Valley | NaN |
| 20653 | Washington | Washington |
| 20654 | Washington | Columbia Valley (WA) |
| 20655 | California | California |
| 20656 | Tuscany | Chianti Superiore |
| 20657 | New York | Finger Lakes |
| 20658 | Vinho Verde | NaN |
| 20659 | California | Santa Barbara County |
| 20660 | California | Yorkville Highlands |
| 20661 | Languedoc-Roussillon | Corbières |
| 20662 | California | Paso Robles |
| 20663 | Northern Spain | Vino de la Tierra de Castilla y León |
| 20664 | California | Lake County |

| | region_2 | winery | variety |
|---|---|---|---|
| 0 | Willamette Valley | Boedecker Cellars | Pinot Noir |
| 1 | NaN | Mendoza Vineyards | Malbec |
| 2 | Napa | Prime | Chardonnay |
| 3 | NaN | Bodega Cuarto Dominio | Malbec |

| | | | |
|---|---|---|---|
| 4 | NaN | SassodiSole | Sangiovese |
| 5 | NaN | Mount Pleasant Winery | Chardonnay |
| 6 | NaN | Château Trolliet Lafite | Bordeaux-style Red Blend |
| 7 | Sonoma | Dry Creek Vineyard | Cabernet Sauvignon |
| 8 | NaN | Viña Casas Patronales | Red Blend |
| 9 | NaN | Contadi Castaldi | Sparkling Blend |
| 10 | Willamette Valley | Spindrift Cellars | Sauvignon Blanc |
| 11 | NaN | Eric Taillet | Champagne Blend |
| 12 | NaN | Bougrier | Pinot Noir |
| 13 | Columbia Valley | Almquist | Pinot Noir |
| 14 | Willamette Valley | Brittan Vineyards | Cabernet Sauvignon |
| 15 | NaN | Bründlmayer | Riesling |
| 16 | Sonoma | Mounts | Cabernet Sauvignon |
| 17 | Central Coast | HammerSky | Pinot Noir |
| 18 | Columbia Valley | Novelty Hill | Cabernet Sauvignon |
| 19 | NaN | Olivier Leflaive | Chardonnay |
| 20 | Sonoma | Schug | Pinot Noir |
| 21 | NaN | Lenné Estate | Pinot Noir |
| 22 | NaN | Cruz | Cabernet Sauvignon |
| 23 | NaN | Wines & Winemakers | Portuguese Red |
| 24 | California Other | Butterfly Kiss | Chardonnay |
| 25 | NaN | Les Collines du Bourdic | Rosé |
| 26 | Sonoma | Medlock Ames | Cabernet Sauvignon |
| 27 | NaN | Pierre Gruber | Chardonnay |
| 28 | Sonoma | Decoy | Sauvignon Blanc |
| 29 | NaN | Montevetrano | Red Blend |
| ... | ... | ... | ... |
| 20635 | Central Coast | Opaque | Red Blend |
| 20636 | Sonoma | Domaine Chandon | Chardonnay |
| 20637 | Sonoma | Iron Horse | Sparkling Blend |
| 20638 | NaN | Château Ponzac | Bordeaux-style Red Blend |
| 20639 | NaN | Joseph Mellot | Chardonnay |
| 20640 | NaN | Rudi Pichler | Chardonnay |
| 20641 | Willamette Valley | Castle Rock | Pinot Noir |
| 20642 | Central Coast | Dierberg | Pinot Noir |
| 20643 | Columbia Valley | Fielding Hills | Syrah |
| 20644 | NaN | Duménil | Champagne Blend |
| 20645 | NaN | Château Jeanguillon | Portuguese White |
| 20646 | Columbia Valley | Ash Hollow | Cabernet Sauvignon |
| 20647 | Central Coast | La Montagne | Pinot Noir |
| 20648 | NaN | Errazuriz | Red Blend |
| 20649 | NaN | Herdade do Esporão | Red Blend |
| 20650 | NaN | Andriano | Pinot Grigio |
| 20651 | NaN | Yealands | Sauvignon Blanc |
| 20652 | NaN | MontGras | Cabernet Sauvignon |
| 20653 | Washington Other | Almquist | Red Blend |
| 20654 | Columbia Valley | J. Bookwalter | Cabernet Sauvignon |
| 20655 | California Other | Huntington | Zinfandel |

```
20656                NaN        Tenuta di Morzano                 Nebbiolo
20657       Finger Lakes          Sheldrake Point                Chardonnay
20658                NaN      Quinta do Cruzeiro           Sparkling Blend
20659      Central Coast                Blackjack                Chardonnay
20660        North Coast       Yorkville Cellars                Pinot Noir
20661                NaN         Château Ribaute                      Rosé
20662      Central Coast                    Daou       Cabernet Sauvignon
20663                NaN                Peñascal                    Malbec
20664                NaN                 Langtry       Cabernet Sauvignon

[20665 rows x 12 columns]
```

```
In [78]: # Finally, let's export the above dataframe to a csv file.
