Ira A. Fulton Schools of Engineering
Arizona State University

CSE 578: Data Visualization

Course Project Final Report

Goals and Business Objective:

The overarching goals of this project encompasses conducting a through analysis of United States Census Bureau data to craft detailing marketing profiles for ASU College, developing an innovative application for predicting individual income based on key demographical variables, and facilitating the tailoring of effective marketing strategies. The ultimate business objective is to empower ASU College with a data-driven approach, optimizing marketing efforts and contributing to increased enrollment by leveraging synthesized demographic insights.

Assumptions:

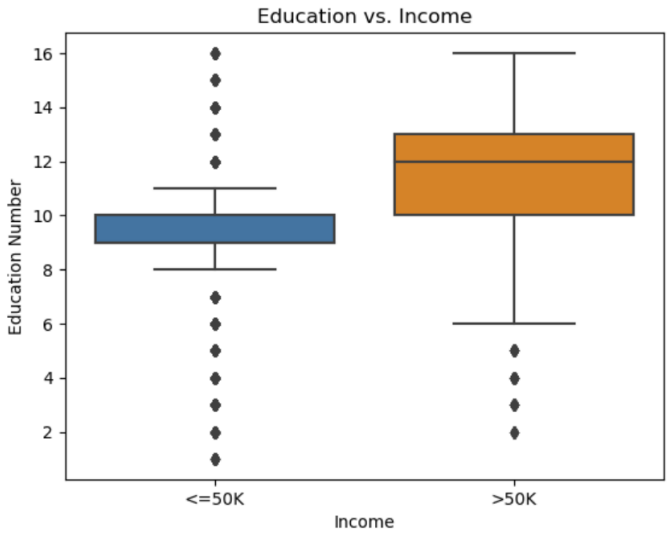
- **Data Source Reliability:** It is assumed that the database obtained from the United States Census Bureau is reliable and accurately represents the demographical characteristics of the population. The project relies on the assumption that the data source provides a comprehensive and trustworthy foundation for analysis.
- **Attribute Relevance:** The assumption is made that the chosen demographic attributes, such as education, occupation, age, hours-per-week, sex, marital status, capital-gain, and workclass, are relevant and sufficient for predicting individual income accurately. The effectiveness of the project hinges on the assumption that these selected attributes encompass the essential factors influencing income.
- **Salary Threshold:** The assumption includes a clear understanding that the salary threshold for analysis is $50k, and it is presumed that this threshold distinguishes between individuals earning above and below this amount. The project's accuracy relies on the assumption that this threshold is a meaningful and relevant criterion for demographic segmentation.
- **Data Cleaning Impact:** It is assumed that data cleaning procedures, specifically addressing missing values marked as '?', have successfully enhanced the dataset's reliability. The reliability of subsequent analysis and visualizations depends on the assumption that the data cleaning process effectively handled any discrepancies, ensuring the dataset's integrity for meaningful conclusions.

User Stories:

- **User Story 1:** Understanding relationship between education and income.
- **User Story 2:** Understanding relationship between sex, marital-status and income.
- **User Story 3:** Understanding relationship between occupation and income.
- **User Story 4:** Understanding relationship between workclass, capital-gain and income.
- **User Story 5:** Understanding relationship between age, hours-per-week and income.

Visualizations:

User Story1: Relationship between education and income.

