CSE578 Data Visualization Portfolio

Xue Yang
Email: xyang234@asu.edu

***Abstract*—This project aims to develop marketing profiles using income more or less than 50K as a key demographic. We used the United States Census Bureau data to explore the relationships between the income and various features which including age, education, marital-status, occupation, race, sex, capital-gain, capital-loss, hours per week, etc. Univariate and multivariate graphs were used to visualize the features which may determine an individual's income.**

Keywords—visualization, correlation, pie chart, mosaic plot

I. Introduction

Data visualization is the visual representations of data to amplify cognition. By using common graphics, such as charts, plots, and maps, data visualizations display the complex data relationships and data-driven insights in a way that is easy to discover and explain[1]. Additionally, it provides an accessible way for the non-technical audiences to see the trends, outliers, and patterns in data, which can help them understand data easily[2].

Univariate analysis is a type of data visualization when we visualize only one variable at a time, which including pie chart, bar chart, line chart, histogram, box and whisker plots and so on. Multivariate analysis involves multiple variables at the same time, like mosaic plot, parallel plot, scatter plot, etc. With different data types or analysis purposes, different graphs will be chosen.

Before data visualization, manipulate data by a computer named Data Processing. Data processing includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output[3]. The main processing stages including data collection, data preparation, data input, processing, data output and data storage[4].

II. Description of Solution

A. Data Processing

After reviewing the initial dataset, I found that some fields contain the value "?", which represents a null or undefined value. To avoid the affect that may be caused by it, I removed all the rows which contain the "?". What's more, I found that there are some duplicate data and one column named "fnlwgt" in the dataset. I dropped the duplicate rows and the column which is not going to be used.

Before doing analysis, I noticed that the dataset includes categorical data (like sex, education, occupation, race, etc.) and quantitative data (age, capital gain, capital loss, hours per week, etc.). For some visualizations, we need to transform the categorical data into quantitative data, so I used the code "astype('category').cat.codes" to pivot the categorical data to numbers.

B. Pearson's correlation and Spearman's correlation

To advanced explore the relationship between each feature and find the best indicators for predicting, Python's pandas' libraries and seaborn tool were used to visualize the Pearson's feature correlation (Fig1.1) and Spearman's feature correlation (Fig1.2). Red color means positive correlation, blue color means negative correlation, white color means no correlation between two features.

According to the Fig1.1 and Fig1.2, I found that some features such as age, education-num, capital gain, hours per week, sex, marital status, and so on, have correlations with income, which means these features may be important when determining or predicting one person's income.

