# Day 2 - Exploratory Data Analysis & Feature Correlation

## Task Description

- **Task**: Analyse feature correlations within the dataset.
- **Description**: Calculate correlation coefficients between numerical features and visualise the correlation matrix using a heat map. Identify highly correlated features and consider their impact on the analysis.

## What is Feature Correlation?

- Feature correlation refers to the degree to which two or more features (**variables**) in a dataset are **related** to each other. In statistical terms, correlation measures the **strength and direction** of the relationship between variables.
- There are several types of correlation measures, but one of the most common is **Pearson's correlation coefficient**, which ranges from -1 to

1. A correlation coefficient of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation

# Why is Feature Correlation Important?

- Feature correlation is particularly important in fields like **machine learning** and **data analysis** because it can provide insights into how different features interact with each other and how they collectively influence the outcome of a model or analysis. High correlations between features can sometimes indicate redundancy, where one feature is highly predictable from another, or multicollinearity, which can cause issues in some statistical models. On the other hand, low or moderate correlations might suggest that the features are more independent of each other, which can be beneficial for model performance.

# Dataset of Choice

## Red Wine Quality

- This dataset was taken from Kaggle.
- URL: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data

# Analysis

## Library Imports

```
# Importing Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Importing the dataset
df = pd.read_csv('winequality-red.csv')
df.head()
```

## Column Insights

- **Fixed acidity**: The amount of non-volatile acids in the wine, typically measured in grams per litre (g/L).
- **Volatile acidity**: The amount of volatile acids in the wine, which can contribute to unpleasant flavours, typically measured in g/L.
- **Citric acid**: The amount of citric acid in the wine, which can impart freshness and flavour, typically measured in g/L.
- **Residual sugar**: The amount of sugar remaining in the wine after fermentation, typically measured in g/L.

- **Chlorides**: The amount of salt in the wine, which can influence taste and mouthfeel, typically measured in g/L.
- **Free sulphur dioxide**: The amount of sulphur dioxide present in free form, which acts as a preservative and antimicrobial agent, typically measured in parts per million (ppm).
- **Total sulphur dioxide**: The total amount of sulphur dioxide present in the wine, including both free and bound forms, typically measured in ppm.
- **Density**: The density of the wine, which can be influenced by sugar content and alcohol level, typically measured in g/cm$^3$.
- **pH**: The acidity level of the wine, which can affect its taste and stability, measured on a scale from 0 to 14, with lower values indicating higher acidity.
- **Sulphates**: The amount of sulphur dioxide in the wine, which can contribute to its antioxidant properties and preservation, typically measured in g/L.
- **Alcohol**: The alcohol content of the wine, typically measured as a percentage of volume.
- **Quality**: The quality rating of the wine, which is often a subjective assessment provided by experts or consumers.

# Data Cleaning

```
# Obtain dataset information
print("\nDataset Information:")
df.info()

# .describe() to obtain the summary statistics
```

```
print("\nSummary Statistics:")
df.describe()

# NaN Values
# Count NaN values in each column
nan_values_count = df.isna().sum()

print("\nCount of NaN values in each column:")

# Print the result
print(nan_values_count)
```

- From here we can see that there are no NaN or missing values in our dataset.
- All data points are represented are numerical, with only quality being an integer whilst every other column features data points as floats.

# Visualisation

```
# Calculate Correlation Matrix

# Calculate correlation matrix
correlation_matrix = df.corr()
# Set up the matplotlib figure
plt.figure(figsize=(10, 8))

# Plot the heatmap
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt=".2f", linewidths=0.5)

# Add title and display the plot
plt.title('Feature Correlation Heatmap')
plt.show()

# Extract the lowest correlation
```

```
correlation_between_features = correlation_matrix.loc['pH',
'fixed acidity']
print("Correlation between pH and fixed acidity:",
correlation_between_features)
```

# Correlation Statistics

```
# Extract the highest correlation
correlation_between_features = correlation_matrix.loc['density',
'fixed acidity']

print("Correlation between density and fixed acidity:",
correlation_between_features)
```

- **Correlation between density and fixed acidity**: The correlation coefficient between density and fixed acidity is approximately 0.668. This indicates a moderately strong positive correlation between these two variables. A positive correlation suggests that as one variable (density) increases, the other variable (fixed acidity) also tends to increase, and vice versa. In this case, it means that wines with higher fixed acidity tend to have higher density, and wines with lower fixed acidity tend to have lower density.

```
# Extract the lowest correlation
correlation_between_features = correlation_matrix.loc['pH',
'fixed acidity']

print("Correlation between pH and fixed acidity:",
correlation_between_features)
```

- **Correlation between pH and fixed acidity**: The correlation coefficient between pH and fixed acidity is approximately -0.683. This indicates a moderately strong negative correlation between these two variables. A negative correlation suggests that as one variable (pH) increases, the other variable (fixed acidity) tends to decrease, and vice versa. In this case, it means that wines with higher fixed acidity tend to have lower pH levels, and wines with lower fixed acidity tend to have higher pH levels.

This markdown script provides an overview of the exploratory data analysis and feature correlation analysis performed on the Red Wine Quality dataset. It includes explanations of feature correlation, dataset description, and the analysis steps along with corresponding Python code snippets.