CSE 578 Project Milestone- Portfolio Report, Summer 2022

Marketing Profiles Using data Visualization

Nigar Sultana
Ira A. Fulton School of Engineering
nsultan5@asu.edu

ABSTRACT

This report aims to document the project completed as part of the Data Visualization (CSE578) course taught in the summer session of the 2022 academic year at Arizona State University. The goal of this project is to develop a marketing profile using individual income as a key demographic in the United State Census Bureau data. Data visualization is the key for this project to visualize the relations between the key individual features in the United States Census Bureau data. The data visualization shows if an individual has income above or below $50,000 with the relative attributes. Also, all the attributes come from the United State Census Bureau data related to the key demographics.

KEYWORDS

Multivariate Analysis, Univariate Analysis, Individual user story, Python Programming, Matplotlib Library, NumPy, panda, Jupiter notebook, bar chart, histogram, donut chart, box and whisker plot, visual variable, color schemes, and design, visualization, mosaic plot, correlation, data exploration.

1. Introduction

The first step of this project is data preprocessing, which removes the irrelevant and incorrect data from the given dataset. Preprocessing has several steps, including data collection, preparation, processing, and interpretation. During the preprocessing, it needs to ensure that this processing does not negatively affect or compromise the outcome of the dataset. Preprocessing step transfers the raw dataset to a more reliable and easily readable dataset for the computer. The data collection step needs to ensure that the source of the data set is authentic and reliable. Usually, the original data set can be collected from the data warehouse. The data cleanup process is part of the data processing step. In this step, the data diligence and error check process can be done to eliminate incorrectly, repeating, and incompatible data before going to the interpretation stage. Data are available in the interpretation stage, generally presented in text, videos, graphs, and other visual aids. A bar chart is generally helpful to visualize the comparison of matric values across the subset of a dataset. A pie chart is a more efficient way to visualize the categorical data. A mosaic plot displays the comparison of the different classes of the subsets of the vast dataset. Finally, to plot a collection of a dataset on an interval scale, a box and whisker plot is a more viable option.

Data visualization is created in this project to answer clients' questions. The project utilizes the given dataset and information to create an individual's marketing profile to boost college enrollment. The client intends to use salary as a vital segment to create this marketing profile.

2. Explanation of Solution

The data visualization course teaches different techniques to visualize the relationship between different features of the given data set. In this project, data visualization tools visualize relationships between different features associated with individual salaries. This course also teaches how to use design techniques and principles and the knowledge to visualize the dataset to find particular relations between the subset of the given dataset.

Table: Data visualization tools and software

Software, Tools	Details (Version,URL,Usage)
Python	Version 3.8, http://www.python.org/, programming language
PyCharm	Version 2020.3.3 community Edition, https://www.jetbrains.com/pycharm/,used IDE for Python
Jupyter Notebook	Version6.1.4, https://jupyter.org/install,used as IDE
Anaconda Navigator	Version 1.10.0, https://www.anaconda.com/ , UI for python library packages.
Conda	https://docs.conda..io/projects/conda/en/latest/used to creat multiple python library packages with varying version of python libraries.

Numpy	Version v1.20.0, https://numpy.org/, for Data Wrangling.
Matplotlib	Version v3.4.1, https://matplotlib.org/, for plotting.
Pandas	Version 0.25.1, https://pandas.pydata.org/, us for Data Wrangling.
Seaborn	Version v0.11.1, https://seaborn.pydata.org/,used for plotting.

Equations

Formula to calculate width and bins:

Sturge's formula

$$k = \lceil \log N + 1 \rceil$$

Freedman-Diaconis rule

$$h = 2IQR(x)N^{-\frac{1}{3}}$$

3. Description of Results

A. Sex and Salary

The mosaic plot shows the size of the percentage distribution of data containing two categories. These categories are put on the length and width of the plot. The different sizes of the rectangles within the mosaic plot then show the percentage amount of the data in the two categories.

The mosaic plot has been used to show the percentage distribution of male and female data. The length of the plot represents the percentages of the proportions less than or equal to 50k, and greater than 50k.The width on the other hand shows the percentages of the female and male genders. There are four proportions each color coded in a unique color to show the distribution of the genders with respect to the categories less than or equal to 50k, and greater than 50k, Males less than or equal to 50k form the largest percentage whereas females greater than 50k form the smallest percentage of the four groups (proportions) in the plot.

