"Dataset Dynamics: A Hands-On Approach to Data Work

Introduction

This introduction gives a quick peek into Section 3.1: Choose, Load, and Inspect Your Data, where we start understanding data. It's like laying the groundwork for our analysis. We'll focus on picking the right data, loading it correctly, and looking closely at it. Using tools like Pandas, Numpy, Matplotlib, and Seaborn, we'll show a step-by-step process to find hidden patterns in a small dataset. It's not just about using technical tools but also making smart decisions about data. We carefully chose our dataset, loaded it with attention to detail, and looked at it closely to set up a strong exploratory data analysis. This report is to see how we handle and understand data, making it all clear and meaningful.

3.1: Choose, Load and Inspect your Data.

1. Drive Mounted:
    - Google Drive has been connected to Colab using the code drive. mount('/content/drive').
    - This setup allows us to access files stored in Google Drive.
2. Datasets Loaded:
    - The dataset named "WineQT.csv" has been loaded into Colab using Pandas with the code sk=pd. read_csv("/content/drive/MyDrive/AIAssignment1/WineQT.csv").
    - The loaded dataset is assigned to the variable sk.
3. Dataset Overview:
    - The dataset, "WineQT.csv," shared by Mr. M YASSER H in 2021, contains key columns like Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol, Quality, and ID.
4. Insights from the Dataset:
    - The dataset offers valuable information on wine quality, pH, density, and more.

- It provides opportunities for exploring basic statistics, distribution, and relationships among different attributes.

5. Data Inspection:
    - Size of the Data Frame:
    - The total number of rows and columns is obtained using sk.shape.
    - Data Types of Each Variable:
    - Data types of each column are checked using sk.dtypes.
    - Check for Missing Values:
    - The count of missing values in each column is examined using sk.isnull().sum().
6. Purpose:
    - The descriptive summary and data inspection steps are crucial for understanding the dataset's structure, dimensions, and initial characteristics.
    - This information forms the basis for further analysis and interpretation of the data.

Data Pre-processing and Statistical Interpretation:

Data Cleaning:

1. Handling Missing Values:
    - We found and filled in missing values using the median, which is like picking the middle value. This helps when there are weird numbers in the data, and it works well for different types of information.

2. Duplicate Removal:
    - We checked for and got rid of any duplicate rows in the dataset.

3. Additional Cleaning Actions:
    - We made the 'fixed acidity' column easier to understand by renaming it to 'Acidity.'
    - We filtered the data to keep only rows where 'residual sugar' is more than 2.
    - We removed the 'density' column from the dataset.

Wine Datasets

Summary Statistics:

 Numeric Columns:

- We looked at basic stats for things like acidity, sugar levels, and more in the wine dataset. This helped us understand the dataset size, average values, and how spread out the data is.

  Observations:

- The dataset covers a lot of info about wine, like acidity types and sugar levels. Our stats analysis helped us understand the dataset better.

 Categorical Columns:

 We figured out the different values in categories and which ones show up the most.

Data Visualization and Explorations

3.3: Visualize the Data: Make Interpret and Save your Charts.

(a) Univariate Analysis:

