CSE 578: Data Visualization

Course Project Final Report – Aasish Tammana (1225545568)

Goals and Business objective

This project intends to produce marketing profiles and data visualizations for UVW College using data from the US Census Bureau. The main goal is to determine the important variables that affect a person's income, particularly those who make more than $50,000, in order to create a program that can forecast a person's income based on various input parameters. 32,561 rows of pertinent demographic information from the given data collection, including age, gender, educational level, marital status, and employment, will be extracted for analysis.

The project will include statistical analysis to find connections and trends in the data and as a future work it could include the creation of prediction models that can precisely anticipate income based on the found variables. Advanced data analytics technologies, techniques, and data visualization tools, will be used to do this.

The analysis's findings will be used for UVW College's marketing plans in order to inform and improve them. UVW College can boost the enrolment rates and their total return on investment by focusing their marketing efforts on the precise demographic groups that are most likely to enrol. The initiative seeks to offer useful insights into the elements that affect people's enrolment choices as well as pointers on how to create successful marketing campaigns that connect with various target audience segments.

Assumptions

We assume that data needs to be processed in order to make it suitable for analysis. This presumption includes data cleansing, feature engineering, and normalizing /standardizing. This is evident because of the missing numbers i.e., "?" values.

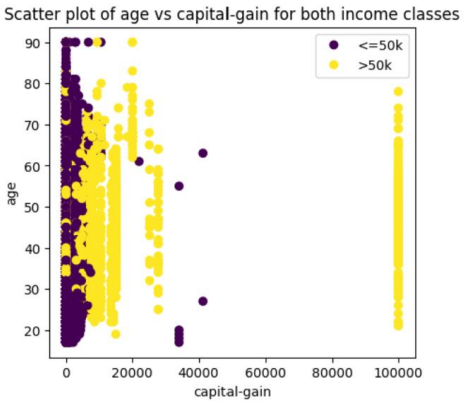
We assume that the United States Census Bureau's dataset is dependable and accurate for identifying the variables that affect a person's income. Also, there aren't any significant problems with the quality of the data after pre-processing, such as missing numbers, outliers, or inconsistencies. Having inaccurate data could result in having erroneous predictions and might result in incorrect conclusions.

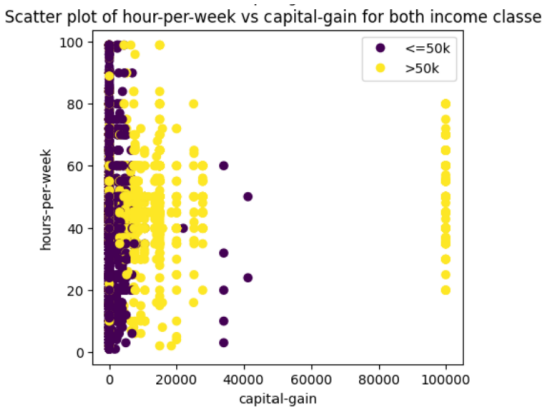
It is assumed that the elements that influence someone's income are independent of one another, which means that changes to one factor do not change the values of the others. However, certain characteristics, like occupation and education level, may interact or depend on one another and have an impact on how accurately the patterns are perceived.

The target demographic for UVW's marketing efforts is interested in pursuing higher education. The marketing profiles developed using the data analysis will be effective in increasing enrolment at UVW

User Stories

1. **Scatter Plot – Age vs Capital Gain and Hour per week vs Capital Gain**





To Determine if Age and Hours worked per week are related to the Capital Gain and check if they influence income.

The scatter plot produced illustrates the link between income classes, age, hours worked per week, and capital gains. Particularly for people who earn more than $50,000, it pinpoints the factors that significantly affect revenue.

It demonstrates the correlation between age / weekly hours and capital gain for both income classes. The dots are coloured according to the income class (<=50k or >50k), and the x-axis represents capital gain; the y-axis represents age/hours per week.