         wine_test_data.to_csv('test_predicted')
```

### 0.0.1 Since, the predictor part is now over. Let's try to extract some actionable insights from the training data

```
In [103]: wine_train_data['variety'].value_counts()
```

```
Out[103]: Pinot Noir                  10587
          Chardonnay                   9403
          Cabernet Sauvignon           7552
          Red Blend                    7166
          Bordeaux-style Red Blend     5497
          Riesling                     4148
          Sauvignon Blanc              4011
          Syrah                        3316
          Rosé                         2831
          Merlot                       2471
          Nebbiolo                     2242
          Zinfandel                    2209
          Sangiovese                   2165
          Malbec                       2119
          Portuguese Red               1969
          White Blend                  1896
          Sparkling Blend              1739
          Tempranillo                  1448
          Rhône-style Red Blend        1182
          Pinot Gris                   1148
          Champagne Blend              1133
          Cabernet Franc               1095
          Grüner Veltliner             1055
          Portuguese White              896
          Pinot Grigio                  873
          Bordeaux-style White Blend    850
          Gewürztraminer                840
          Gamay                         816
          Name: variety, dtype: int64
```

In [105]:
```
# Let's look at the average points for different varieties
varieties = list(wine_train_data['variety'].value_counts().index)
mean_points = []
for variety in varieties:
    mean_points.append(wine_train_data[wine_train_data['variety']==variety]['points']
mean_points
```

Out[105]:
```
[89.4172097855861,
 88.32564075295119,
 88.58130296610169,
 88.40720066982975,
 89.1026014189558,
 89.46528447444551,
 87.4619795562204,
 89.27834740651387,
 86.82232426704344,
 87.22622420072845,
 90.28367528991971,
 87.8492530556813,
 88.5325635103926,
 87.99056158565361,
 88.78059928897918,
 87.38132911392405,
 88.06210465784933,
 87.53729281767956,
 89.17174280879864,
 88.49912891986062,
 89.72374227714033,
 88.1917808219178,
 90.05118483412322,
 87.00558035714286,
 86.19587628865979,
 88.69294117647058,
 88.57857142857142,
 88.05882352941177]
```

In [115]:
```
# Similarly let's look at the average price for different varieties
mean_price = []
for variety in varieties:
    mean_price.append(wine_train_data[wine_train_data['variety']==variety]['price'].m
mean_price
```

Out[115]:
```
[47.63177496814662,
 34.71859126312225,
 47.84141616566466,
 36.11030927835051,
 47.988673902784335,
 32.431549861843756,
```

```
 20.316688227684345,
 39.127522935779815,
 18.56988416988417,
 29.633046330463305,
 66.00053390282969,
 29.356009070294785,
 45.412294647588766,
 30.28447444551591,
 24.797842135150482,
 23.341505131128848,
 30.058572300183037,
 30.812894183601962,
 35.344552701505755,
 22.91992720655141,
 71.59301130524152,
 34.93270142180095,
 28.251116071428573,
 15.348226018396847,
 16.394927536231883,
 34.422867513611614,
 25.847029077117572,
 20.762406015037595]
```

```
In [128]: countries = wine_train_data['country'].value_counts().index
          countries = countries[:11]
          countries
```

```
Out[128]: Index(['US', 'France', 'Italy', 'Portugal', 'Chile', 'Spain', 'Argentina',
                 'Austria', 'Germany', 'New Zealand', 'Australia'],
                dtype='object')
```

```
In [129]: country_mean_price = []
          for country in countries:
              country_mean_price.append(wine_train_data[wine_train_data['country']==country]['
          country_mean_price
```

```
Out[129]: [37.46294849763058,
           43.37165354330708,
           46.57380982447279,
           21.46588486140725,
           20.84254723582925,
           27.773884175580967,
           25.398896334253056,
           31.662900188323917,
           43.59173126614987,
           26.504587155963304,
           29.482107355864812]
```

```
In [136]: wine_train_data[wine_train_data['winery']=='Testarossa']
```

Out[136]:

| | user_name | country | \ |
|---|---|---|---|
| 828 | @mattkettmann | US | |
| 1306 | NaN | US | |
| 1833 | NaN | US | |
| 1976 | NaN | US | |
| 2060 | NaN | US | |
| 2110 | @mattkettmann | US | |
| 2419 | NaN | US | |
| 3063 | @mattkettmann | US | |
| 3544 | NaN | US | |
| 3586 | @mattkettmann | US | |
| 4214 | NaN | US | |
| 4494 | @mattkettmann | US | |
| 5777 | @mattkettmann | US | |
| 5857 | @mattkettmann | US | |
| 6764 | @mattkettmann | US | |
| 6841 | NaN | US | |
| 7664 | @mattkettmann | US | |
| 8407 | NaN | US | |
| 9107 | @mattkettmann | US | |
| 9276 | NaN | US | |
| 9370 | @mattkettmann | US | |
| 9474 | @mattkettmann | US | |
| 9595 | NaN | US | |
| 10096 | NaN | US | |
| 11855 | NaN | US | |
| 11960 | @vboone | US | |
| 11964 | NaN | US | |
| 12123 | @mattkettmann | US | |
| 13403 | @mattkettmann | US | |
| 14296 | NaN | US | |
| ... | ... | ... | |
| 71164 | @mattkettmann | US | |
| 71807 | @mattkettmann | US | |
| 71841 | NaN | US | |
| 71961 | @mattkettmann | US | |
| 72101 | NaN | US | |
| 72163 | @mattkettmann | US | |
| 72168 | NaN | US | |
| 72267 | @mattkettmann | US | |
| 72316 | NaN | US | |
| 73027 | NaN | US | |
| 73480 | NaN | US | |
| 73585 | @mattkettmann | US | |
| 73700 | @mattkettmann | US | |
| 73996 | NaN | US | |
| 74150 | @mattkettmann | US | |
| 74207 | NaN | US | |

```
74248            NaN      US
74643  @mattkettmann      US
74833  @mattkettmann      US
74879  @mattkettmann      US
75715  @mattkettmann      US
76685            NaN      US
77241  @mattkettmann      US
77915            NaN      US
78585            NaN      US
78639            NaN      US
80224  @mattkettmann      US
80300            NaN      US
80822  @mattkettmann      US
82074            NaN      US

