Analyzing Wine Reviews

This project aims use Natural Language Processing to analyze Wine Reviews by 18 wine tasters. Both supervised and unsupervised learning methods were used to identify the best model for identifying the texts according to reviewer.

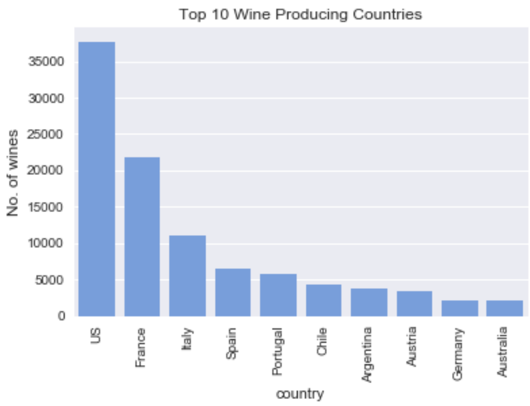
Original Data

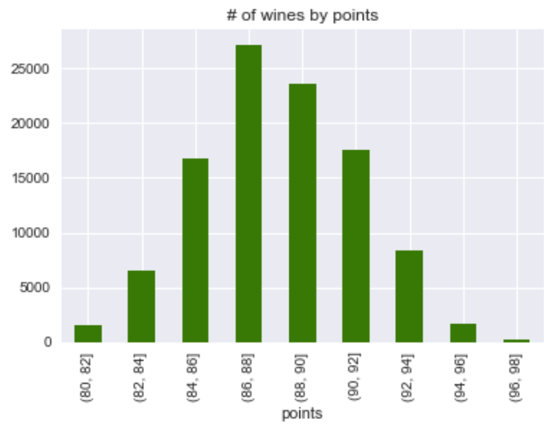
- Data contained features such as wine description, points, price, province, taster_name, wine variety.

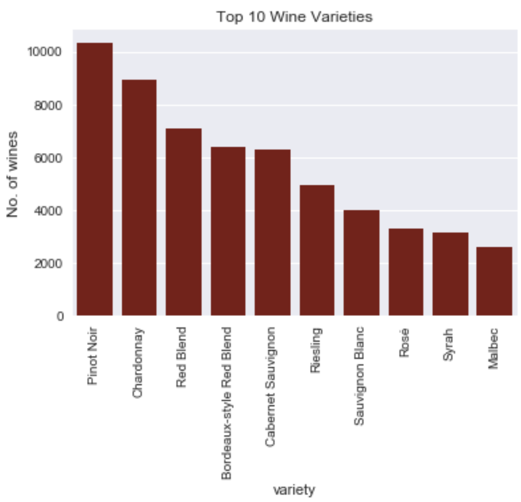
	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitte	title	variety	winery
0	Italy	Aromas include tropi	VulkÃ Bianc	87		Sicily & Sardi	Etna		Kerin Oâ€™Keefe	@kerinokeef	Nicosia 2013	White Blend	Nicosia
1	Portugal	This is ripe and fruity,	Avidagos	87	15	Douro			Roger Voss	@vossroger	Quinta dos A	Portuguese R	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime		87	14	Oregon	Willamette V	Willamette V	Paul Gregutt	@paulgwine.	Rainstorm 20	Pinot Gris	Rainstorm
3	US	Pineapple rind, lemor	Reserve Late	87	13	Michigan	Lake Michigan Shore		Alexander Peartree		St. Julian 201	Riesling	St. Julian
4	US	Much like the regular	Vintner's Res	87	65	Oregon	Willamette V	Willamette V	Paul Gregutt	@paulgwine.	Sweet Cheek	Pinot Noir	Sweet Cheeks
5	Spain	Blackberry and raspb	Ars In Vitro	87	15	Northern Spa	Navarra		Michael Schachner	@wineschac	Tandem 201	Tempranillo-	Tandem
6	Italy	Here's a bright, inforr	Belsito	87	16	Sicily & Sardi	Vittoria		Kerin Oâ€™Keefe	@kerinokeef	Terre di Giur	Frappato	Terre di Giurfo
7	France	This dry and restrained wine offers		87	24	Alsace	Alsace		Roger Voss	@vossroger	Trimbach 20	GewÃ¼rztrar	Trimbach
8	Germany	Savory dried thyme n	Shine	87	12	Rheinhessen			Anna Lee C. Iijima		Heinz Eifel 2(	GewÃ¼rztrar	Heinz Eifel
9	France	This has great depth (	Les Natures	87	27	Alsace	Alsace		Roger Voss	@vossroger	Jean-Baptiste	Pinot Gris	Jean-Baptiste Adam
10	US	Soft, supple plum env	Mountain Cu	87	19	California	Napa Valley	Napa	Virginie Boone	@vboone	Kirkland Sign	Cabernet Sau	Kirkland Signature