• Chart 1

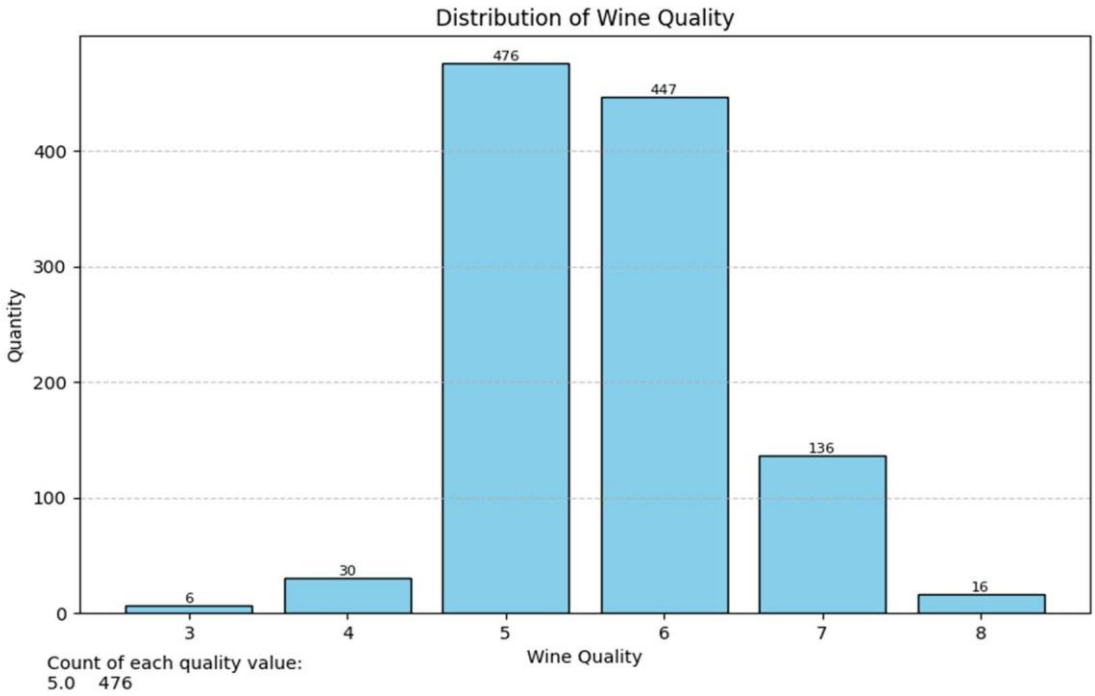

Figure 1: Distribution of Wine Quality

The bar graph, titled "Distribution of Wine Quality," illustrates how wine quality is rated on a scale from 3 to 8. Here's a breakdown of the information:

- Wines with a quality rating of 3: Very few instances (only 6 occurrences).
- Quality rating of 4: A relatively low quantity, with 30 instances.
- Quality rating of 5: The highest quantity, with 476 instances.
- Quality rating of 6: The second-highest quantity, with 447 instances.
  Quality rating of 7: Moderate quantity, observed in 136 instances.
- Wines with a quality rating of 8: Very few instances, specifically 16.

The majority of wines fall into the quality ratings of 5 and 6, each having over 400 instances. On the other hand, wines with quality ratings of 3, 4, 7, and 8 are significantly less common. This data provides valuable insights into the distribution of wine quality in the dataset, highlighting that wines with ratings of 5 and 6 are more prevalent. Understanding this distribution can be useful for assessing the overall quality trends in the given dataset.