                                                 review_title  \
828     Testarossa 2014 Niclaire Pinot Noir (California)
1306   Testarossa 2007 Garys' Vineyard Pinot Noir (Sa...
1833    Testarossa 2007 Niclaire Pinot Noir (California)
1976   Testarossa 2001 Sleepy Hollow Vineyard Pinot N...
2060   Testarossa 2005 Garys' Vineyard Syrah (Santa L...
2110   Testarossa 2015 Cuvée 107 Pinot Noir (Santa Lu...
2419    Testarossa 2011 Niclaire Pinot Noir (California)
3063   Testarossa 2013 Pisoni Vineyard Pinot Noir (Sa...
3544   Testarossa 2009 Sleepy Hollow Vineyard Pinot N...
3586   Testarossa 2014 Doctor's Vineyard Pinot Noir (...
4214   Testarossa 2010 Cuvée 107 Pinot Noir (Santa Ba...
4494   Testarossa 2012 Brosseau Vineyard Pinot Noir (...
5777   Testarossa 2012 Rincon Vineyard Chardonnay (Ar...
5857   Testarossa 2015 Fogstone Vineyard Pinot Noir (...
6764   Testarossa 2014 Rosella's Vineyard Chardonnay ...
6841   Testarossa 2009 Garys' Vineyard Syrah (Santa L...
7664   Testarossa 2014 Brosseau Vineyard Chardonnay (...
8407   Testarossa 2010 Cuvée 107 Pinot Noir (Santa Ba...
9107   Testarossa 2014 Rincon Vineyard Chardonnay (Ar...
9276   Testarossa 2011 Bien Nacido Vineyard Chardonna...
9370   Testarossa 2014 Cuvée Los Gatos Rob's Red (Cen...
9474   Testarossa 2015 La Encantada Vineyard Pinot No...
9595   Testarossa 2009 Pisoni Vineyard Pinot Noir (Sa...
10096   Testarossa 2008 Niclaire Pinot Noir (California)
11855  Testarossa 2001 Rosella's Vineyard Chardonnay ...
11960  Testarossa 2013 Graham Family Vineyard Pinot N...
11964  Testarossa 2006 Bien Nacido Vineyard Pinot Noi...
12123   Testarossa 2012 Niclaire Pinot Noir (California)
13403  Testarossa 2012 Fogstone Vineyard Pinot Noir (...
14296  Testarossa 2011 Sierra Madre Vineyard Chardonn...
...                                                  ...
71164  Testarossa 2012 Rosella's Vineyard Chardonnay ...
```

```
71807  Testarossa 2012 Sierra Madre Vineyard Pinot No...
71841  Testarossa 2010 Sierra Madre Vineyard Pinot No...
71961  Testarossa 2012 Novitiate Late Harvest Sauvign...
72101  Testarossa 2010 Dos Rubios Vineyard Pinot Noir...
72163  Testarossa 2014 Rosemary's Vineyard Pinot Noir...
72168  Testarossa 2012 Dahlia Dijon Clones Chardonnay...
72267  Testarossa 2012 La Rinconada Vineyard Chardonn...
72316  Testarossa 2012 Pinot Noir (Santa Lucia Highla...
73027  Testarossa 2011 Soberanes Vineyard Chardonnay ...
73480  Testarossa 2011 Lone Oak Vineyard Chardonnay (...
73585   Testarossa 2012 Niclaire Pinot Noir (California)
73700  Testarossa 2013 Sierra Madre Vineyard Chardonn...
73996  Testarossa 2010 Brosseau Vineyard Pinot Noir (...
74150  Testarossa 2014 Soberanes Vineyard Pinot Noir ...
74207  Testarossa 2009 Bien Nacido Vineyard Chardonna...
74248  Testarossa 2008 Rosella's Vineyard Pinot Noir ...
74643    Testarossa 2012 Diana's Chardonnay (California)
74833  Testarossa 2013 Cuvee 107 Pinot Noir (Santa Lu...
74879  Testarossa 2014 Garys' Vineyard Pinot Noir (Sa...
75715  Testarossa 2014 Guidotti Vineyard Pinot Noir (...
76685  Testarossa 2010 Fogstone Vineyard Chardonnay (...
77241  Testarossa 2013 Lone Oak Vineyard Chardonnay (...
77915  Testarossa 2011 Graham Family Vineyard Pinot N...
78585  Testarossa 2012 Zinnia Grandiflora Pinot Noir ...
78639  Testarossa 2001 Bien Nacido Vineyard Pinot Noi...
80224  Testarossa 2014 Pisoni Vineyard Pinot Noir (Sa...
80300  Testarossa 2010 Garys' Vineyard Pinot Noir (Sa...
80822  Testarossa 2012 Sierra Madre Vineyard Chardonn...
82074  Testarossa 2011 Chardonnay (Santa Lucia Highla...

