# Red Wine Quality Analysis: Presentation of Findings

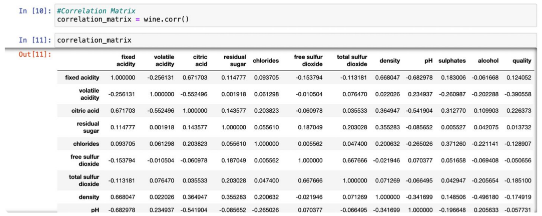
- Bhavya Reddy Patlolla

# Agenda

- Exploring Data
- Visualizations
- Identifying Additional Data/Variables
- Summary of Analysis

# Exploring Data

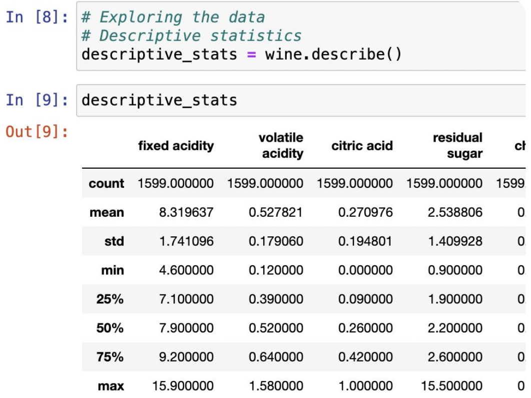
1. Loaded 'Red Wine Quality' data using pd.read_csv(), initiating the analysis process.
2. Utilized descriptive statistics (mean, median, etc.) for better comprehension of data distribution and central tendencies.
3. Employed a correlation matrix to pinpoint features strongly correlating with 'quality' and each other.
4. Descriptive stats and correlation matrix identified key features, aiding in hypothesis generation for wine quality factors.
5. These statistical measures provided a robust foundation for subsequent analysis, model building, and decision-making in the project.



```
In [10]: #Correlation Matrix
         correlation_matrix = wine.corr()

In [11]: correlation_matrix
Out[11]:
```

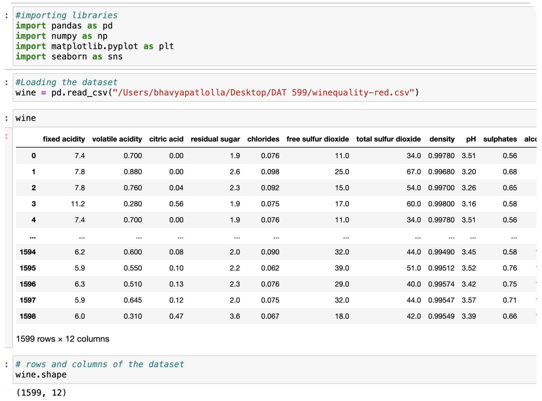
| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **fixed acidity** | 1.000000 | -0.256131 | 0.671703 | 0.114777 | 0.093705 | -0.153794 | -0.113181 | 0.668047 | -0.682978 | 0.183006 | -0.061668 | 0.124052 |
| **volatile acidity** | -0.256131 | 1.000000 | -0.552496 | 0.001918 | 0.061298 | -0.010504 | 0.076470 | 0.022026 | 0.234937 | -0.260987 | -0.202288 | -0.390558 |
| **citric acid** | 0.671703 | -0.552496 | 1.000000 | 0.143577 | 0.203823 | -0.060978 | 0.035533 | 0.364947 | -0.541904 | 0.312770 | 0.109903 | 0.226373 |
| **residual sugar** | 0.114777 | 0.001918 | 0.143577 | 1.000000 | 0.055610 | 0.187049 | 0.203028 | 0.355283 | -0.085652 | 0.005527 | 0.042075 | 0.013732 |
| **chlorides** | 0.093705 | 0.061298 | 0.203823 | 0.055610 | 1.000000 | 0.005562 | 0.047400 | 0.200632 | -0.265026 | 0.371260 | -0.221141 | -0.128907 |
| **free sulfur dioxide** | -0.153794 | -0.010504 | -0.060978 | 0.187049 | 0.005562 | 1.000000 | 0.667666 | -0.021946 | 0.070377 | 0.051658 | -0.069408 | -0.050656 |
| **total sulfur dioxide** | -0.113181 | 0.076470 | 0.035533 | 0.203028 | 0.047400 | 0.667666 | 1.000000 | 0.071269 | -0.066495 | 0.042947 | -0.205654 | -0.185100 |
| **density** | 0.668047 | 0.022026 | 0.364947 | 0.355283 | 0.200632 | -0.021946 | 0.071269 | 1.000000 | -0.341699 | 0.148506 | -0.496180 | -0.174919 |
| **pH** | -0.682978 | 0.234937 | -0.541904 | -0.085652 | -0.265026 | 0.070377 | -0.066495 | -0.341699 | 1.000000 | -0.196648 | 0.205633 | -0.057731 |



```
In [8]: # Exploring the data
        # Descriptive statistics
        descriptive_stats = wine.describe()

In [9]: descriptive_stats
Out[9]:
```

| | fixed acidity | volatile acidity | citric acid | residual sugar | ch |
|---|---|---|---|---|---|
| **count** | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599 |
| **mean** | 8.319637 | 0.527821 | 0.270976 | 2.538806 | 0 |
| **std** | 1.741096 | 0.179060 | 0.194801 | 1.409928 | 0 |
| **min** | 4.600000 | 0.120000 | 0.000000 | 0.900000 | 0 |
| **25%** | 7.100000 | 0.390000 | 0.090000 | 1.900000 | 0 |
| **50%** | 7.900000 | 0.520000 | 0.260000 | 2.200000 | 0 |
| **75%** | 9.200000 | 0.640000 | 0.420000 | 2.600000 | 0 |
| **max** | 15.900000 | 1.580000 | 1.000000 | 15.500000 | 0 |



```
: #importing libraries
  import pandas as pd
  import numpy as np
  import matplotlib.pyplot as plt
  import seaborn as sns

: #Loading the dataset
  wine = pd.read_csv("/Users/bhavyapatlolla/Desktop/DAT 599/winequality-red.csv")

: wine
:
```

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 7.4 | 0.700 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.99780 | 3.51 | 0.56 | |
| **1** | 7.8 | 0.880 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.99680 | 3.20 | 0.68 | |
| **2** | 7.8 | 0.760 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.99700 | 3.26 | 0.65 | |
| **3** | 11.2 | 0.280 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.99800 | 3.16 | 0.58 | |
| **4** | 7.4 | 0.700 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.99780 | 3.51 | 0.56 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1594** | 6.2 | 0.600 | 0.08 | 2.0 | 0.090 | 32.0 | 44.0 | 0.99490 | 3.45 | 0.58 | |
| **1595** | 5.9 | 0.550 | 0.10 | 2.2 | 0.062 | 39.0 | 51.0 | 0.99512 | 3.52 | 0.76 | |
| **1596** | 6.3 | 0.510 | 0.13 | 2.3 | 0.076 | 29.0 | 40.0 | 0.99574 | 3.42 | 0.75 | |
| **1597** | 5.9 | 0.645 | 0.12 | 2.0 | 0.075 | 32.0 | 44.0 | 0.99547 | 3.57 | 0.71 | |
| **1598** | 6.0 | 0.310 | 0.47 | 3.6 | 0.067 | 18.0 | 42.0 | 0.99549 | 3.39 | 0.66 | |

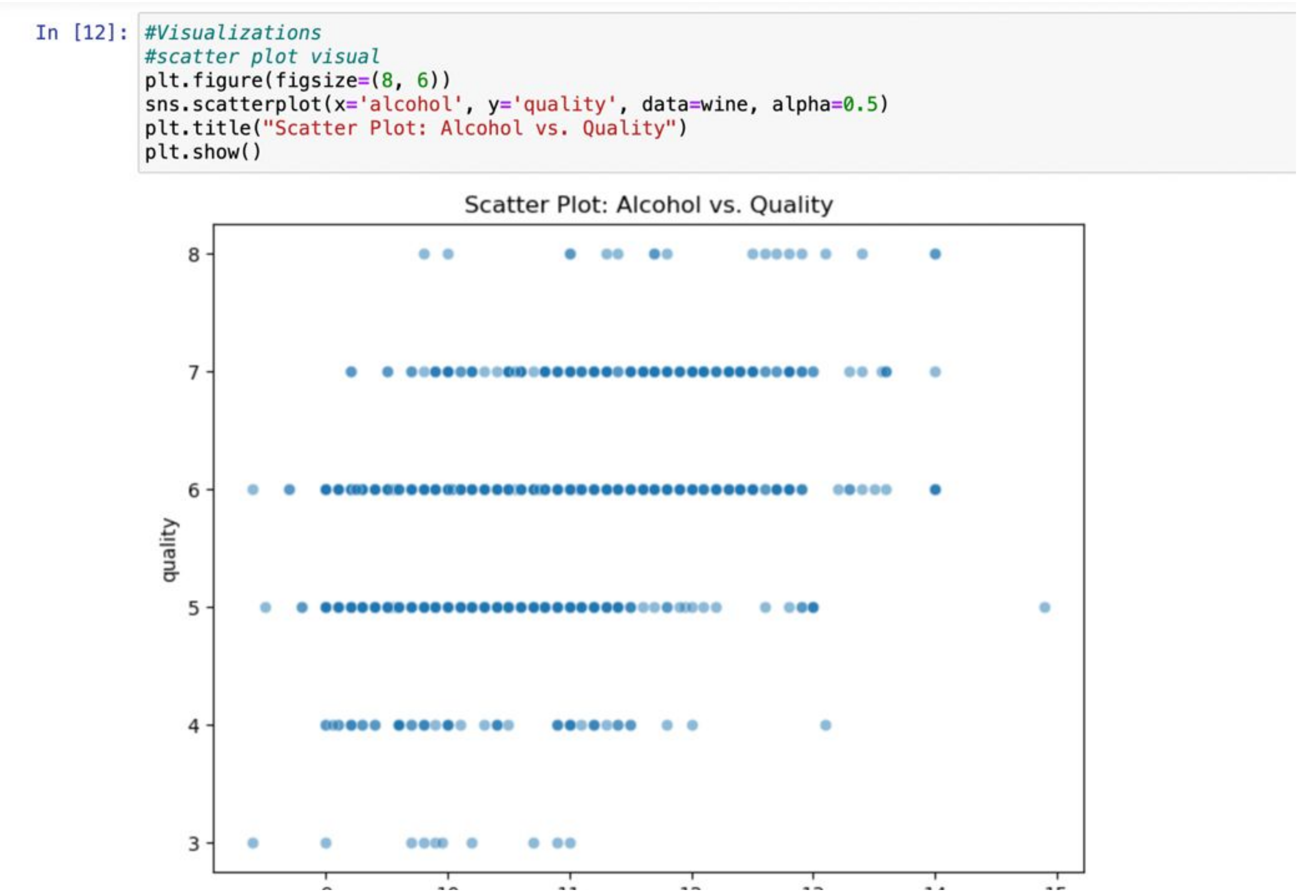
1599 rows × 12 columns

```
: # rows and columns of the dataset
  wine.shape

  (1599, 12)
```

