

Exploratory Data Analysis

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.preprocessing import Imputer

In [2]: df = pd.read_csv("http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-
red.csv", sep = ';')

In [3]: df.head()
```

Out[3]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Describe() function in Order to Analyze Total Count , mean , standard deviation minimum , maximum and percentile values along the rows

```
In [4]: df.describe()
#sns.heatmap(df.isnull())

Out[4]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density				
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747				3.311
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887				0.154
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070				2.740
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600				3.210
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750				3.310
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835				3.400
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690				4.010

In order to do some analysis regarding null values we purposely make some values with np.NaN

```
In [18]: df.replace (2.3, np.NaN, inplace = True )

In [19]: df.head()

Out[19]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	NaN	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

HeatMap Analysis of Null values

In this heatmap we could provide different colors by using argument cmap and its various options such as icefire , inferno , pastel1 , viridis etc

```
In [20]: sns.heatmap(df.isnull())

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x252ad7b7390>
```

isnull().sum() gives the total count of Null values along particular coulmns

```
In [30]: df.isnull().sum()

Out[30]: fixed acidity      0
volatile acidity      0
citric acid           0
residual sugar       109
chlorides             0
free sulfur dioxide   0
total sulfur dioxide  0
density              0
pH                   0
sulphates            0
alcohol              0
quality              0
dtype: int64
```

Using Replace function to handle null values

```
In [31]: df.replace(np.NaN , 2 , inplace = True)

In [32]: df.isnull().sum()

Out[32]: 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
dtype: int64

In [34]: df.replace (2, np.NaN, inplace = True )

In [35]: df.isnull().sum()

Out[35]: fixed acidity      0
volatile acidity      0
citric acid           0
residual sugar       265
chlorides             0
free sulfur dioxide   1
total sulfur dioxide  0
density              0
pH                   0
sulphates            1
alcohol              0
quality              0
dtype: int64

In [55]: df[df['residual sugar'].isnull()].head()

Out[55]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
2	7.8	0.760	0.04	NaN	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
8	7.8	0.580	0.02	NaN	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
21	7.6	0.390	0.31	NaN	0.082	23.0	71.0	0.9982	3.52	0.65	9.7	5
23	8.5	0.490	0.11	NaN	0.084	9.0	67.0	0.9968	3.17	0.53	9.4	5
29	7.8	0.645	0.00	NaN	0.082	8.0	16.0	0.9964	3.38	0.59	9.8	6

```
In [53]: sns.heatmap(df.isnull(), cmap = 'viridis')

Out[53]: <matplotlib.axes._subplots.AxesSubplot at 0x252baba7eb8>
```

```
In [77]: df.head()

Out[77]:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	NaN	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Using dropna() function to handle missing values

```
In [93]: df.dropna(how = 'any' , axis = 0 , inplace = True )

In [94]: df.shape

Out[94]: (1332, 12)

In [95]: df.isnull().sum()

Out[95]: fixed acidity      0
volatile acidity      0
citric acid           0
residual sugar        0
chlorides             0
free sulfur dioxide   0
total sulfur dioxide  0
```