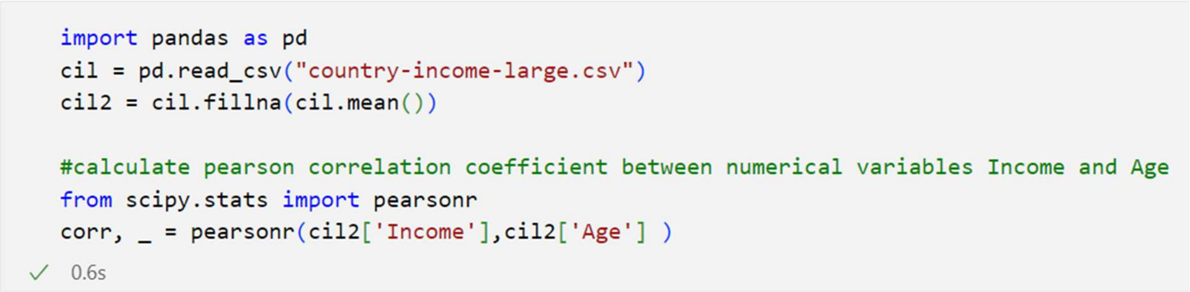
ASSIGNMENT -1

Q3 A. Performed the Pearson correlation coefficient calculation and obtained the correlation value for numerical attributes Income and Age which is 0.98. The pearson correlation coefficient is used to measure linear correlation between two variables by performing ratio of covariance and the product of their standard deviation.



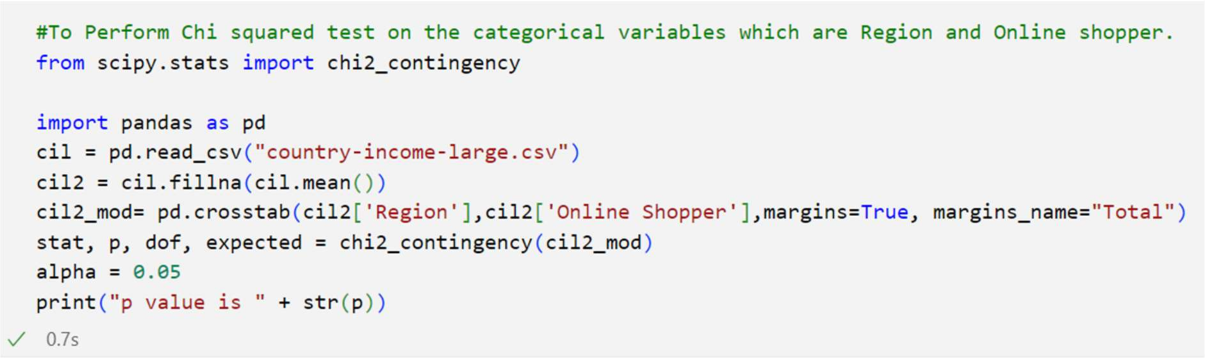
```
import pandas as pd
cil = pd.read_csv("country-income-large.csv")
cil2 = cil.fillna(cil.mean())

#calculate pearson correlation coefficient between numerical variables Income and Age
from scipy.stats import pearsonr
corr, _ = pearsonr(cil2['Income'],cil2['Age'] )
```

✓ 0.6s

The both attributes from the linear coefficient and data analysis can be inferred to be strongly correlated.

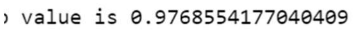
The Pearson correlation coefficient is not required to be calculated to evaluate the correlation between numerical attributes Income and Age, as it is inherently visible from analysing their data points. That is said to be correlated if a change or deviation in one variable causing the simultaneous similar magnitude of change or deviation in the other variable.

Q4 A. Chi squared test is implemented to evaluate the correlation between categorical variables, hence helpful in finding the magnitude of change in distribution of variable. Not in correlation that can be the contingency table.



```
#To Perform Chi squared test on the categorical variables which are Region and Online shopper.
from scipy.stats import chi2_contingency

import pandas as pd
cil = pd.read_csv("country-income-large.csv")
cil2 = cil.fillna(cil.mean())
cil2_mod= pd.crosstab(cil2['Region'],cil2['Online Shopper'],margins=True, margins_name="Total")
stat, p, dof, expected = chi2_contingency(cil2_mod)
alpha = 0.05
print("p value is " + str(p))
```

✓ 0.7s

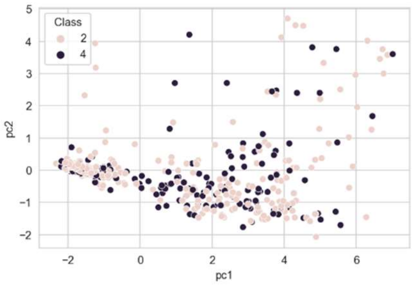


```
› value is 0.9768554177040409
```

Degrees of freedom = no of rows -1 * no of col -1 = 60-1 x 2-1 = 59.