# Visualizations - Scatter Plot

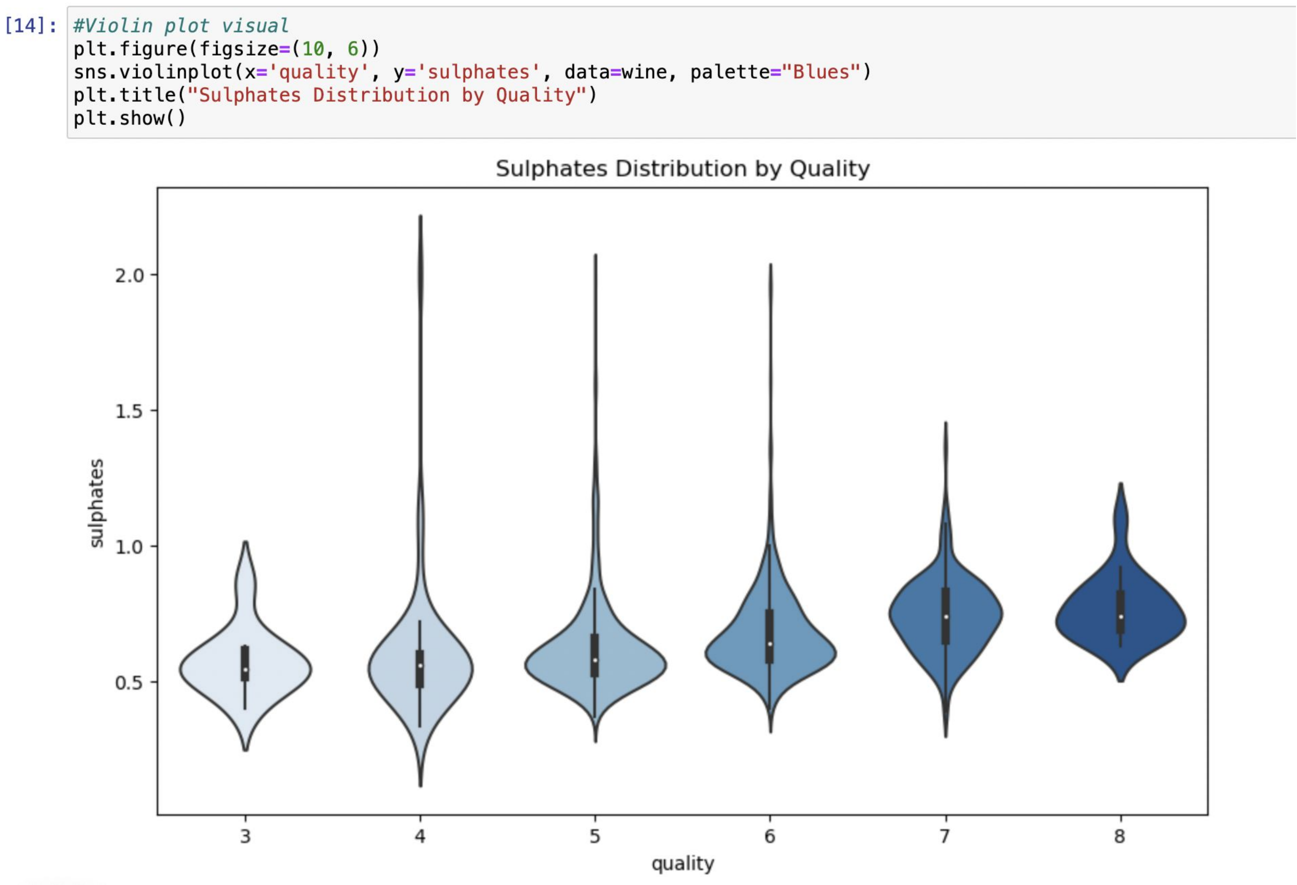
1. Worked on a scatter plot visual to explore relationships between variables.
2. Utilized the plot to reveal different features during data analysis.
3. Chose to represent alcohol quality in the scatter plot to uncover linear or non-linear patterns.



```
In [12]: #Visualizations
         #scatter plot visual
         plt.figure(figsize=(8, 6))
         sns.scatterplot(x='alcohol', y='quality', data=wine, alpha=0.5)
         plt.title("Scatter Plot: Alcohol vs. Quality")
         plt.show()
```