Wine Datasets

• Chart 2

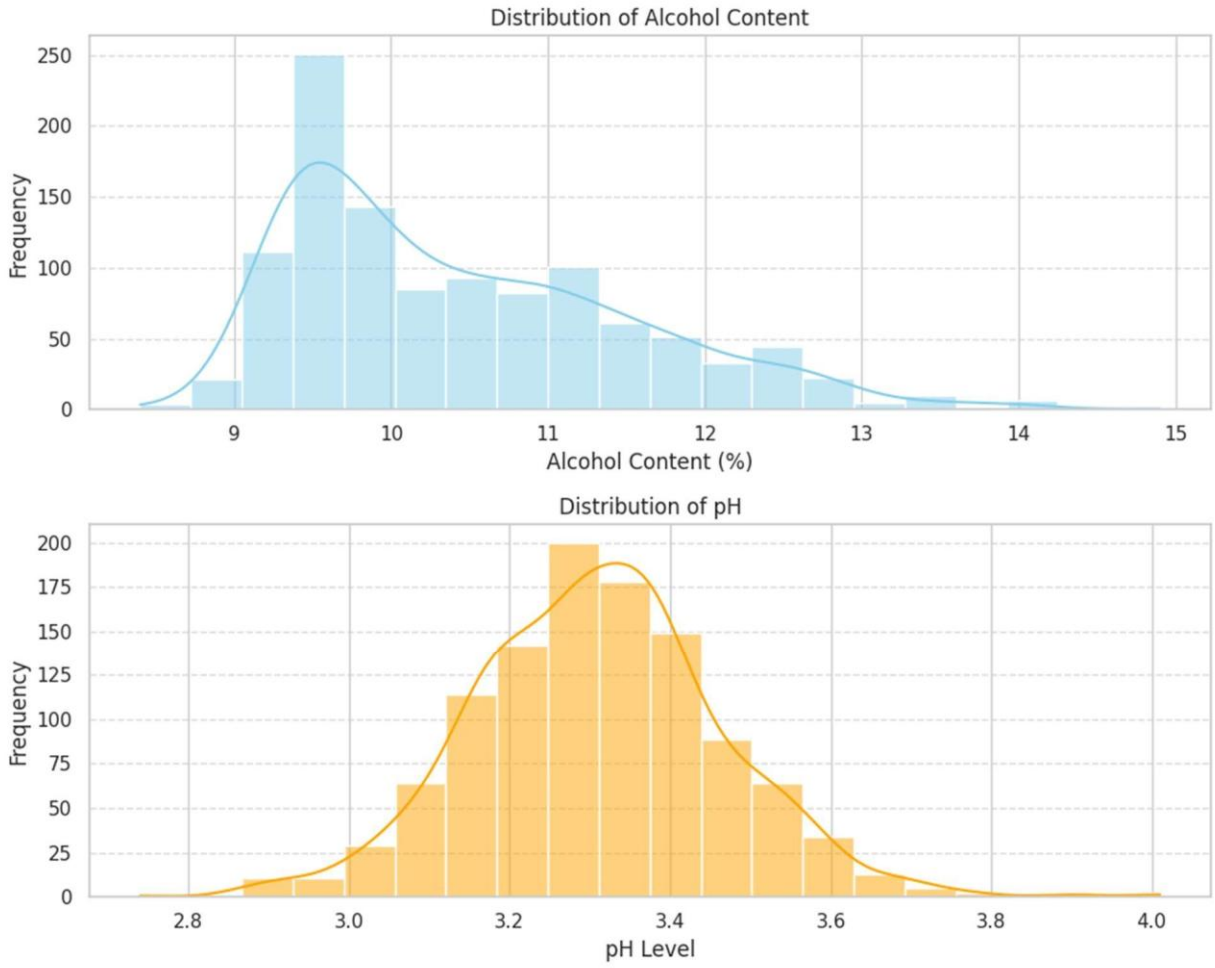

Figure 2: Histogram for alcohol and pH.

The image contains two histograms, one for alcohol content and the other for pH levels.

1. Alcohol Content Histogram (Blue): This histogram shows the distribution of alcohol content in various samples, ranging from 9 to 15. There is a peak around an alcohol content level of 10, indicating that many samples have this alcohol level. The line graph overlaid on the bars indicates the trend in the distribution.
2. pH Level Histogram (Yellow): This histogram shows the distribution of pH levels in the samples, ranging from approximately 2.8 to 4.0. There's a prominent peak at around a pH level of 3.2, indicating that many samples have this pH level. The line graph overlaid on the bars shows the trend in the distribution.

These histograms above could be useful for understanding the general alcohol content and pH levels in a given dataset of wine samples.

