Q1

volatile acidity : Volatile acidity is the gaseous acids present in wine. fixed acidity : Primary fixed acids found in wine are tartaric, succinic, citric, and malic residual sugar : Amount of sugar left after fermentation. citric acid : It is weak organic acid, found in citrus fruits naturally. chlorides : Amount of salt present in wine. free sulfur dioxide : So2 is used for prevention of wine by oxidation and microbial spoilage. total sulfur dioxide pH : In wine pH is used for checking acidity density sulphates : Added sulfites preserve freshness and protect wine from oxidation, and bacteria. alcohol : Percent of alcohol present in wine. Rather than chemical features, you can see that there is one feature named Type it contains the types of wine we here discuss on red and white wine, the percent of red wine is greater than white. Here's the use of Machine Learning comes, yes you are thinking to write we are using machine learning to check wine quality.

Q2

In the dataset, there is so much notice data present, which will affect the accuracy of our ML model. In machine learning, there are many ways to handle null or missing values. Now, we will use them to handle our unorganized data. We see that there are not many null values are present in our data so we simply fill them with the help of the fillna() function. Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. Advantages:- These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

Disadvantages: 1.Incompatible with most of the Python libraries used in Machine Learning:- Yes, you read it right. While using the libraries for ML(the most common is skLearn), they don't have a provision to automatically handle these missing data and can lead to errors. 2.Distortion in Dataset:- A huge amount of missing data can cause distortions in the variable distribution i.e it can increase or decrease the value of a particular category in the dataset. 3.Affects the Final Model:- the missing data can cause a bias in the dataset and can lead to a faulty analysis by the model.

Q3 These factors include pre-schooling background, family background, personal characteristics, college environment and learning habits etc. These factors may be broadly classified into academic and non-academic factors. Some of the previous research works related with the students' performance are discussed here. Some statistical techniques to analyze the data is:- 1.mean(),var(),std() 2.corr() of the data 3.also use describe() to see the full graph of minimum,maximum,count,percentile 25,75 and 100.

Q4 Feature Engineering Steps performance:- 1.Feature Creation 2.Feature Transformation 3.Feature Extraction 4.Feature selection 5.Feature scaling 6.Feature iteration 7.Feature split

Some Machine Learning models, like Linear and Logistic regression, assume that the variables follow a normal distribution. More likely, variables in real datasets will follow a skewed distribution. Sklearn has three Transformations-

1 Function Transformation 2 Power Transformation 3 Quantile transformation

In [2]:
```
#Q5
# import pandas
import pandas as pd

# import numpy
import numpy as np

# import seaborn
import seaborn as sb

# import matplotlib
import matplotlib.pyplot as plt
```

In [4]:
```
df=pd.read_csv("winequality-red.csv")
```

In [5]:
```
df.head(5)
```

Out[5]:

	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
0	7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
1	7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
2	7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;...
3	11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58...
4	7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5

In [8]:
```
df.describe()
```

Out[8]:

	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
count	1599
unique	1359
top	7.2;0.36;0.46;2.1;0.074;24;44;0.99534;3.4;0.85...
freq	4

In [9]:
```
df.isnull()
```

Out[9]:

	fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
0	False
1	False
2	False
3	False
4	False
...	...
1594	False
1595	False
1596	False
1597	False
1598	False

1599 rows × 1 columns

Q6

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process. Step 1: Standardization. Step 2: Covariance Matrix computation. Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components. Step 4: Feature vector. Step 5: Recast the data along the principal components axes

In [ ]: