PROJECT FINAL REPORT

Goals and business objectives:

Certain college named UVW's marketing department is planning to increase the enrollment in their college. They need to derive various insights from the given data, so that they can target specific individuals with their marketing strategy and become successful in boosting the admissions.

I have produced data-visualized reports in my position as a data analyst for UVW College and assist them in choosing the appropriate demographic groups to target for enrollment growth. An individual's income is determined by a number of important criteria, with the $50,000 mark serving as a vital reference point in the statistics. I will look at the variables that affect income and utilize some evaluation of input parameters to estimate income to help us accomplish our enrollment goals.

Assumptions:

For this project, in order to achieve the intended business objectives, there were multiple assumptions taken into consideration. I was provided with the data of a certain college to boost their enrollment. The initial assumption I relied upon was that the data I was working with is accurate, complete and reliable for the goal of business. So, I believe that the choice of certain variables aligns with our objective.

One assumption was to establish the standards for marketing its programs as UVW College decided on a salary of $50,000 as a key demographic. The marketing department of the college intends to develop profiles based on variables like gender, education level, age, ethnicity, income, etc. A skewed extract from the 1994 US Census database serves as the study's data source.

After thorough analysis of the data, certain assumptions were made, including the assumption that pre-processing of the data would not impact the decision-making process or alter the distribution of the data. As a part of pre-processing, duplicate rows were removed, and the '?' values were replaced with the column mean values or dropped for certain rows. There exists a net value for capital gains or losses, as the two values are not combined and are not equal to zero.

User Stories:

In the initial attempt to understand distribution of income, I generated a countplot for the income variable, which provided insights that were further visualized in a pie chart. The pie chart revealed that the proportion of individuals with incomes below $50,000 is three times higher than those with incomes exceeding $50,000. On behalf of the marketing team, I have prioritized five user stories. Two of these stories involve univariate analysis, while the other three involve multivariate analysis. For these analyses, I have utilized eight variables, namely Age, Education-num, Sex, Race, Capital-gain, Capital-loss, hours-per-week, and relationship, in order to gain insights and identify relationships among the data cases.

User Story #1:
The marketing team wants to find the relationship between age and income.

User Story #2:

The executive in charge of marketing takes interest in the education-number because education level affects a person's salary.

User Story #3:
The marketing director is curious about the influence of race and sex on a human being's income.

User Story #4:
A marketing intern wants to know the ratio of capital gains to capital losses in changing a person's income.

User Story #5:
The marketing group is interested in investigating the effectiveness of hours-per-week and relationship of a person on his/her income.

Visualizations:

Beginning with the first user story, it was simple to find and understand the relationship between a person's age and salary. Age is a continuous variable while income is a categorical variable with two values: < $50K and > $50K. With $50K being the important criteria for the college, I created bar plots for groups with income <= $50K and income > $50K. I found it significant because it could help the marketing team to target specific age groups and increase the enrollment. Upon analysis, I found that individuals with income less than or equal to $50,000 tend to be younger compared to those with income over $50,000. The mean count for the former group falls between 30 to 40, while for the latter group falls between 40 to 50. Furthermore, the box plot helped me understand that the former group exhibits a wider range and greater variability.
Hence, the targeted individuals must be having age less than 40 years, in order to get them enrolled into the university.

[Refer to page no. 5 for the visualization of the bar plot.]

For the second story, I considered the relationship between the number of levels of education and income. I'm trying to help increase enrollment at UVW College, so it was important to plot against income using education numbers. The value of education level ranges from 1 to 16, where 1 is the lowest and 16 being the highest level of education. I created boxplots for both income groups. The results were astonishing. The median for the low-income group is 10, and the median for the rest of the group is 12. The former group has a minimum age of 1 and an outlier range of up to level 16, the highest level, which indicated a narrower boxplot range and greater variability. Another difference is that the group with income above $50,000 has 10 in the lower quartile, 8 and 11 in the other group, showing that those with higher levels of education have higher income, concluded to be high.

Hence, the individuals who are asked to join the UVW college, must be having their education level between 8 and 11, to be successfully enrolled.

[Refer to page no. 6 for the visualization of the box plot.]

After age, education I found sex and race to be an important parameter for income. So, I decided to understand the income relationships between different racial groups and different genders. The criteria for gender is Male or Female, and for race we have White, Black, Asian Pacific Islander, Amer Indian Eskimo and Others. This is an essential task, as targeting low income minorities can be easy and fruitful. As all the three variables are categorical, I decided to use a mosaic plot. There were similarities

between groups across races, with a higher percentage of women earning less than $50,000 than the men earning less than $50,000 in each race. Therefore, it is possible to attract women especially to go on to university. The groups, in descending order of race, were White, Black, Pacific Islander Asian, Amer Indian Eskimo, and Other, with incomes greater than $50,000. There are more men with higher income who are White and Asian Pacific Islander. Also, there are almost nil women having race as Others.

