RED -WINE QUALITY ANALYSIS

Using Exploratory Data Analysis

Presented by MD. SHAKIB

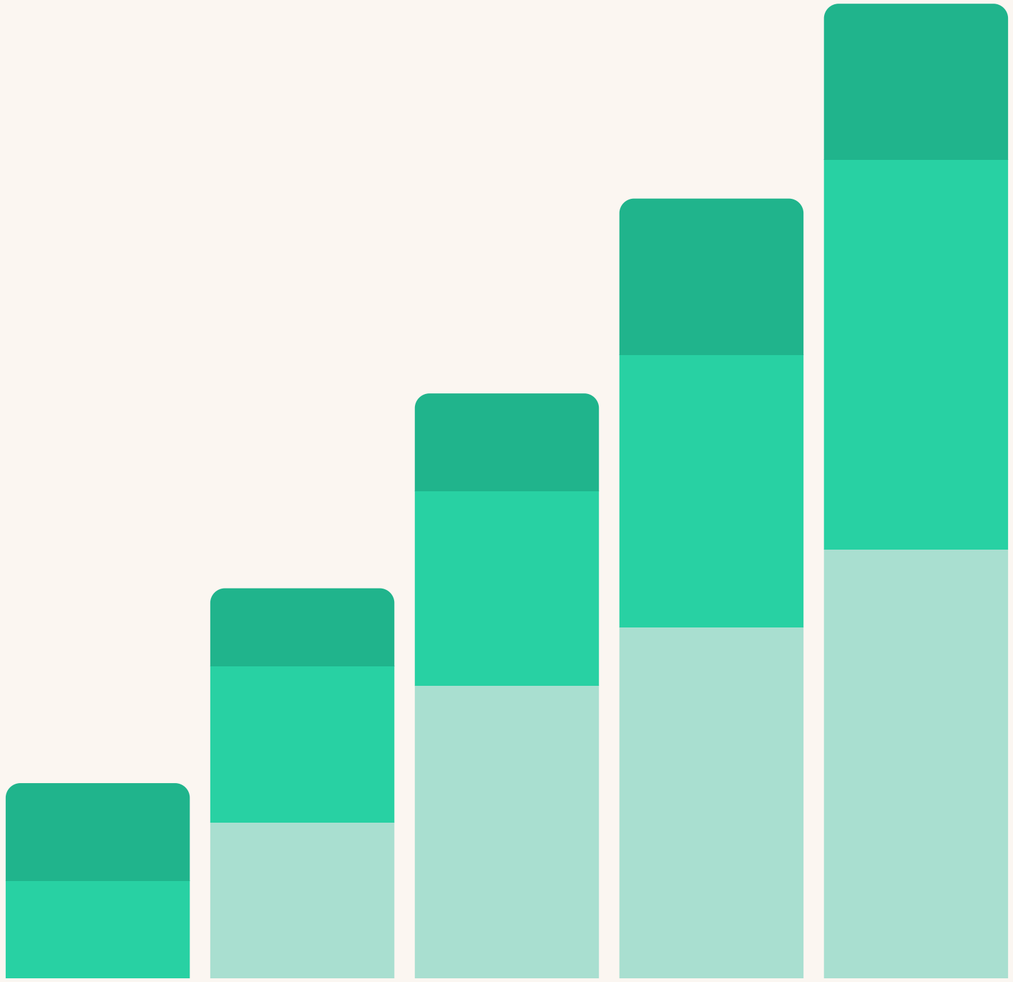
Dataset Description

Input variables (based on physicochemical tests):

- 1 - fixed acidity
- 2 - volatile acidity
- 3 - citric acid
- 4 - residual sugar
- 5 - chlorides
- 6 - free sulfur dioxide
- 7 - total sulfur dioxide
- 8 - density
- 9 - pH
- 10 - sulphates
- 11 - alcohol

Output variable (based on sensory data):

- 12 - quality (score between 0 and 10)



Data Exploratory Analysis

- Performing exploratory data analysis (EDA) in Red-Wine Quality Data involves several steps to understand the data, identify patterns, detect anomalies, and form hypotheses for further analysis.

```
import pandas as pd
df = pd.read_csv('winequality-red.csv', delimiter=';')
df.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
```

Data Cleaning

✅ Checked for missing values using df.isnull().sum()

➤ Result: No missing values found in the dataset

✅ Checked and removed duplicates using df.duplicated().sum()

