Project 2

Connor Pawlowski | Sima Shafaei | CS – 150

Presentation Agenda

Dataset Selection

Data Initial Analysis

Coding Walkthrough

- Import and Load Data
- Statistical Summary
- Data Interpretation

Visualizations

- Scatterplot
- Histograms (Alc. & pH)
- Histogram by Quality

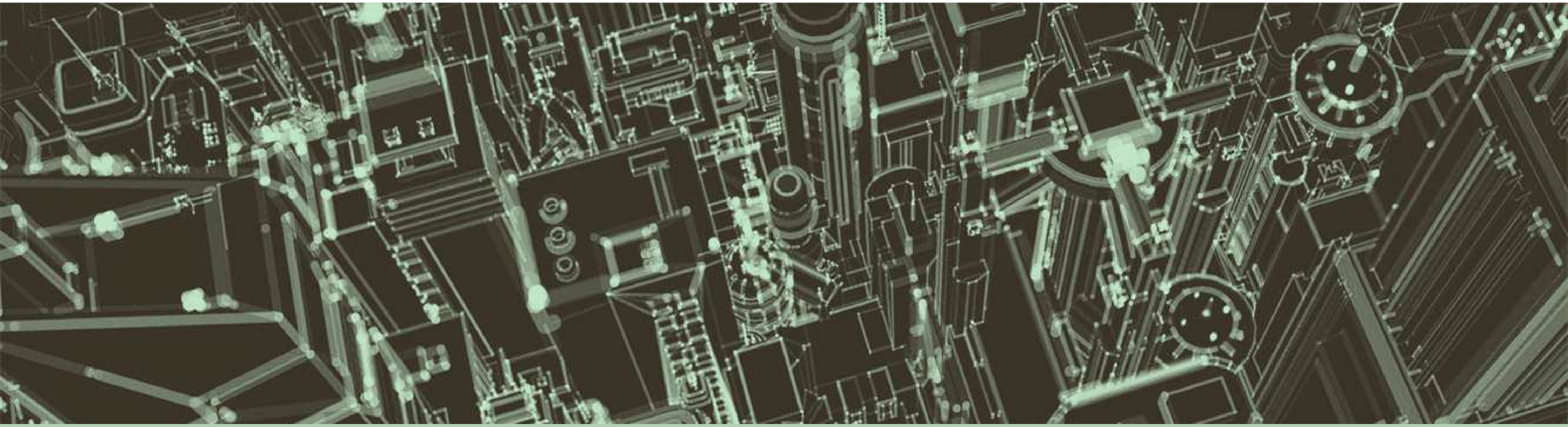
Dataset Selection

- Visit Kaggle to download set of data
- Wine Quality
  - 1599 rows of samples
  - 11 features
  - Mix of many measurements (pH, acidity, alcohol etc.)
  - Target Variable – Quality Score

Data Initial Analysis

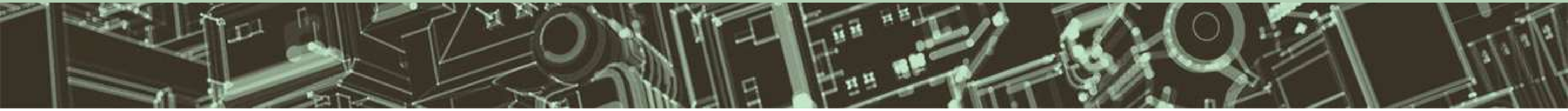
- 11 possible variables to analyze
- The project highlights two main variables
  - Alcohol Content
  - pH Levels
- The two variables are compared across quality ranges for the samples

Fixed Acidity	Volatile Acidity	Citric Acid
Residual Sugar	Chlorides	Free Sulfur Dioxide
Total Sulfur Dioxide	Density	pH
Sulphates	Alcohol	Quality