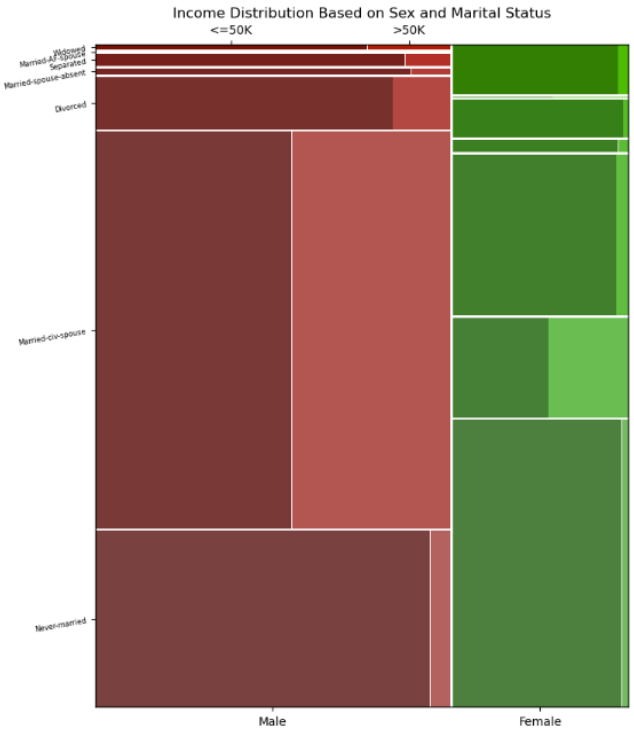


- **Type of Visualization:** I have created a box plot to visually analyze the relationship between education and income. A boxplot is chosen for its ability to effectively showcase the distribution of a continuous variable, in this case, 'education-num', across different income levels. This type of plot provides insights into the central tendency, spread, and potential outliers within each income category.
- **Design Process:** To create this visualization, I utilized 'seaborn' library to generate boxplot. The x-axis is explicitly labeled with two categories '<=50k, >50k', representing the income groups. The y-axis is dedicated to education-num variable. The design process involves organizing the data into two income categories, followed by the creation of two distinct whisker boxes side by side on the same graph. One box signifies income <=50k, while the other represents income >50k. This arrangement enables a clear visual comparison of education-num distributions for individuals earning less than or equal to $50k and those earning more than $50k.
- **Conclusion:** From this visualization I can conclude that individuals with education levels above 11 are more likely to have an income exceeding $50k. Notably, having education level 10 or below is an indicator that income is less than $50k.

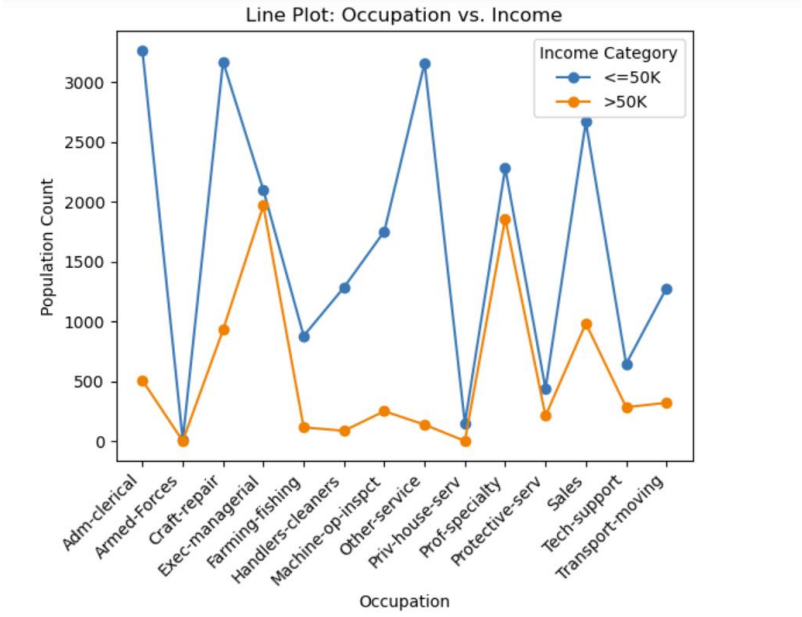
User Story 2: Relationship between sex, marital-status and income.

- **Type of visualization:** I have chosen 'mosaic' plot to visually analyze the relationship between sex, marital status, and income. The mosaic plot is an ideal choice for representing the distribution of categorical variables and their interactions.
- **Design process:** For this visualization, I utilized 'mosaic' function from the plotting library. The x-axis represents 'sex', the y-axis represents 'marital-status', and each box is subdivided to show the distribution if income categories (<=50k, >50k).

- **Conclusion:** The varying sizes of the segments within each box provides a clear visual comparison of how income is distributed within different sex and marital status groups. Stakeholders can readily discern patterns, making it clear which combination of sex and marital status are associated with higher or lower income levels. For instance, it is notable that within the 'Married-civ-spouse' marital-status category, males exhibit a higher count for income more than $50k.