# Violin plot Visual

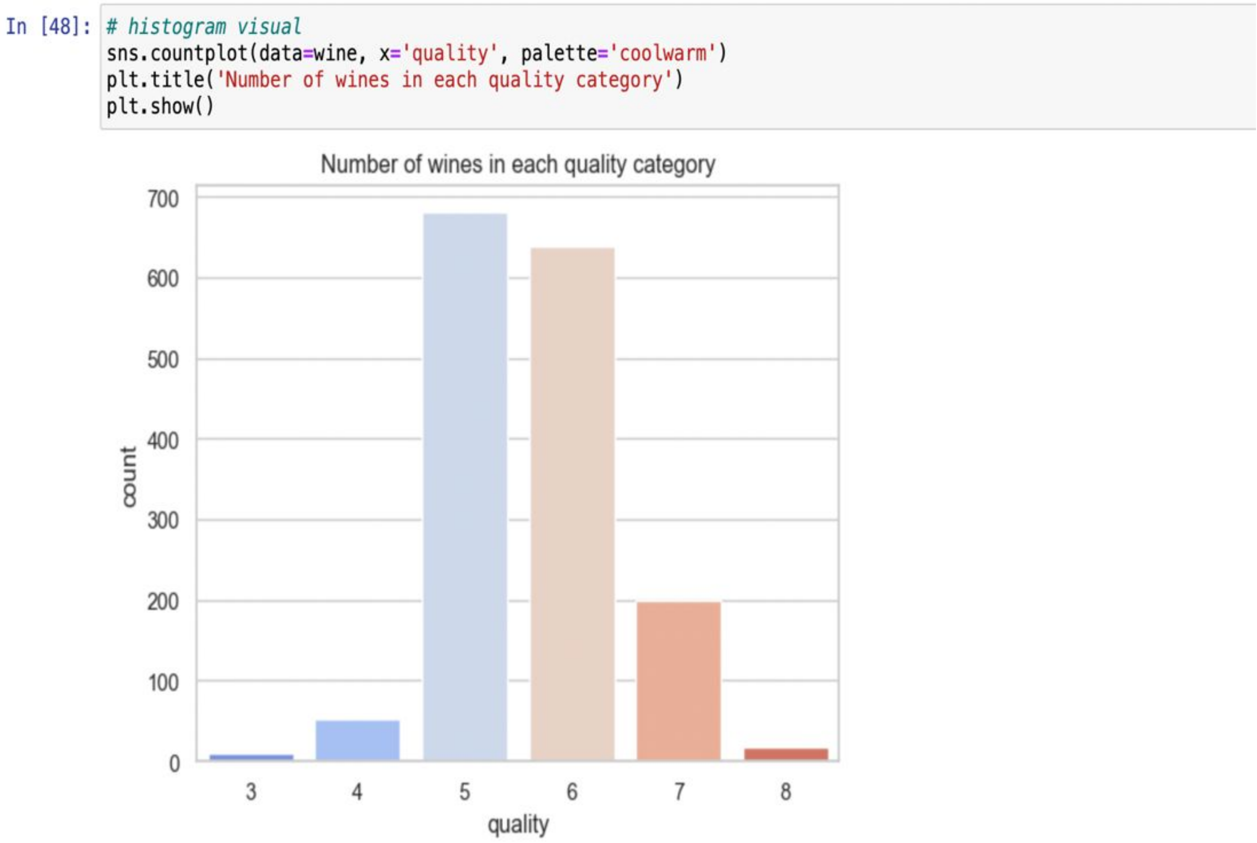
- Employed a violin plot to depict the distribution of sulphate quality levels.
- Chose this plot for its clear display of detailed feature distributions.
- Gained insights into the range, central tendency, and shape of feature distributions.
- Utilized the plot to assess variations across wine quality levels.
- Investigated for significant differences in feature distributions between different quality categories.



[14]:
```
#Violin plot visual
plt.figure(figsize=(10, 6))
sns.violinplot(x='quality', y='sulphates', data=wine, palette="Blues")
plt.title("Sulphates Distribution by Quality")
plt.show()
```