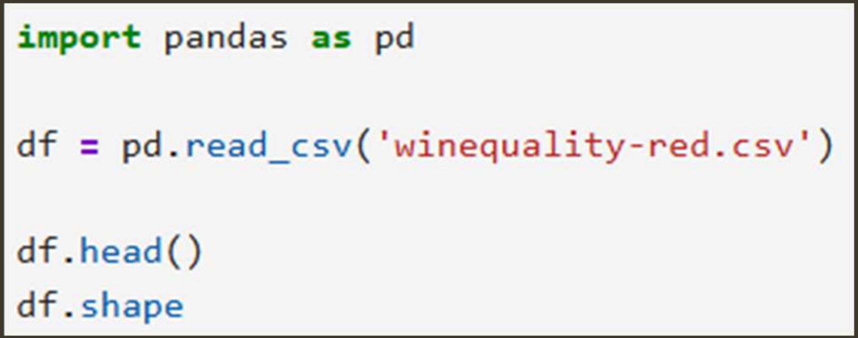


Coding Walkthrough

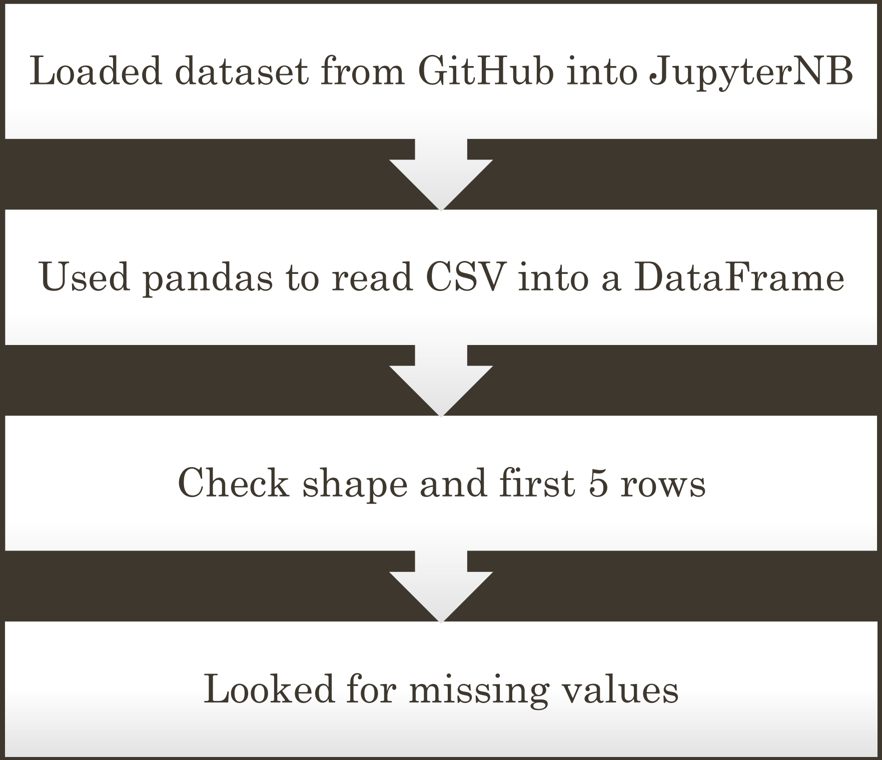
Step by Step Process



Import and Load Data



```
import pandas as pd

df = pd.read_csv('winequality-red.csv')

df.head()
df.shape
```