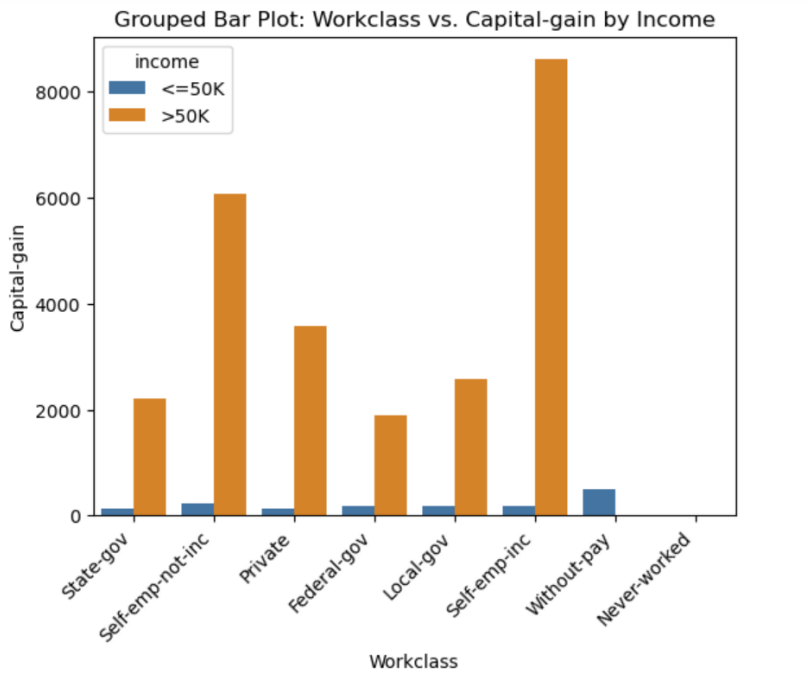


User Story 3: Relationship between occupation and income.



- **Type of Visualization:** For exploring the relationship between occupation and income, I have used line plot. This type of plot is chosen to visually represent how the population count varies for each occupation in relation to income.
- **Design Process:** To construct this visualization, I utilized the 'crosstab' function from the pandas library to create a contingency table of occupation and income. The resulting table is then used to generate a line plot. Each line in the plot corresponds to a specific income category (<=50k, >50k). The x-axis represents different occupations, while the y-axis indicates the population count.
- **Conclusion:** From this line plot, it can be deduced how the distribution of population count across various occupations correlates with different income categories. Notably, occupations such as 'Exec-managerial' and 'Prof-specialty' exhibit a substantial count in the '>50k' income category while occupations like 'Adm-clerical' and 'craft-repair' have a predominant count in the '<=50k' income category.

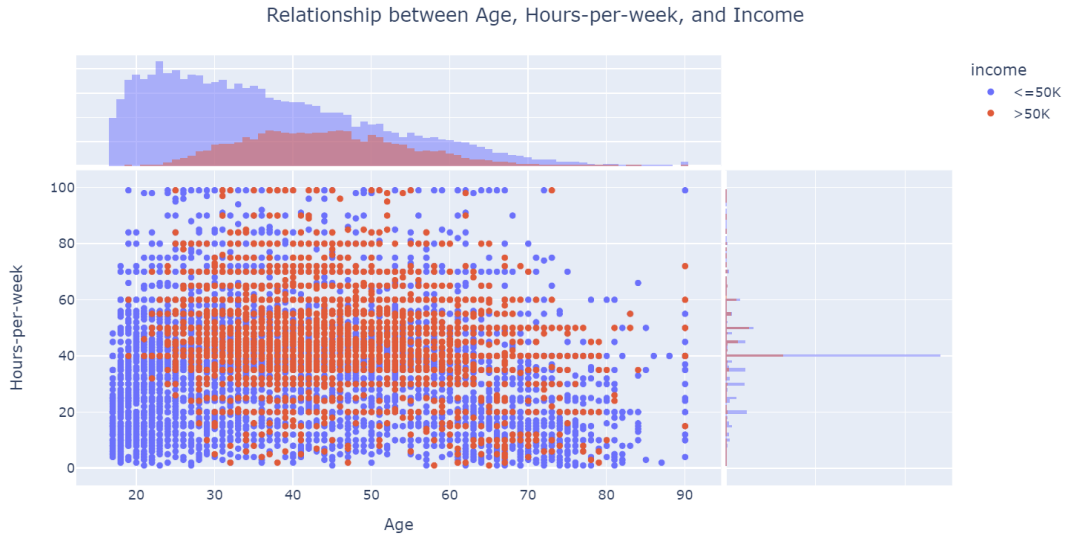
User Story 4: Understanding relationship between workclass, capital-gain and income.