For this specific group, applying the marketing strategies for any race of women is fruitful. While for men, who are Asian Pacific Islander they are easy to target for college enrollment.

[Refer to page no. 6 for the visualization of the mosaic plot.]

Another user story consists of depicting the parameters capital-gain and capital-loss against the income groups. Capital-gain and capital-loss are two continuous variables. This was an interesting plot and I figured out more information than I had expected. Apart from other characteristics, it is important to consider the financial gain or loss to get a person enrolled in college. One significant derivation was that the values of capital-gain and capital-loss are not zero not together. Hence, I plot a scatter plot with color coding. One simple and easy prediction was that the group with income <= $50,00 had more capital-losses counts. The group with income > $50,000 had significantly more capital-gains.

As per the derivation from the charts, the group of people with more capital-gain are more likely to enroll and get admitted to the university, as they have income <= $50,000.

[Refer to page no. 7 for the visualization of scatter plots with color coding.]

For the final story, I was having a dilemma in selecting the parameters. After multiple trials, many visualizations and combinations of features, I was firmly convinced to use hours-per-week and relationship with income to select a particular group of individuals for college. Hours-per-week is a continuous variable and relationship is a categorical variable with different values as Husband, Not-in-family, Wife, Own-child, Unmarried, and Other-relative. Hence, I plot grouped bar plots for groups with different incomes. For each relationship, people with higher working hours had higher income than those with lower working hours in a week. It was quite noticeable from the plot. In this case, they had 45 or more than 45 hours-per-week. The distribution of income and hours-per-week was equivalent in individuals who were unmarried or were not-in-family. This explains how being husband, wife or having family impacts on income as well as the working hours for the week. So, to derive individuals who were wives or had their own-child had remarkably less working hours than 40.

Individuals being a wife or having a child, and working less than 40 hours per week has a high probability of having income less than $50K, so they can be easier to convince and get an admission in the college.

[Refer to page no. 7 for the visualization of grouped bar charts with color coding.]

Based on the results, certain factors appear to be strongly associated with individual income, and these factors appear to be strongly associated with individual income, and these factors can be used to target specific demographics. After analyzing the data, I was able to identify eight of the parameters that greatly influenced a person’s income.

So, for those parameters I have the subset of the data chosen, which can be encouraged with marketing strategy to join the college and help them increase their income. Hence, to combine an individual with age less than 40 years, having education level between 8 and 11, an individual being a woman and working less than 40 hours per week, apart from being wife or having a child, are the narrowed down criterias for the effortless profiles interested in joining the college UVW, after a successful campaign by the marketing team.

Questions:

I took some time to familiarize myself with the data in the adult.data file and the parameter information in the adult.names file. I first decided to convert the data file into an excel (.xlsx) file with the aim to present the data in tabular format. As a result, it is simpler to grasp and assess the data. Another important question was about having data errors which need to be addressed before visualizations and drawing the conclusions. So, it was regarding the special character '?' and dealing with the missing values. As part of the cleaning process, I removed some rows. While for the columns with int64 data type, adding column mean to other rows was reasonable.

I encountered a challenge in determining how variables interact with each other and also had a significant impact on the income parameter, and ultimately relied on the insights gained from data visualizations to make my selection. Hence, I ended up dropping the id, finalwgt, and native-country columns. Also, ensuring that the visualizations are appealing and appropriate for the intended audience.

Visualizations and Charts:

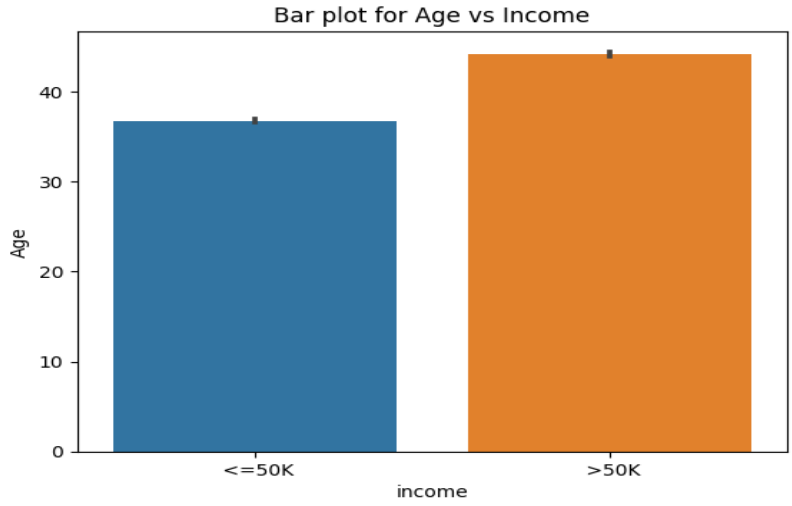
<u>USER STORY #1</u>

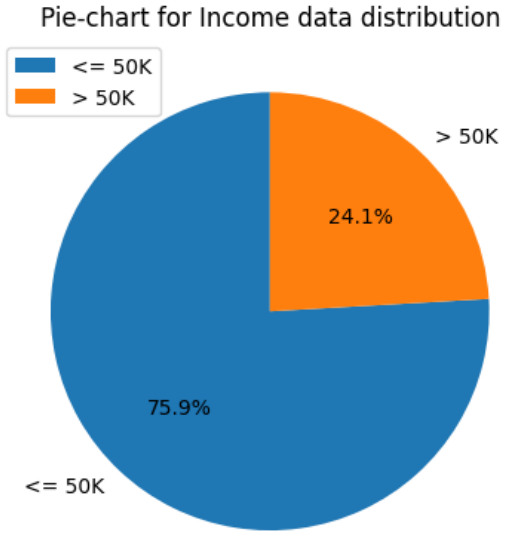
Not doing:

I decided not to mess with the data more, though I thought a lot more could be done. After studying the data, I found that some features can be combined to create one feature. This can help lower the dimension. Considering the example of education and education-num. These features provide redundant information, as they explain similar information. Likewise, by calculating the net-capital using positive values for capital-gain and negative values for capital-loss, we can reduce dimensionality and better understand the correlation between variables, uncovering patterns and structures in the data.

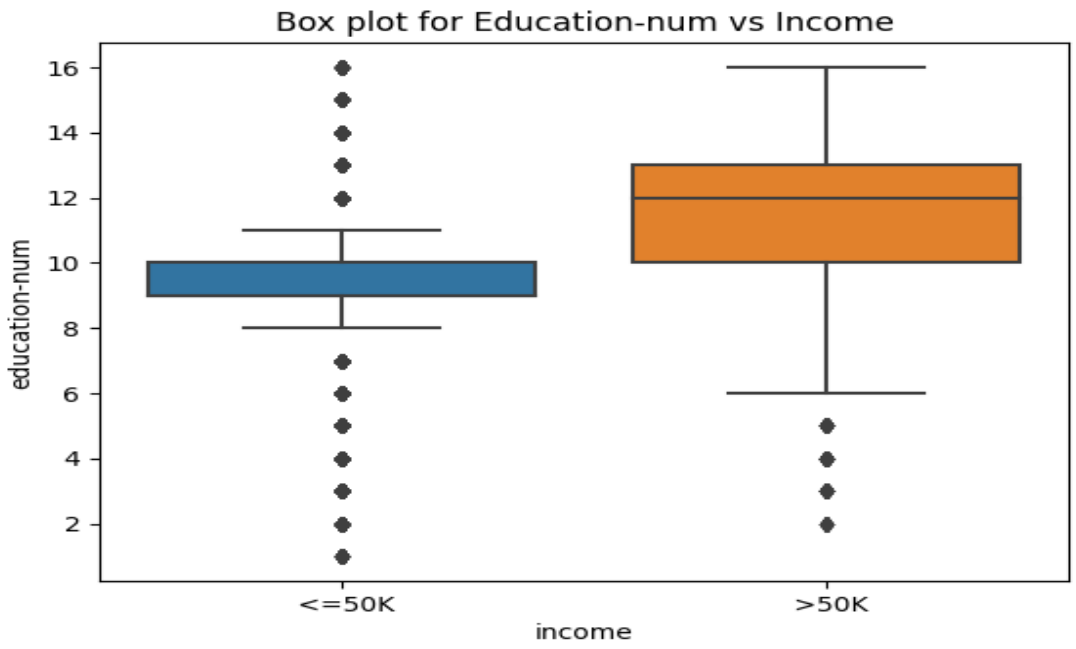
Another task I am looking forward to is developing a machine learning model that can precisely forecast a person's salary based on the data provided excites me. To optimize the performance of the machine learning model, several steps would be taken. Firstly, thorough data cleaning would be performed to rectify any errors or inconsistencies. Next, variable selection would be conducted based on their correlation with income values. Subsequently, various machine learning algorithms such as logistic regression, support vector machines, decision trees, and random forests would be implemented. The accuracy of these models would then be evaluated, and the one with the best results would be selected for further refinement and deployment.