Loaded dataset from GitHub into JupyterNB

Used pandas to read CSV into a DataFrame

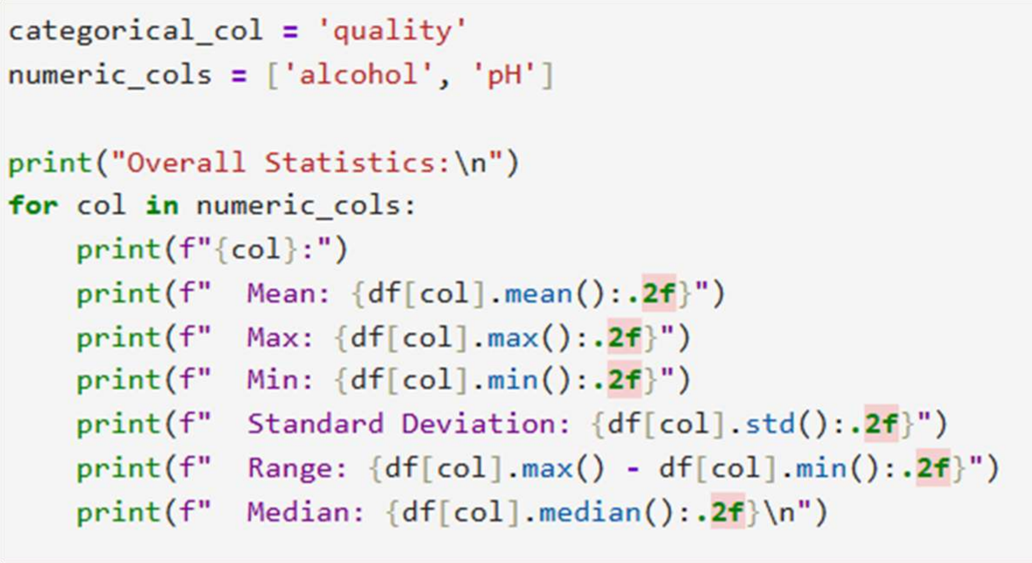
Check shape and first 5 rows

Looked for missing values

Statistical Summary

Code Function/Purpose

- First Loop: calculates stats for the whole dataset
- Stats Calculated
  - Mean
  - Median
  - Max
  - Min
  - Standard Deviation
  - Range



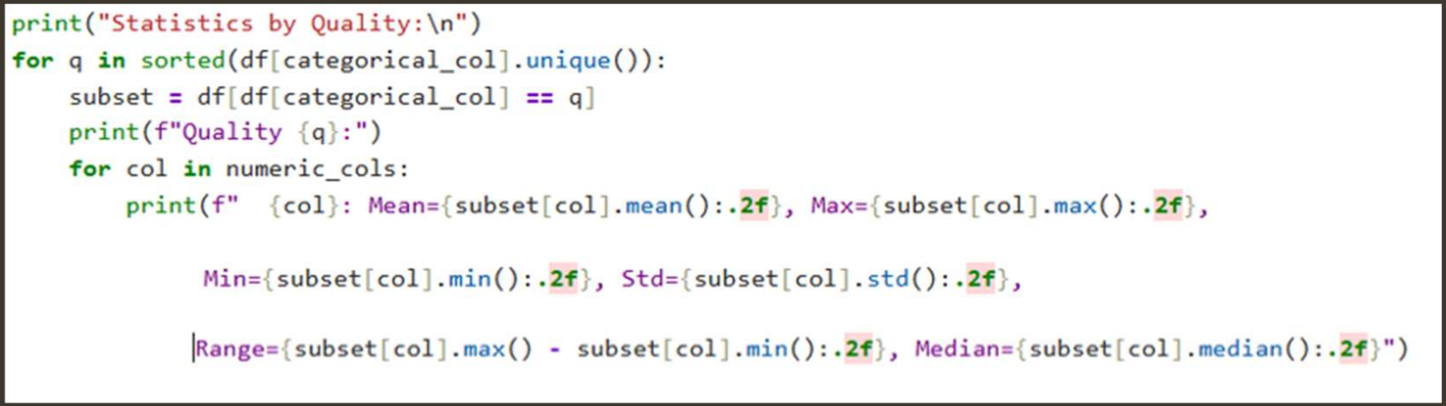
```
categorical_col = 'quality'
numeric_cols = ['alcohol', 'pH']

print("Overall Statistics:\n")
for col in numeric_cols:
    print(f"{col}:")
    print(f"  Mean: {df[col].mean():.2f}")
    print(f"  Max: {df[col].max():.2f}")
    print(f"  Min: {df[col].min():.2f}")
    print(f"  Standard Deviation: {df[col].std():.2f}")
    print(f"  Range: {df[col].max() - df[col].min():.2f}")
    print(f"  Median: {df[col].median():.2f}\n")
```

Statistical Summary

Code Function/Purpose

- Second Loop: groups rows by quality and compares averages

- Shows how wine chemistry changes across all of the quality levels



```
print("Statistics by Quality:\n")
for q in sorted(df[categorical_col].unique()):
    subset = df[df[categorical_col] == q]
    print(f"Quality {q}:")
    for col in numeric_cols:
        print(f"  {col}: Mean={subset[col].mean():.2f}, Max={subset[col].max():.2f},

              Min={subset[col].min():.2f}, Std={subset[col].std():.2f},

             Range={subset[col].max() - subset[col].min():.2f}, Median={subset[col].median():.2f}")
```