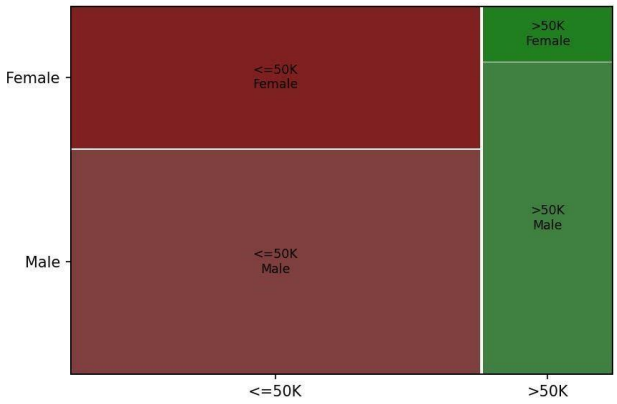

Fig.1 Mosaic plot of Sex versus Salary.

B. Education vs Salary

The univariate plot is best suitable for identifying the relationship between the continuous variable of years of education and income. Violin plot provides insights into the density of the kernel estimation to display the distribution and the central tendencies. The Violin plot indicates a strong correlation of the individual's income with the years of education. Individuals with a bachelor's degree have about two and a half times possibility to make more than $50K.

The plot shows the relation between the variable salary and the highest level of education. Those with higher levels of education tend to earn more than those with lower levels of education.

This relationship is not always linear, however. There appears to be a jump in earnings between those who have a college degree and those who do have some bachelor's degree. The individual who holds a higher degree than the undergrad also tends to earn more than those with a bachelor's degree.
Overall, this graph shows that higher levels of education tend to lead to higher earnings.

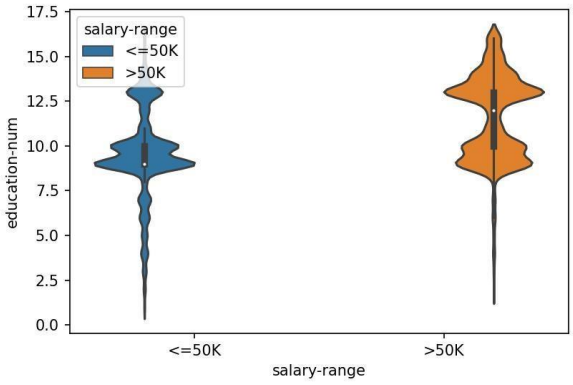

Fig 2: Violin plot of Education versus Salary

C. Relationship vs Salary

The pie plot here represents different relation and their salaries. This plot compares the asymmetry between the different parts highlighted with different colors. This plot also shows that 20% who have at least one child makes less than 50 thousand a year.

A pie chart is a circle, also called a 360 graph where 360=100%.

In graph 1:

For husband = 30.13% so 30.13% of 360 = 108,468

Other-relative = 3.33% so 3.33% of 360 = 11.988

Own child = 3.82% so 3.82% of 360 = 13.752

Unmarried = 13.06 % so 13.06% of 360 = 47.016

Wife = 20.23% so 20.23% of 360 = 72.828

Not in family = 29.43% so 29.43 % of 360 = 106.948

In graph 2:

For husband =75.48% so 75.48 % of 360 = 271.728

Other relative =0.47% so 0.47% of 360 = 1.692

Own child = 0.85% do 0.85% of 360 = 3.06

Unmarried= 2.78% so 2.78% of 360 = 10.008

Wife= 9.50% so 9.50% of 360 = 34.2

Not in family=10.92% so 10.92 % of 360 = 39.312

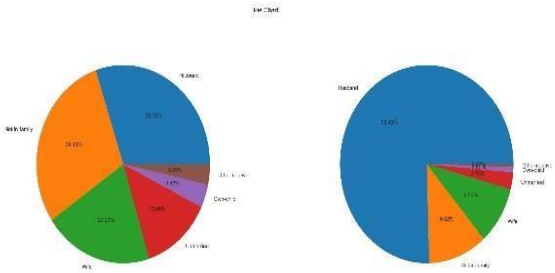

Fig 3: Pie Plot of Relationship versus Salary

D. Hours per week and Salary

This chart represents the population fall between a certain age and makes above and below $50k. This plot also represents the correlation between salary and work hours per week for an individual. It is clear from the chart that people who work 20 to 60 hours per week make more than $50k. on the other hand, people who work 30 to 40 hours per week likely make less than $50k. The second plot shows that individuals in the ages of 25 to 60 make more than $50k, and those between the ages of 20 to 30 years make less than $50k.