Therefore, as a result of the project, I was able to derive insights from the data and able to help the marketing team of UVW college in creating certain target profiles and elevate the admissions significantly.

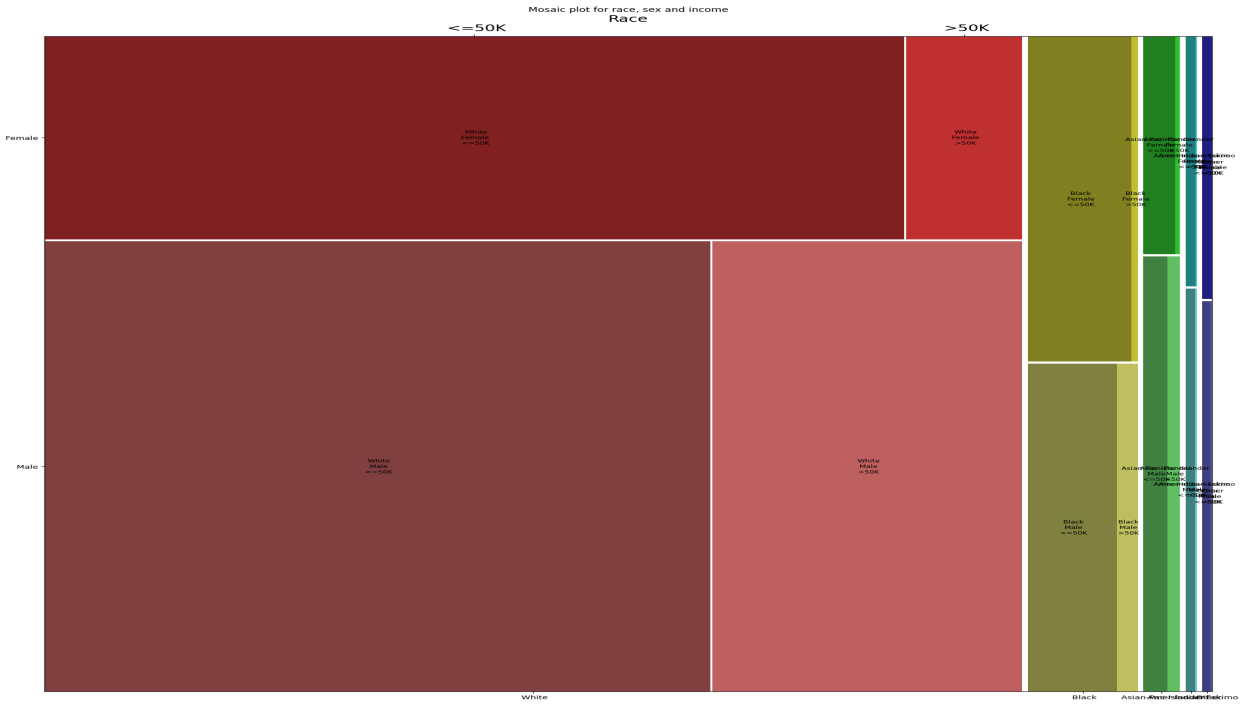

Bar plot for Age vs Income
40
30
20
10
0
Age
<=50K
>50K
income


Pie-chart for Income data distribution