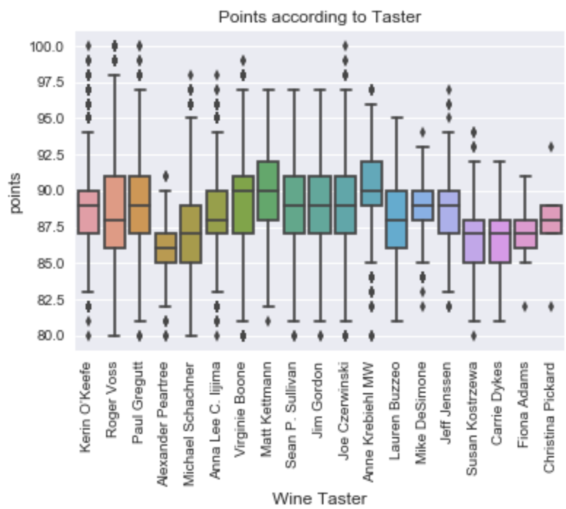
Data Exploration

- Top 3 countries of wines reviewed were from US, France and Italy.
- Most wines reviewed scored between 84-92 points.
- 4 out of the 5 top popular wine varieties were red wines. Pinot Noir and Chardonnay were the most popular red and white wines respectively.









Unsupervised Learning

Term Frequency / Inverse document frequency (TF-IDF)

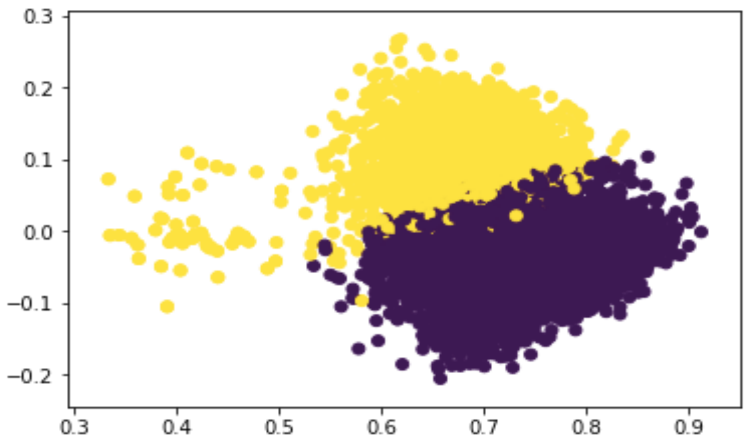
TF-IDF was implemented to identify how important a word is in the reviews. From this, I reduced the number of features to 500 using Truncated Singular Value Decomposition (SVD).

```
svd = TruncatedSVD(500)
lsa = make_pipeline(svd, Normalizer(copy=False))
X_train_lsa = lsa.fit_transform(X_train_tfidf)
variance_explained = svd.explained_variance_ratio_
total_variance = variance_explained.sum()
Percent variance captured by all components: 14.6%
```

Due to the nature of the reviews which mostly contained similar terms, reducing to 500 features could only capture 14.6% of total variance in the dataset.

Identifying Clusters

1. K-Means was used to separate the reviewers into 2 clusters.



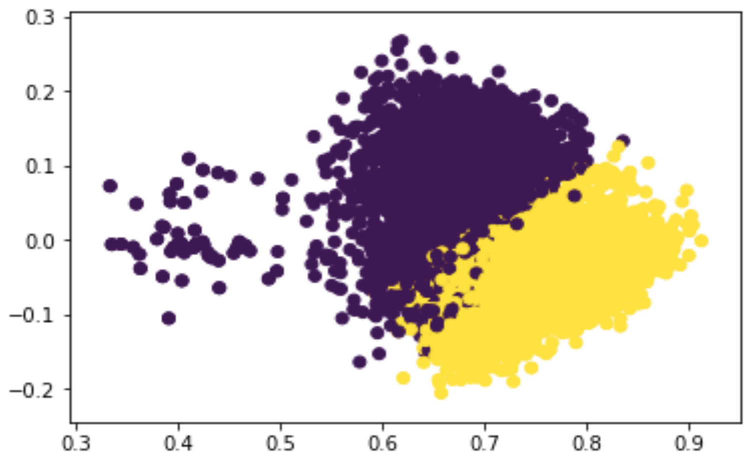
Comparing k-means and pred solutions:		
	0	1
Alexander Peartree	29	2
Anna Lee C. Iijima	300	8
Anne Krebiehl MW	261	9
Carrie Dykes	10	0
Fiona Adams	1	0
Jeff Jenssen	30	1
Jim Gordon	300	40
Joe Czerwinski	387	12
Kerin O'Keefe	760	8
Lauren Buzzeo	147	5
Matt Kettmann	464	4
Michael Schachner	1174	8
Mike DeSimone	41	0
Paul Gregutt	650	29

