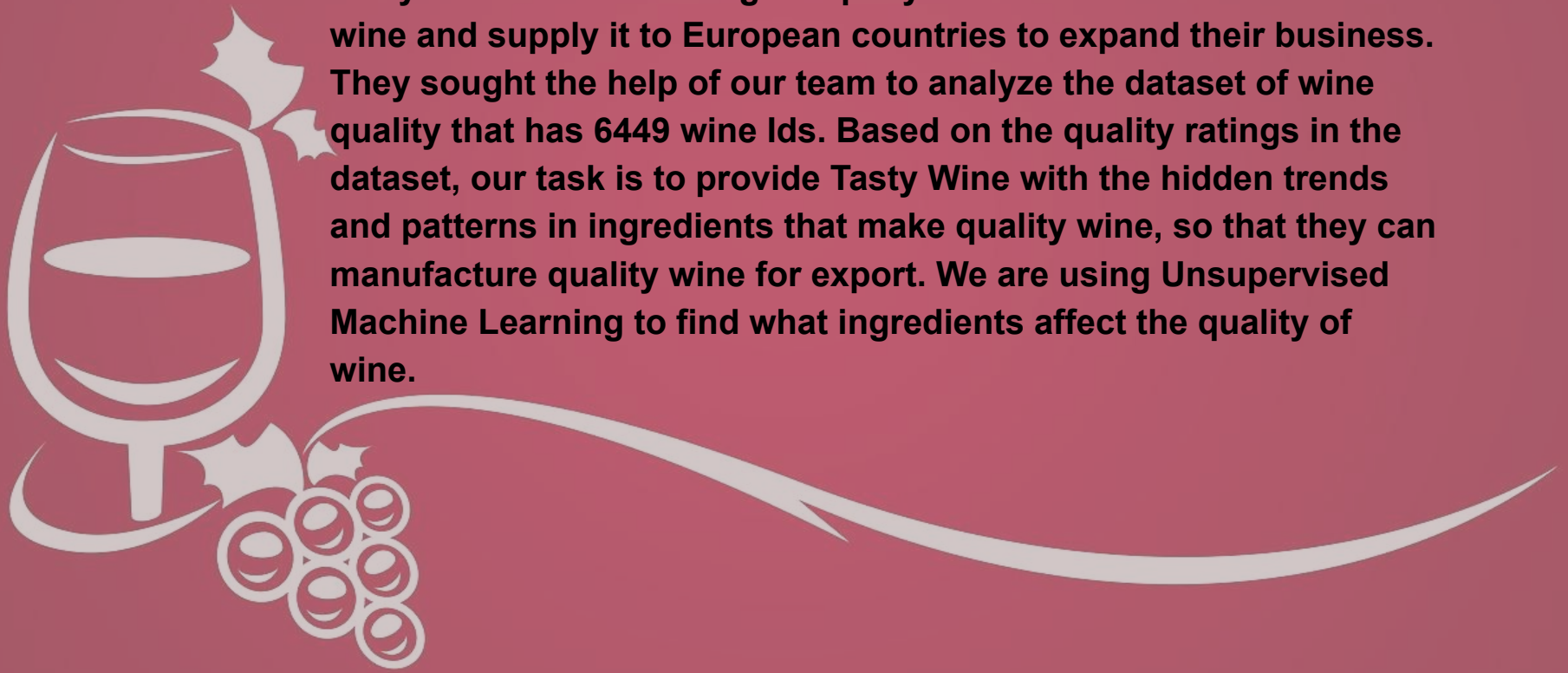
Wine Quality Analysis

By: Purnima Pathak, Anthony Pino, Emily Washburn, Sergio Roberto de Souza Junior

Background



Tasty Wine manufacturing company wants to enhance the taste of wine and supply it to European countries to expand their business. They sought the help of our team to analyze the dataset of wine quality that has 6449 wine Ids. Based on the quality ratings in the dataset, our task is to provide Tasty Wine with the hidden trends and patterns in ingredients that make quality wine, so that they can manufacture quality wine for export. We are using Unsupervised Machine Learning to find what ingredients affect the quality of wine.

Setting up the database

```python
# SQLAlchemy
from sqlalchemy import create_engine
from config import password
```

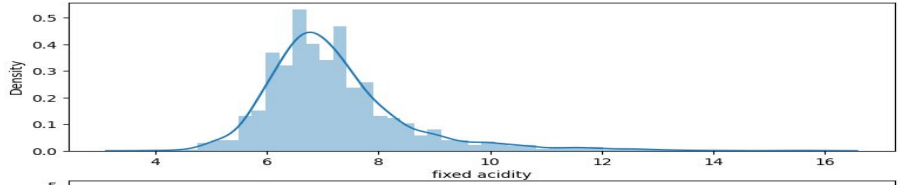
```python
# Create an engine that can talk to the database
engine = create_engine(f"postgresql://postgres:{password}@localhost/Wine_quality")

```

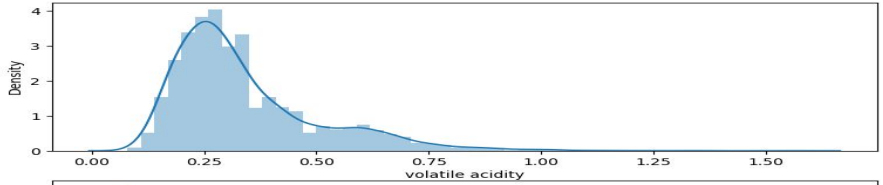
```python
# Query All Records in the the Database
data = engine.execute("SELECT * FROM wines")

for record in data:
    print(record)
```

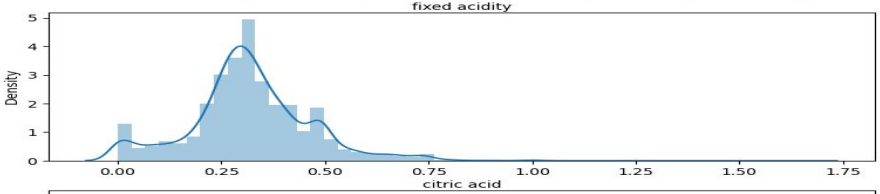
```python
df = pd.read_sql_table('wines', con = engine)
df.head()
```

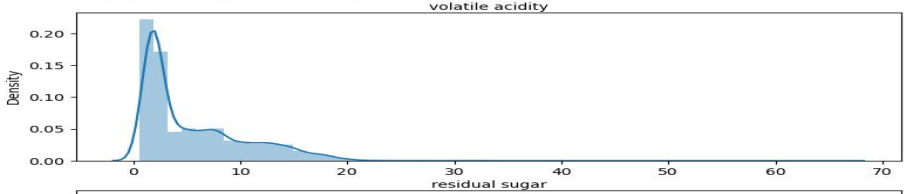
Python

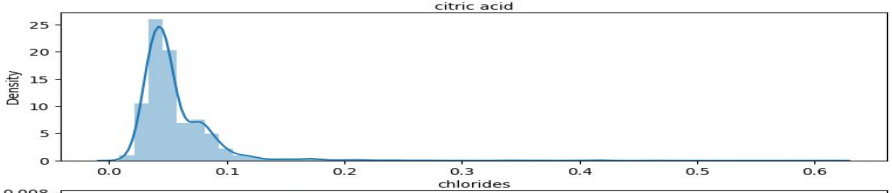
	wine_id	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	ph	sulphates	alcohol	quality	type
0	2732	7.4	0.170	0.29	1.4	0.047	23.0	107.0	0.99390	3.52	0.65	10.4	6	White Wine
1	2607	5.3	0.310	0.38	10.5	0.031	53.0	140.0	0.99321	3.34	0.46	11.7	6	White Wine
2	1653	4.7	0.145	0.29	1.0	0.042	35.0	90.0	0.99080	3.76	0.49	11.3	6	White Wine
3	3264	6.9	0.260	0.29	4.2	0.043	33.0	114.0	0.99020	3.16	0.31	12.5	6	White Wine
4	4931	6.4	0.450	0.07	1.1	0.030	10.0	131.0	0.99050	2.97	0.28	10.8	5	White Wine