<= 50K
> 50K
> 50K
24.1%
75.9%
<= 50K

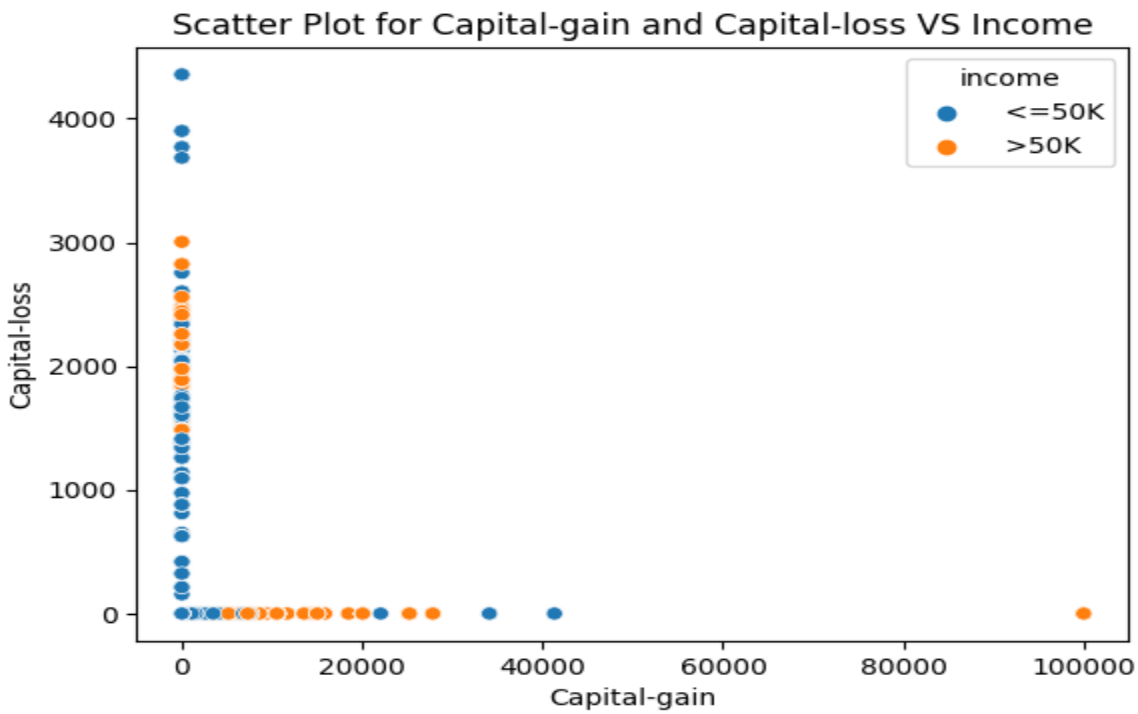
USER STORY #2



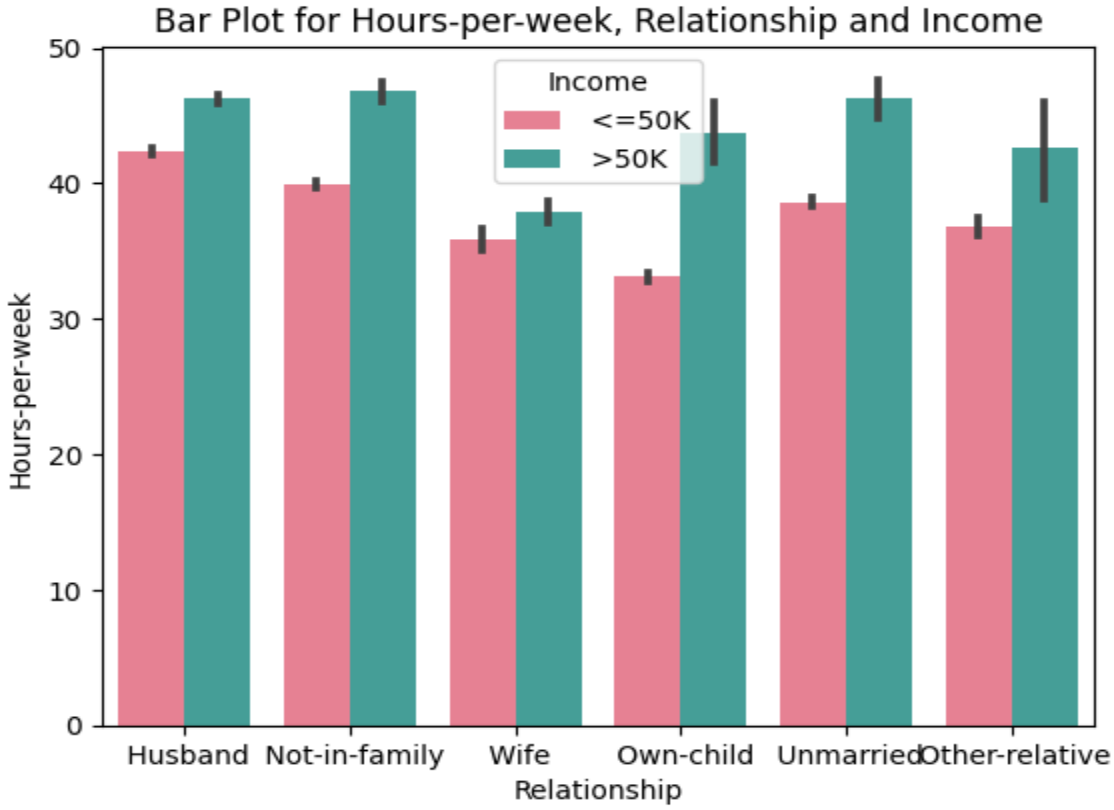
USER STORY #3 Race [White, Black, Asian Pacific Islander, Amer Indian Eskimo, Other]



USER STORY #4



USER STORY #5



Appendix:

Python code for Data visualization is in another PDF.