CSE 578- DATA VISUALIZATION
Visualizing and forecasting the marketing strategies based on key factors
Systems Documentation Report

By
SHERIN SEBASTIAN
ASU ID: 1222858338

Organization: XYZ Corporation
Customer: UVW College

As part of a team of data analysts at XYZ Corporation, UVW College is tasked with developing an application to categorize the characteristics that determine an individual's income. In order to increase enrolment, UVW College has determined that income is a crucial component in marketing its degree programs. Therefore, using this forecast can help marketing experiments find consumers who share their interests. The United States Census Bureau will offer statistics for the application's dataset, with $50,000 serving as the application's primary wage figure. The current main goals include understanding the dataset, identifying the variables that affect an individual's income, and constructing machine learning models that use the identified variables to forecast an individual's income.

Goals and Business Objective

This project's goal is to create maps of the visuals that can be used to find the factors that determine a person's income and then provide those factors to UVW executives so they may spot trends in the dataset. The following phase is to develop machine learning models based on the preliminary analysis that accurately forecast individual income. This will facilitate the usage of this information by the marketing team that was previously indicated.

Assumptions

- **The dataset is comprehensive, with no missing values.**

The risk and equivalent of utilizing the incorrect dataset is using incomplete data. Therefore, we assume that the dataset is full and that there are no data collecting inconsistencies that could skew the overall prediction and thus result in a poor visualization. Uninformed actions may take place if one doesn't have a clear understanding of how operations are progressing. The full collection of requirements that make up a complete set of data must be understood in order to determine whether or not the requirements are being met.

- **Genuine and Reliable Dataset**

Features like gender, for example, are often restricted to a set of alternatives in the dataset and are not based on the open answers available outside. The dataset is accurate and well-validated. All other responses will not be taken into account as genuine or legitimate based on the specifications of the dataset.

- **Dataset has no inconsistencies.**

Given that it should not deceive and must convey the correct information without being misleading, assuming that the data provided is precise and reliable, this dataset does not contain any inaccuracies that would cause it to be inaccurate. It's likely that keeping precision and accuracy will be off-target or cost more than is necessary.

- **Current and Reliable Data**

The information was gathered at the right moment since information gathered too early or too late could cause a scenario to be misinterpreted and result in incorrect decisions. For instance, if the incomes in the dataset were collected during a period of deflation, which is not the appropriate time to gather information and utilize it in a dataset, it would be unjustifiable and the inaccurate results would have a negative impact on accuracy.

User Stories

- **Multivariate Analysis of different variables using Tree Map**

Using a tree map, the effects of several factors, such as sex, workclass, and occupation, were assessed for the two income categories—those earning more than $50,000 and those earning less than $50,000.

- **Multivariate Analysis of different variables using 3D- Scatter Plot**

Age, number of years of education, and hours worked per week were among the many factors analyzed for the two groups of income ranges: less than $50,000 and more than $50,000. The outcomes were shown using 3D-Scatter plots. Age, education level, and hours worked per week are correlated.

- **Univariate Analysis of single variable using Histogram**

Analysis of race against salary range was done using histogram

- **Multivariate Analysis of different variables using Scatter Plot**

Analysis of a number of variables, including age, capital gain, and occupation, was done for two categories: income range less than $50,000 and salary range more than $50,000. Scatter plots were used to display the results. The relationship between age and capital gain, hours worked each week and age, and capital gain relative to hours worked each week is plotted.

- **Multivariate Analysis of different variables using Scatter Matrix**

The impacts of a number of variables, including fnlwgt, hours worked per week, and age, on the income categories of more than $50,000 and less than $50,000 were examined using scatter matrices.

- **Multivariate Analysis of different variables using Parallel Plot**

The impacts of several variables, including age, capital gain, and the number of years of education on the two income categories—those earning more than $50,000 and those earning less than $50,000—were evaluated using a parallel plot.