- Chart 3

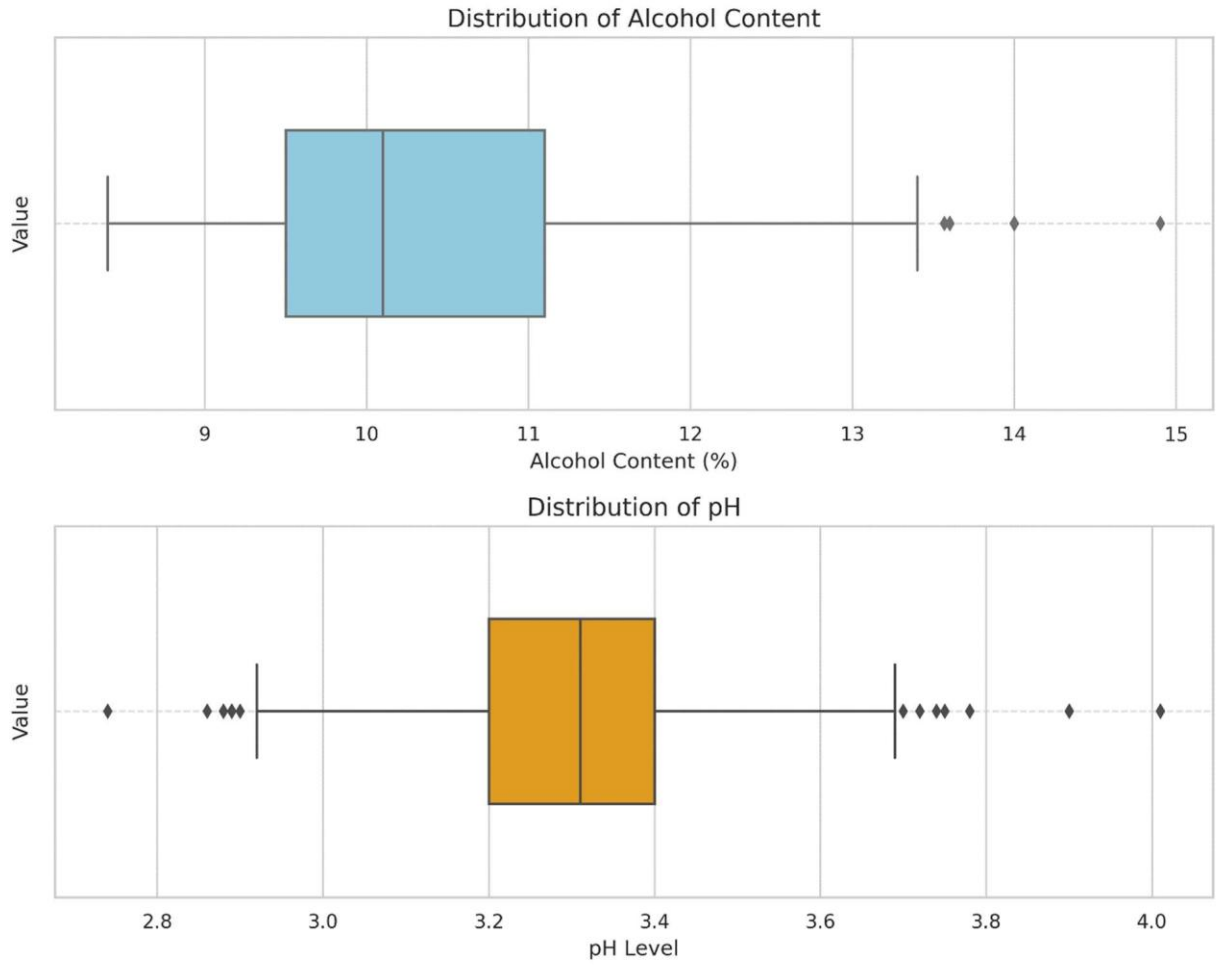

Figure 3: Boxplot

The image contains two boxplots, one for alcohol content and the other for pH levels.

1. Alcohol Content Boxplot (Blue): This boxplot shows the distribution of alcohol content in various samples. The median alcohol content is around 10.5, with a majority of the data points falling between approximately 9.5 and 11.5. There are also some outliers present beyond an alcohol content of 12.
2. pH Level Boxplot (Orange): This boxplot shows the distribution of pH levels in the samples. The median pH is about 3.3, with most data points falling between approximately 3.2 and 3.4. There are several outliers present below a pH of 3 and above a pH of 3.4.

Wine Datasets

These boxplots could be useful for understanding the general alcohol content and pH levels in a given dataset of wine samples.

(b) Bi-variate Analysis:

• Chart 4

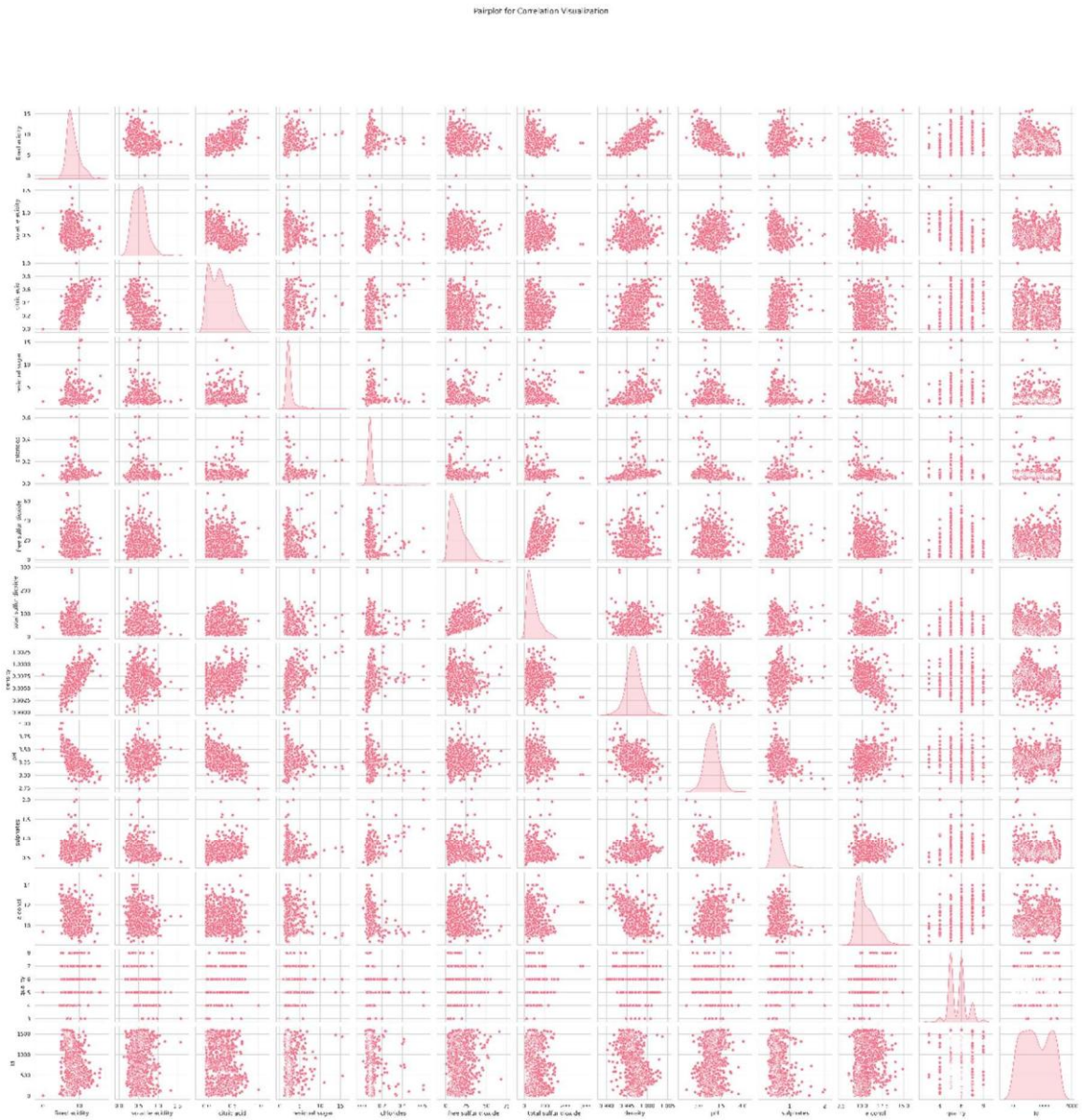

Figure 4: Pair plot Correlation Visualization

The image above is of pair plot using the Seaborn library in Python. A pair plot is a grid of scatter plots and histograms that visualize the pairwise relationships and distributions of variables in a dataset.

As for the output image, it's a pair plot that shows the relationships between different pairs of variables in your dataset. Here's what the image represents:

- Each cell in the grid represents a plot between two variables of the dataset.
- The diagonal cells contain histograms showing the distribution of single variables.
- The off-diagonal cells contain scatter plots showing relationships between pairs of variables.
- There are multiple clusters and patterns visible in various scatter plots, indicating correlations or groups within the data.