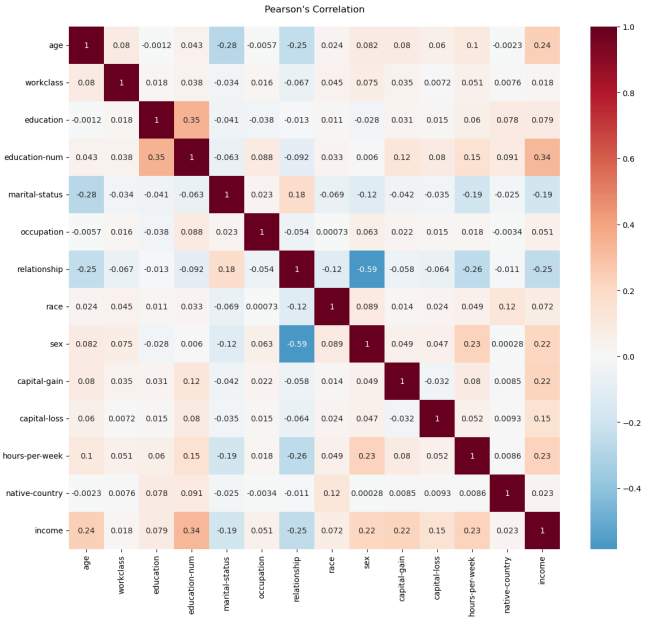

Fig 1.1 Pearson's Features Correlation

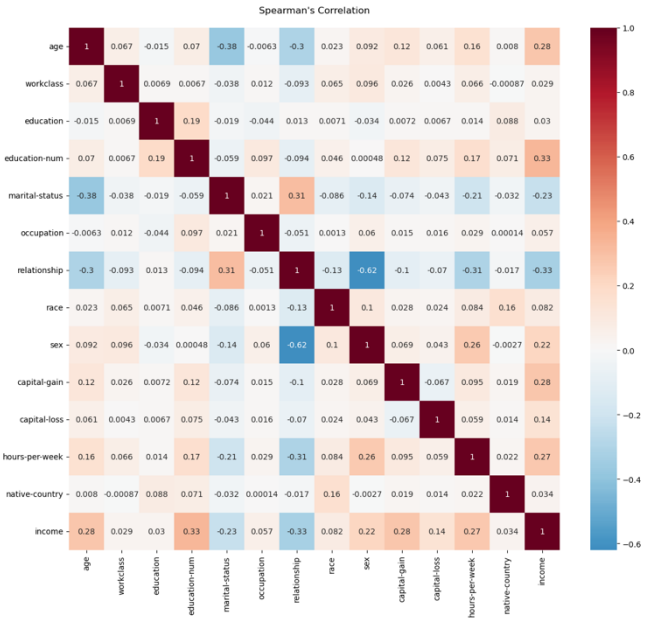

Fig 1.2 Spearman's Features Correlation

C. Visualizations

I used univariate and multivariate visualization to show the relationships between features and income, including age,

education, education-num, hours per week, occupation, capital gain and loss, sex, race, marital status.

- Bar charts can help us to find the most significant values and differences between groups, Box and Whisker plots can provide us more information, such as median, maximum, and minimum. So, the distribution of age based on income was shown by the bar chart and the Box and Whisker Plot.
- As a kind of categorical data, I want to show the component parts of a whole, so the relationship between education and income was shown by pie charts.
- Scatter plots can be used to observe relationships between multiple variables, so I used a scatter plot to visualize the relationships between "education-num" and "hours per week" based on income.
- For two or more categorical variables, mosaic plots should be chosen for the visualization. I used a mosaic plot to show the distribution of "occupation" based on income.
- The numerical data "capital gain" and "capital loss" and their relationship with income were visualized by the Parallel Coordinate Plot. A parallel coordinate plot allows us to compare the features of several observations on a set of numeric variables.
- For the qualitative data, a mosaic plot was used to visualize the relationship among "marital status", "sex" and income, and "gender", "race" and income.

III. Results

We used multiple visualizations to make sense out of data, tried to find key features and draw the conclusions for each visualization.

A. *Age vs Income*

- In Fig1.1, the blue bars represent the income less than 50K, the orange bars represent the income more than 50K. We noticed that younger people are more likely to have an income less than 50K (blue bars).
- The distribution of people who have income more than 50K (orange bars) has a trend of increasing first and then decreasing with age.
- At a very young age (17~ 22 years old) or a very old age (older than 80 years old), almost no one can have income more than 50K.
- In Fig1.2, we found that most people who earn more than 50K are 35 to 50 years old. Most people whose income is less than 50K are 25 to 45 years old.
- The median age for people whose income is more than 50K is around 45 years old, for people whose income is less than 50K is around 35 years old.
- We also can find some outliers between the ages 72 to 90 years old.

