

Report Of Clustering and Fitting Analysis on Wine Dataset

Name: Alvin James

Student Number: 23111190

GitHub Repository: [RoronoaJames/Clustering-and-Fitting](https://github.com/RoronoaJames/Clustering-and-Fitting)

Introduction

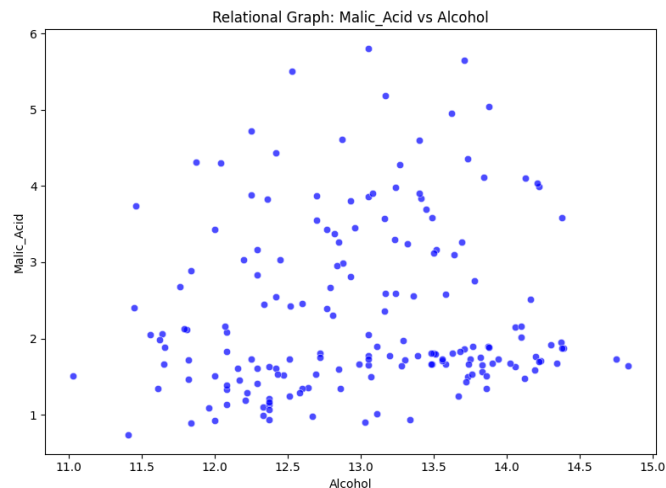
This report analyzes a Wine dataset from Kaggle using clustering and fitting techniques to uncover patterns and relationships. The key objectives include:

1. **Relational Analysis:** Scatter plot for variable relationships.
2. **Categorical Analysis:** Bar plot for comparisons across categories.
3. **Statistical Analysis:** Heatmap for correlations and descriptive statistics.
4. **Clustering:** Identify groups using k-means with optimal clusters via the elbow method.
5. **Fitting:** Analyze and model trends using linear regression.

1. Relational Analysis

- **Scatter Plot: Alcohol vs Malic_Acid**
 - Weak positive correlation.
 - Higher alcohol content does not significantly predict malic acid levels.
 - Outliers may slightly affect the data distribution.

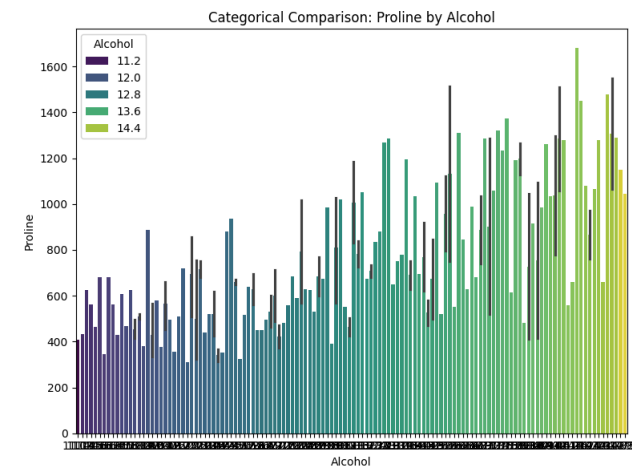
- **Graph Details:**



2. Categorical Analysis

- **Bar Plot: Alcohol and Proline**
 - Proline levels tend to increase with higher alcohol content.
 - Highlights distinct variations among wines with different alcohol levels.

- **Graph Details:**



3. Statistical Analysis

- **Heatmap:**

- **Key Correlations:**

- Alcohol positively correlates with Proline .
- Malic_Acid negatively correlates with Hue.