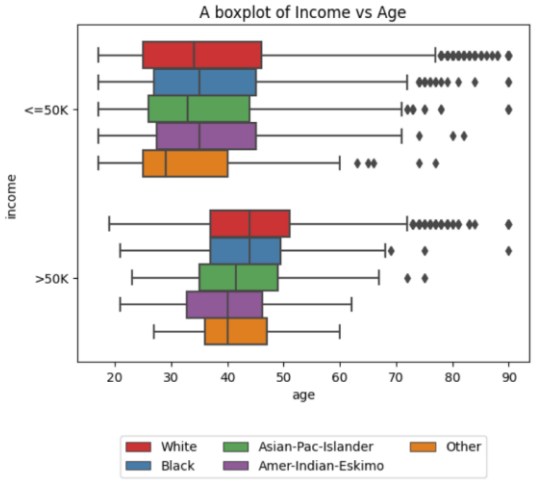
The following steps were used in the scatter plot's design process:

- Set the capital-gain, age, and hours-per-week variables that will be plotted.
- Choose a visualization style that is appropriate for the variables and use scatter function for implementing the scatter plot.
- Select the right colour scheme for the different income levels.
- Set the plot's axis labels and titles.
- To differentiate across income classes, the plot should include a legend.

Inference: It demonstrates that those with bigger capital gains tend to be older and work more hours per week, and that there is a higher concentration of people with incomes over $50,000 in those categories. This implies that these factors may serve as markers of a person's income level and be helpful in making income predictions.

We can also probably say that very high capital gains pushed people's incomes greater than $50000 and this can be observed with a clear distinction on the visuals.

2. Box Plot – Income vs Age with Race



To identify the pattern between and age Race and determine which ethnicities have higher incomes.

The relationship between a person's age, income, and race is illustrated by the box plot. It demonstrates a link between age and income that is positive, indicating that as age increases, so does the median income. The picture also demonstrates that some racial groups have higher median incomes than others, with Black people having the lowest median income and Asian-Pacific Islanders having the highest. This visualization was chosen because it offers a clear grasp of how income varies by age and ethnicity, which could be useful in pinpointing the key factors that influence someone's income.

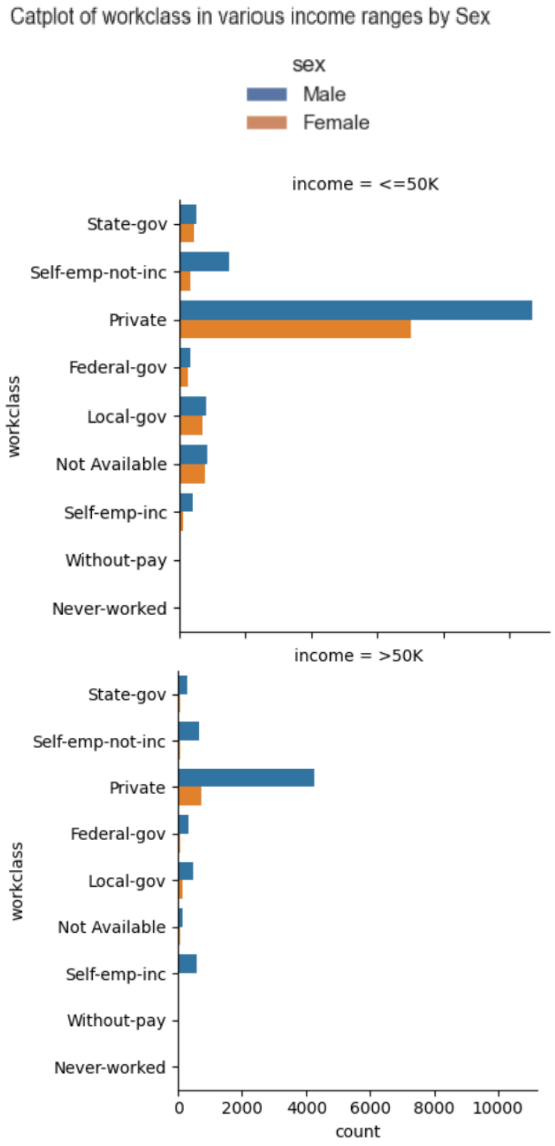
The following phases make up the design process for producing the box plot visual:

- Import the required libraries, Seaborn and Matplotlib. Create a Pandas data frame by loading the dataset.
- Determine the box plot's x-axis, y-axis, and colour, with age on the x-axis, income on the y-axis, and race serving as the grouping variable.
- Use the "Set1" palette to specify the colour scheme for the plot and set title.

- Decide where and how the legend will appear, ideally at the plot's base.