Visualizations

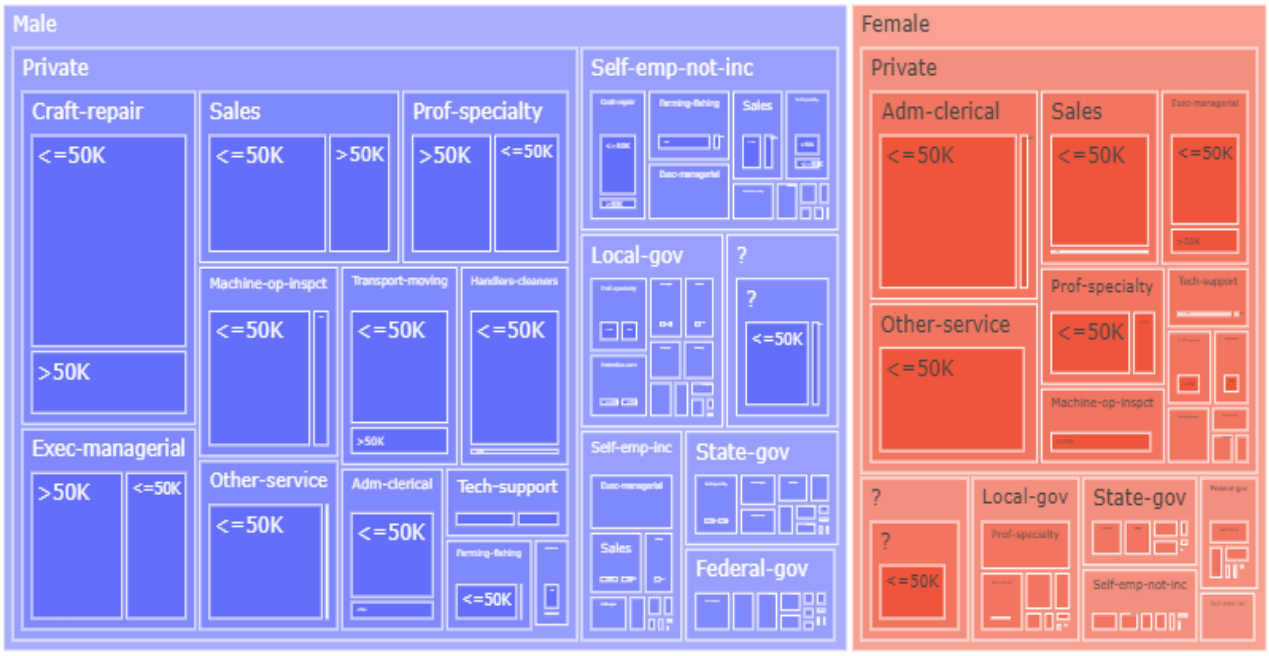
A variety of visualization techniques, including tree map, scatter plots, scatter matrix, and parallel coordinate plots, are used to assess attributes. Utilized statistical metrics when visualizing in the proper places. The impact of several groups of variables on salary was examined using multivariate analysis on various sets of variables. Examined and depicted the impact of several sets of variables on the rise in salary. The use of multivariate analysis helps to clarify how a particular combination of variables is impacting the growth and decline of salary.

Multivariate Analysis

1. TREE MAP

- **Attributes:** Sex, workclass, Occupation, Salary-Range
- **Values:** Male, Female, Private, State-gov, Without-pay, Federal-gov, Craft, Machine-op, Adm-clerical, Transport-moving, Armed-Forces, Other services etc.,

Graph:



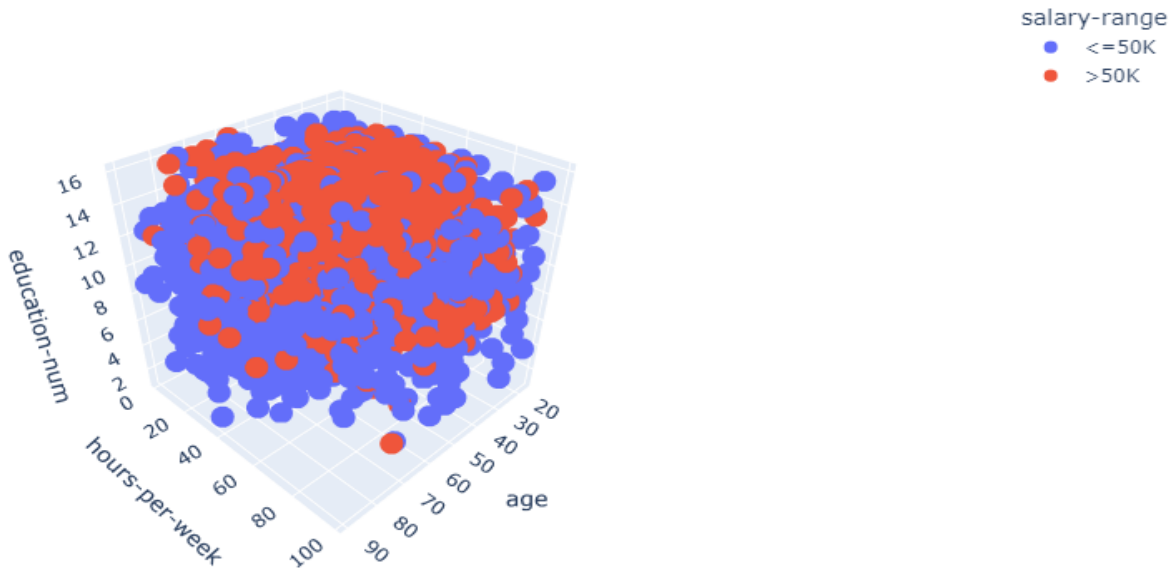
Observation:

- The category male in private sector in the stream of “craft repair” and category female in private sector in the stream of “admin-clerical” earning **less** than 50k are the targeted audience.
- The category male in private sector in the stream of “executive-managerial” and category female in private sector in the stream of “executive-mangerial” earning **more** than 50k are the targeted audience.

