**Inferential Statistics – for capstone 1**

1. *Check collinearity between independent variables*

A heatmap was plotted to visualize the correlation between all the variables. The plot showed that there is strong relationship between tuition fee and cost of attendance. In order to further substantiate the relationship, we used scatter plot (sns.plot). Finally, we used bootstrap method to check the statistical significance of the collinearity between tuition fee and the cost of attendance.

* Correlation coefficient between the two variables was estimated using numpy package (np.corrcoef(x,y)). The estimated value was 0.97
* We also tested the significance (for alpha =0.05) of the value using bootstrap method:

1. Null hypothesis: Cost of attendance is completely uncorrelated to tuition fee.
2. Alternate hypothesis: Cost of attendance and tuition fee are correlated.

* The p-value for the test was 0.0, this confirmed that the cost of attendance and tuition fee are statistically strongly correlated. Also, since the estimated value of correlation coefficient is high we can say that it is also practically significant.

1. *Compare mean income of students graduating from colleges that gives predominantly graduate degrees against other colleges*

Violin plot was generated to visualize the spread of income for students from different colleges based on the type of degree predominantly awarded by the college (1. Certificate, 2. Associate, 3. Bachelor, 4. Graduate). The graph showed that mean income was higher for colleges that awarded graduate degrees. We further checked it using statistical analysis

* We separated the colleges in two groups: colleges that predominantly give graduate degrees and colleges that do not predominantly give graduate level degrees.
* The empirical difference between the means of the groups was around $ 61,684, with mean income of colleges giving predominantly graduate degrees around $94, 447.
* Bootstrap method along with the t-test using stats package were performed to check if the difference was statistically significant.

1. Null hypothesis: students from graduate level colleges have similar salaries as the student from other colleges
2. Alternate hypothesis: students from graduate level colleges have salaries significantly higher as compared to the students from other colleges.

* Both the tests suggested that the salaries of the student from colleges offering graduate degrees were significantly higher.

1. *Check the normality of the Income data*

Histograms and empirical cumulative distribution functions were plotted for the Income data. Also, to confirm the normality we performed chi-square test.

* Both the plots suggested that, while the data in the middle looks close to normally distributed, the data at lower and higher extremes (could be outliers) deviate the plot from being normally distributed.
* We also tested the normality of the income data using chi-square test. Finally, we concluded that the data is not normally distributed.
* We also checked the normality of log transformed Income data, and it was observed that the log transformed income data is normally distributed.

**Recommendations for further analysis and building the model:**

* We can eliminate one of the variables, either tuition fee or the cost of attendance.
* We can divide the dataset into two groups based on if they awarded the graduate degrees or not.
* We can use log transformations of the income data for further analysis.