Data Interpretation

Overall Statistics

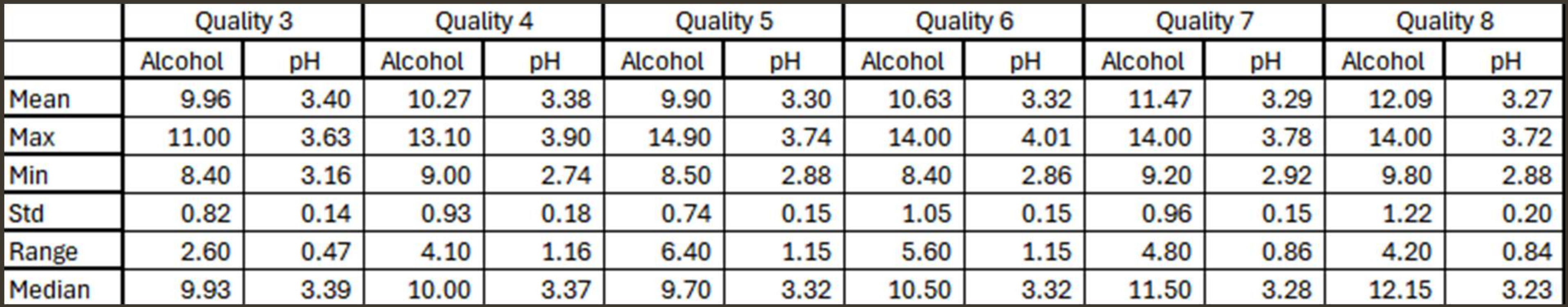
Alcohol

- Mean: 10.42
- Max: 14.90
- Min: 8.40
- Standard Deviation: 1.07
- Range: 6.50
- Median: 10.20

pH

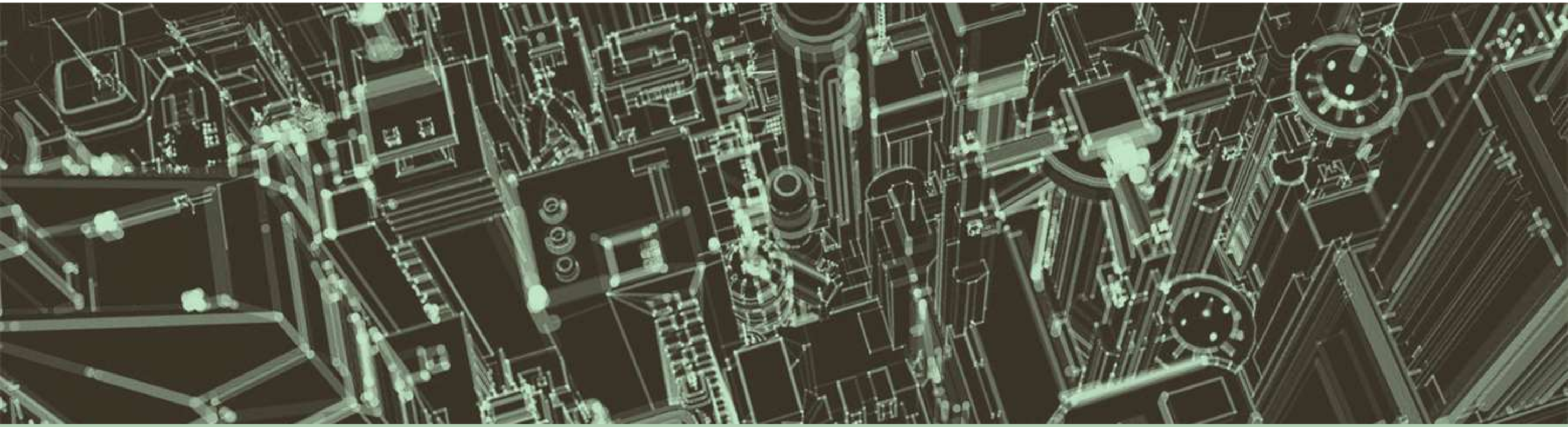
- Mean: 3.31
- Max: 4.01
- Min: 2.74
- Standard Deviation: 0.15
- Range: 1.27
- Median: 3.31