➤ Found and removed 240 duplicate rows.

```
]: #Finding Duplicate Records
   df[df.duplicated()]
```

]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
4	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
11	7.5	0.500	0.36	6.10	0.071	17.0	102.0	0.99780	3.35	0.80	10.5	5
27	7.9	0.430	0.21	1.60	0.106	10.0	37.0	0.99660	3.17	0.91	9.5	5
40	7.3	0.450	0.36	5.90	0.074	12.0	87.0	0.99780	3.33	0.83	10.5	5
65	7.2	0.725	0.05	4.65	0.086	4.0	11.0	0.99620	3.41	0.39	10.9	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1563	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.54	10.1	5
1564	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.54	10.1	5
1567	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.54	10.1	5
1581	6.2	0.560	0.09	1.70	0.053	24.0	32.0	0.99402	3.54	0.60	11.3	5
1596	6.3	0.510	0.13	2.30	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6

240 rows × 12 columns

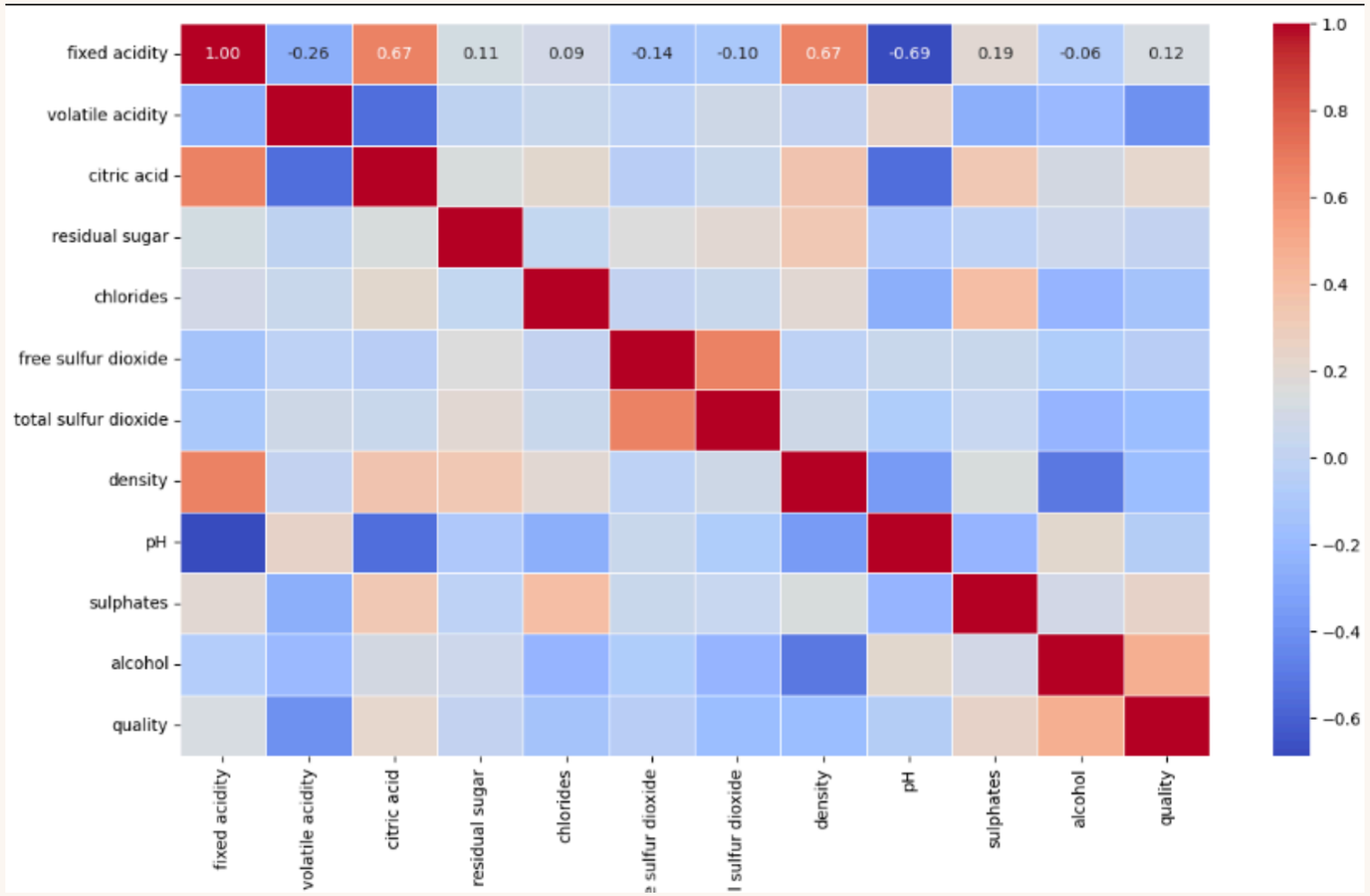
```
]: #Removing Duplicate Records
   df.drop_duplicates(inplace=True)
```

```
]: df.shape #after removing duplicates
```

```
]: (1359, 12)
```