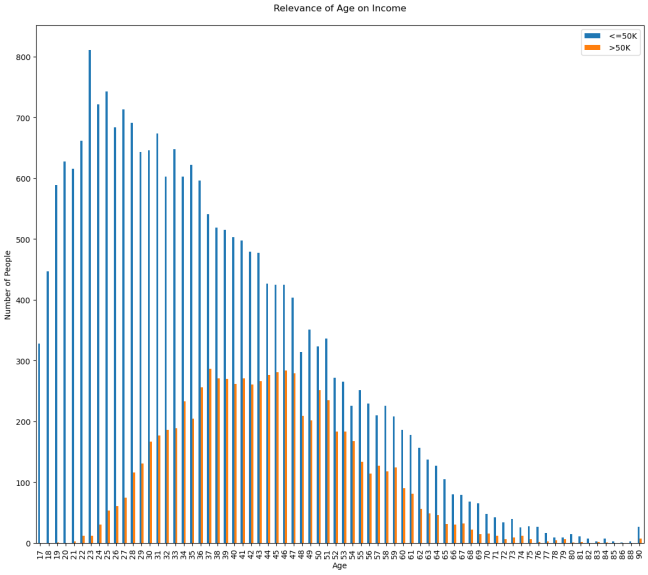

Fig2.1 Bar chart of Age vs Income

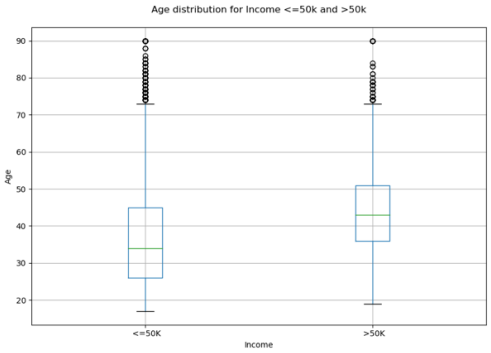

Fig2.2 B&W plot of Age vs Income

- **Conclusion**: Age is a relevant feature in determining a person's income. Most people who earn more than 50K are 35-50 years old.

B. *Education vs Income*

- In Fig2.1, most people who earn more than 50K hold a college or higher degree. Only a very small part of people (about 10%) who hold an associate or lower degree can have income more than 50K. This strongly indicates that a higher degree is related to higher income.

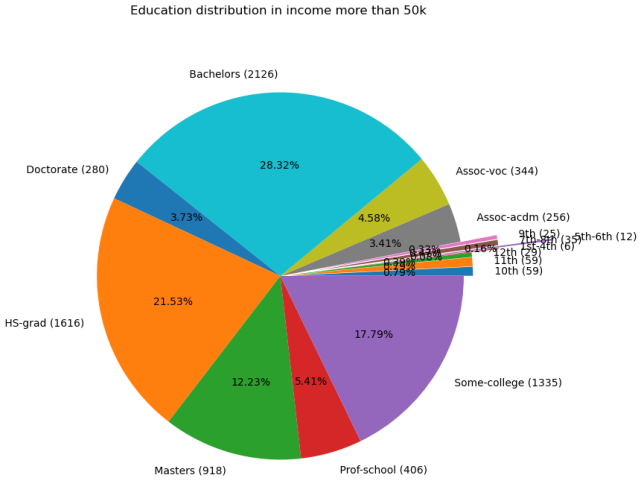

Fig3.1 Pie chart of Education vs Income>50K

- In Fig2.2, most people who have an income less than 50K hold a college or HS-grad degree, and nearly 25% people have an associate or lower degree. Only 15% of people who have bachelors or higher degrees earn less than 50K.

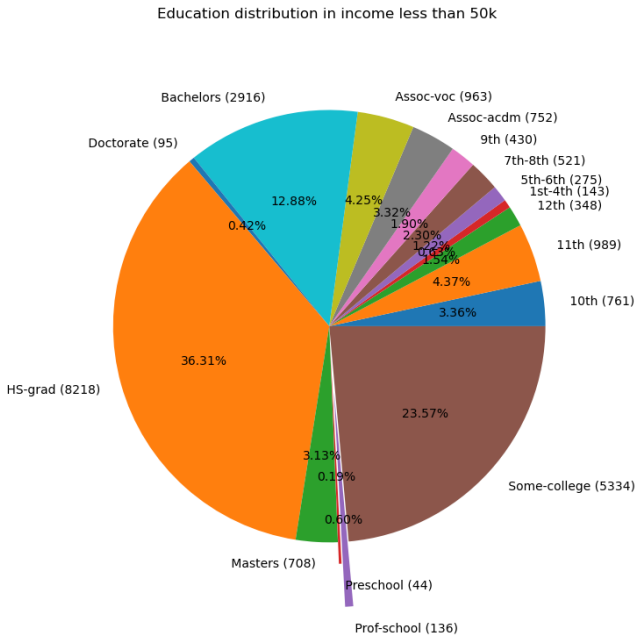

Fig3.2 Pie chart of Education vs Income<=50K

- **Conclusion**: An individual's education level is associated with his/her income. Higher education level indicates a higher probability of higher income.

C. *Education-num and Hours Per Week vs Income*

- The education levels higher than 9 have more people who have an income more than 50K, while people whose education level lower than 9 are more likely to earn income less than 50K.
- Most people who earn more than 50K work about 40 hours per week.
- People having higher education levels are more likely to have an income more than 50K when working less than 40 hours per week.