Inference: The box plot visualization offers important insights into how age, income, and race interact. It shows that age and income are positively associated, with older people often earning more than younger people. It also demonstrates that race is a crucial determinant income. In particular, white people have the highest median income, followed by the Asian-Pacific Islander group, while black people have the lowest median income.

3. Categorical Plot – Work-class vs Income with Gender



Marketing wants to know the relationship between the Work-Class of an individual, their gender and income.

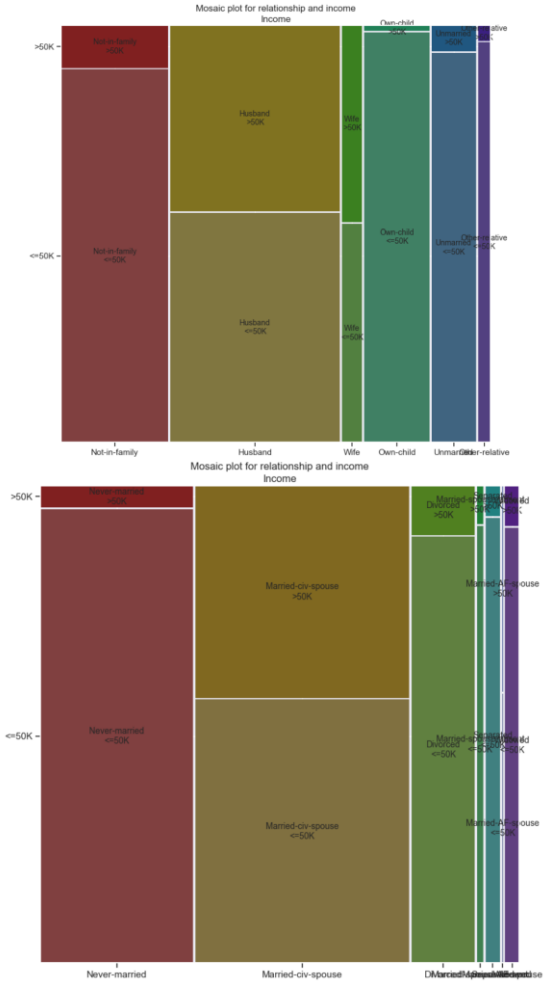
A categorical plot that displays the distribution of categorical data is called a catplot. The catplot visualization will help UVW College marketing team to understand the distribution of work-class by sex and income levels.
The graphic in this instance displays the distribution of work-class by men/women across various income brackets. The graphic is helpful because it enables us to compare the distributions between sexes and visualize how each category is distributed throughout each income level.

To create the visualization, we followed the following steps:

- Import the required libraries for data manipulation and visualizing. We imported the Pandas and Seaborn libraries.
- We used the Pandas read_csv function to load the dataset.
- We used Pandas tools to remove any missing values and change categorical into numerical columns.
- To visualize work-class by sex in various income brackets, we utilized Seaborn's catplot tool. We designated the work-class y-axis, the income row, the sex as hue, and the kind as count.
- We used Seaborn's fig.suptitle function to add a title to the plot and fig.subplots_adjust to change the layout.

Inference: According to the visualization, the vast majority of people making under $50,000 have the work-class "Private," while people making above $50,000 have a variety of work-classes, including "Self-emp-inc" and "Federal-gov." Additionally, it shows that the majority of people across all income brackets and genders belong to the "Private" work-class.

4. Mosaic Plot – Relationship vs Income and Marital-Status vs Income



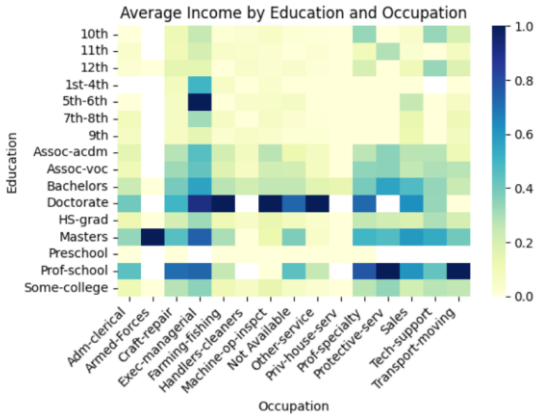
A mosaic plot in the first visual shows the distribution of income levels based on a person's relationship status. A second mosaic plot shows the distribution of income levels based on a person's marital status. If there is a large disparity in income levels depending on either type of connection, the plot can reveal it.