                                      review_description  \
828    Welcoming on the nose, this top-end bottling o...
1306   Released at less than one year of age, this wi...
1833   Testarossa reserves their most expensive wines...
1976   A very good wine while somehow failing to rise...
2060   This is a big wine that's not ready nowit's t...
2110   A nod to this winery being the 107th ever bond...
2419   Made from selected lots of Testarossa's variou...
3063   Pretty notes of black-cherry blossom and fruit...
3544   Although this Pinot is captivatingly rich in r...
3586   One of the more engaging single-vineyard expre...
4214   Heavy and seemingly sweet, this has the taste ...
4494   This begins with a reserved nose that unfolds ...
5777   Salted peach, vanilla-smeared French toast and...
5857   One of four monopole Central Coast bottlings f...
6764   Light, clean and citrusy aromas of lemon, lime...
6841   Testarossa is better known for Pinot Noir, but...
7664   Winemaker Bill Brousseau taps his family viney...
```

```
8407   Heavy and seemingly sweet, this has the taste ...
9107   Very stony and chalky sea-shell aromas arise o...
9276   Rich and creamy in pineapple, orange and lime ...
9370   Plump plum, blackberry and a bit of Christmas ...
9474   Earthy aromas of rust, loam and dark red plum ...
9595   A first-rate Pinot Noir, dry and firm in struc...
10096  A bit of a letdown after a string of riper, mo...
11855  From a cool appellation, a wine that smells br...
11960  This is a generous, tightly wound vineyard-des...
11964  Silky, supple and easy to like, this great win...
12123  Testarossa's annual best-barrel offering start...
13403  A savory nose of seared beef crust gets lift f...
14296  Rich and ripe in tropical fruit, apricot and p...
...                                                  ...