Correlation HeatMap

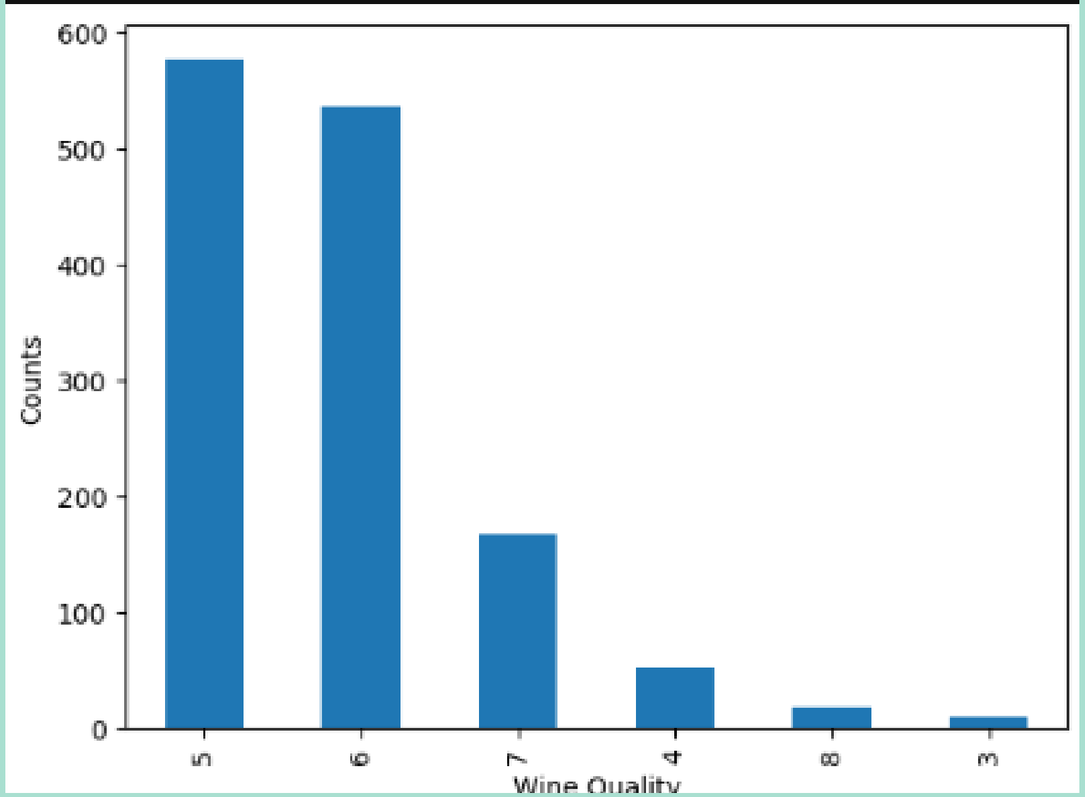
- isualized the correlation matrix using Seaborn heatmap with annotations.
- Identified how each feature correlates with the wine quality score.
- 🔼 Positively Correlated with Quality:
- Alcohol – strongest positive correlation with quality
- Citric Acid – slight positive correlation
- 🔽 Negatively Correlated with Quality:
- Volatile Acidity – strong negative correlation
- Density – slight negative correlation
- Total Sulfur Dioxide – slight negative impact



Countplot for Wine Quality Distribution

- Displays the frequency of each wine quality category.
- Helps identify the most common quality ratings in the dataset.

```
df.quality.value_counts().plot(kind = 'bar')
plt.xlabel("Wine Quality")
plt.ylabel("Counts")
plt.show()
```