The steps in the design process were as follows:

- Determine the important factors: The relationship, marital status, and other demographic characteristics were selected as the predictor variables, with the income level variable as the target.
- The data was cleaned and put into a format that could be used for analysis. The required variables were chosen for analysis, and missing data and outliers were handled.
- Because the mosaic plot offers a clear and succinct portrayal of the relationship between categorical variables, it was chosen as the visualization type. The income level is represented by the x-axis, while the predictor is represented by the y-axis.
- The 'mosaic' function in the 'statsmodels.graphics.mosaicplot' pack was used to construct the mosaic plot. The figure was updated with the required labels and captions, and the x-axis tick labels were rotated for improved visibility.

Inference: According to both of the visualizations, there is a sizable variance in income levels depending on an individual's marital status and relationship status. Most people who are married earn more than those who are single or divorced. Clearly both visuals indicate husband's / married status individuals dominate both the categories of incomes.

5. Heatmap – Average Income by Education and Occupation



Note: The income values have been scaled to 0 and 1 to generate the heatmap where 0 indicates <=50k and 1 indicates >50k.

The staff would like to understand more about the group of people based on their education and occupation.

The visualization displays the average salary for each combination of education and occupation using a pivot table and heatmap. UVW College can utilize this information to sell its degree programs to people with various income levels and educational backgrounds.

The following phases are part of the design process for the pivot table and heatmap:

- The information is tidied up and put into an analysis-ready format.
- Using the pandas pivot_table() function, a pivot table is built with education as the index, employment as the columns, and scaled income as the values.
- Using the pivot table as input and the YlGnBu colormap, a heatmap is produced using the seaborn heatmap () function.
- Using the tight_layout() function, the plot is personalized by including a title, x-axis and y-axis labels, rotating the x-axis labels, and changing the layout.

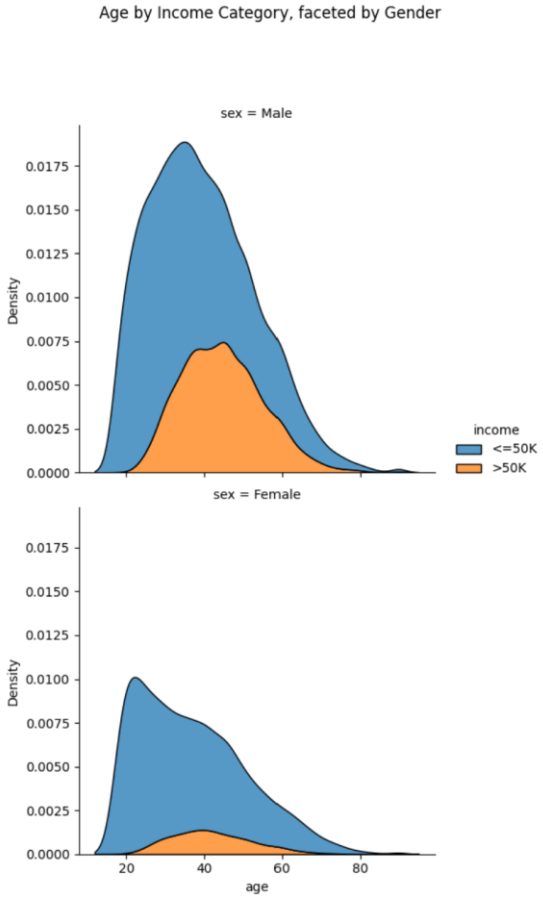
Inference: The pivot table and heatmap visualizations show that there is a distinct link between education, occupation, and income based on the provided dataset. For instance, people with a bachelor's degree or higher often earn more money on average than those with only a high school diploma or less. Furthermore, compared to other occupations like farming/fishing, private home services, and other service roles, some occupations like executive/managerial roles, professional specialties, and technical roles typically have higher average wages. It is also observed that people with doctorate as education, generally have higher incomes irrespective of the occupation.

6. Additional User Stories and Visualizations

A) Violin Plot – Hours per Week vs Sex

Inference: According to the violin plot, there are some disparities in the number of hours that men and women labor each week. Males often put in more hours per week than females, especially for those with incomes under $50,000, when the spread of hours worked is broader. The distribution of hours worked for people making over $50,000 is narrower for both men and women, indicating that there is less variation in the number of hours worked each week for people in this income range.