71164  Testarossa's single-vineyard expertise shines ...
71807  Straightforward Pinot Noir scents of red fruit...
71841  Ripe, savory cherry, cola and currant flavors ...
71961  Luscious golden in color, this dessert wine sh...
72101  A difficult Pinot Noir to evaluate. It's very ...
72163  Comprised solely of the richer 667 clone, this...
72168  Ripe, jammy orange, lime, apricot and mango fr...
72267  The single-vineyard experts at Testarossa hit ...
72316  Lots of ripe raspberry, cherry, cola and sanda...
73027  This delicious Chardonnay combines brisk acidi...
73480  This wine is opulent with flavors of orange, p...
73585  Testarossa's annual best-barrel offering start...
73700  Toast, iodine, steel and kumquat-skin aromas s...
73996  With overripe flavors, this Pinot is extremely...
74150  Stark minerality, from hot gravel to chipped s...
74207  Just delicious. Superrich in pineapple, white ...
74248  Not the most successful Rosella's Pinot ever. ...
74643  There's a touch of Riesling-like petrol to the...
74833  This is a breadwinner bottling for the otherwi...
74879  Winemaker Bill Brousseau works to restrain the...
75715  One of the four vineyards that this winery ent...
76685  A leesy, young wine, with the slightly sour, c...
77241  Warm vanilla notes are laid across ripe nectar...
77915  This wine, from one of the coolest parts of th...
78585  This new line from the well-regarded Testaross...
78639  Lots of flavor packed into this light-bodied, ...
80224  Crushed slate, plush rose petals, red cherry, ...
80300  What a delicious Pinot Noir. The cool, long vi...
80822  Vanilla-soaked French toast and caramelized ap...
82074  While Testarossa specializes in single-vineyar...

                designation  points  price    province  \
828                Niclaire      92   80.0  California
1306         Garys' Vineyard      87   59.0  California
```

| | | | | |
|---|---|---|---|---|
| 1833 | Niclaire | 94 | 75.0 | California |
| 1976 | Sleepy Hollow Vineyard | 89 | 50.0 | California |
| 2060 | Garys' Vineyard | 91 | 54.0 | California |
| 2110 | Cuvée 107 | 88 | 23.0 | California |
| 2419 | Niclaire | 92 | 79.0 | California |
| 3063 | Pisoni Vineyard | 92 | 75.0 | California |
| 3544 | Sleepy Hollow Vineyard | 93 | 59.0 | California |
| 3586 | Doctor's Vineyard | 93 | 66.0 | California |
| 4214 | Cuvée 107 | 85 | 18.0 | California |
| 4494 | Brosseau Vineyard | 88 | 61.0 | California |
| 5777 | Rincon Vineyard | 90 | 45.0 | California |
| 5857 | Fogstone Vineyard | 92 | 65.0 | California |
| 6764 | Rosella's Vineyard | 93 | 48.0 | California |
| 6841 | Garys' Vineyard | 94 | 54.0 | California |
| 7664 | Brosseau Vineyard | 92 | 48.0 | California |
| 8407 | Cuvée 107 | 85 | 18.0 | California |
| 9107 | Rincon Vineyard | 92 | 48.0 | California |
| 9276 | Bien Nacido Vineyard | 90 | 41.0 | California |
| 9370 | Cuvée Los Gatos Rob's | 87 | 35.0 | California |
| 9474 | La Encantada Vineyard | 91 | 65.0 | California |
| 9595 | Pisoni Vineyard | 93 | 65.0 | California |
| 10096 | Niclaire | 88 | 75.0 | California |
| 11855 | Rosella's Vineyard | 91 | 35.0 | California |
| 11960 | Graham Family Vineyard | 88 | 64.0 | California |
| 11964 | Bien Nacido Vineyard | 92 | 56.0 | California |