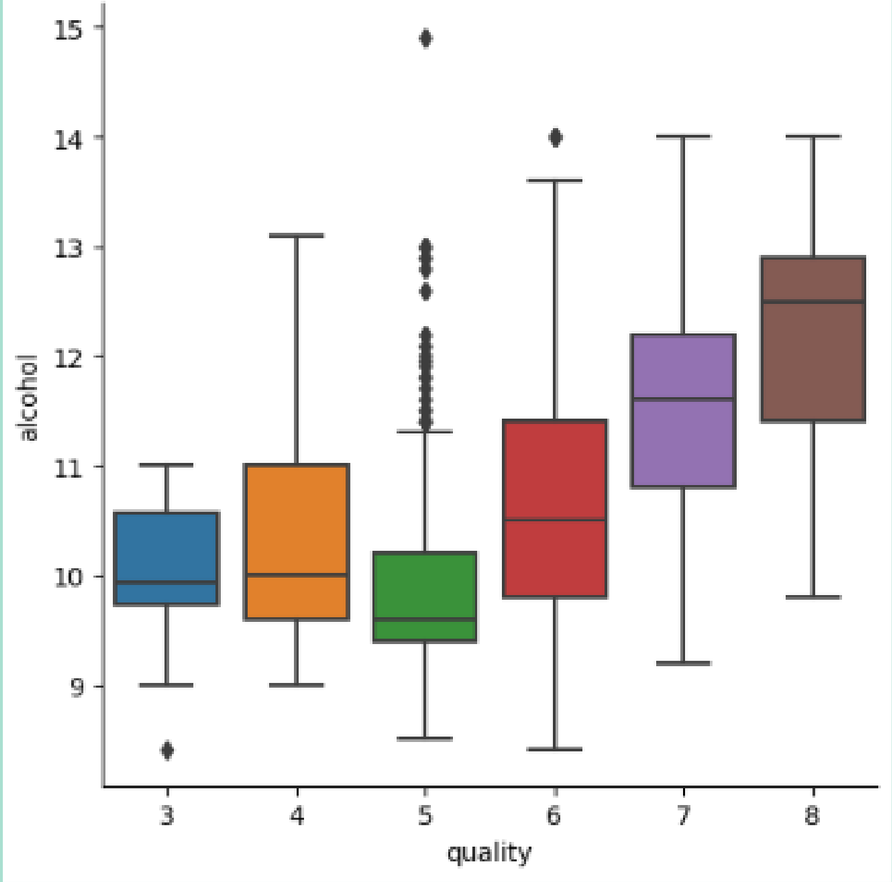


Boxplot of Quality vs Other Features

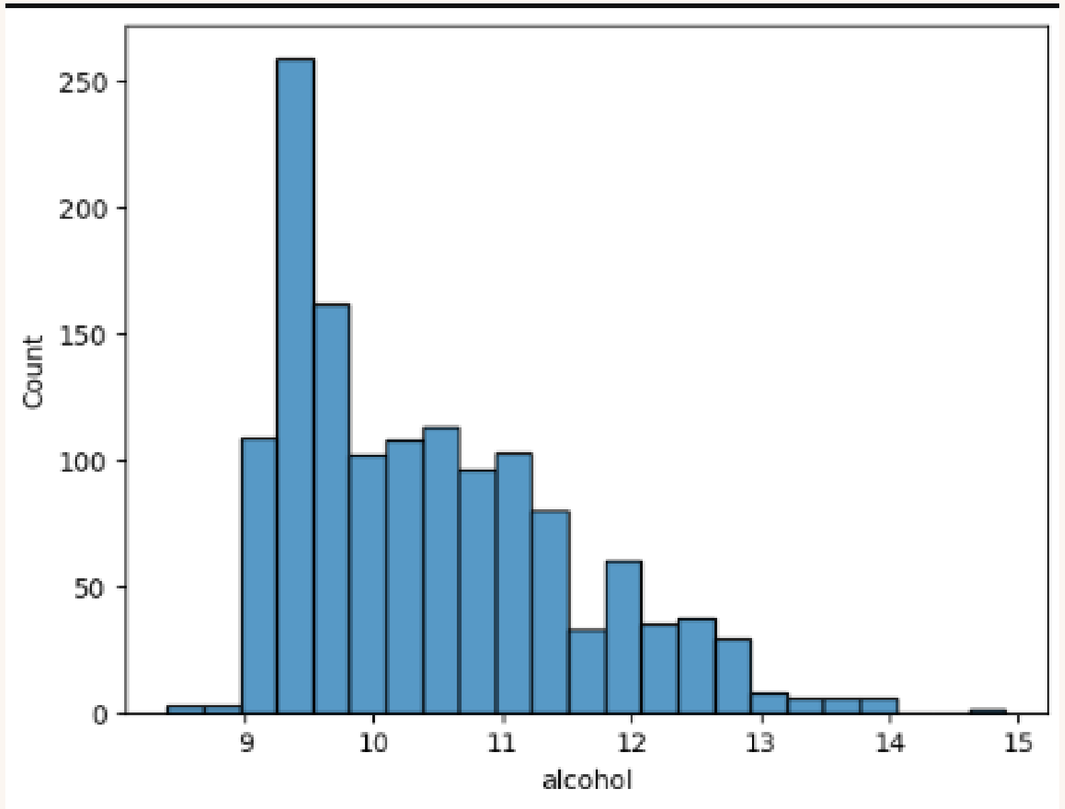
- Visualizes the distribution of features across different quality ratings.
- Highlights variations, medians, and spread for each quality group.