Data Interpretation



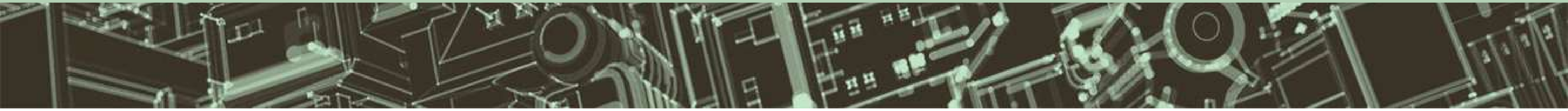
	Quality 3		Quality 4		Quality 5		Quality 6		Quality 7		Quality 8	
	Alcohol	pH	Alcohol	pH	Alcohol	pH	Alcohol	pH	Alcohol	pH	Alcohol	pH
Mean	9.96	3.40	10.27	3.38	9.90	3.30	10.63	3.32	11.47	3.29	12.09	3.27
Max	11.00	3.63	13.10	3.90	14.90	3.74	14.00	4.01	14.00	3.78	14.00	3.72
Min	8.40	3.16	9.00	2.74	8.50	2.88	8.40	2.86	9.20	2.92	9.80	2.88
Std	0.82	0.14	0.93	0.18	0.74	0.15	1.05	0.15	0.96	0.15	1.22	0.20
Range	2.60	0.47	4.10	1.16	6.40	1.15	5.60	1.15	4.80	0.86	4.20	0.84
Median	9.93	3.39	10.00	3.37	9.70	3.32	10.50	3.32	11.50	3.28	12.15	3.23

This table shows Alcohol and pH statistics by Quality



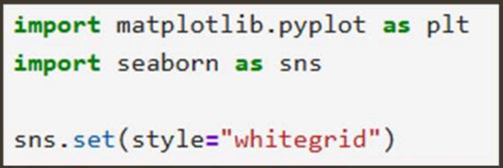
Visualizations

Scatterplot and Histograms

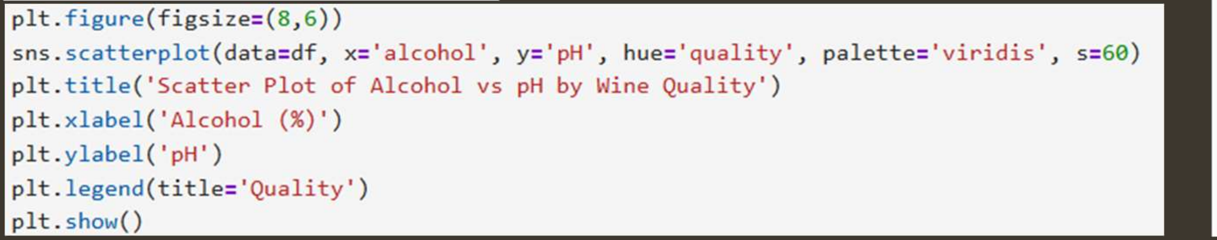


Scatterplot

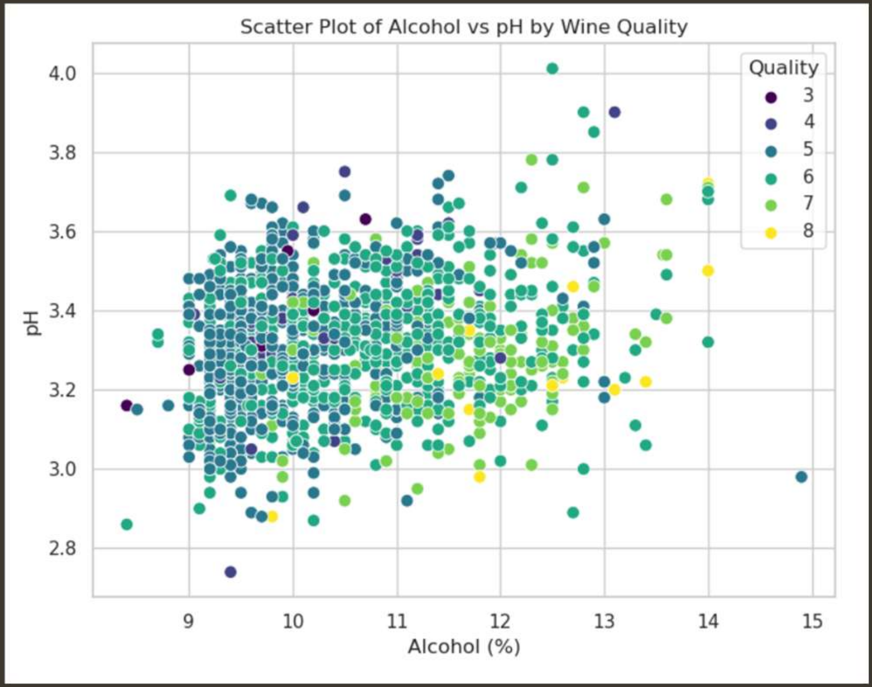
- Created using matplotlib
- X-Axis: Alcohol | Y-Axis: pH
- Slight Downward trend – higher alcohol comes with slightly lower pH



```
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid")
```