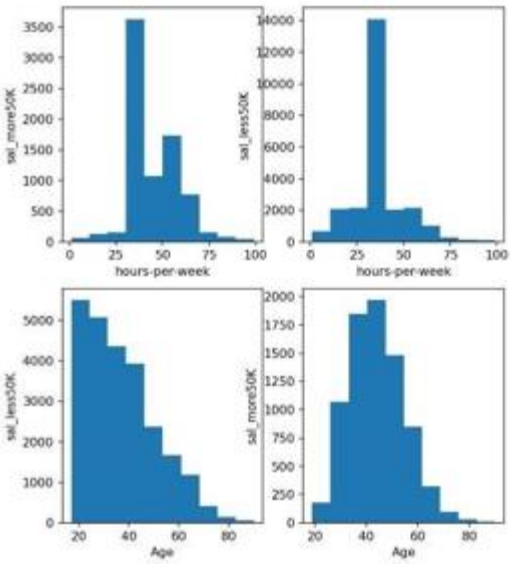

Fig 4: Histogram plot of Hours per Week versus Salary

E. Age vs Salary

This plot is used to compare datasets when the given dataset has a different sample size. Here, a box and whisker plot display the relation between age and income of different individuals. In this plot, an individual's income is determined by age. According to the plot, the median income group is about 44 years of age.

The first box and whisker plot show the relationship between age and the salary which is less than or equal to 50k.Here, the minimum age is around 15 and the maximum age of 77. The first quartile that means 25% of people that have ages below 25 and they income less than or equal to 50k. The median of this plot is almost 34 that means 50% of people's income less than or equal to 50k. Age below 46 have salary below 50k.

The second box and whisker plot describe minimum age 21 and maximum age 72 who's income is greater than or equal to 50k. Here the median age is almost 45. People are ages below 51 that contribute e to 75% for incoming greater than or equal to 50k.

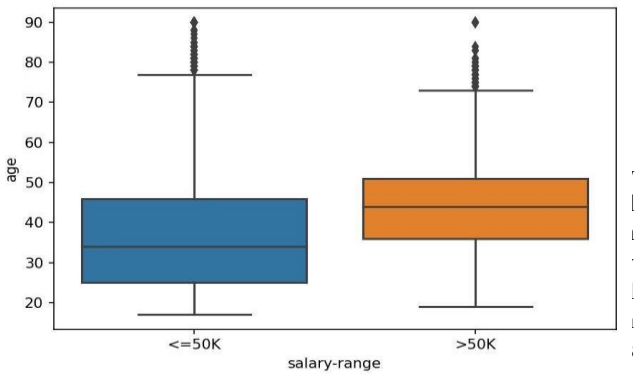

Fig 5: Box and Whisker plot of Age versus Salary

F. Marital status and Salary

Based on the graph it can be said that whose income is 50k or below 50k they are either 4 never married or married -civ-spouse. And whose income is above 50k they are married - civ-spouse. I can choose this chart because without it I wouldn't be able to display this much information. A bar chart is used to show how a marital status affects the salary of individuals. From this bar chart I can infer that the single status individuals are more likely to make less than 5.

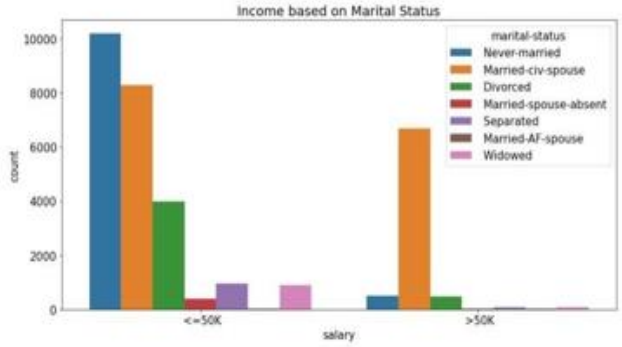

Fig 6: A Bar plot of Marital status versus Salary.