- **Type of visualization:** To analyze the relationship between workclass, capital-gain, and income, I have created a grouped bar plot. This type of plot enables effective comparison of multiple categorical variables within each workclass simultaneously.
- **Design Process:** I used the 'barplot' function from seaborn library for this visualization. With 'workclass' on x-axis, 'Average capital-gain' on y-axis, and the hue parameter set to 'income' to distinguish between the two income categories. Each workclass is represented by two bars, one for income <=50k and the other for >50k, providing a clear comparison of average capital gain.
- **Conclusion:** This grouped bar plot provides stakeholders with insights into how average capital gain varies across different workclasses for both income categories. By comparing the heights of the bars within each workclass, one can identify patterns and trends in capital gain distribution. Notably, individuals in the 'Self-emp-inc' workclass with capital-gain exceeding $8000 are more

likely to surpass the $50k income threshold, suggesting a significant correlation between high capital gains in the workclass and achieving higher incomes.

User Story 5: Relationship between age, hours-per-week and income.



- **Type of Visualization**: For this user story, I used a scatter plot for visualizing the relationship between age, hours-per-week and income. Scatter plot is appropriate for this user story as it allows the examination of the relationship between two continuous variables, age and hours-per-week, in the context of income.
- **Design Process:** To create this visualization, I utilized 'Ploty Express' with 'scatter' function. The x-axis represents age, the y-axis represents hours-per-week, and color is employed to distinguish between income categories. Additionally, marginal histograms are included on both axes, providing insights into the univariant distributions of age and hours-per-week.
- **Conclusion:** From this scatter plot, stakeholders can gain insights into the distribution and potential relationships between age, hours-per-week and income. The color-coded points allow for a quick assessment of income categories, while the position of points in the scatter plot reveals potential trends or clusters. Notably, the individuals within the age group of 35 to 50, working an average of 40 hours per week, exhibit a higher tendency to have an income exceeding $50k.

Questions:

This section addresses critical questions that arose during the project progression, each accompanied by effective solutions implemented for a successful data analysis.

Q1: Are there any missing or inconsistent values in the dataset that could impact the analysis?

Solution: Performed comprehensive data cleaning and preprocessing procedures to handle missing values, ensuring data integrity.

Q2: Which variables are most relevant for the analysis?

Solution: Recognized that not all features contribute equally to predictions. Analyzed each attribute separately and chose the most informative and visually prominent visualizations to inform feature selection.

Q3: How to select appropriate visualizations to analyze the data effectively?

Solution: Identified the nature of data (continuous or categorical) and number of variables in the user story (univariant or multivariant) and selected visualizations accordingly.

Q4: How to make plots visually look appealing and informative?

Solution: Considered aesthetics in plot design, including color schemes, font sizes, label rotations on axes to enhance readability, visual appeal while ensuring clarity and informativeness in conveying insights.

Future Plans (Not Doing Now):

- **Exploring interactions of Multiple Variables:** Craft user stories that involve the interplay of four or more variables to provide a more comprehensive understanding of the factors influencing income.
- **Advanced Visualization Techniques:** Explore and implement advanced visualization techniques to enhance the interpretability of the analysis. Incorporate interactive visualizations to provide a more engaging and user-friendly experience for end users.
- **Creating Predictive Model:** Develop a predictive model to estimate the income of an individual based on various input parameters. Explore different machine learning algorithms and techniques to create an accurate and reliable prediction model aligned with the marketing team's objectives.
- **Automated Data Pipelines:** Establish automated data pipelines for seamless data preprocessing, ensuring efficiency in handling datasets.