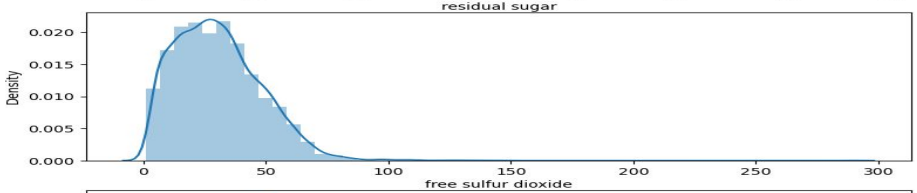

Density
fixed acidity

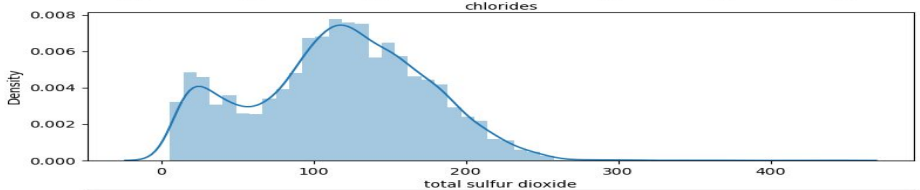

Density
volatile acidity

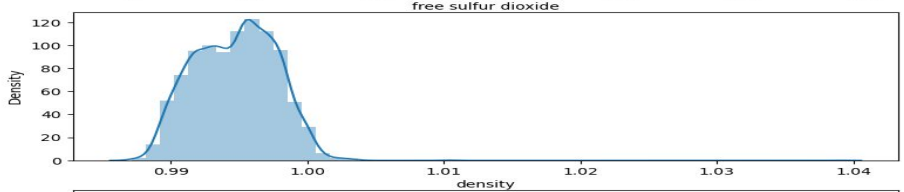

Density
citric acid

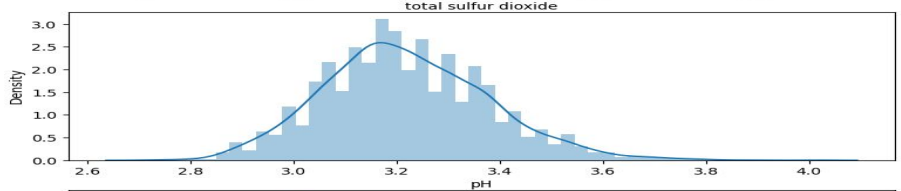

Density
residual sugar

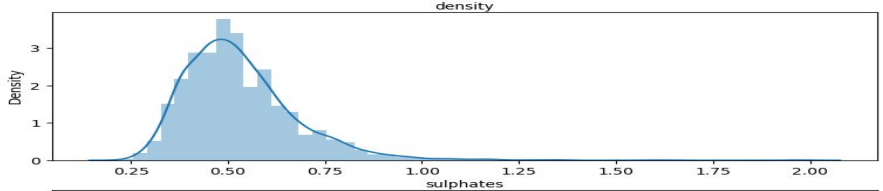

Density
chlorides

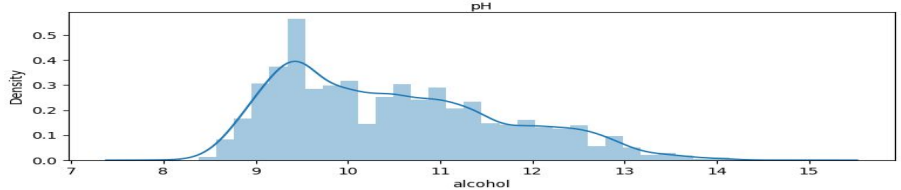

Density
free sulfur dioxide

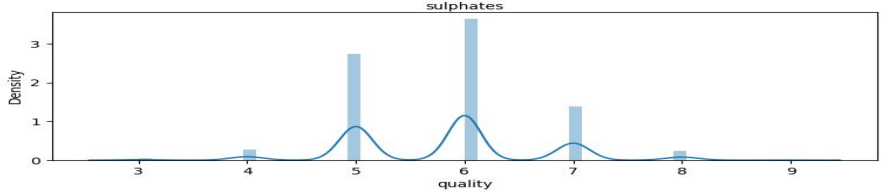