G. Race and Salary

Visualization for different attributes of race and salary: A donut chart is used to show proportion of individuals that fall into different categories of race for the category of above and below 50k.From these 2 space efficient charts I can infer that: Above 50k chart: 91% of people who makes more than 50k are white. 5% of individuals who makes more than 50k are Black. 4% of people who makes more than 50k are Asian -Pac-Islander. Below 50k chart: 84% of people who makes more than 50k are white. 11 % of individuals who makes more than 50k are black. 3% of people who makes more than 50k are Asian -Pac-Islander. 1% of people who makes more than 50k are American Indian -Eskimo,1% others.

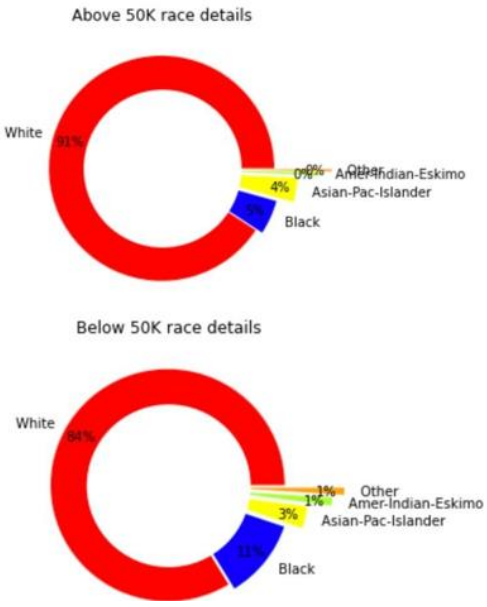

Fig 7: Donut chart for Salary above and below 50k.

H. Age and Hours per week versus Salary

Based on the graph younger people work less and they also income less compared to older people. When peoples age increase, they work more hour to get high salary. This scatterplot is best to show the correlation between features that's why I choose this graph to visualize the information. To generate scatterplots, at first, I divided the data into two parts, one for each income - group.

Then scatterplot has been constructed by mapping age to hours -per-week for all values. 5 This graph does not show any correlation between age and working hours per week.

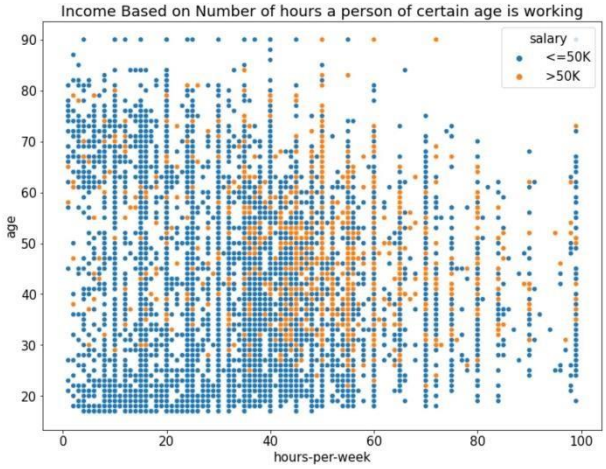

Fig 8: Scatter plot of Age versus Hours per week

4. Contributions and lesson learned

As this is an individual project so everything was done by me. From this project the skill I learned, is how and when to use different Data visualization plots. I learned how to plot them using python libraries such as Matplotlib and Seaborn.

I learned many analysis techniques that can be used for Data visualization, namely univariate analysis, multivariate analysis specially on categorical data, Hierarchical analysis, Temporal analysis, time series and so on.

Also, I learned about many useful concepts that can be used to make more intuitive plots. I learned about Data exploration components, Design Principles for different types of plots, text visualization, supervised and unsupervised learning.

REFERENCES

[1] Danyel Fisher & Miriah Meyer, "Making Data Visual", A Practical Guide to using Visualization for Insight, 1st Edition, January 2018, O'Reilly Media.

[2] Daniel Nelson, "Data Visualization in Python", Explore and Manipulate Data and Create Engaging Interactive Plots with 9 Python Libraries, 2020, Stack Abuse.

[3] Tamara Munzer, "Visualization Analysis & Design", 2015, CRC Press.

[4] Chun-houh Chen, Wolfgang Hardle, Antony Unwin, "Handbook of Data Visualization", 2008, Springer.