Read_Between_the_Wine

Presented by:
Andrew Taylor
Fidelia Akparu
Heather Lea
Mia Goodwin

Our Topic

Wine Quality

We chose to analyze data on wine quality because it seemed interesting and was an emotionally lighter and happier topic than some of the others that we considered.

Data

For this analysis we downloaded two datasets from Kaggle (https://kaggle.com/).

1. Red Wine Quality (winequality-red.csv)

fixed acidi	volatile ac	citric acid	residual su	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7

Data Cont.

2. White Wine Quality (winequality-white.csv)

fixed acidi	volatile ac	citric acid	residual su	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	quality
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6
8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5
8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5
7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5
6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7
8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5
6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7

Data Cont.

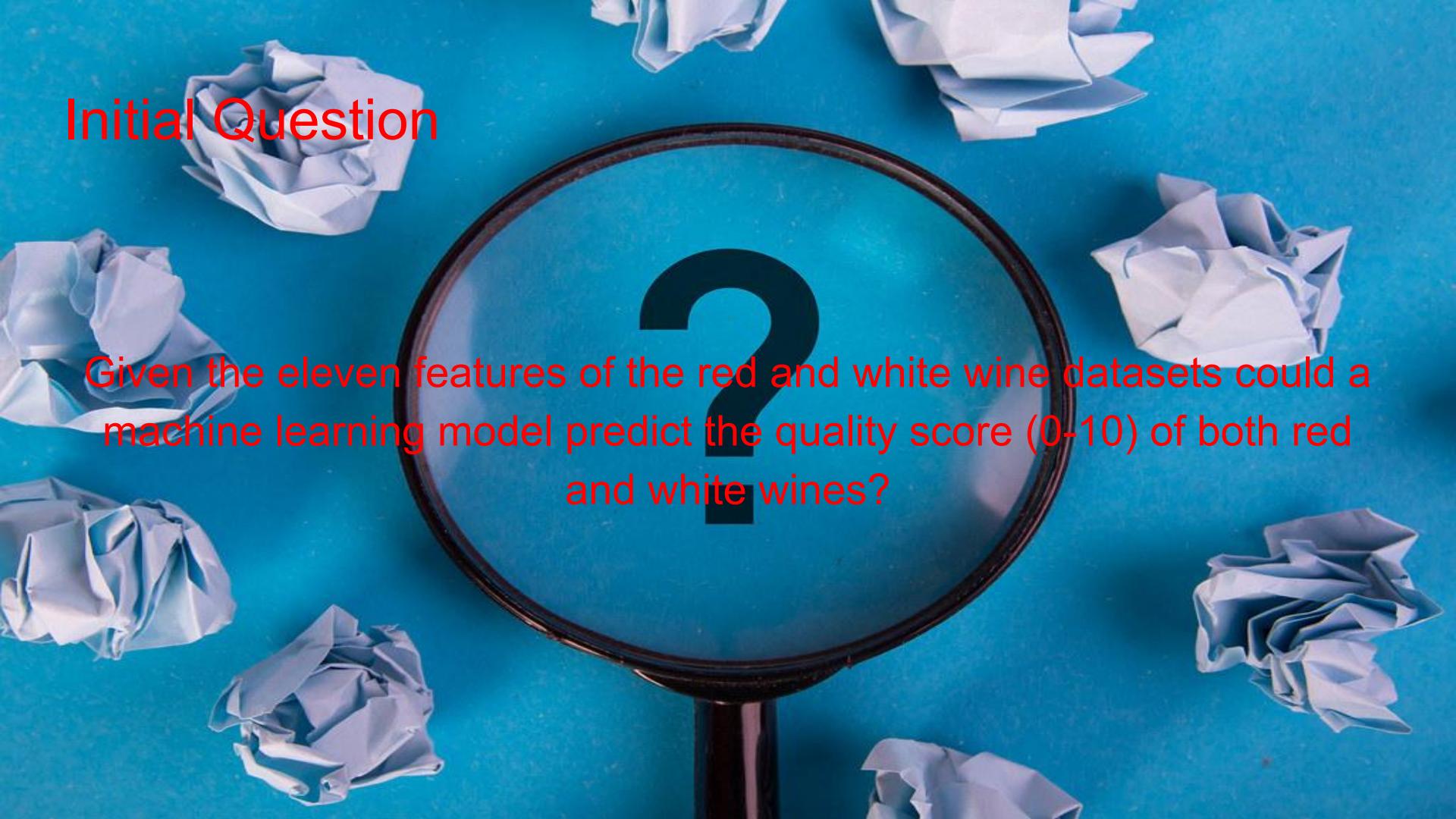
Both datasets contained eleven features:

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulfur dioxide
- Total sulfur dioxide
- Density
- pH
- Sulphates
- Alcohol

Technologies

- Python
- SQL
- Matplotlib
- Tableau
- During EDA we used Kmeans and Clustering Algorithms
- Our models were fitted on a RandomForest Classifier

Initial Question



Given the eleven features of the red and white wine datasets could a machine learning model predict the quality score (0-10) of both red and white wines?

Data Exploration

1. First we loaded in our two csv files as data frames. Then for each data frame we:
    - Checked dtypes.
    - Checked for any null values
    - Checked for and dropped duplicates
    - Checked the number of data points in each table

2. We added a type column to each data frame because each table did not have an identifying label for the type of wine. Red wines were identified as type 0 and white wine was identified as type 1.

3. We merged the two dataframes into wine_df.

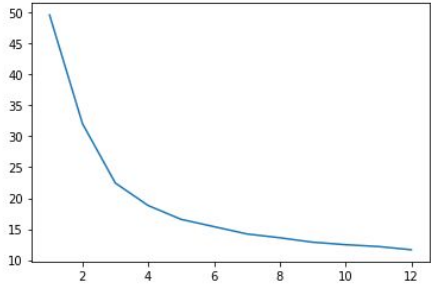
4. Checked for unique values in each feature of the merged data frame.

5. Using a for loop, we iterated through the features to get a K-means model and plotted our distortions and inertias.

Data Exploration

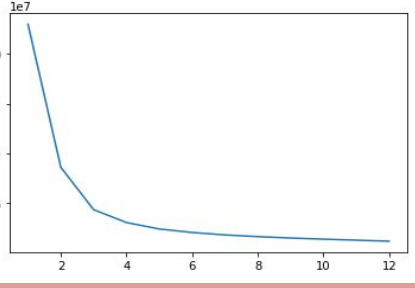
```
plt.plot(range(1,13), distortions)
```

```
[<matplotlib.lines.Line2D at 0x7f5c5002df10>]
```




```
plt.plot(range(1,13), inertias)
```

```
[<matplotlib.lines.Line2D at 0x7f5c4fa98690>]
```



From these plots we see that we should be able to have three distinct clusters. So we printed out a correlations table to determine which three features have the most effect on quality.

```
wine_df.corr()
```

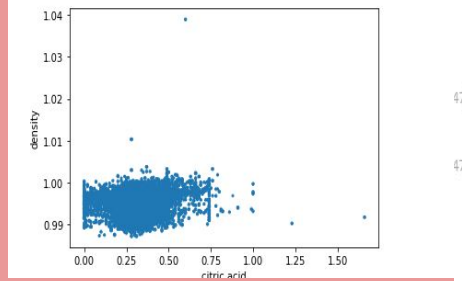
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
fixed acidity	1.000000	0.219008	0.324436	-0.111981	0.298195	-0.282735	-0.329054	0.458910	-0.252700	0.299568	-0.095452	-0.076743	-0.486740
volatile acidity	0.219008	1.000000	-0.377981	-0.196011	0.377124	-0.352557	-0.414476	0.271296	0.261454	0.225984	-0.037640	-0.265699	-0.653036
citric acid	0.324436	-0.377981	1.000000	0.142451	0.038998	0.133126	0.195242	0.096154	-0.329808	0.056197	-0.010493	0.085532	0.187397
residual sugar	-0.111981	-0.196011	0.142451	1.000000	-0.128940	0.402871	0.495482	0.552517	-0.267320	-0.185927	-0.359415	-0.036980	0.348821
chlorides	0.298195	0.377124	0.038998	-0.128940	1.000000	-0.195045	-0.279630	0.362615	0.044708	0.395593	-0.256916	-0.200666	-0.512678
free sulfur dioxide	-0.282735	-0.352557	0.133126	0.402871	-0.195045	1.000000	0.720934	0.025717	-0.145854	-0.188457	-0.179838	0.055463	0.471644
total sulfur dioxide	-0.329054	-0.414476	0.195242	0.495482	-0.279630	0.720934	1.000000	0.032395	-0.238413	-0.275727	-0.265740	-0.041385	0.700357
density	0.458910	0.271296	0.096154	0.552517	0.362615	0.025717	0.032395	1.000000	0.011686	0.259478	-0.686745	-0.305858	-0.390645
pH	-0.252700	0.261454	-0.329808	-0.267320	0.044708	-0.145854	-0.238413	0.011686	1.000000	0.192123	0.121248	0.019506	-0.329129
sulphates	0.299568	0.225984	0.056197	-0.185927	0.395593	-0.188457	-0.275727	0.259478	0.192123	1.000000	-0.003029	0.038485	-0.487218
alcohol	-0.095452	-0.037640	-0.010493	-0.359415	-0.256916	-0.179838	-0.265740	-0.686745	0.121248	-0.003029	1.000000	0.444319	0.032970
quality	-0.076743	-0.265699	0.085532	-0.036980	-0.200666	0.055463	-0.041385	-0.305858	0.019506	0.038485	0.444319	1.000000	0.119323
type	-0.486740	-0.653036	0.187397	0.348821	-0.512678	0.471644	0.700357	-0.390645	-0.329129	-0.487218	0.032970	0.119323	1.000000