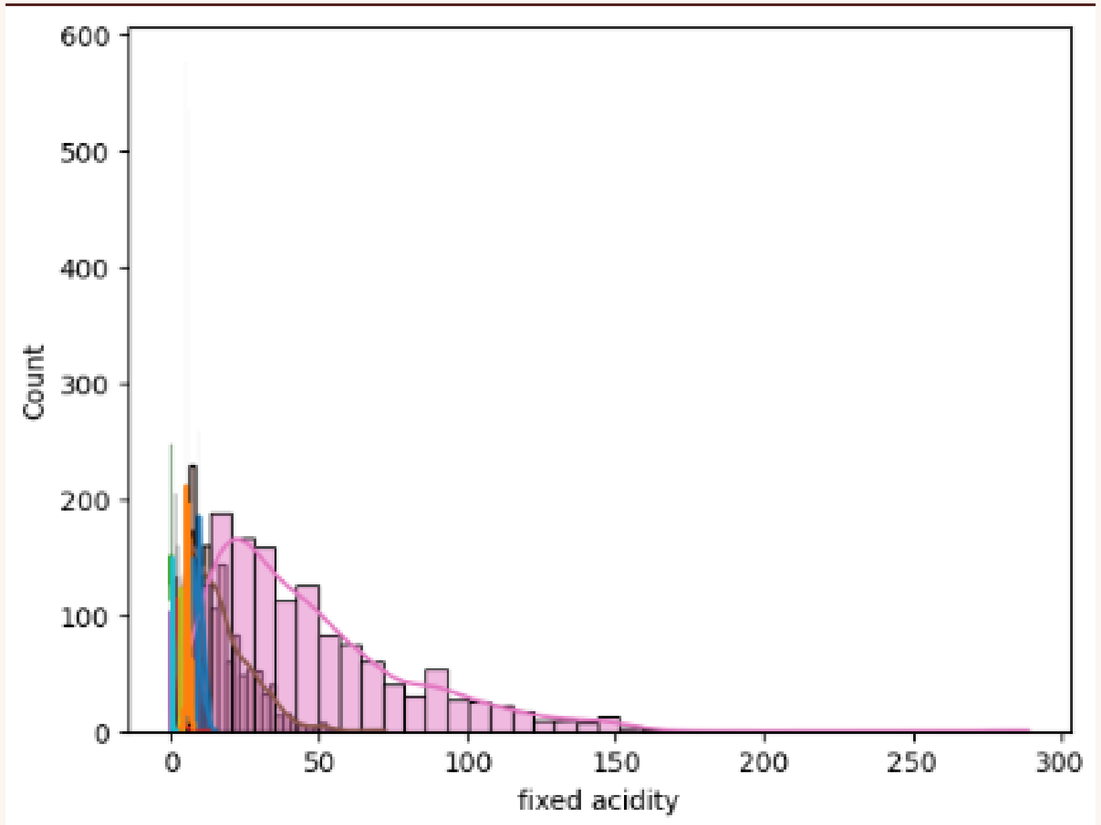
```
##categorical Plot
sns.catplot(x='quality', y='alcohol', data=df, kind="box")
```

```
<seaborn.axisgrid.FacetGrid at 0x236a1dd48d0>
```