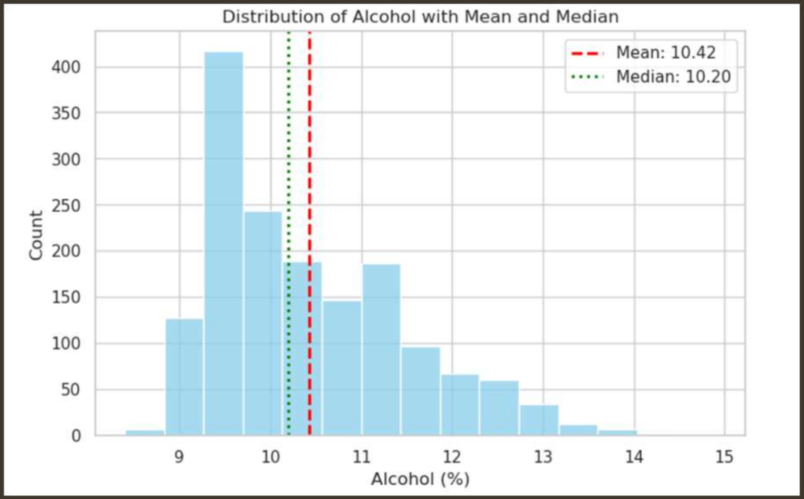


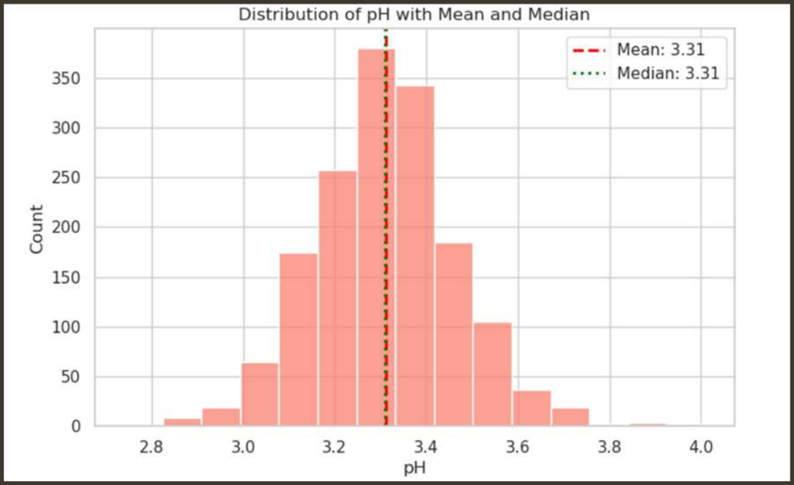
```
plt.figure(figsize=(8,6))
sns.scatterplot(data=df, x='alcohol', y='pH', hue='quality', palette='viridis', s=60)
plt.title('Scatter Plot of Alcohol vs pH by Wine Quality')
plt.xlabel('Alcohol (%)')
plt.ylabel('pH')
plt.legend(title='Quality')
plt.show()
```