Density
total sulfur dioxide


Density
density


Density
pH

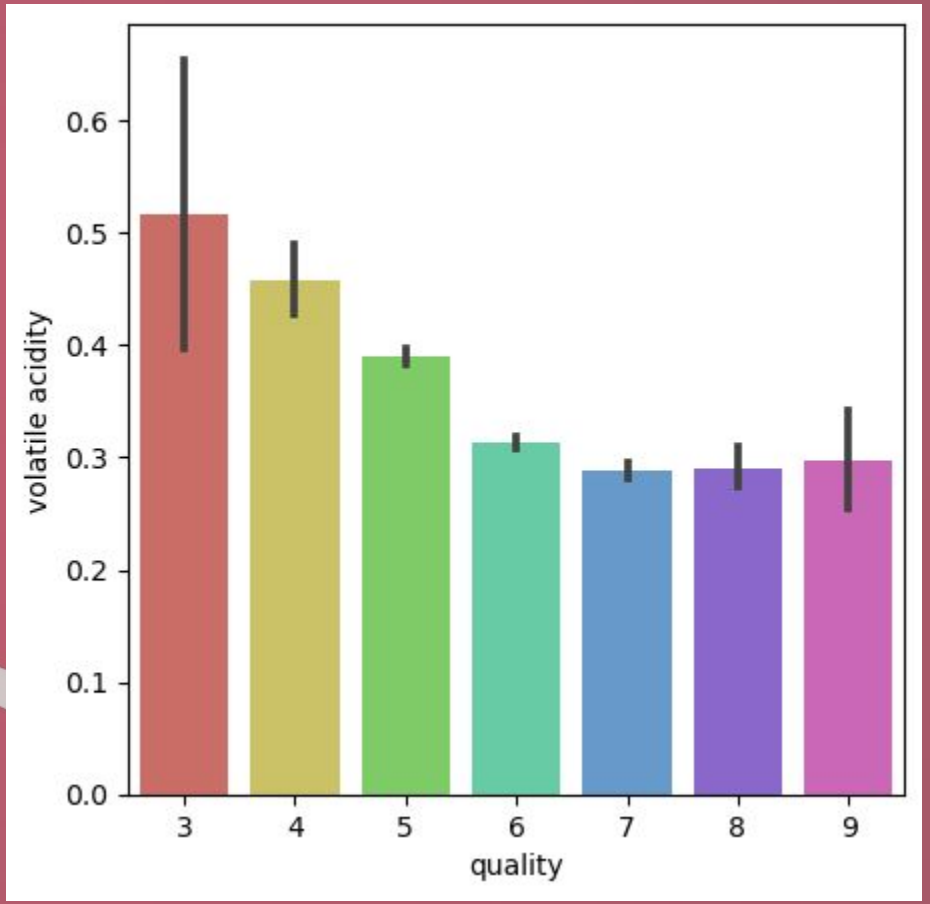

Density
sulphates


Density
alcohol

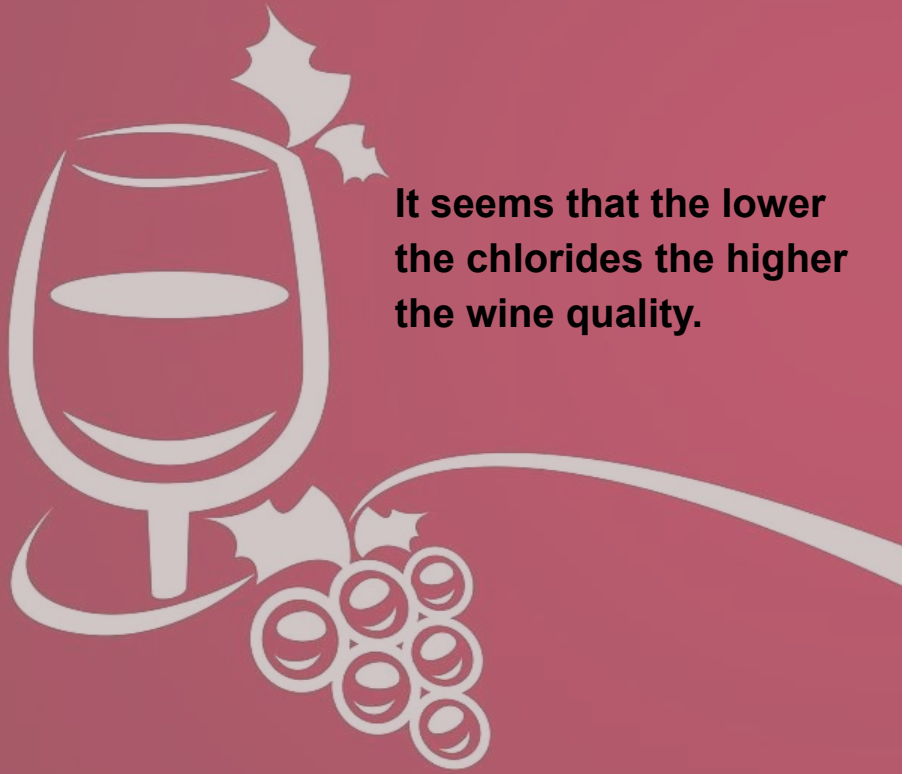

Density
quality

What makes a quality wine

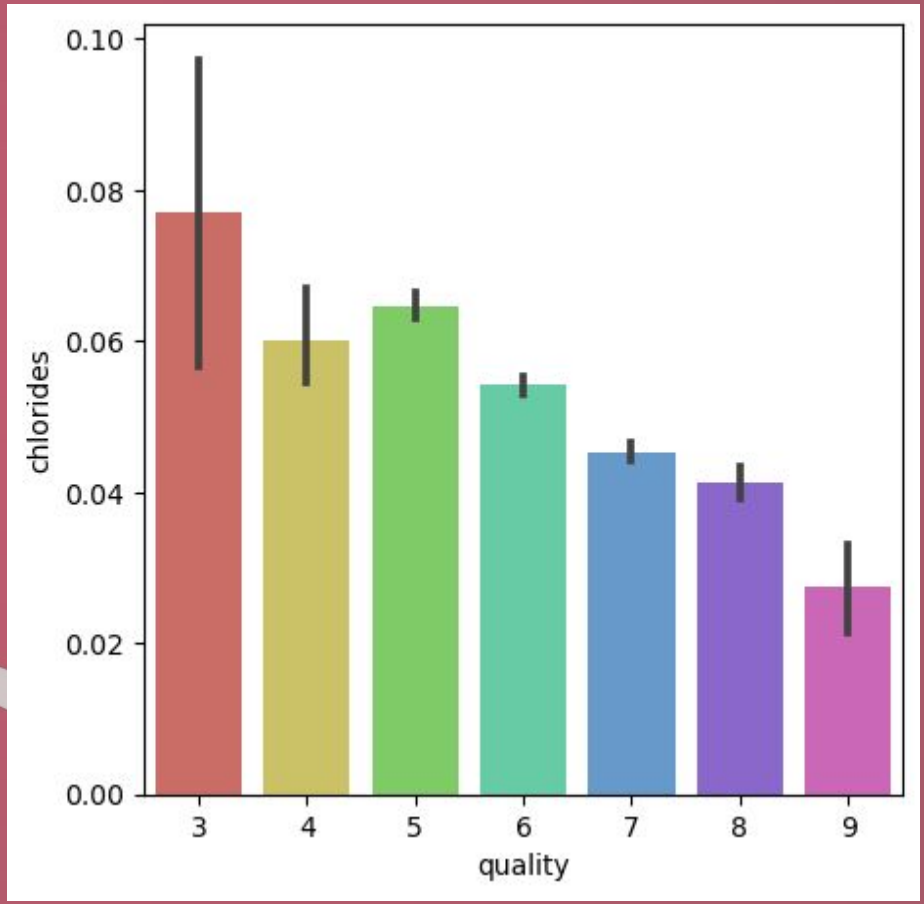
- Through our data exploration we found 4 areas that affected quality.
- This chart shows that the lower the volatile acid, the higher the wine quality.



