Assignment 1

Daniel Mehta

Question 1

In [75]:

```
import pandas as pd
import numpy as np
```

Question 2

In [77]:

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

Question 3

In [79]:

```
df = pd.read_csv("WineQT.csv")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1142 non-null   float64
 1   volatile acidity      1139 non-null   float64
 2   citric acid           1140 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1141 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1141 non-null   float64
 12  Id                    1143 non-null   int64  
dtypes: float64(12), int64(1)
memory usage: 116.2 KB
```

Question 4

In [13]:

```
df.head()
```

Out[13]:

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

Question 5

In [24]:

```
top_5_wine_alc = df.sort_values(by="alcohol",ascending=False).head(5)
print(f"The top 5 alcohol for wine are: \n{top_5_wine_alc}")
```

```
The top 5 alcohol for wine are: 
     fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
462           15.9              0.36         0.65             7.5      0.096   
329            8.8              0.46         0.45             2.6      0.065   
98             5.2              0.34         0.00             1.8      0.050   
898            5.0              0.38         0.01             1.6      0.048   
419            5.0              0.42         0.24             2.0      0.060   

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
462                 22.0                  71.0  0.99760  2.98       0.84   
329                  7.0                  18.0  0.99470  3.32       0.79   
98                  27.0                  63.0  0.99160  3.68       0.79   
898                 26.0                  60.0  0.99084  3.70       0.75   
419                 19.0                  50.0  0.99170  3.72       0.74   

     alcohol  quality    Id  
462     14.9      5.0   652  
329     14.0      6.0   467  
98      14.0      6.0   144  
898     14.0      6.0  1270  
419     14.0      8.0   588  
```

Question 6

In [31]:

```
print(df["density"].dtype)
```

```
float64
```

Question 7

In [68]:

```
missing_vals = df.isnull().sum()
print(f"Missing Values by Column: \n {missing_vals}")
print("-"*40)
print(f"Total Missing Values: {missing_vals.sum()}")
```

```
Missing Values by Column:
 fixed acidity           1
volatile acidity        4
citric acid             3
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 2
pH                      0
sulphates               0
alcohol                 0
quality                 2
Id                      0
dtype: int64
----------------------------------------
Total Missing Values: 12
```

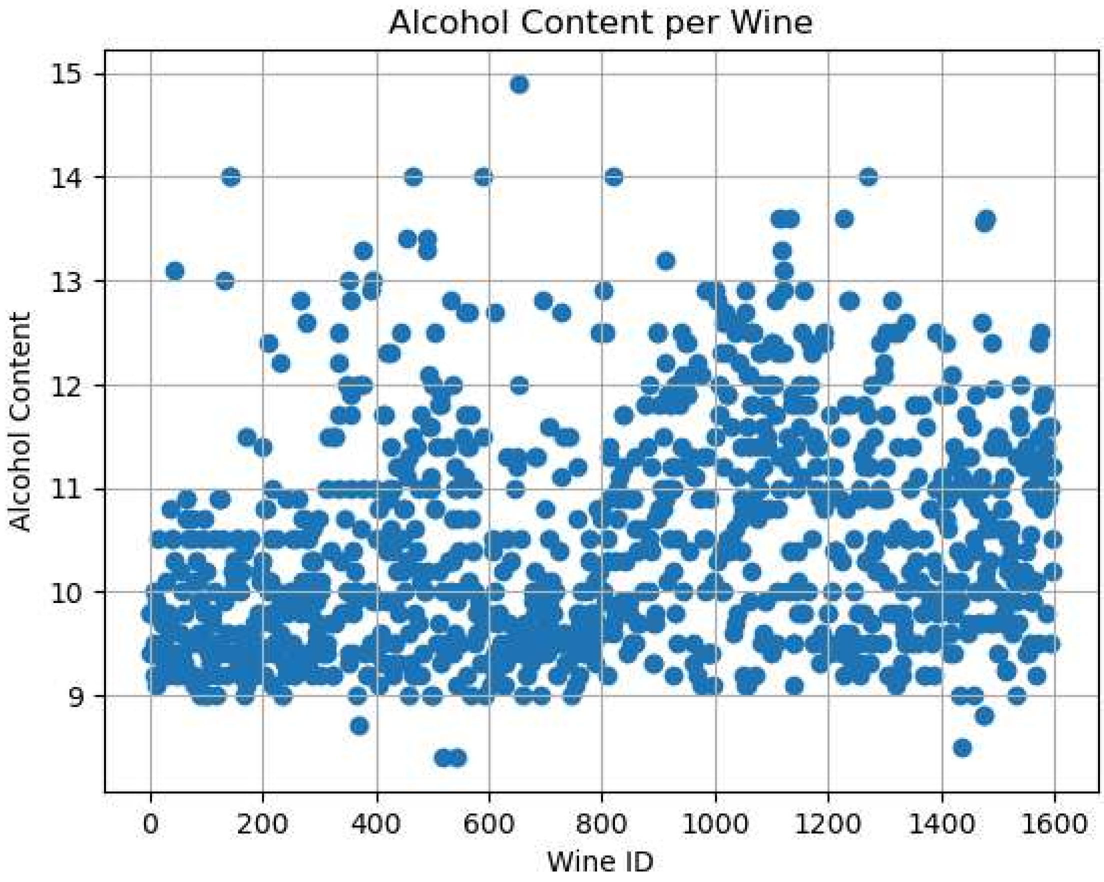
Question 8

```
In [85]: df.fillna({'fixed acidity': df['fixed acidity'].mean(),
                    'volatile acidity': df['volatile acidity'].median(),
                    'citric acid': df['citric acid'].median(),
                    'density': df['density'].mean(),
                    'quality': df['quality'].mode()[0]}, inplace=True)
         print(f"Missing Values by Column: \n {df.isnull().sum()}")
```

```
Missing Values by Column:
 fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
Id                      0
dtype: int64
```