2. 3-D SCATTER PLOT

- **Attributes:** Age, hours-per-week, Education-num, Salary-Range
- **Values:** Range of age, range of hours, range in no.of education years

Graph:



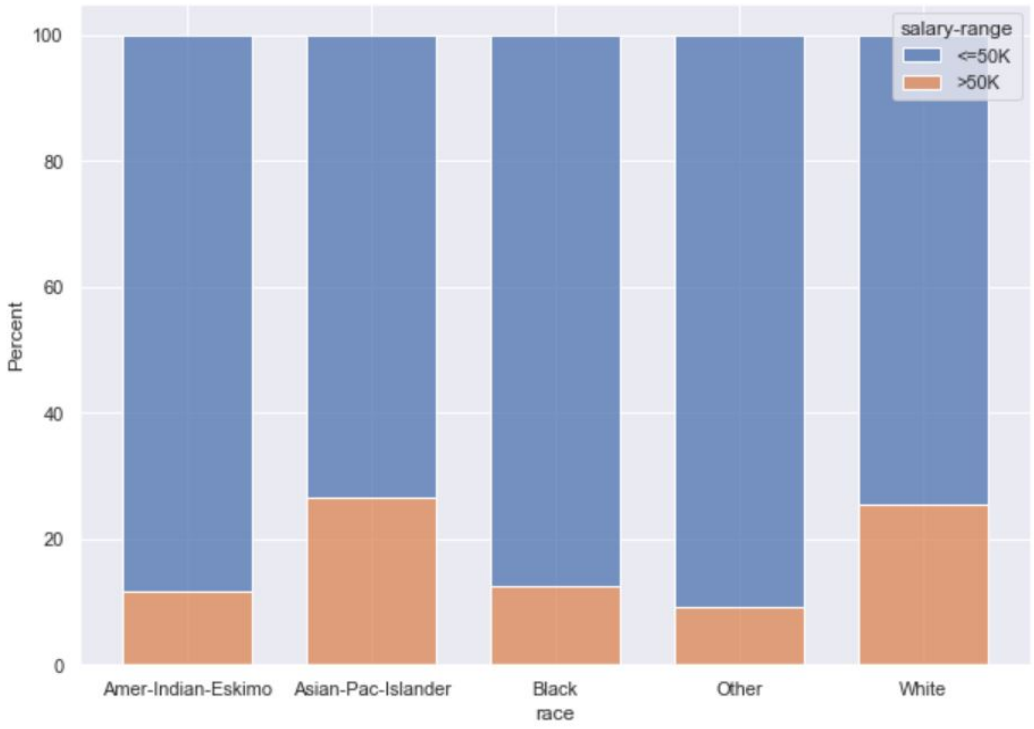
Observation:

- The category “education-num” from 6 to 10, “age” ranging from 17 to 25 sector working 10 to 20 “hours per week” earning **less** than 50k are the targeted audience.
- The category “education-num” from 9 to 12, “age” ranging from 30 to 45 sector working 35 to 40 “hours per week” earning **more** than 50k are the targeted audience.

3. HISTOGRAM PLOT

- **Attributes:** Race, Salary-Range
- **Values:** Black, White, Asian Pacifican Islander, Amer Indian Eskimo, Education range from 10th - Masters

Graph:



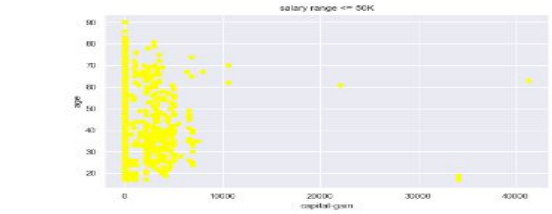
Observation:

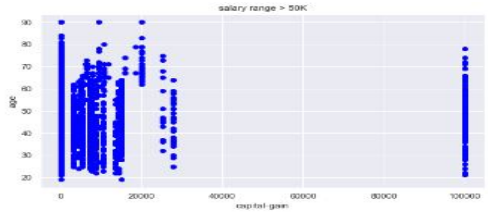
- The category race “other” earning **less** than 50k are the targeted audience.
- The category race “Asian-Pcific-Islander” earning **more** than 50k are the targeted audience.