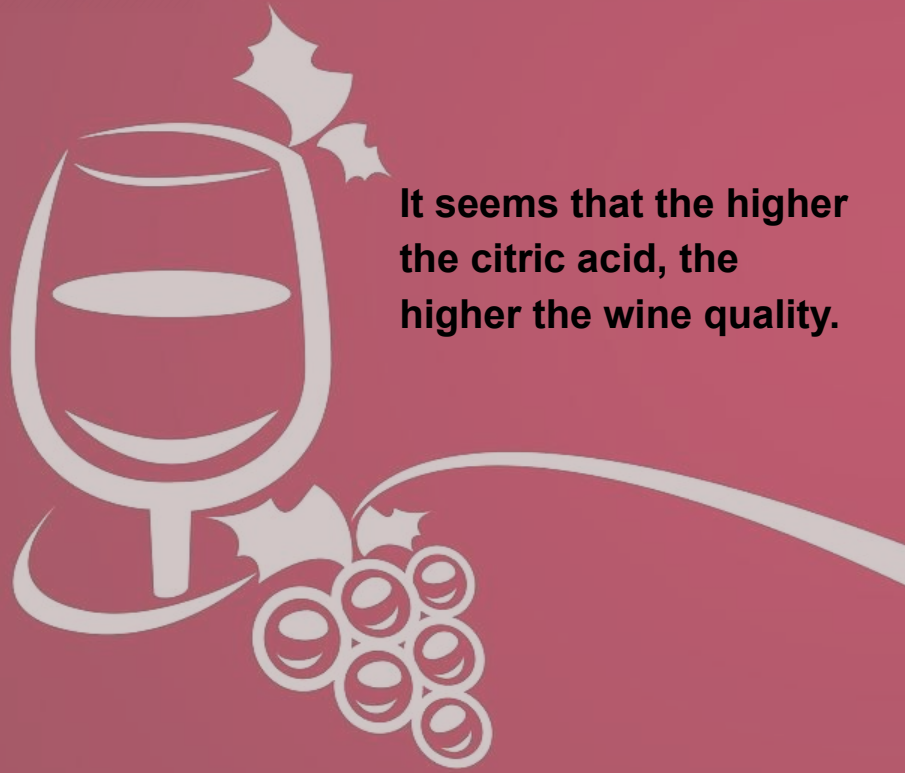
What makes a quality wine



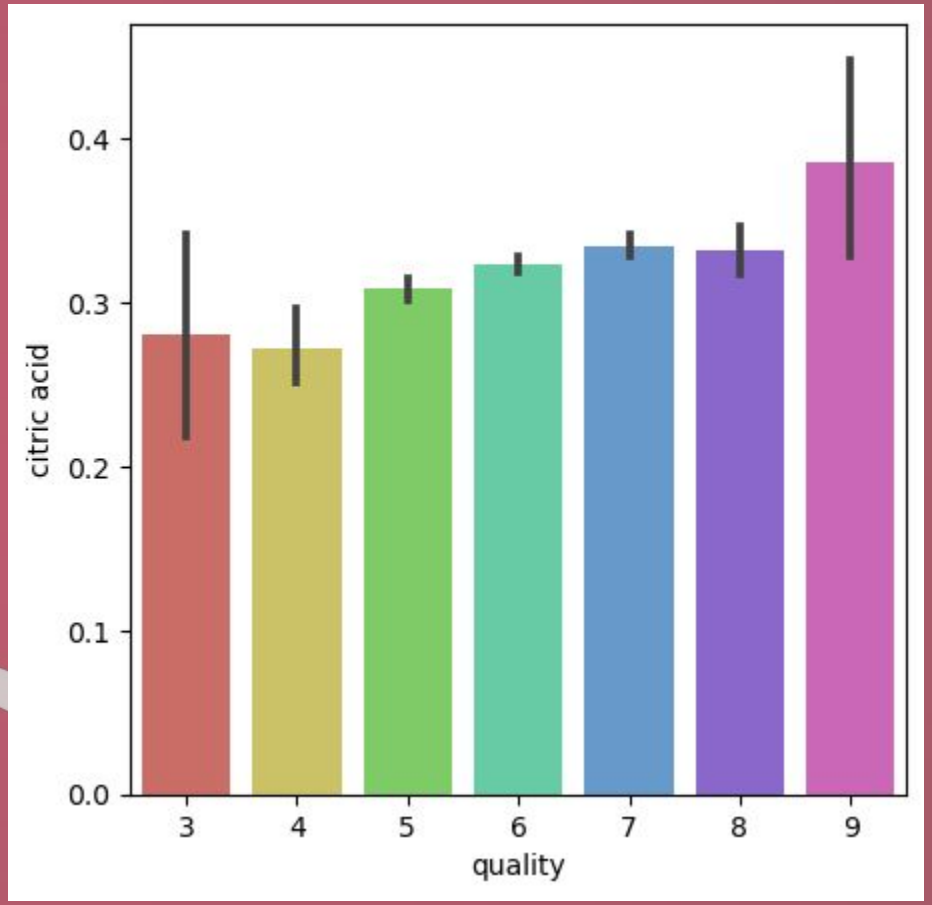
It seems that the lower the chlorides the higher the wine quality.



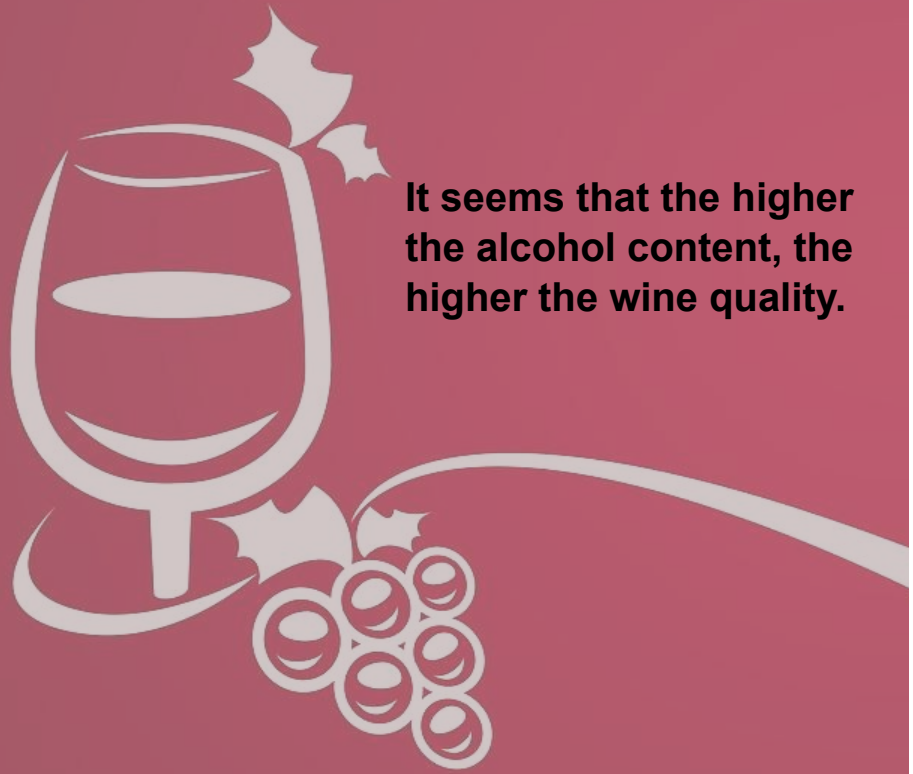
What makes a quality wine



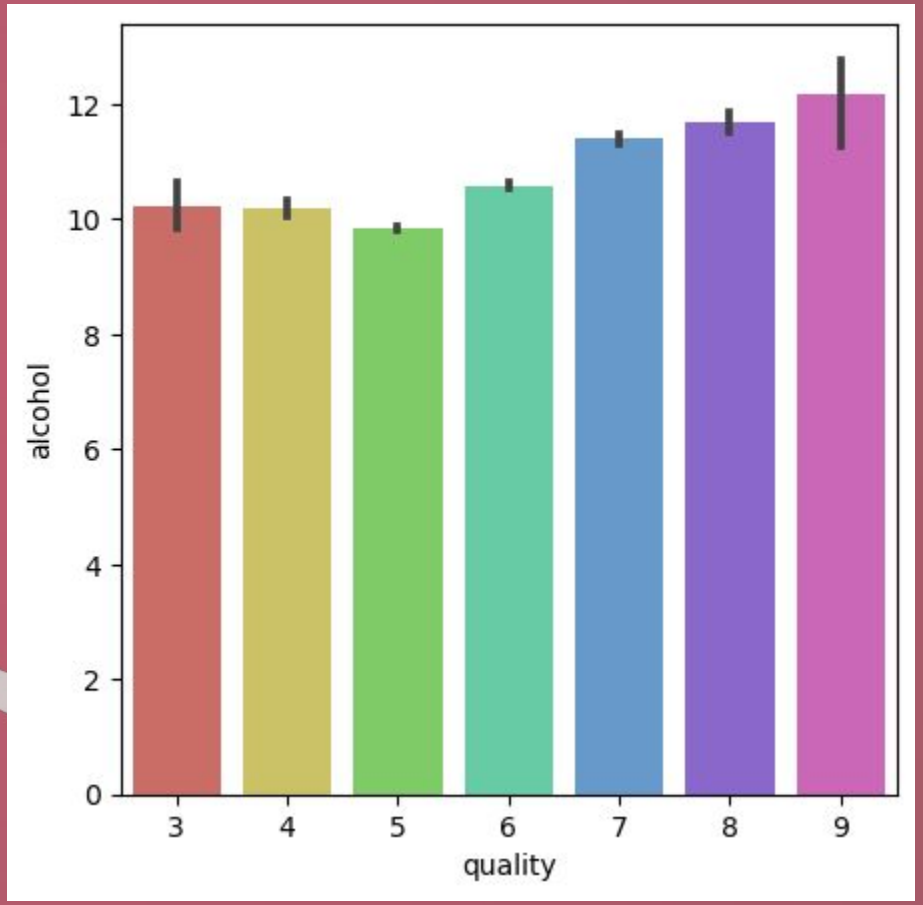
It seems that the higher the citric acid, the higher the wine quality.



