Wine Quality

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
#load the data set
df = pd.read_csv
('https://raw.githubusercontent.com/dsrscientist/DSData/master/winequality-red.csv')
df

#shows top 15 values
df.head (15)

#shows top 15 values
df.tail (15)

#shows the shpae of data type
df.shape

df.columns

df.columns.tolist()

#checking Data type of the columns
df.dtypes
df.isnull().sum()
```

```
df.info()
```

```
df.describe()
```

Key Observation

1. The Mean value is more than the median
2. There is an large difference in residual sugar, free sulfur dioxide, total sulfur dioxide for 75% and max.

```
a=sorted(df.quality.unique())

a
```

Key Observation

1. Quality score range from 3 to 8

```
df.quality.value_counts()
```

Key Observation

1. Quality has the most values in range of 4,5 and 6
2. very few observation found in 3 and 8

```
#for cheking the missing value

sns.heatmap(df.isnull())
```

```
# TO check co-relation

dfcor=df.corr()

dfcor
```

```
sns.heatmap(dfcor)
```

```
plt.figure(figsize=(6,4))

sns.heatmap(dfcor,cmap='Blues',annot=True)
```

Key Observation

1. Dark shade are highly co-related

```
plt.figure(figsize=(10,6))

sns.heatmap(dfcor,cmap='YlOrRd',annot=True)
```

Key Observation

1. Light shades are highly correleted
2. Quality is highly coreleted to alcohol
3. alcohol is negative correleted with density
4. Density is positively correleted with residual sugar
5. Volatile acidity is negatively correleted with the quality
6. Free sulfur dioxide is correleted with the total sulfer dioxide

```
df.columns.tolist()

df['fixed acidity'].plot.box()

df['volatile acidity'].plot.box()

df['citric acid'].plot.box()

df['alcohol'].plot.box()

df['free sulfur dioxide'].plot.box()

df['total sulfur dioxide'].plot.box()

df.plot(kind='box',subplots=True,layout=(3,4),figsize=(10,10))

sns.distplot(df['density'])

sns.distplot(df['citric acid'])

df.plot(kind='kde',subplots=True,layout=(2,6),figsize=(15,6))

plt.scatter(df['pH'],df['quality'])

sns.pairplot(df)
```

```
plt.scatter(df['volatile acidity'],df['quality'])
plt.show()
df.drop('volatile acidity',axis=1,inplace=True)
df.head()
df.shape
from scipy.stats import zscore
z=np.abs(zscore(df))
z
threshold =3
print(np.where (z>3))
df_new=df[(z<3).all(axis=1)]
df_new
```