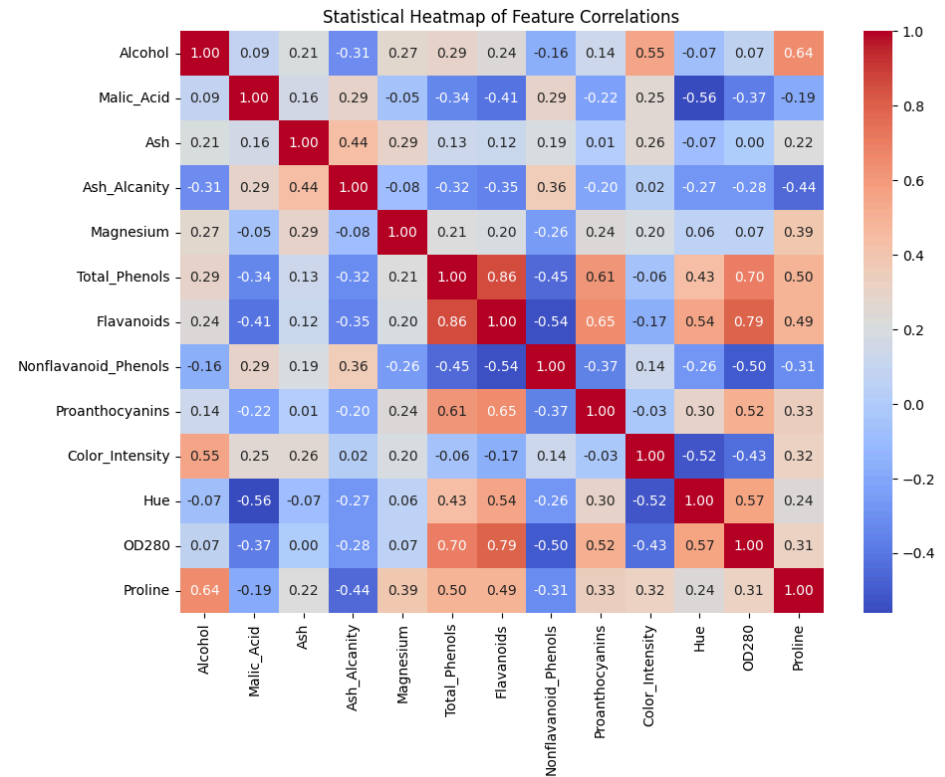
- **Key Moments (Selected Features):**

- **Alcohol:**

- Mean: 13.00, Std Dev: 0.81, Skewness: -0.05, Kurtosis: -0.86.
- Distribution is symmetric with moderate variation.

- **Malic_Acid:**

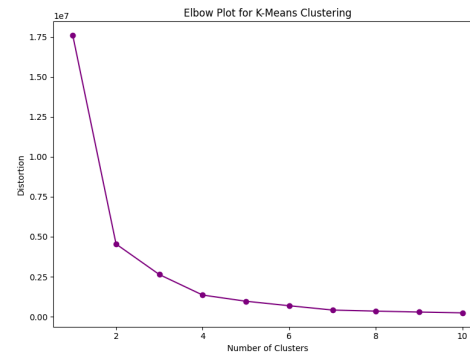
- Mean: 2.34, Std Dev: 1.12, Skewness: 1.03, Kurtosis: 0.26.
- Positively skewed distribution with occasional extreme values.



4. Clustering Analysis

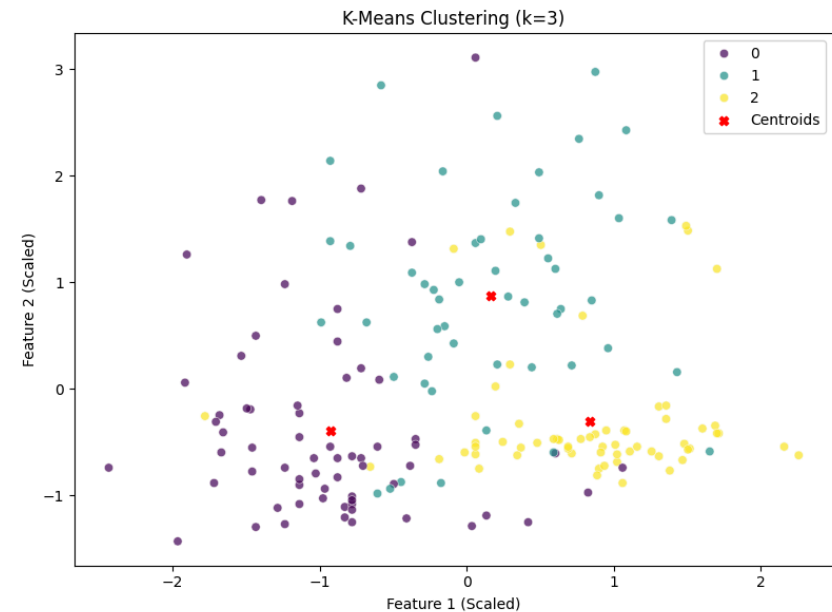
- **Elbow Plot:**

- Optimal clusters: k=3.



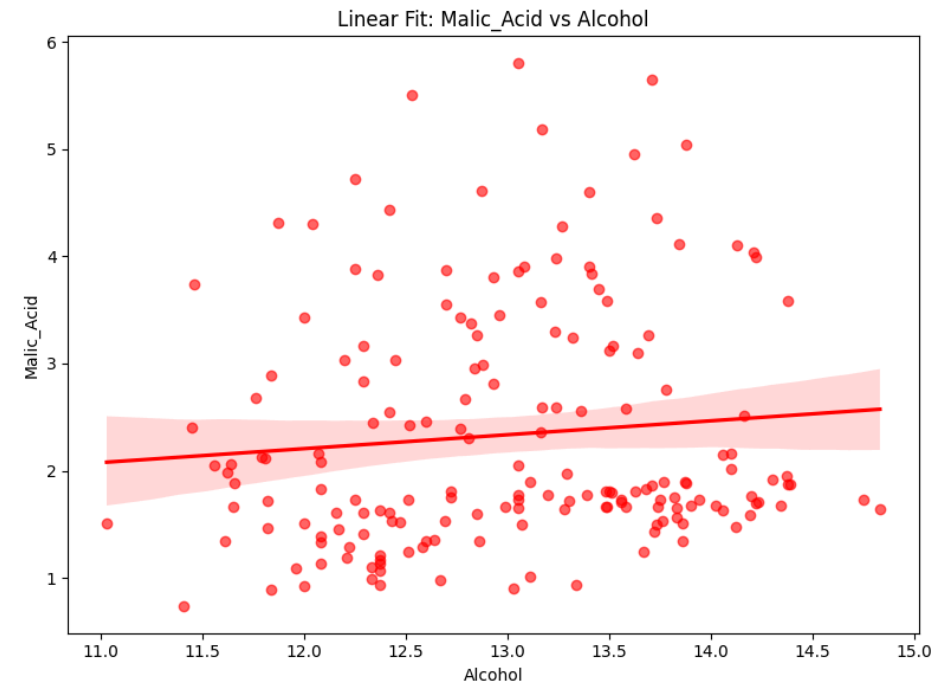
- **Cluster Visualization:**

- Data normalized for clustering & clusters are well-defined.
- Centroids marked & silhouette score: 0.28.
- Clusters are distinguishable but moderately cohesive.



5. Fitting Analysis

- **Linear Regression: Alcohol vs Malic_Acid**
 - **Regression Model:**
 - Slope suggests a weak increase in malic acid with higher alcohol levels.
 - **Residual Analysis:**
 - Weak predictive capability due to low correlation.



Conclusion

1. **Relational Trends:** Weak correlation between Alcohol and Malic_Acid indicates limited linear dependency.
2. **Categorical Insights:** Higher alcohol content is linked with increased proline levels, distinguishing wine types.
3. **Statistical Observations:**
 - Alcohol and Proline have strong relationships, while Malic_Acid negatively impacts Hue.
 - Statistical moments highlight symmetric and skewed feature distributions.
4. **Clustering:** Optimal clustering ($k=3$) shows moderate cohesion with potential for refining metrics.
5. **Fitting:** Linear regression demonstrates weak predictability, suitable for initial trend exploration.