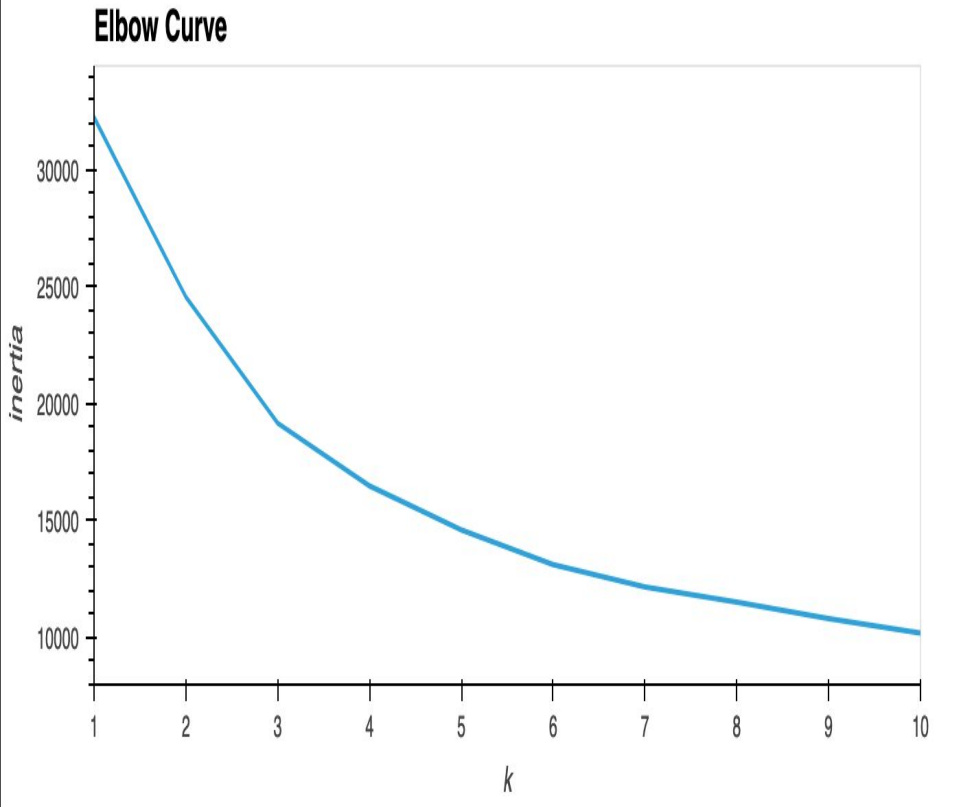
What makes a quality wine

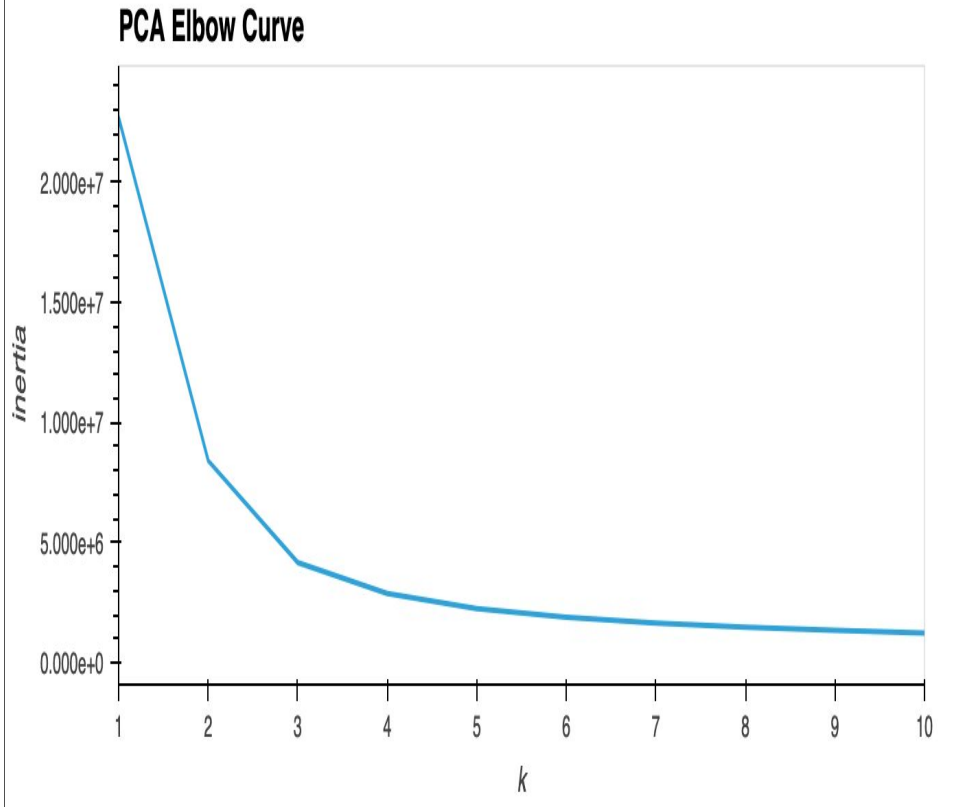


It seems that the higher the alcohol content, the higher the wine quality.