Roger Voss	123	1708
Sean P. Sullivan	397	18
Susan Kostrzewa	78	0
Virginie Boone	690	53

I also experiemented using mini batch K-Means using 2 clusters and batch-sizes of 200 to identify new predicted clusters.

Comparing k-means and mini-batch k-means:		
	0	1
0	5818	185
1	24	1720

2. Spectral Clustering is based on quantifying similarity between the data points but not necessarily compact or clustered within convex boundaries.



Comparing Spectral Clustering to the ones in the data		
	0	1
Alexander Peartree	2	29
Anna Lee C. Iijima	43	265
Anne Krebiehl MW	36	234
Carrie Dykes	0	10
Fiona Adams	0	1
Jeff Jenssen	12	19
Jim Gordon	135	205
Joe Czerwinski	67	332
Kerin O'Keefe	171	597
Lauren Buzzeo	18	134
Matt Kettmann	8	460
Michael Schachner	51	1131
Mike DeSimone	4	37
Paul Gregutt	130	549

Roger Voss	1720	111
Sean P. Sullivan	125	290
Susan Kostrzewa	9	69
Virginie Boone	161	582

In general, the two clustering methods were unable to identify clear clusters by each reviewer, indicating that reviewers often used the same words to describe wines, identifying clusters based on their writing styles was ineffective. Both methods showed similar clusters, though K-Means surprisingly showed a clearer distinction versus Spectral Clustering.

Supervised Learning

Bag of Words was used to process the reviews of the 18 wine tasters. After identifying the 300 most common words for each reviewer, new features such as length of sentences, punctuation length and length of unique words were also added to the dataframe. I then ran Multi Layer Processing Classifier (MLP), Random Forest, Logistic Regression as well as Naïve Bayes classifier to identify the best performing model on the test set (25% of data). 5 fold cross validation was also carried out to assess how the results of the analysis will perform on the test.

1) **Multi Layer Processing Classifier**

```
mlp = MLPClassifier(hidden_layer_sizes=(800,800),
max_iter=100, batch_size=500, learning_rate_init=0.001,
alpha=0.6)
mlp.fit(X_train, y_train)
Training set score:  0.825353
Test set score: 0.766401
```

Overfitting was also observed for MLP Classifier, which is also a common problem for neural networks. I adjusted the alpha, the parameter for regularization which helps in combating overfitting by constraining the size of the weights. Learning rate used was also important, and a lower rate performed better. After experimenting with different number of layers and hidden nodes, I found that increasing the hidden layer size improved the performance of test set.

2) **Random Forest**

```
Training set score: 0.972130
Test set score: 0.632429
Mean cross validation score is: 0.630422
```

While random forest is a robust and versatile method, it's clearly not the best choice for high-dimensional sparse data such as BoW representation, which showed overfitting of the data. This seems to be an inherent challenge that bag of words models will always face as it routinely runs into tasks where number of features is much bigger than number of examples.

3) **Logistic Regression**

```
Training set score: 0.797498
```

```
Test set score: 0.768519
Mean cross validation score is: 0.766267
```

Logistic Regression was a better performing model, with less overfitting than random forest.

4) **Naïve Bayes (Bernoulli Classifier)**

```
Training set score: 0.729619
Test set score: 0.723079
Mean cross validation score is: 0.721031
```

Naïve Bayes is flexible enough to capture imbalance in the frequency of sparse and dense data, and showed relatively good results for text classification inn this case. It performed worse than Logistic Regression though probably due to the assumption of conditional independence of its features.

Conclusion

Supervised learning regression methods were better at predicting the texts according to reviewers with Logistic Regression showing the best performance. Unsupervised learning failed to identify clear clusters of the reviewers due to similarity of words used among the reviewers.