| 12123 | Niclaire | 90 | 81.0 | California |
| 13403 | Fogstone Vineyard | 93 | 61.0 | California |
| 14296 | Sierra Madre Vineyard | 90 | 41.0 | California |
| ... | ... | ... | ... | ... |
| 71164 | Rosella's Vineyard | 92 | 45.0 | California |
| 71807 | Sierra Madre Vineyard | 88 | 61.0 | California |
| 71841 | Sierra Madre Vineyard | 92 | 59.0 | California |
| 71961 | Novitiate Late Harvest | 90 | 25.0 | California |
| 72101 | Dos Rubios Vineyard | 87 | 54.0 | California |
| 72163 | Rosemary's Vineyard | 92 | 64.0 | California |
| 72168 | Dahlia Dijon Clones | 85 | 30.0 | California |
| 72267 | La Rinconada Vineyard | 93 | 45.0 | California |
| 72316 | NaN | 87 | 23.0 | California |
| 73027 | Soberanes Vineyard | 89 | 41.0 | California |
| 73480 | Lone Oak Vineyard | 91 | 41.0 | California |
| 73585 | Niclaire | 90 | 81.0 | California |
| 73700 | Sierra Madre Vineyard | 90 | 48.0 | California |
| 73996 | Brosseau Vineyard | 82 | 54.0 | California |
| 74150 | Soberanes Vineyard | 94 | 64.0 | California |
| 74207 | Bien Nacido Vineyard | 93 | 39.0 | California |
| 74248 | Rosella's Vineyard | 86 | 59.0 | California |
| 74643 | Diana's | 93 | 61.0 | California |
| 74833 | Cuvee 107 | 88 | 20.0 | California |

| | | | | |
|---|---|---|---|---|
| 74879 | Garys' Vineyard | 93 | 66.0 | California |
| 75715 | Guidotti Vineyard | 92 | 64.0 | California |
| 76685 | Fogstone Vineyard | 92 | 39.0 | California |
| 77241 | Lone Oak Vineyard | 91 | 48.0 | California |
| 77915 | Graham Family Vineyard | 91 | 60.0 | California |
| 78585 | Zinnia Grandiflora | 91 | 60.0 | California |
| 78639 | Bien Nacido Vineyard | 86 | 50.0 | California |
| 80224 | Pisoni Vineyard | 93 | 75.0 | California |
| 80300 | Garys' Vineyard | 92 | 59.0 | California |
| 80822 | Sierra Madre Vineyard | 90 | 45.0 | California |
| 82074 | NaN | 91 | 34.0 | California |

| | region_1 | region_2 | winery | variety |
|---|---|---|---|---|
| 828 | California | California Other | Testarossa | Pinot Noir |
| 1306 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 1833 | California | California Other | Testarossa | Pinot Noir |
| 1976 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 2060 | Santa Lucia Highlands | Central Coast | Testarossa | Syrah |
| 2110 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 2419 | California | California Other | Testarossa | Pinot Noir |
| 3063 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 3544 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 3586 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 4214 | Santa Barbara County | Central Coast | Testarossa | Pinot Noir |
| 4494 | Chalone | Central Coast | Testarossa | Pinot Noir |
| 5777 | Arroyo Grande Valley | Central Coast | Testarossa | Chardonnay |
| 5857 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 6764 | Santa Lucia Highlands | Central Coast | Testarossa | Chardonnay |
| 6841 | Santa Lucia Highlands | Central Coast | Testarossa | Syrah |
| 7664 | Chalone | Central Coast | Testarossa | Chardonnay |
| 8407 | Santa Barbara County | Central Coast | Testarossa | Pinot Noir |
| 9107 | Arroyo Grande Valley | Central Coast | Testarossa | Chardonnay |
| 9276 | Santa Maria Valley | Central Coast | Testarossa | Chardonnay |