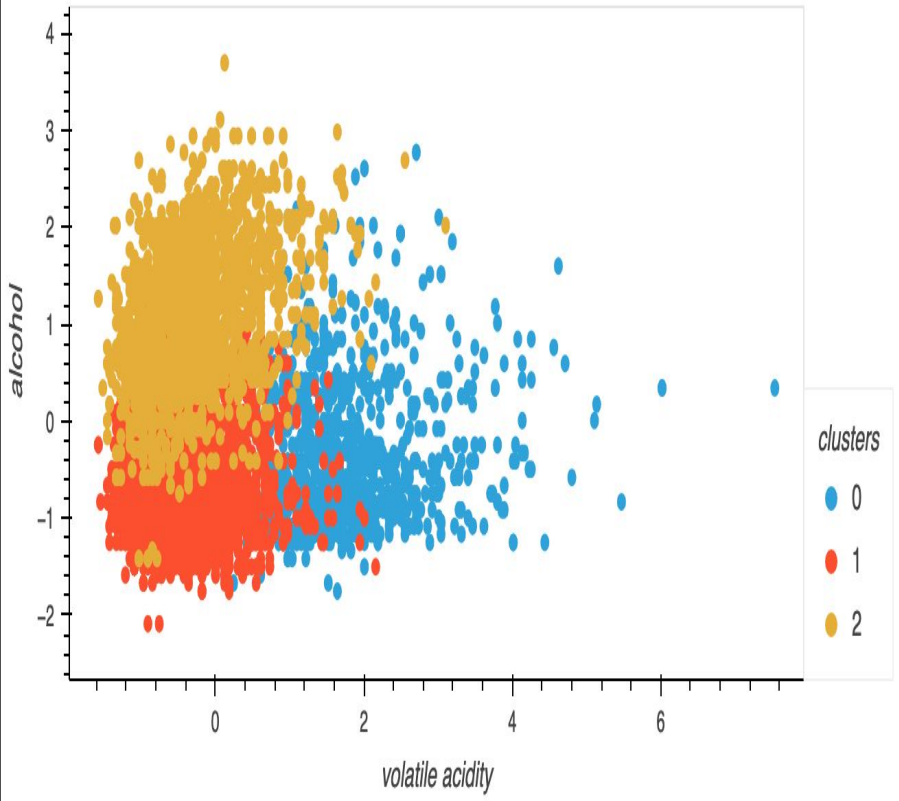
Elbow Curves



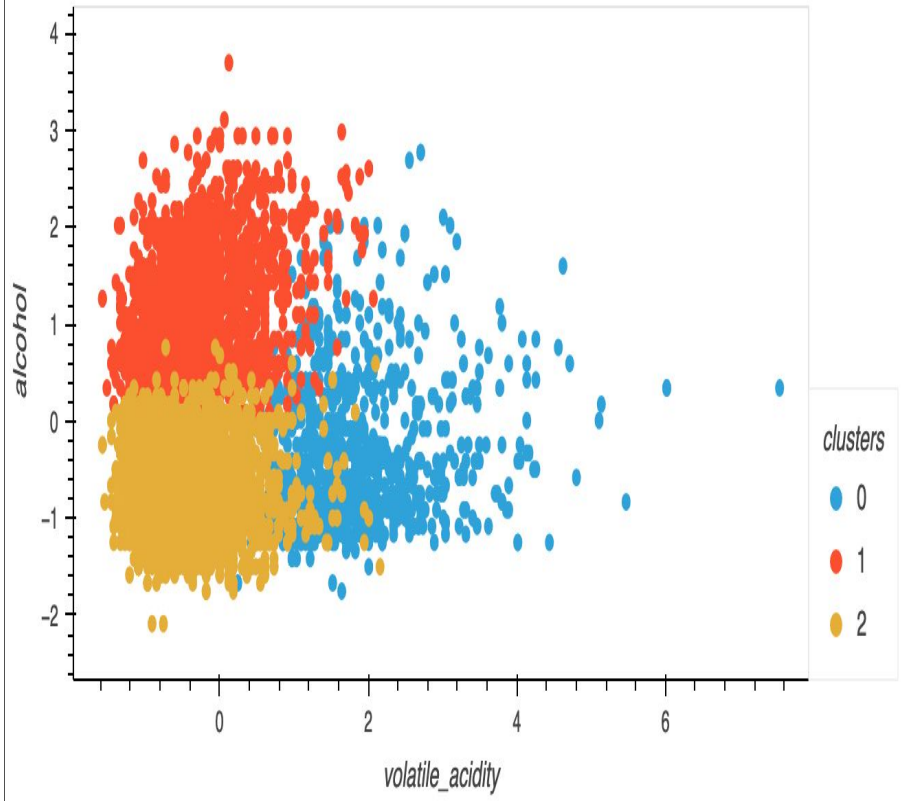


Volatile acidity with and without quality

With quality

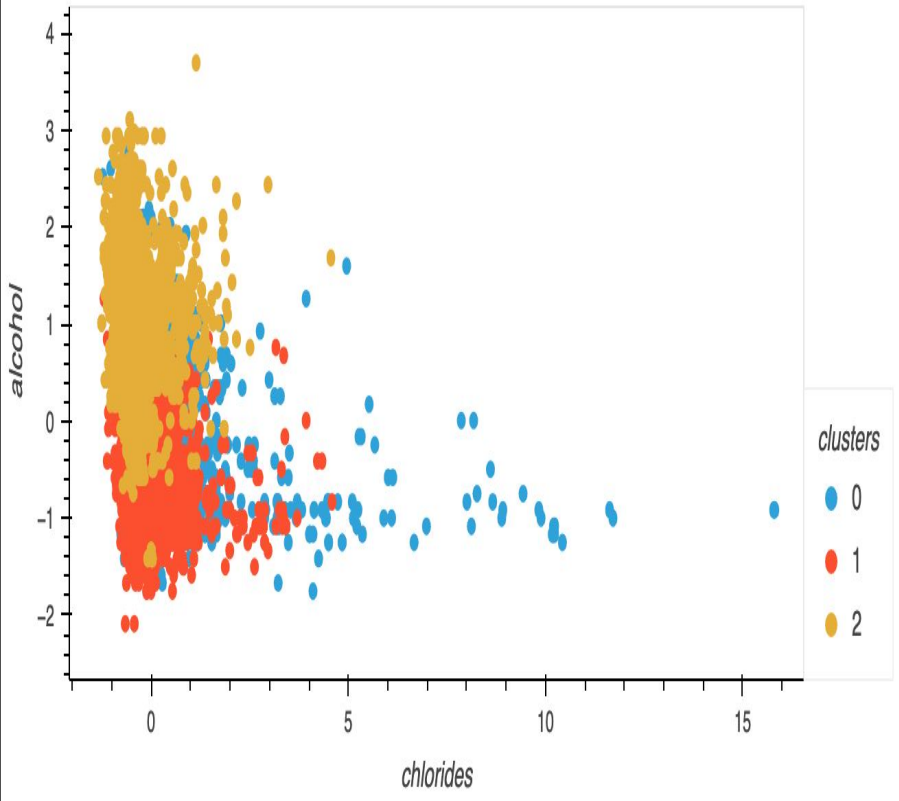


Without quality

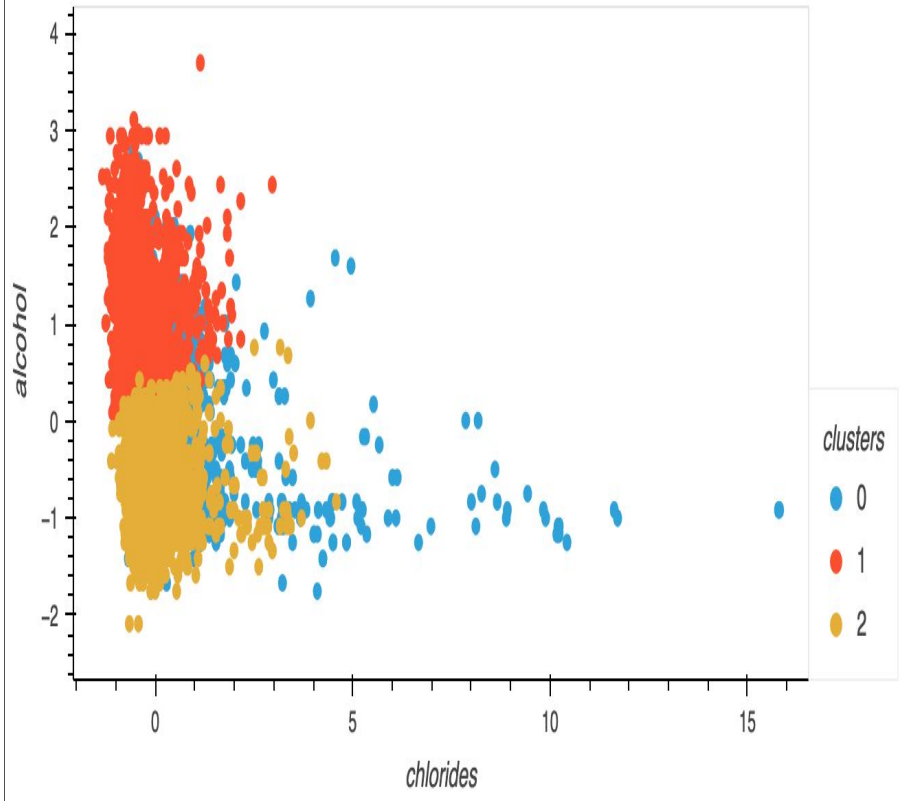


Chlorides with and without quality

With quality

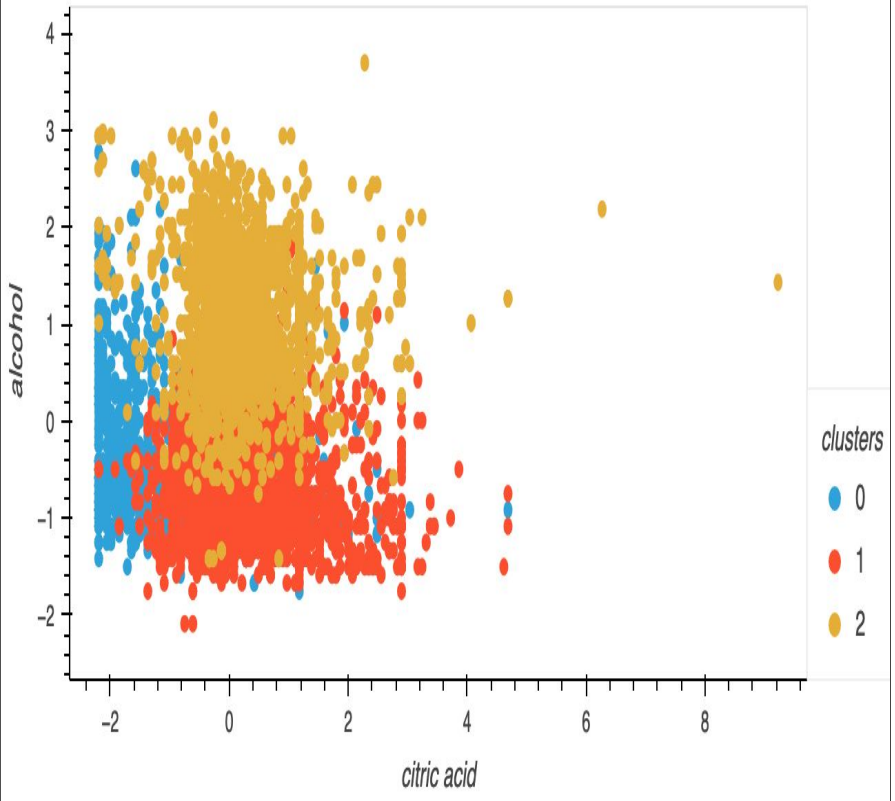


Without quality

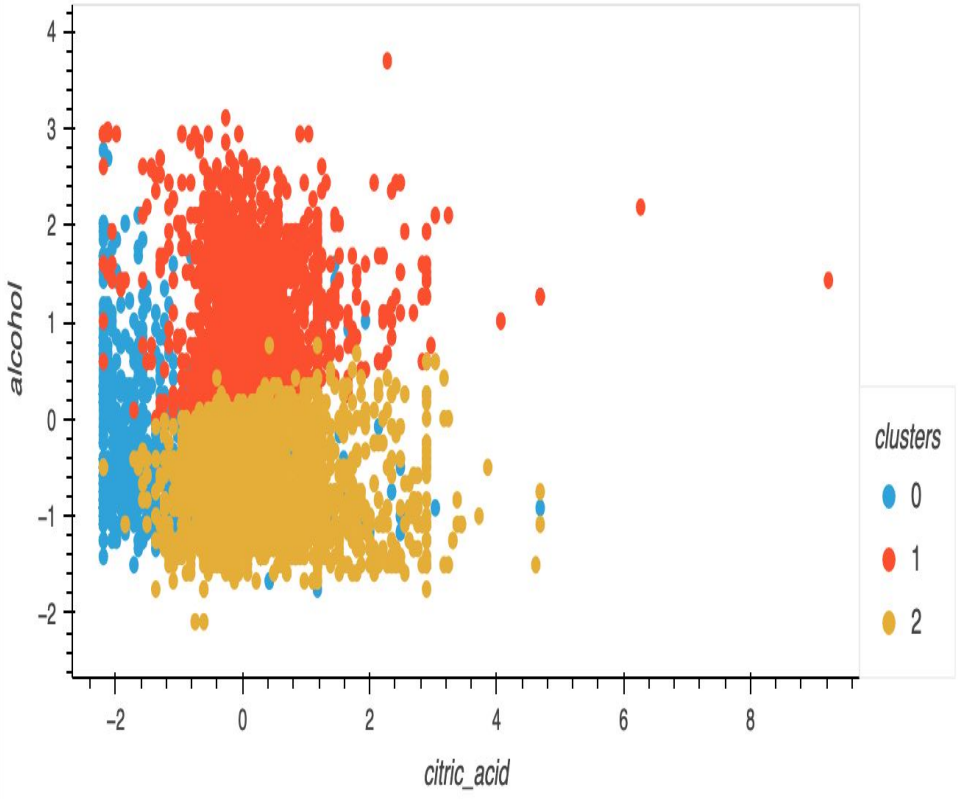


Citric acid with and without quality

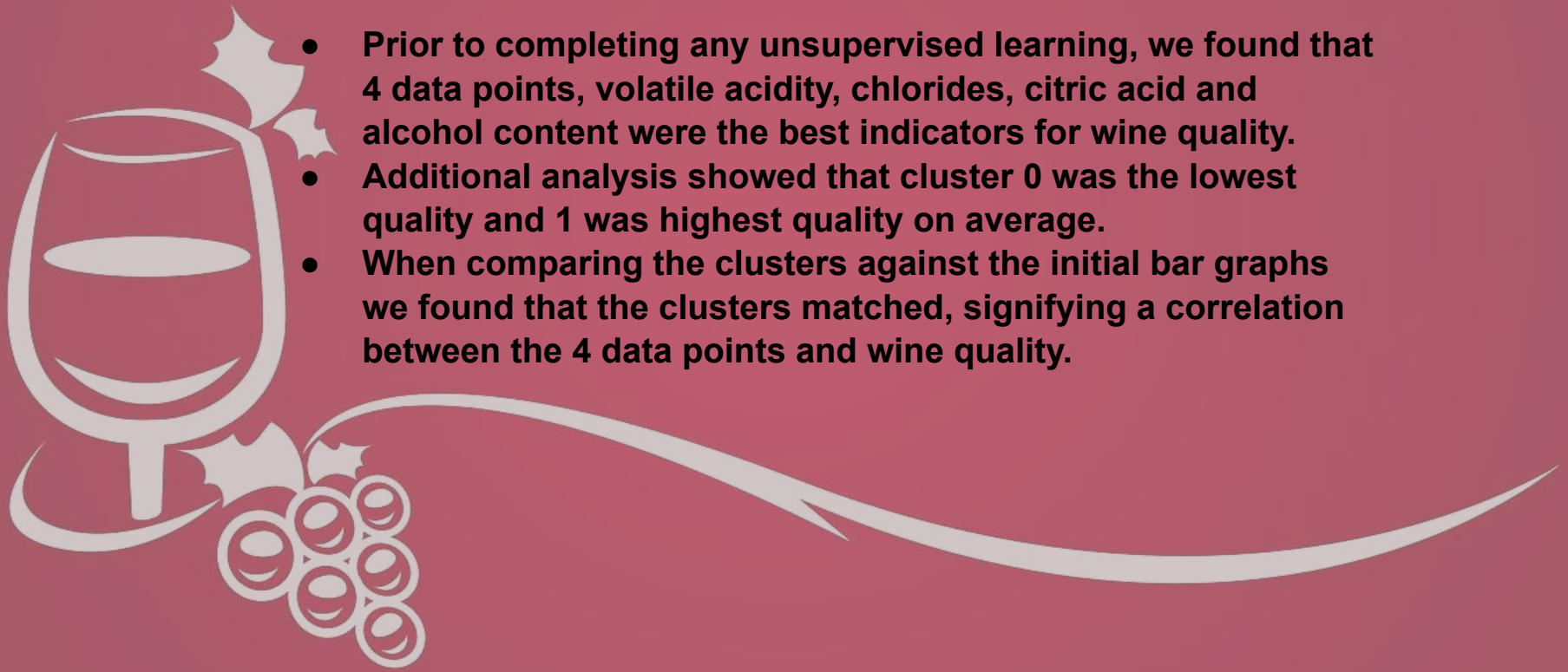
With quality



Without quality



Findings



- Prior to completing any unsupervised learning, we found that 4 data points, volatile acidity, chlorides, citric acid and alcohol content were the best indicators for wine quality.
- Additional analysis showed that cluster 0 was the lowest quality and 1 was highest quality on average.
- When comparing the clusters against the initial bar graphs we found that the clusters matched, signifying a correlation between the 4 data points and wine quality.

Averages and their significance

```