Data Exploration

We then plotted scatter plots of the the features with the highest correlations to quality to find the top three features for our clusters.

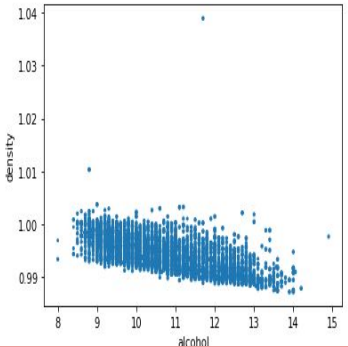
```
wine_df[['citric acid','density', 'quality']].plot.scatter(x='citric acid',y='density', s='quality')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5c4f6ed150>
```



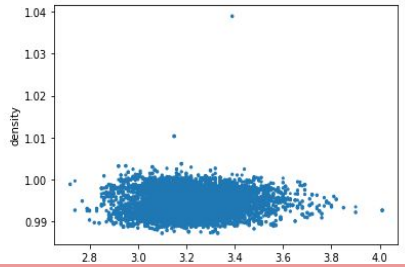

```
wine_df[['alcohol','density', 'quality']].plot.scatter(x='alcohol',y='density', s='quality')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5c4f5aba10>
```




```
wine_df[['pH', 'density', 'quality']].plot.scatter(x='pH',y='density', s='quality')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5c4f6b8c90>
```



Data Exploration

Our scatter plots were compacted and did not show a clear picture of what three features affected the quality of a wine the most. So we performed a PCA and plotted clusters.

We could not get three distinct clusters and we determined that the features important for wine quality was different for each type of wine (red or white). We used violin plots to depict the features importance to quality.

New Question

Given the current datasets and features can we predict a “Good” wine with a machine learning model?

Preprocessing

- We combined the two datasets
- Dropped duplicates
- Encoded our quality feature:
    - A quality rating greater than or equal to 7 was classified as “Good”
    - Anything less than 7 was classified as “Bad”.
- Our features and target (quality) were defined.
- Data was split into training and testing.
- A standard scaler was applied to the data.

Model

The scaled data was fitted to a RandomForest Classifier. We were able to predict a good wine approximately 85% of the time.

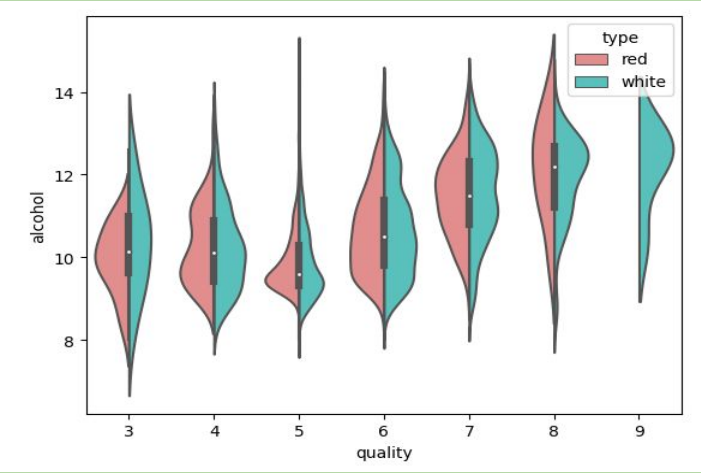
	Predicted 'Good'	Predicted'Bad'
Actual 'Good'	100	153
Actual 'Bad'	52	1025