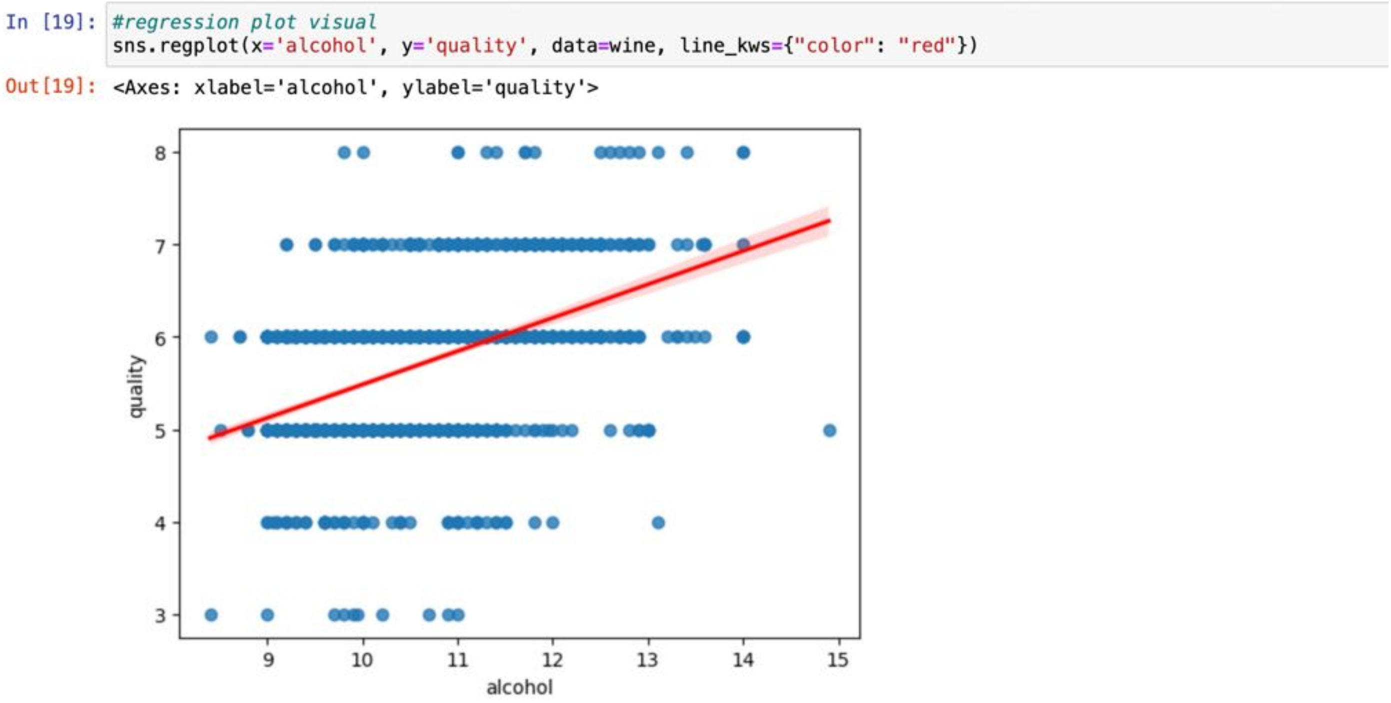
# Histogram Visual

- Created a histogram visual to display the distribution of alcohol content in the dataset.
- Used the histogram to represent the frequency of alcohol content occurrences.
- Utilized the plot to gain insights into the shape of the feature's distribution.
- Examined central tendencies of the alcohol content distribution through the histogram.
- Investigated the presence of outliers to understand characteristics and their association with wine quality.



```
In [48]: # histogram visual
         sns.countplot(data=wine, x='quality', palette='coolwarm')
         plt.title('Number of wines in each quality category')
         plt.show()
```

# Regression Plot Visual



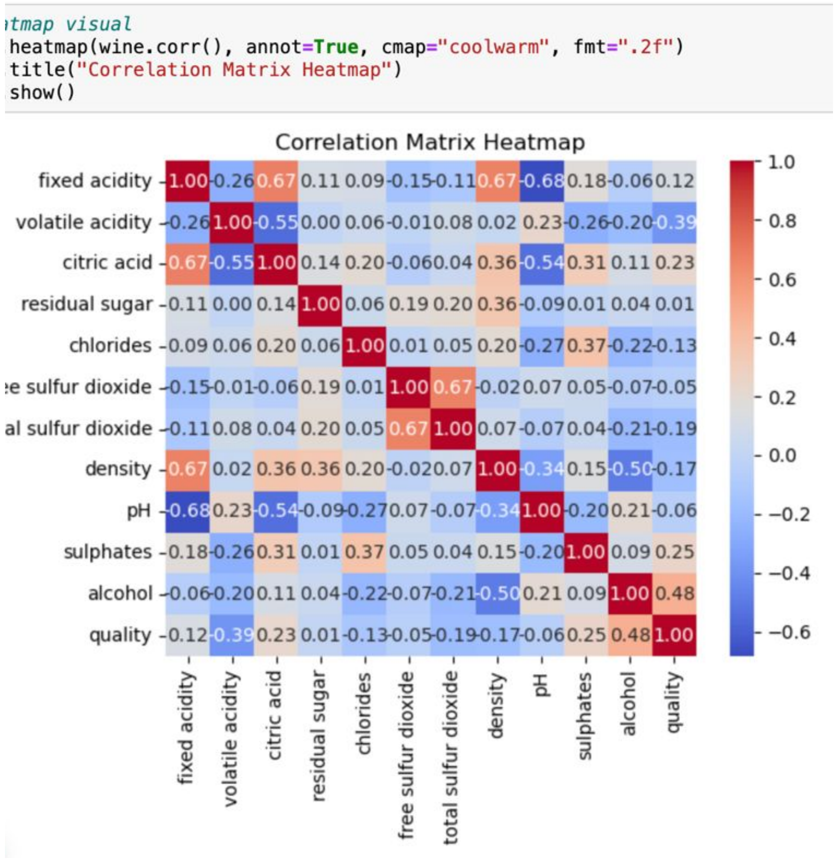
```
In [19]: #regression plot visual
         sns.regplot(x='alcohol', y='quality', data=wine, line_kws={"color": "red"})

Out[19]: <Axes: xlabel='alcohol', ylabel='quality'>
```

1. Explored the relationship between alcohol content and wine quality using the red wine quality dataset.
2. Utilized a regression plot to visually represent the relationship between the two variables.
3. Determined whether a positive or negative correlation exists between alcohol content and wine quality.
4. Evaluated the strength and direction of the feature and wine quality relationship.
5. Incorporated a regression line to illustrate the trend and facilitate predictions, aiding in understanding how changes in the predictor variable impact the target variable.

# Heat Map Visual

- Visualized the correlation matrix as a heatmap for enhanced representation.
- Used color coding to distinguish and signify the strength and direction of correlations.
- Utilized the heatmap to identify both strong and weak correlations among features.
- Applied the heatmap as a valuable tool for guiding feature selection.
- Leveraged the plot to uncover patterns and inform subsequent analysis and modeling decisions in the exploration of the red wine quality dataset.