cluster0[["volatile_acidity", "citric_acid", "chlorides", "alcohol", "quality"]].describe()
```

✓ 0.0s

	volatile_acidity	citric_acid	chlorides	alcohol	quality
count	1125.000000	1125.000000	1125.000000	1125.000000	1125.000000
mean	0.614778	0.166462	0.092913	10.071274	5.344889
std	0.151973	0.145659	0.059399	0.868560	0.717215
min	0.200000	0.000000	0.028000	8.400000	3.000000
25%	0.520000	0.040000	0.069000	9.400000	5.000000
50%	0.600000	0.140000	0.080000	9.800000	5.000000
75%	0.690000	0.250000	0.094000	10.600000	6.000000
max	1.580000	1.000000	0.611000	13.800000	8.000000

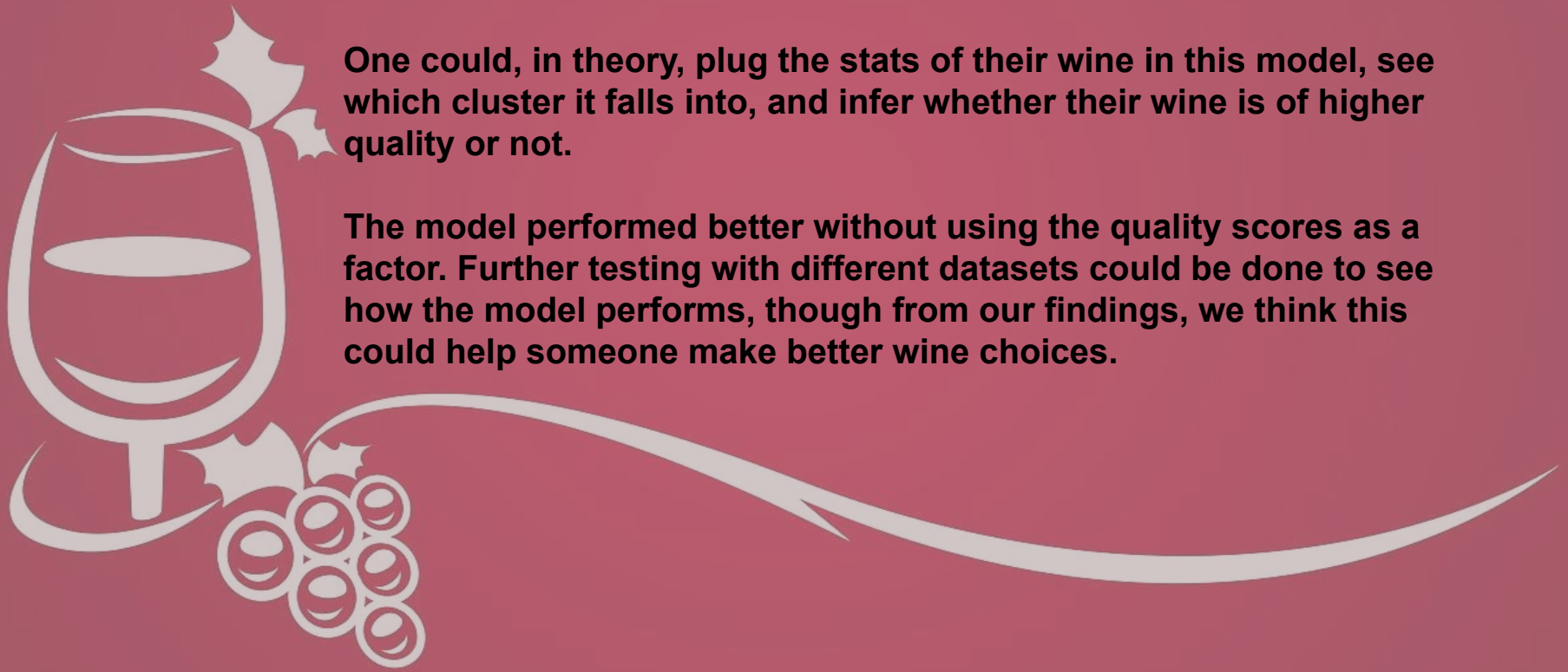
```
cluster1[["volatile_acidity", "citric_acid", "chlorides", "alcohol", "quality"]].describe()
```

✓ 0.0s

	volatile_acidity	citric_acid	chlorides	alcohol	quality
count	2266.000000	2266.000000	2266.000000	2266.000000	2266.000000
mean	0.289788	0.337683	0.041273	11.784406	6.317299
std	0.097400	0.109840	0.016708	0.762097	0.851142
min	0.080000	0.000000	0.009000	10.500000	3.000000
25%	0.220000	0.280000	0.031000	11.100000	6.000000
50%	0.280000	0.320000	0.037000	11.700000	6.000000
75%	0.340000	0.380000	0.046000	12.400000	7.000000
max	0.680000	1.660000	0.160000	14.900000	9.000000

- Our lowest average quality cluster, cluster 0, had more volatile acidity and chlorides, on average than cluster 1, our highest average quality cluster.
- Cluster 0 also had a lower average alcohol and citric acid content, which also matches up to our initial findings.
- The KMeans algorithm was able to cluster wines that match up to the original dataset without using the quality column
- Cluster 2 ended up being in between the two other clusters.

How can the model be used?



One could, in theory, plug the stats of their wine in this model, see which cluster it falls into, and infer whether their wine is of higher quality or not.

The model performed better without using the quality scores as a factor. Further testing with different datasets could be done to see how the model performs, though from our findings, we think this could help someone make better wine choices.

Source

https://www.kaggle.com/datasets/subhajournal/wine-quality-data-combined