Question 9

```
In [92]: plt.scatter(df['Id'], df['alcohol'])
         plt.xlabel('Wine ID')
         plt.ylabel('Alcohol Content')
         plt.title('Alcohol Content per Wine')
         plt.grid(True)
         plt.show()
```



Question 10

```
In [97]: df.rename(columns={'residual sugar': 'sugar'}, inplace=True)
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   sugar                 1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   float64
 12  Id                    1143 non-null   int64  
dtypes: float64(12), int64(1)
memory usage: 116.2 KB
```

Question 11

In [102…

```
df.sort_values(by=['quality', 'alcohol'], ascending=[False, False], inplace=True)
df.head()
```

Out[102…

	fixed acidity	volatile acidity	citric acid	sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	a
419	5.0	0.42	0.24	2.0	0.060	19.0	50.0	0.99170	3.72	0.74	
321	11.3	0.62	0.67	5.2	0.086	6.0	19.0	0.99880	3.22	0.69	
793	7.9	0.54	0.34	2.5	0.076	8.0	17.0	0.99235	3.20	0.72	
271	5.6	0.85	0.05	1.4	0.045	12.0	88.0	0.99240	3.56	0.82	
190	7.9	0.35	0.46	3.6	0.078	15.0	37.0	0.99730	3.35	0.86	