The categorical variables hence can be observed to be not in any kind of correlation with expected and observed being in higher difference.

Q5 A. The observation of the resulting scatter plot from the pre-processed PCA is not that fetching as the classification of the two component class projections can be seen of have an unclear line of classification. Hence, it can be used to strengthen the statement that PCA low interpretability.

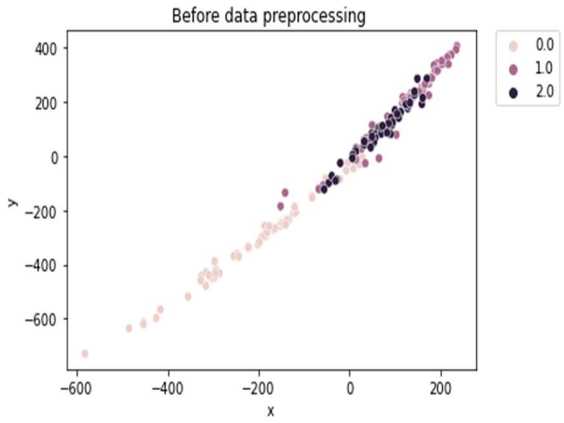


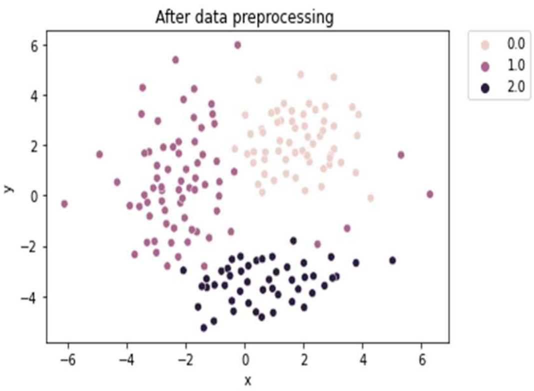
Part 2

Q1 A. The quantile-quantile plot of parental income being below the 45 line stating that the it is not a normal distribution. Since the data points are below 45 line more towards the x -axis quantile which is the parental income of individuals of people graduate over high school. It signifies that data has more distribution towards that.

Whereas, in the years to graduate quantile-quantile plot it can be observed that some in normal distribution and others not being normally distributed and distributed across both quantiles.

Q3 A. The plot before pre-processing is can be seen below.

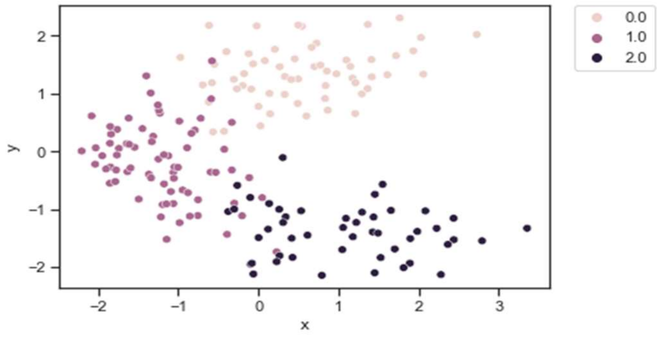




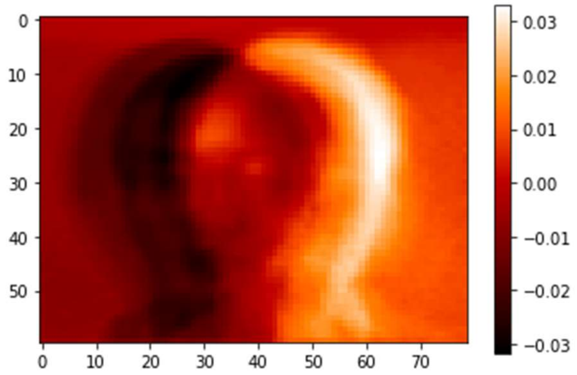
The plot was not satisfactory and needed pre-processing. The use of data normalization pre-processing step before performing MDS, has significantly helped in defining clusters of classes for the data being easily distinguishable and quality of the plot is improved.

Q4 A) On the inspection of the scatter matrix plot, the classes can be seen to distinguishable clusters for features such as color intensity and od280/od315_of_diluted_wines. The plot is obtained after performing standard scaler pre-processing and MDS to the data with the above two features for the three classes.

The advantage of feature selection is to get better analysis and functionality usage for model. Whereas the disadvantage would be dimensionality reduction and loss of data information considering the final output deployment of model.



Q5 A) The mystery.csv is a numerical array which is the decomposed data of an image which can be seen to be a portrait of a male from the projection PCA heatmap. So the dataset contains the numerical representation decomposition of an image.


0
10
20
30
40
50
0
10
20
30
40
50
60
70
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03