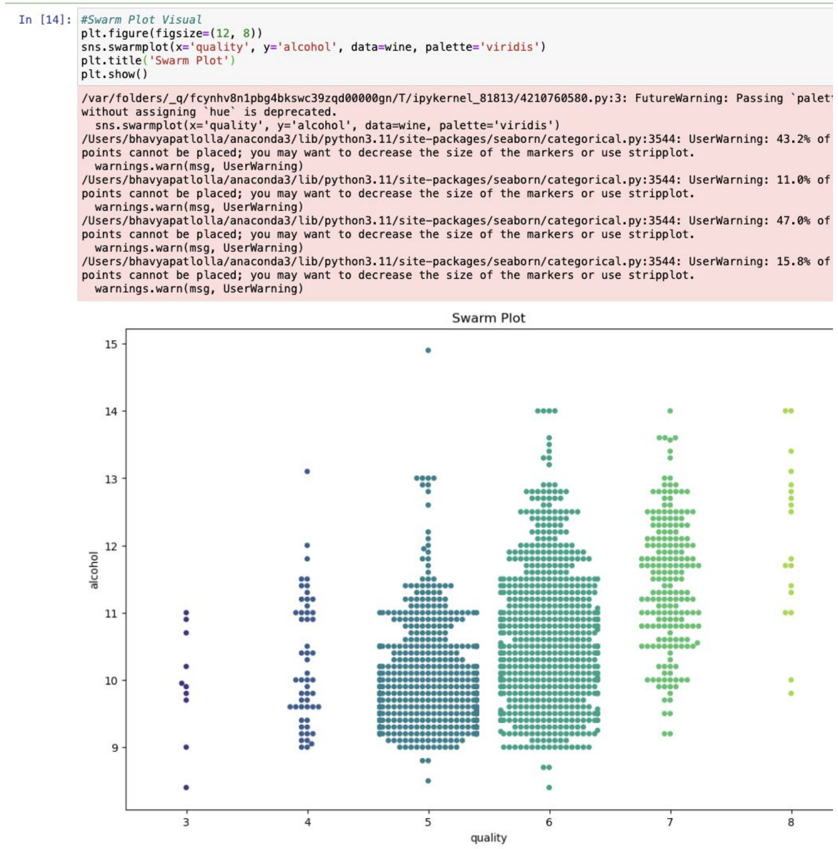
```
tmap visual
heatmap(wine.corr(), annot=True, cmap="coolwarm", fmt=".2f")
title("Correlation Matrix Heatmap")
show()
```

Correlation Matrix Heatmap

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fixed acidity | 1.00 | -0.26 | 0.67 | 0.11 | 0.09 | -0.15 | -0.11 | 0.67 | -0.68 | 0.18 | -0.06 | 0.12 |
| volatile acidity | -0.26 | 1.00 | -0.55 | 0.00 | 0.06 | -0.01 | 0.08 | 0.02 | 0.23 | -0.26 | -0.20 | -0.39 |
| citric acid | 0.67 | -0.55 | 1.00 | 0.14 | 0.20 | -0.06 | 0.04 | 0.36 | -0.54 | 0.31 | 0.11 | 0.23 |
| residual sugar | 0.11 | 0.00 | 0.14 | 1.00 | 0.06 | 0.19 | 0.20 | 0.36 | -0.09 | 0.01 | 0.04 | 0.01 |
| chlorides | 0.09 | 0.06 | 0.20 | 0.06 | 1.00 | 0.01 | 0.05 | 0.20 | -0.27 | 0.37 | -0.22 | -0.13 |
| ee sulfur dioxide | -0.15 | -0.01 | -0.06 | 0.19 | 0.01 | 1.00 | 0.67 | -0.02 | 0.07 | 0.05 | -0.07 | -0.05 |
| al sulfur dioxide | -0.11 | 0.08 | 0.04 | 0.20 | 0.05 | 0.67 | 1.00 | 0.07 | -0.07 | 0.04 | -0.21 | -0.19 |
| density | 0.67 | 0.02 | 0.36 | 0.36 | 0.20 | -0.02 | 0.07 | 1.00 | -0.34 | 0.15 | -0.50 | -0.17 |
| pH | -0.68 | 0.23 | -0.54 | -0.09 | -0.27 | 0.07 | -0.07 | -0.34 | 1.00 | -0.20 | 0.21 | -0.06 |
| sulphates | 0.18 | -0.26 | 0.31 | 0.01 | 0.37 | 0.05 | 0.04 | 0.15 | -0.20 | 1.00 | 0.09 | 0.25 |
| alcohol | -0.06 | -0.20 | 0.11 | 0.04 | -0.22 | -0.07 | -0.21 | -0.50 | 0.21 | 0.09 | 1.00 | 0.48 |
| quality | 0.12 | -0.39 | 0.23 | 0.01 | -0.13 | -0.05 | -0.19 | -0.17 | -0.06 | 0.25 | 0.48 | 1.00 |

# Swarm Plot Visual

- Utilized a Swarm Plot to showcase the distribution of alcohol content in various quality categories within the wine dataset.
- Achieved a detailed analysis of variations in alcohol content within each quality category through the plot's point-by-point correspondence with individual wine samples.
- Found the Swarm Plot particularly effective in identifying patterns and trends in the relationship between alcohol content and wine quality.
- The visualization is especially user-friendly, aiding in the easy comprehension of the dataset's complexities.
- Enhanced the visual presentation by using a "viridis" color scheme, providing a clear display of alcohol levels for wines across different quality classifications.

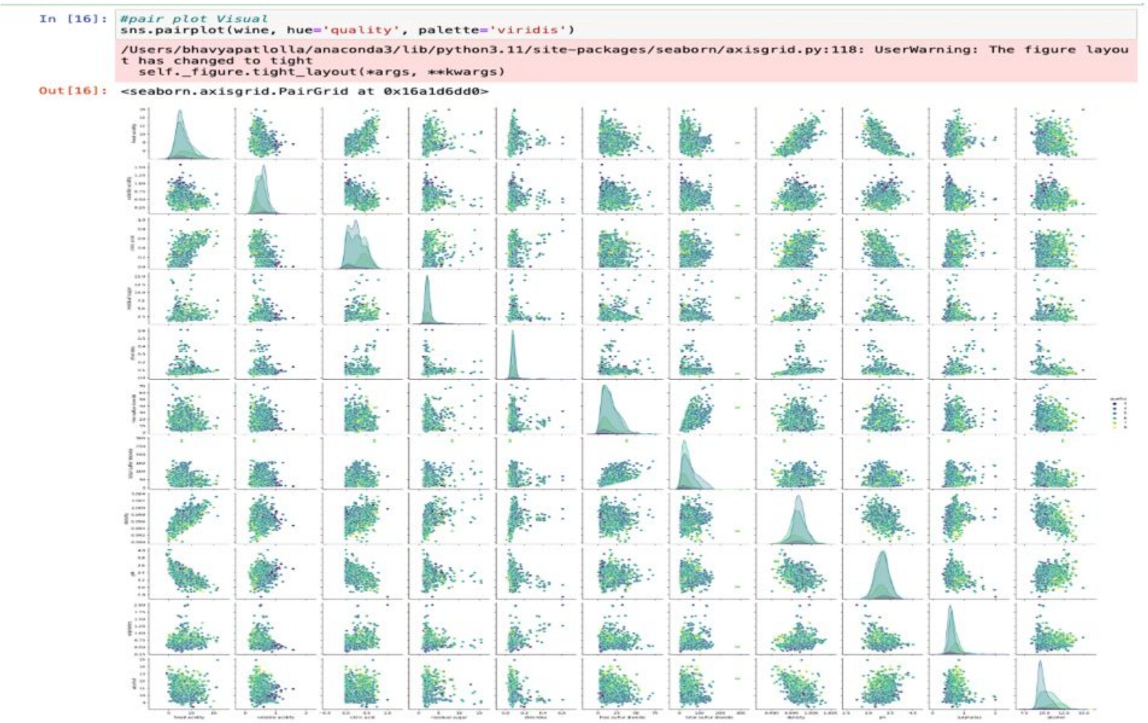


```
In [14]: #Swarm Plot Visual
         plt.figure(figsize=(12, 8))
         sns.swarmplot(x='quality', y='alcohol', data=wine, palette='viridis')
         plt.title('Swarm Plot')
         plt.show()
```

```
/var/folders/_q/fcynhv8n1pbg4bkswc39zqd00000gn/T/ipykernel_81813/4210760580.py:3: FutureWarning: Passing `palet
without assigning `hue` is deprecated.
  sns.swarmplot(x='quality', y='alcohol', data=wine, palette='viridis')
/Users/bhavyapatlolla/anaconda3/lib/python3.11/site-packages/seaborn/categorical.py:3544: UserWarning: 43.2% of
points cannot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/Users/bhavyapatlolla/anaconda3/lib/python3.11/site-packages/seaborn/categorical.py:3544: UserWarning: 11.0% of
points cannot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/Users/bhavyapatlolla/anaconda3/lib/python3.11/site-packages/seaborn/categorical.py:3544: UserWarning: 47.0% of
points cannot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
/Users/bhavyapatlolla/anaconda3/lib/python3.11/site-packages/seaborn/categorical.py:3544: UserWarning: 15.8% of
points cannot be placed; you may want to decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```

# Pair Plot Visual

- Employed a pair plot to gain an in-depth understanding of connections between various features in the wine dataset.
- Diagonal subplots were utilized to display the distribution of individual components within the dataset.
- Each subplot presented a scatterplot, illustrating the interaction between pairs of variables in the dataset.
- Enhanced visual clarity by implementing the 'viridis' color scheme, facilitating the distinction between different groups and aiding in pattern recognition.
- Identified two key benefits of using the pair plot: investigating correlations between variables and gaining insights into how different dataset attributes interact with each other.

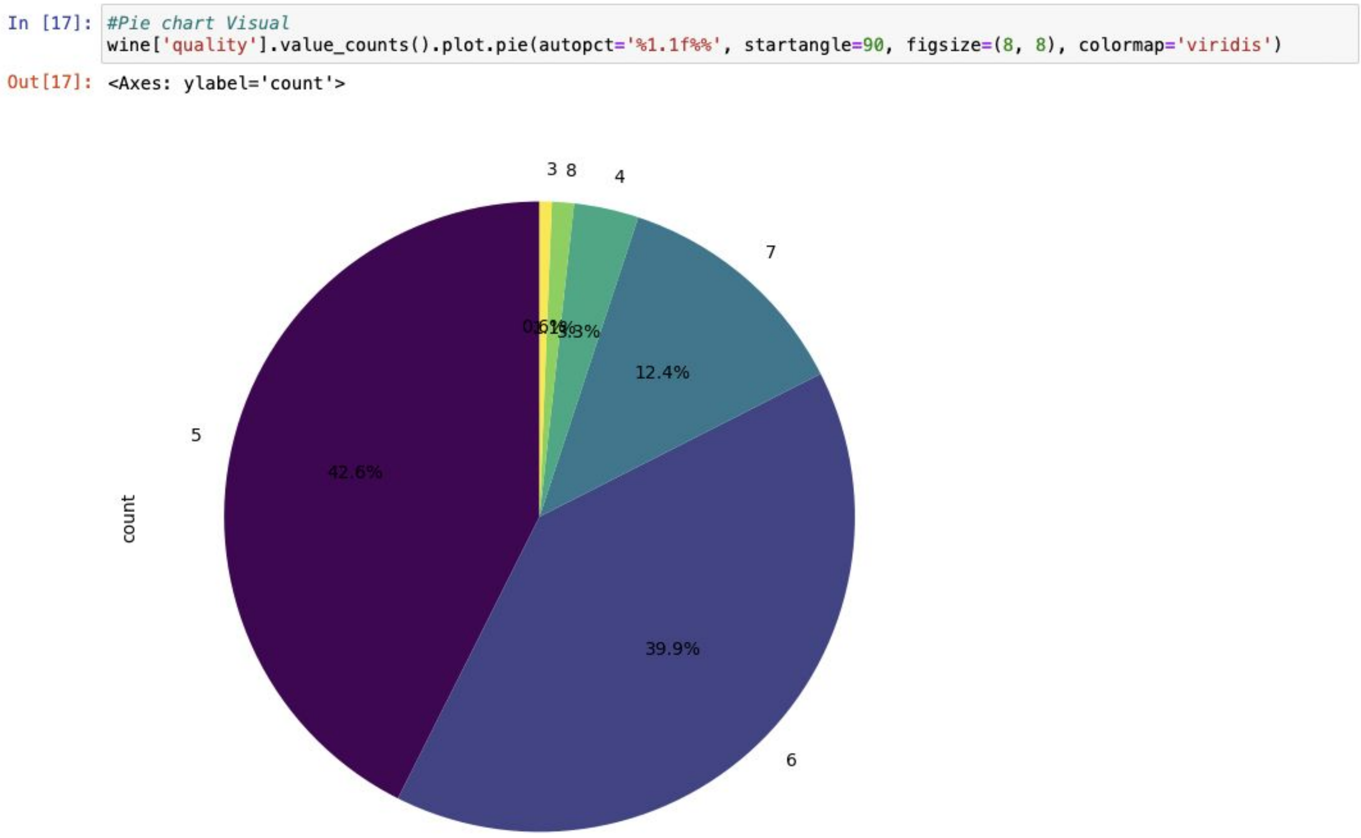