| 9370 | Central Coast | Central Coast | Testarossa | Red Blend |
| 9474 | Sta. Rita Hills | Central Coast | Testarossa | Pinot Noir |
| 9595 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 10096 | California | California Other | Testarossa | Pinot Noir |
| 11855 | Santa Lucia Highlands | Central Coast | Testarossa | Chardonnay |
| 11960 | Russian River Valley | Sonoma | Testarossa | Pinot Noir |
| 11964 | Santa Maria Valley | Central Coast | Testarossa | Pinot Noir |
| 12123 | California | California Other | Testarossa | Pinot Noir |
| 13403 | Santa Lucia Highlands | Central Coast | Testarossa | Pinot Noir |
| 14296 | Santa Maria Valley | Central Coast | Testarossa | Chardonnay |
| ... | ... | ... | ... | ... |
| 71164 | Santa Lucia Highlands | Central Coast | Testarossa | Chardonnay |
| 71807 | Santa Maria Valley | Central Coast | Testarossa | Pinot Noir |
| 71841 | Santa Maria Valley | Central Coast | Testarossa | Pinot Noir |
| 71961 | Arroyo Seco | Central Coast | Testarossa | Sauvignon Blanc |

```
72101  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
72163   Arroyo Grande Valley     Central Coast  Testarossa       Pinot Noir
72168  Santa Lucia Highlands     Central Coast  Testarossa       Chardonnay
72267        Sta. Rita Hills     Central Coast  Testarossa       Chardonnay
72316  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
73027  Santa Lucia Highlands     Central Coast  Testarossa       Chardonnay
73480  Santa Lucia Highlands     Central Coast  Testarossa       Chardonnay
73585             California  California Other  Testarossa       Pinot Noir
73700     Santa Maria Valley     Central Coast  Testarossa       Chardonnay
73996                Chalone     Central Coast  Testarossa       Pinot Noir
74150  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
74207     Santa Maria Valley     Central Coast  Testarossa       Chardonnay
74248  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
74643             California  California Other  Testarossa       Chardonnay
74833  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
74879  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
75715  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
76685  Santa Lucia Highlands     Central Coast  Testarossa       Chardonnay
77241  Santa Lucia Highlands     Central Coast  Testarossa       Chardonnay
77915   Russian River Valley            Sonoma  Testarossa       Pinot Noir
78585  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
78639     Santa Maria Valley     Central Coast  Testarossa       Pinot Noir
80224  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
80300  Santa Lucia Highlands     Central Coast  Testarossa       Pinot Noir
80822     Santa Maria Valley     Central Coast  Testarossa       Chardonnay
82074  Santa Lucia Highlands     Central Coast  Testarossa       Chardonnay

[175 rows x 12 columns]
```

```
In [137]: wine_train_data[wine_train_data['winery']=='Louis Latour']['price'].describe()
```

```
Out[137]: count    146.000000
          mean     131.506849
          std      134.833192
          min       12.000000
          25%       49.250000
          50%       80.000000
          75%      140.000000
          max      650.000000
          Name: price, dtype: float64
```