Univariate & Multivariate Analysis



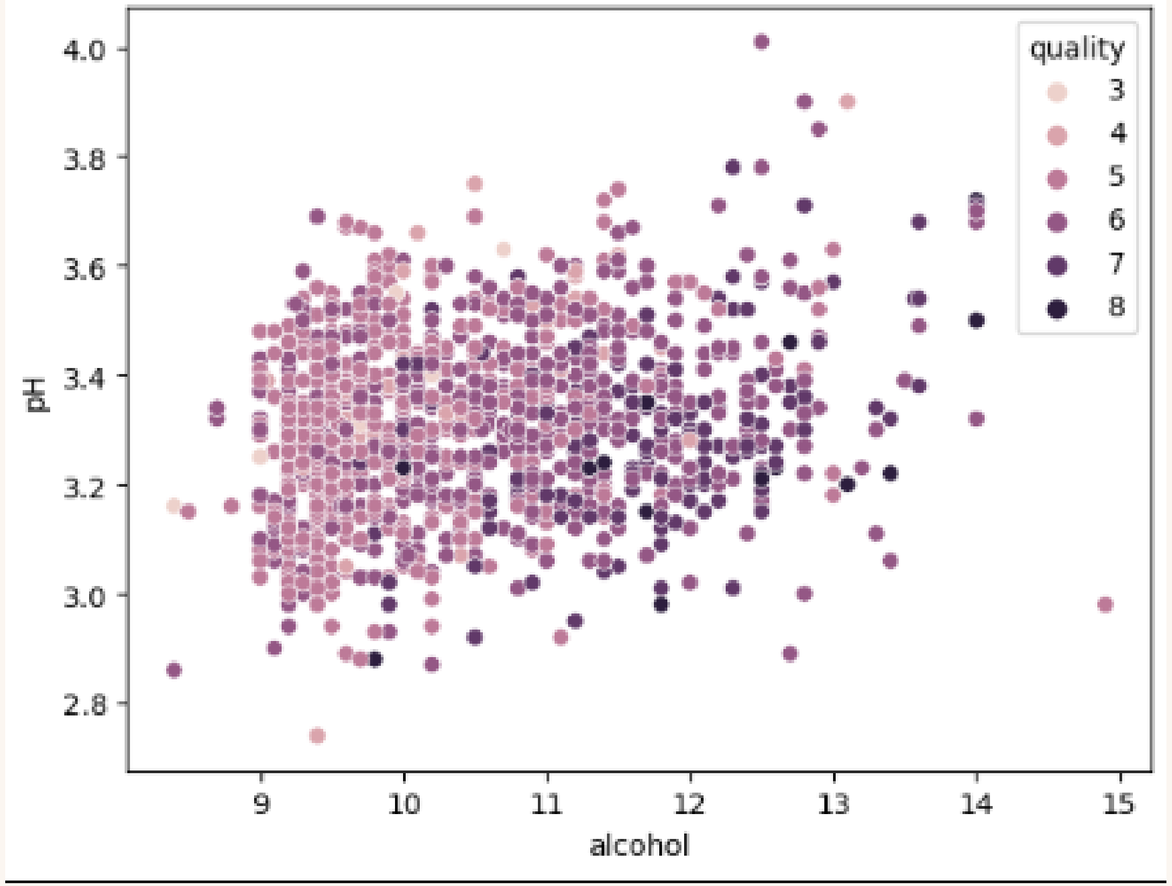


Multivariate Analysis

- Exploring relationships between features using pairplots.
- Example: Scatterplot of alcohol vs pH reveals how alcohol content influences wine quality ratings

```
sns.scatterplot(x = 'alcohol', y ='pH', hue = 'quality', data = df)
```

```
<Axes: xlabel='alcohol', ylabel='pH'>
```



Summary of Findings

- Alcohol Content shows a positive correlation with wine quality — higher alcohol tends to mean better-rated wine.
- Volatile Acidity is negatively correlated with quality — wines with higher acidity often have lower quality scores.
- Most wines in the dataset are of average quality (scores 5 and 6), as seen in the countplot.
- Boxplots/Violinplots revealed clear differences in feature distributions across quality levels, especially for alcohol and sulphates.
- Pairplots helped uncover key interactions between features and confirmed correlations visually.

THANKYOU

For joining me on this journey of exploring Red-Wine Quality Analysis.
Our knowledge and skills will continue to evolve with practice and experimentation.

FOLLOW ME

https://www.linkedin.com/in/md-shakib-6283a7239/

https://github.com/shaky1405