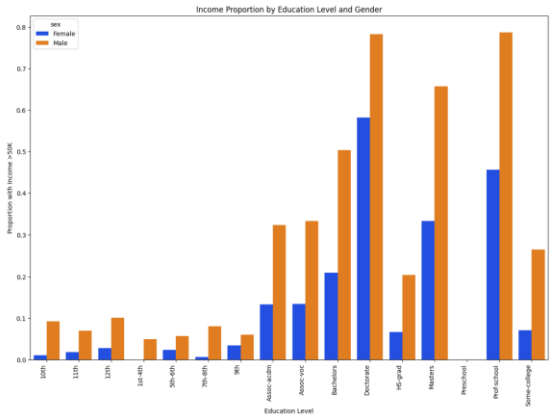
B) Density Plot–Age by Income, Gender



The graphic is faceted by sex and income group to facilitate comparison and employs a kernel density estimate to provide a smooth approximation of the density of age values.

Inference: The plot implies that both male and female age distributions are very comparable, with a minor lean towards younger ages. However, the age distributions for people with incomes above and below $50,000 clearly differ from one another. For both males and females, the density of people with incomes below $50,000 is higher when they are younger, whereas the density of people with incomes above $50,000 is higher when they are older.

C) Barplot - Income Proportion by Education Level and Gender



The percentage of people with incomes over $50,000 is displayed on the graph, with education and gender distinctions. Education is shown on the x-axis, and people earning more than $50,000 are shown on the y-axis. The bars are stacked according to gender, with blue denoting men and orange denoting women.

Inference: The graph demonstrates a definite positive link between education level and the percentage of people making above $50,000 per year. As education level rises, the percentage of people with incomes over $50,000 rises for both men and women.

The graph also reveals that men are more likely than women to have incomes of at least $50,000 regardless of education level. At higher levels of education, the proportional gap is very noticeable. This shows that, even after taking education level into account, gender may still play a role in determining income level.

Questions and Solutions

1) What are the source and quality of the dataset?

The United States Census Bureau provided the dataset for the research. For determining the factors that influence a person's income, it is presumable that the dataset be trustworthy and precise. We have cleaned the missing values in the dataset and handled them by replacing with appropriate values. Feature engineering has been used to create new variables that can better capture the relationship between variables.

2) What tools and technologies should be used for data visualization and analysis?

The dataset will be analysed using cutting-edge data analytics tools, methods, and technology. The particular technology and techniques used will depend on the requirements of the project and the use case. However, famous Python packages like Pandas, Matplotlib, and Seaborn have been used for data visualization and analysis of this project.

3) How are the variables being analysed to determine their relationship with income?

To decide this, we create multiple visuals using all the features (14 provided) and try to analyse any patterns in the data. We can observe that some of them more distinctly impact the outcomes and we can rank the features appropriately to decide which ones to use. We finally come up with visuals to add to user stories.

4) What types of visuals should be used for data analysis?

For categorical data, mosaic plots (multivariate) and heatmaps were utilized since they accurately depict categorical data.

Visualizations including the histogram, box plots, scatter plots, and parallel coordinate plots were employed for continuous data.

5) How to handle implementation errors while working on user stories?

Several mistakes made while implementing and testing the different visuals. resolved after a thorough internet search that included references to the materials and thorough testing.

6) How to handle the complexity of Prediction?

For the purpose of this project, we have not worked on prediction and concentrated solely on implementing the user stories.

Not Doing

- All though exploratory analysis was performed on the provided dataset, no machine learning models were involved for visualising the data. This could be a valuable future prospect to generate some meaningful insights.
- Although we have pre-processed and cleaned the dataset, there is always a possibility that attributes of data were not handled and this can only be resolved with through Quality Assurance testing.
- Implementation of Prediction based models for machine learning analysis on the provided data. This would be a very important task to achieve more satisfactory results of developing marketing profiles. We can come up with suitable algorithms to train and test the data.
- Using a data visualization software such as Microsoft Power BI/Tableau in order to create a dynamic interactive dashboard of the visuals.

Appendix

Provided as a separate attachment which consists of codes for all the visuals developed for the purpose of this project. It also includes visuals which were created as a part of exploratory data analysis.