• Chart 5

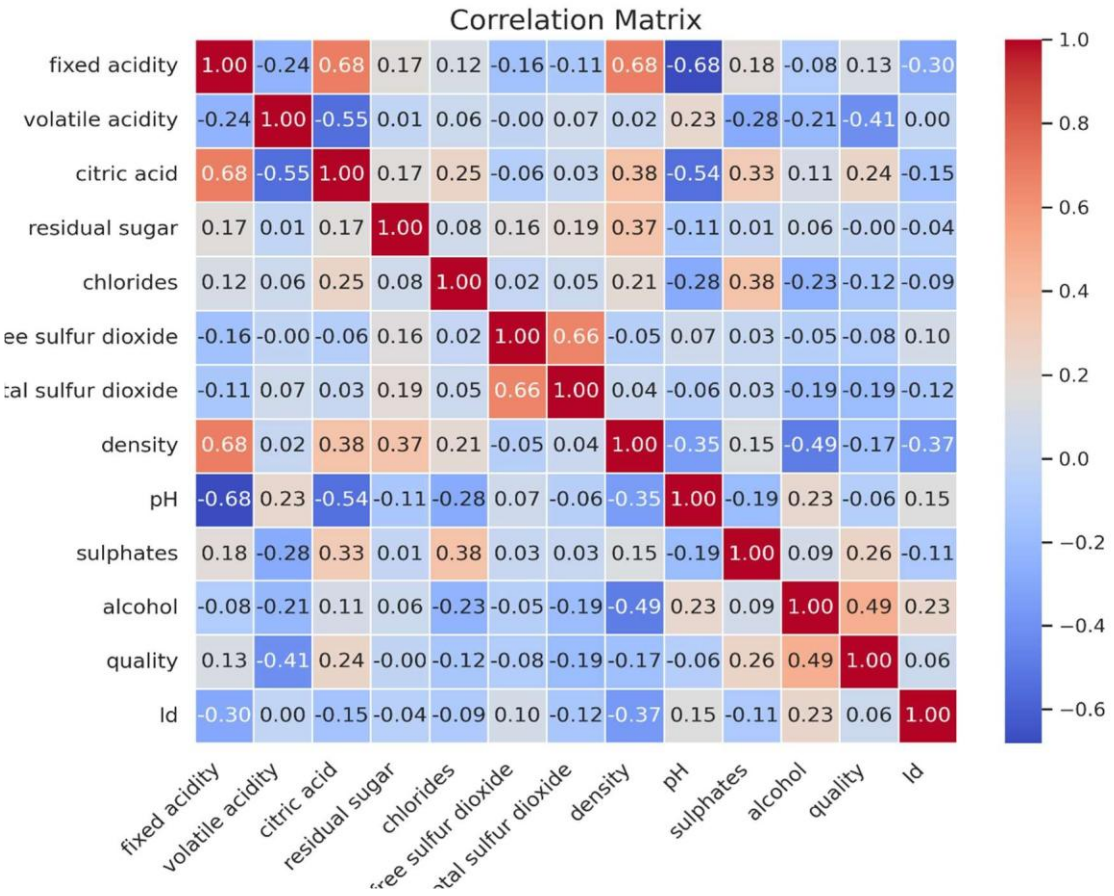

Figure 5: Correlation Matrix

The Heatmap image is of a Correlation Matrix. It's a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The value is in the range of -1 to 1. If two variables have high correlation, it means they are likely to increase or decrease together.

Here's a breakdown of the chart:

- The chart is showing correlations between various attributes of wine such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.
- Each cell in the matrix contains a numerical value representing the strength and direction of correlation. A positive number represents a positive correlation (as one variable increases, the other does too), while a negative number represents a negative correlation (as one variable increases, the other decreases).
- The color of the cells indicates the strength and direction of the correlation. Red indicates a positive correlation; blue indicates a negative correlation. The darker the color, the stronger the correlation.
- The diagonal from the top left to the bottom right represents the correlation of the variables with themselves, which is why they are all 1.00 and colored red.

For example, ‘density’ and ‘fixed acidity’ have a high positive correlation as indicated by the dark red color, meaning they tend to increase or decrease together. On the other hand, ‘pH’ and ‘fixed acidity’ have a high negative correlation as indicated by the dark blue color, meaning as one increase, the other tends to decrease.

This kind of chart is useful for identifying relationships between different variables. In this case, it could help in understanding which factors most influence the quality of wine.

Conclusion:

In conclusion, we've learned about data by laying a strong foundation for analysis. We focused on picking the right data, loading it correctly, and using tools like Pandas, NumPy, Matplotlib, and Seaborn. We looked through finding hidden patterns in a small dataset step by step. It's not just about technical tools but also about smart decisions with data. We carefully chose our dataset, loaded it with attention, and looked at it closely for a solid exploratory data analysis. This report helps to see how we handle and understand data, making it clear and meaningful.