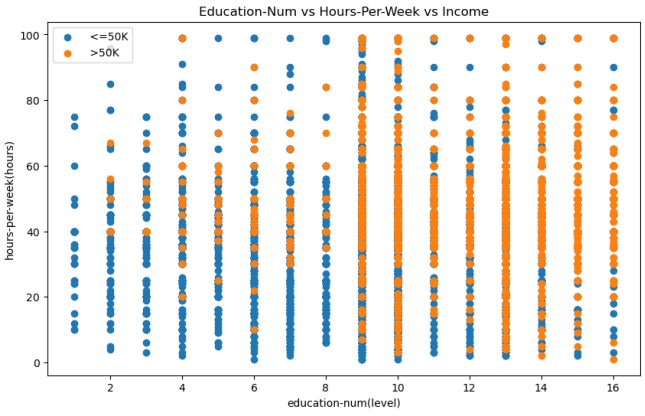

Fig4 Scatter Plot of Education-Num vs Hours Per Week vs Income

- **Conclusion**: Education-Num (level) and Hours Per Week are bearing on an individual's income. People having higher education levels, working longer hours per week are more likely to have an income more than 50K.

D. *Occupation vs Income*

- People whose occupations are "exec-managerial" or "prof-specialty" have a higher proportion to get income more than 50K.
- Some occupations such as "priv-house-serv", "other-service" and "handlers-cleaners" are more likely to have an income less than 50K.

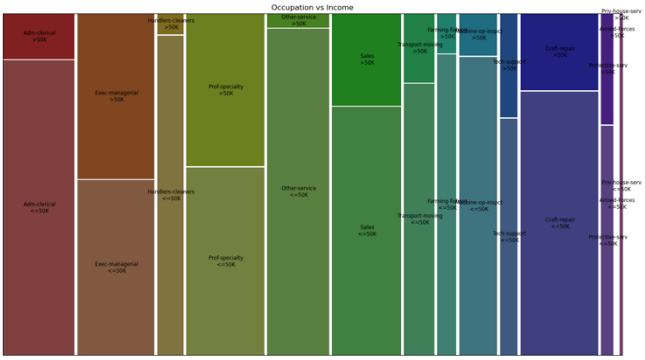

Fig5 Mosaic Plot of Occupation vs Income

- **Conclusion**: The probability of a person's income is relevant to his/her occupations. Managerial and professional occupations have more possibility to get an income more than 50K.

E. *Capital Gain and Capital Loss vs Income*

- Orange lines represent people's income more than 50K, and blue lines mean less than 50K. People having nearly 100K capital gain all have income more than 50K.
- Most people who can earn more than 50K have 5K~20K capital gain and almost zero capital loss.
- Most people who earn less than 50K almost have no capital gain and capital loss.

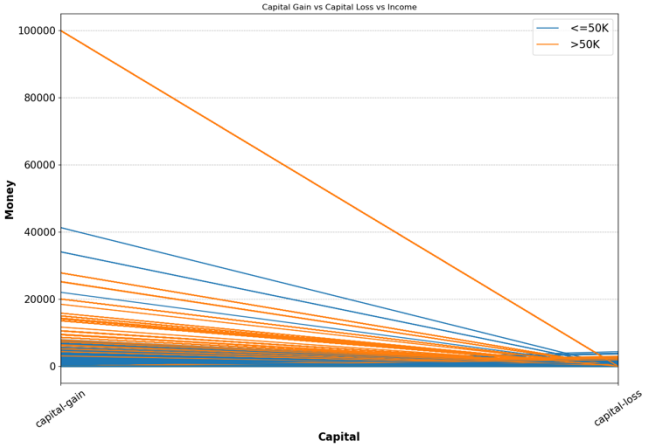

Fig6 Parallel Coordinates Plot of Capital Gain or Loss vs Income

- **Conclusion**: Capital gain and capital loss are related factors in determining a person's income. People who have about 5K~20K capital gain are more likely to have income more than 50K, people who don't have capital gain may have less income.

F. *Sex and Race vs Income*

- "Males" are more likely to have an income more than 50K than "females".
- "White" and "Asian-Pac-Islander" are more likely to earn more income both in male and female.
- "White" "Male" is the largest proportion of people whose income is more than 50K.