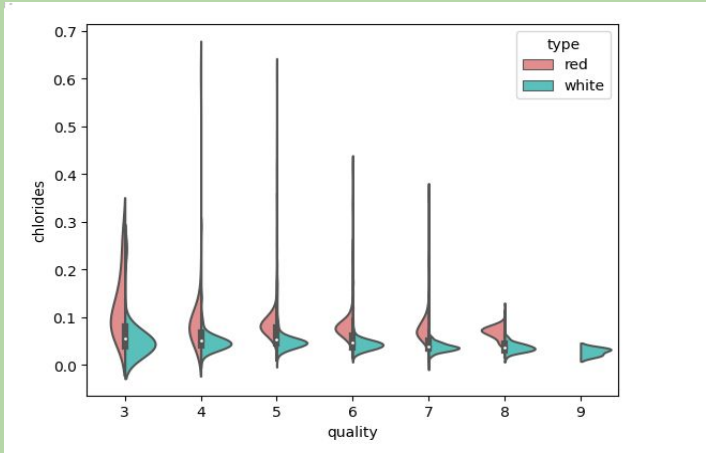
Accuracy Score : 0.8458646616541353
Classification Report

	precision	recall	f1-score	support
0	0.66	0.40	0.49	253
1	0.87	0.95	0.91	1077
accuracy			0.85	1330
macro avg	0.76	0.67	0.70	1330
weighted avg	0.83	0.85	0.83	1330

Model cont.

Utilizing violin plots, we visualized the features importance to quality as it pertained to wine type.





Additional Analysis

Based upon the physiochemical feature differences, can we predict wine type?

Preprocessing

- We combined the two datasets
- Dropped duplicates
- Our features and target (type) were defined.
- Data was split into training and testing.
- A standard scaler was applied to the data.

Model

The scaled data was fitted to a RandomForest Classifier. We were able to predict a good wine approximately 99% of the time.

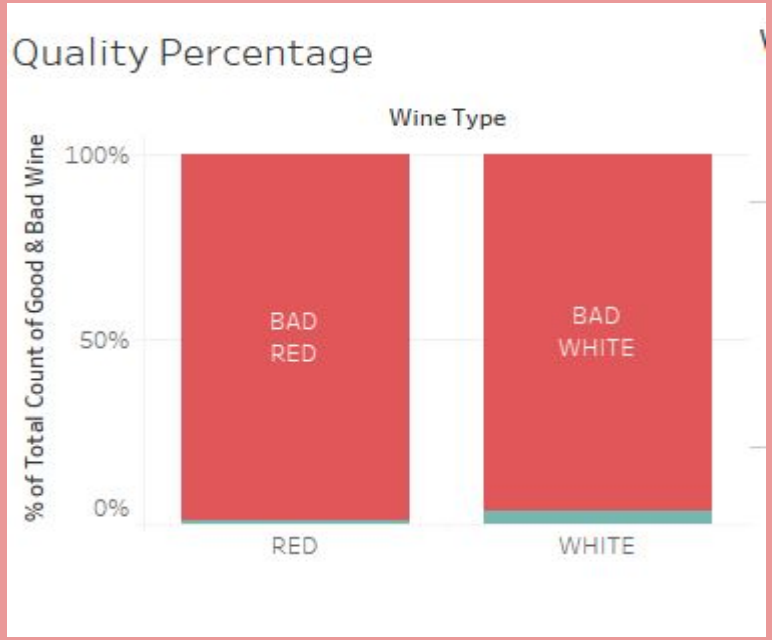
	Predicted Red	Predicted White
Actual Red	349	7
Actual White	1	973

Accuracy Score : 0.9939849624060151
Classification Report

	precision	recall	f1-score	support
0	1.00	0.98	0.99	356
1	0.99	1.00	1.00	974
accuracy			0.99	1330
macro avg	0.99	0.99	0.99	1330
weighted avg	0.99	0.99	0.99	1330

Model cont.

The majority of the wines (both red and white) were classified as bad wines.



Results

-

Future Analysis

-

Dashboard

Link to our dashboard.

https://public.tableau.com/app/profile/fidelia1205/viz/WineQuality_16646390058510/WineQualityPrediction?publish=yes

Questions