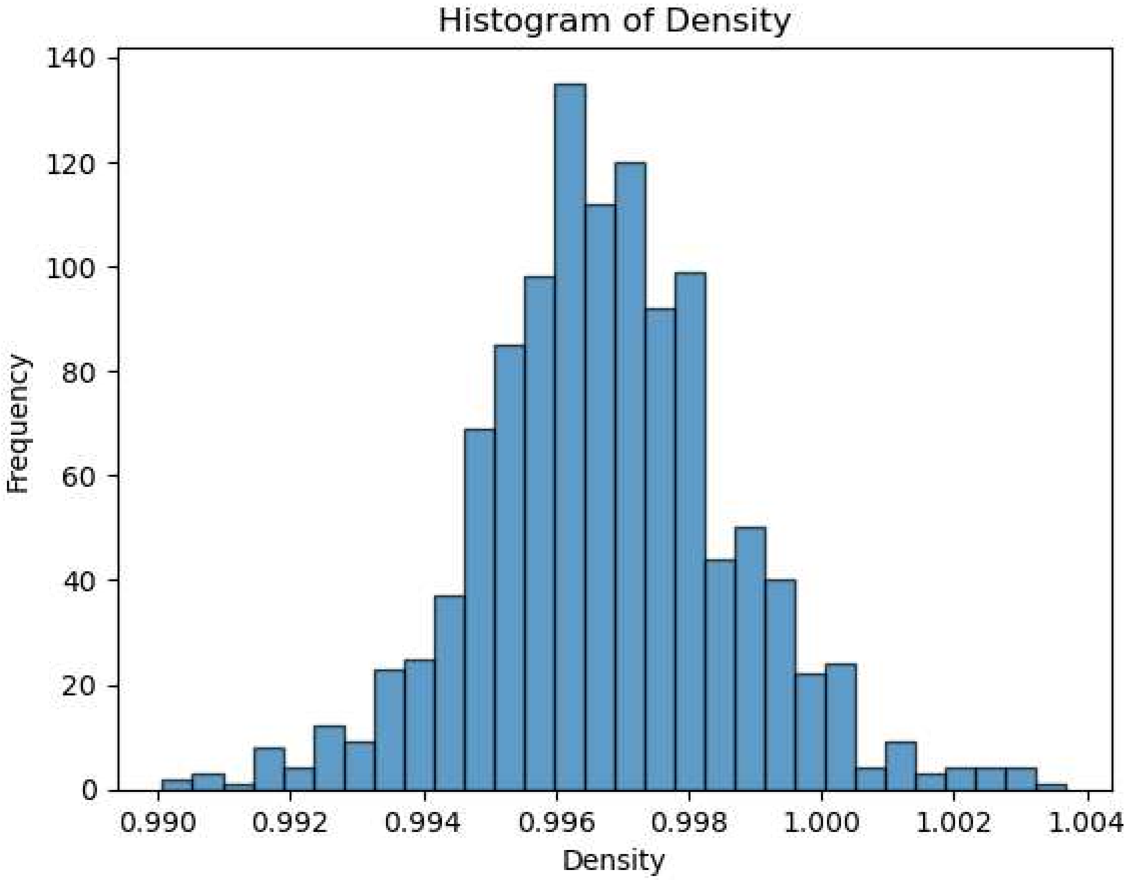
Question 12

In [118…

```
plt.hist(df['density'], bins=30, edgecolor='black', alpha=0.7)

plt.xlabel('Density')
plt.ylabel('Frequency')
plt.title('Histogram of Density')

plt.show()
```

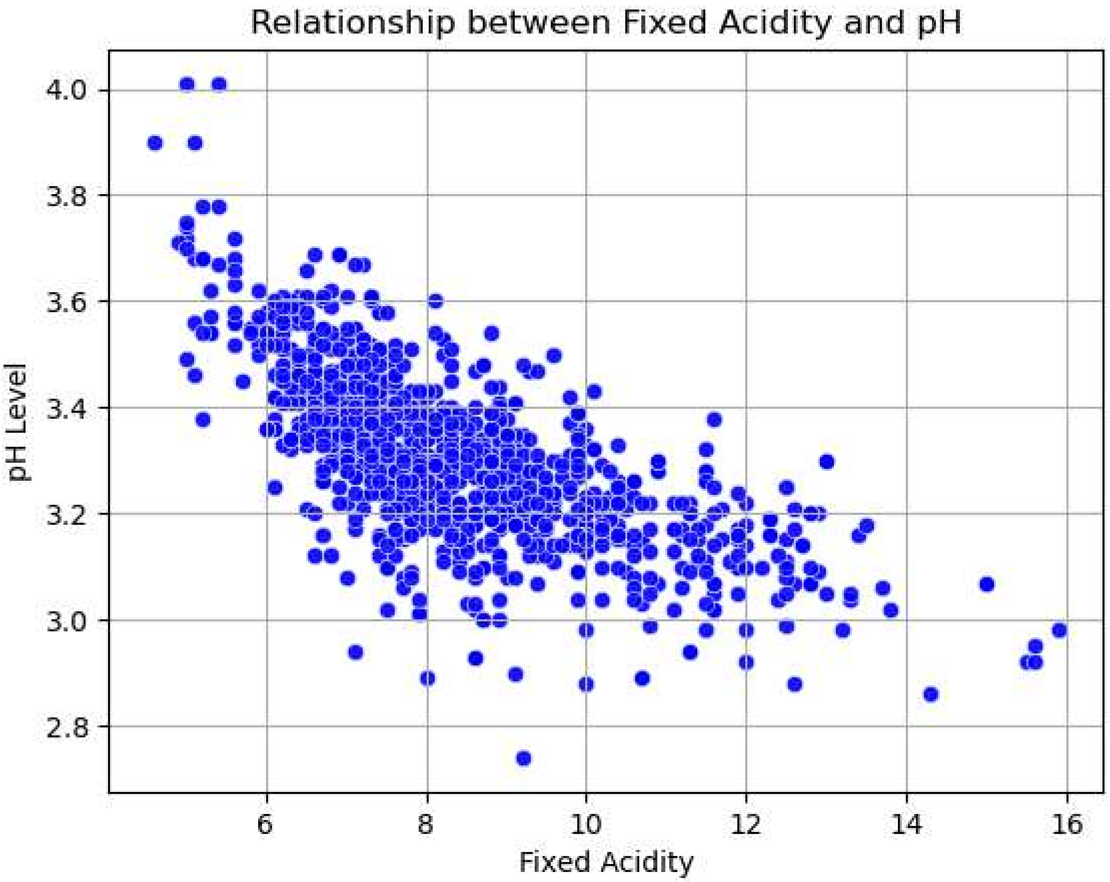


Question 13

In [122…

```
sns.scatterplot(x=df['fixed acidity'], y=df['pH'], alpha=0.7, color='blue')
plt.xlabel('Fixed Acidity')
plt.ylabel('pH Level')
plt.title('Relationship between Fixed Acidity and pH')
plt.grid(True)

plt.show()
```


Relationship between Fixed Acidity and pH
pH Level
4.0
3.8
3.6
3.4
3.2
3.0
2.8
Fixed Acidity
6
8
10
12
14
16

```
In [ ]:
```