Histograms

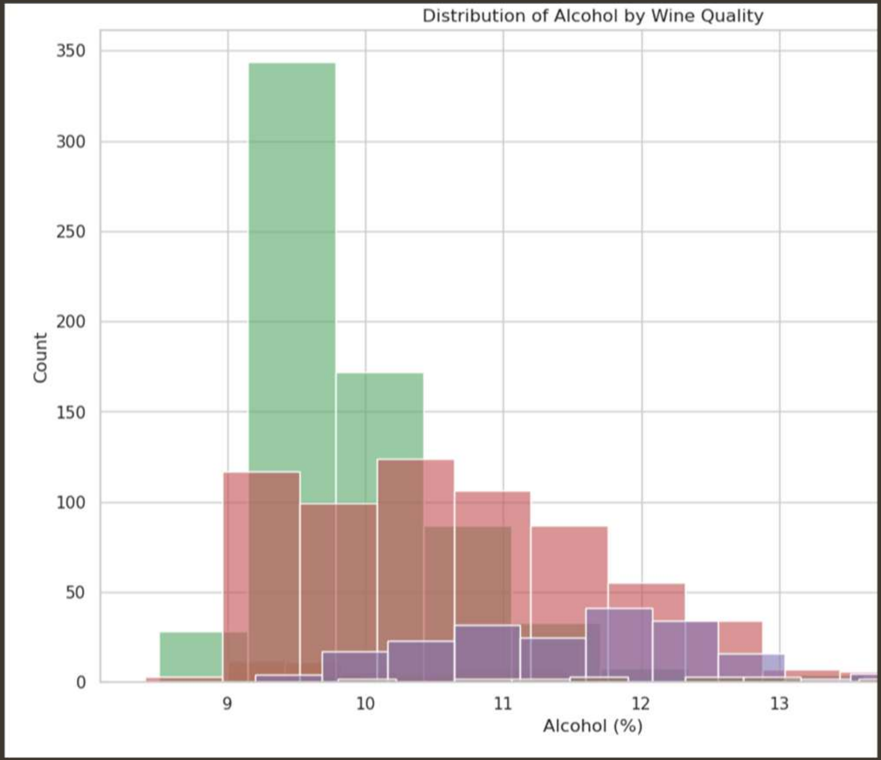
- Two histograms:
  - 1 – Alcohol
  - 2 – pH
- Most wines around 9-12% Alc
- pH centers around 3.2-3.4



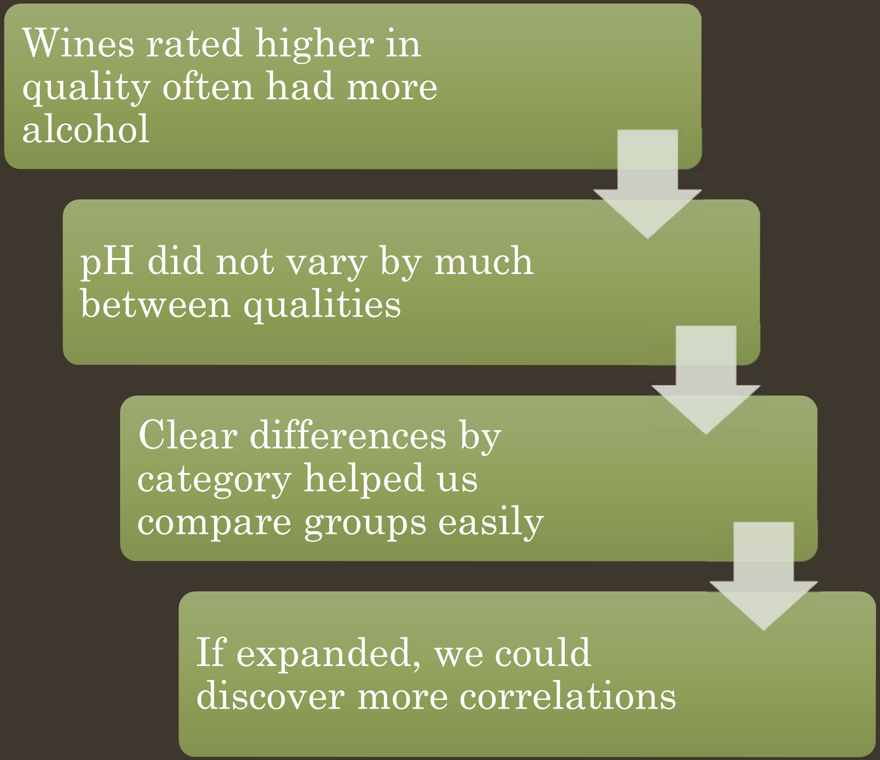


Histograms by Quality

- One histogram per wine quality
- Depicts how alcohol levels differ across categories
- Higher quality = shift rightward (more alcohol)
- Lower quality wines cluster lower
- Most grouped in quality 5&6



Conclusion



Wines rated higher in quality often had more alcohol

pH did not vary by much between qualities

Clear differences by category helped us compare groups easily

If expanded, we could discover more correlations