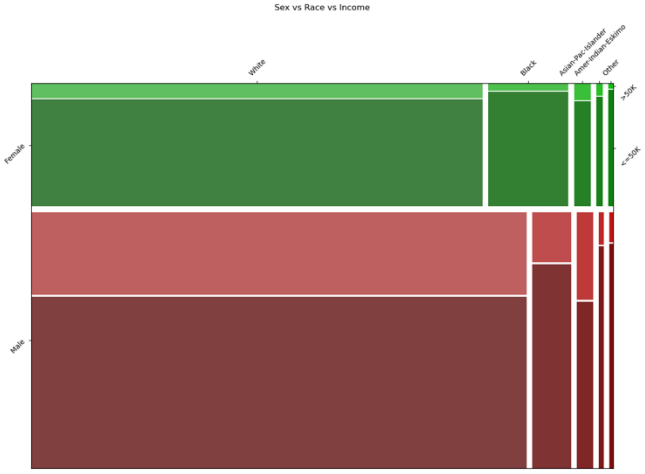

Fig7 Mosaic Plot of Sex vs Race vs Income

- **Conclusion**: Sex and race are associated with a person's income. "White" "Male" has more chances to get a higher income.

G. *Marital Status and Sex vs Income*

- Both in "male" and "female", "Married-civ-spouse" are more likely to have an income more than 50K.
- For "male", "Windowed" status has the second largest proportion to have income more than 50K, while "Never-married" status has the least proportion to have a higher income.
- For females, those who have "Married-AF-spouse" status are more likely to have an income more than 50K, while "Never-married" and "Separated" status are less likely to have a higher income.

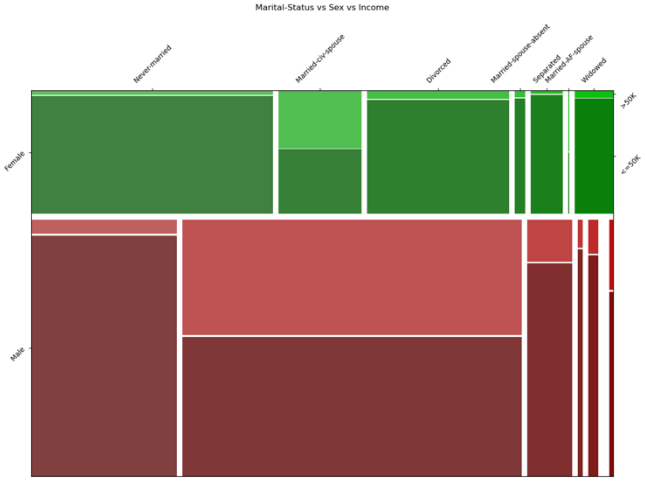

Fig8 Mosaic Plot of Marital Status and Sex vs Income

- **Conclusion**: Marital status and sex are relevant indicators in predicting one's income. "Male" and "female" who have an "Married-civ-spouse" status are more likely to earn more than 50K income.

IV. Lesson Learned

- I learned the concepts of the Data exploration and Visualization, which is not simply making graphs, the goal is to transform the information to thoughts or insights.
- Through the lectures and assignments, I learned a lot of visualization tools which helped me to accomplish this project, such as numpy, matplotlib, seaborn, etc.
- While doing the project, I learned about characteristics of univariate and multivariate plots, and how to decide to choose these visualizations.
- I also learned about the necessity of data processing before starting the analysis, including removing the null and duplicate value.
- Spearman's correlation and Pearson's correlation were used to determine the relationships between income and other features. The selection of algorithms or models was based on the data and accuracy.
- Different kinds of graphs can help us solve various problems. Categorical data should use pie charts, bar charts or mosaic plots, etc. Quantitative data should use line charts, bar charts, parallel coordinates plots, scatter plots to analyze. However, we can give each value of category a number to transform the qualitative data to quantitative data, which can help us make visualizations easier.
- Apart from the various graphs, I learned a lot of visual variables to represents the different categories, like position, size, shape, color, rotation, and so on. In this project, I used different color to show income more than 50K and income less than 50K.
- After completing this project, I have had opportunity to practice the learned skills and draw the important conclusions through data visualizations.

V. References

[1] https://www.ibm.com/topics/data-visualization
[2] https://www.tableau.com/learn/articles/data-visualization
[3] https://www.britannica.com/technology/data-processing.
[4] https://www.talend.com/resources/what-is-data-processing/.