```
In [16]: #pair plot Visual
         sns.pairplot(wine, hue='quality', palette='viridis')

         /Users/bhavyapatlolla/anaconda3/lib/python3.11/site-packages/seaborn/axisgrid.py:118: UserWarning: The figure layou
         t has changed to tight
           self._figure.tight_layout(*args, **kwargs)

Out[16]: <seaborn.axisgrid.PairGrid at 0x16a1d6dd0>
```

# Pie Chart Visual



```
In [17]: #Pie chart Visual
         wine['quality'].value_counts().plot.pie(autopct='%1.1f%%', startangle=90, figsize=(8, 8), colormap='viridis')

Out[17]: <Axes: ylabel='count'>
```

- Utilized a pie chart to represent the distribution of wine quality categories throughout the dataset.
- Each wedge in the pie chart corresponds to a different quality level.
- Emphasized the 'high-quality' category by subtly separating it using the explode parameter.
- Applied the 'Viridis' color palette for visual appeal and clarity in distinguishing categories.
- Displayed percentage labels using the 'autopct' parameter, offering a concise and easily understandable summary of the categorical distribution of wine quality levels in the dataset.

# Identifying Additional Data/Variables

- Identified key variables such as Grape Variety, Winery Information, Geographic Origin, Vintage Year, Wine Aging, Climatic Data, pH Levels, Yeast Strain, Sensory Evaluation, and Chemical Analysis.
- Emphasized the importance of understanding grape variety, winery details, geographic origin, vintage year, wine aging, climatic data, pH levels, yeast strain, sensory evaluation, chemical analysis, food pairings, market and pricing information, and consumer reviews.
- Highlighted how these variables can significantly influence wine quality and contribute to a more thorough analysis.
- Acknowledged that the availability of such data may vary and suggested careful consideration of including additional variables based on project objectives.
- Concluded that incorporating these additional data points could lead to a more comprehensive assessment of factors influencing red wine quality, ultimately aligning with the study's objectives.

# Summary of Analysis

- Conducted a comprehensive analysis of the red wine quality dataset, revealing valuable insights into its characteristics and potential influencing factors.
- Utilized statistical measures and visualizations for an in-depth understanding, highlighting key features and relationships contributing to wine quality.
- Identified the importance of considering additional variables for future research, enhancing the assessment of factors influencing red wine quality.
- Proposed integrating diverse data on grape variety, winery details, geographic origin, vintage year, wine aging, climatic data, pH levels, yeast strain, sensory evaluation, chemical analysis, food pairings, market dynamics, and consumer reviews.
- Established the analysis as a foundation for informed decision-making in red wine quality, suggesting potential solutions and future research directions.

# Conclusion

- The analysis lays the foundation for informed decision-making in red wine quality.
- Proposed solutions include incorporating identified variables and conducting further research to validate their impact.
- Advocated careful consideration of data availability, alignment with objectives, and a balanced approach to including diverse data points.
- The study serves as a model for approaching wine quality analysis, encouraging researchers to explore the complexities of winemaking for a comprehensive understanding.

# Thank you