4. SCATTER PLOT

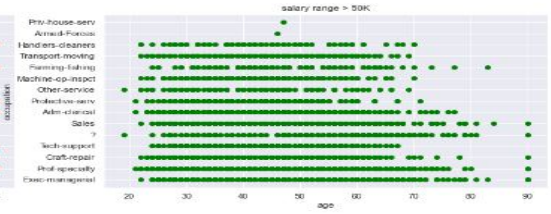
- **Attributes:** Capital-gain, Age, Occupation, Salary-Range
- **Values:** Range of capital-gain, Range of Age, Transport-moving, Machine-op-inspect, Tech-support, Craft-repair etc.,

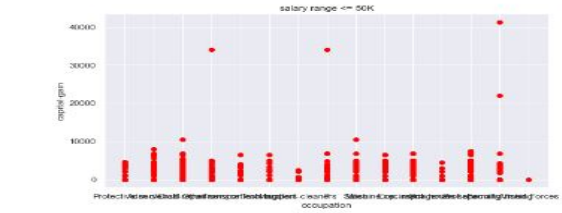
Graph:


salary range <= 50K
age
capital-gain

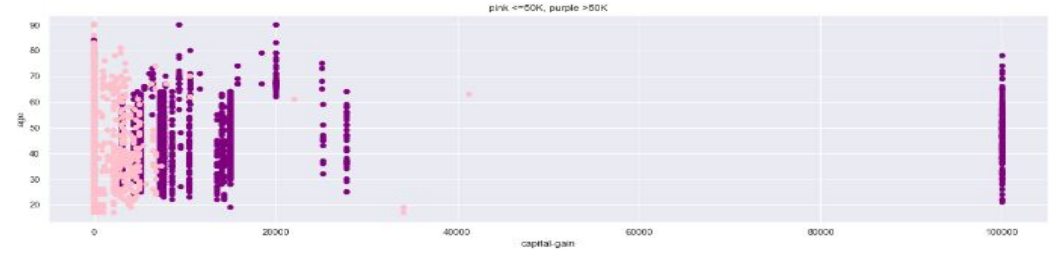

salary range > 50K
age
capital-gain

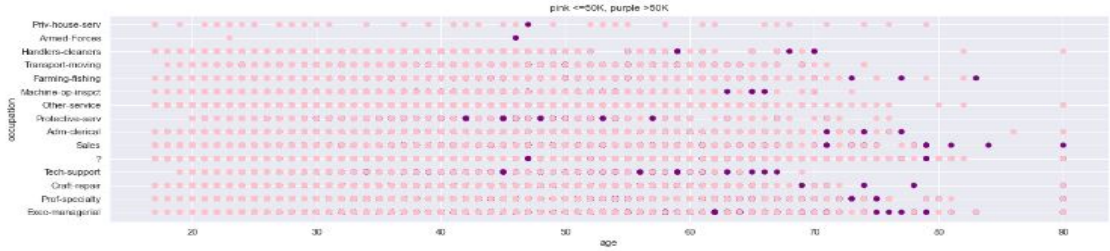

salary range <= 50K
Armed-Forces
Farming-fishing
Prof-specialty
Priv-house-serv
Exec-managerial
Machine-op-inspct
Sales
?
Handlers-cleaners
Tech-support
Transport-moving
Other-service
Craft-repair
Adm-clerical
Protective-serv
occupation
age

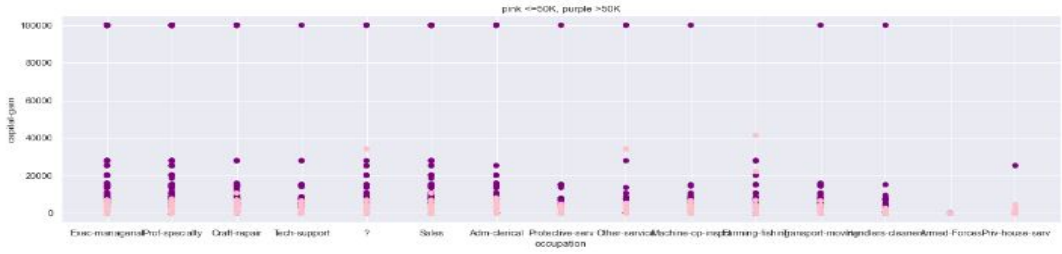

salary range > 50K
Priv-house-serv
Armed-Forces
Handlers-cleaners
Transport-moving
Farming-fishing
Machine-op-inspct
Other-service
Protective-serv
Adm-clerical
Sales
?
Tech-support
Craft-repair
Prof-specialty
Exec-managerial
occupation
age


salary range <= 50K
capital-gain
occupation


salary range > 50K
capital-gain
occupation


pink <=50K, purple >50K
age
capital-gain


pink <=50K, purple >50K
Priv-house-serv
Armed-Forces
Handlers-cleaners
Transport-moving
Farming-fishing
Machine-op-inspct
Other-service
Protective-serv
Adm-clerical
Sales
?
Tech-support
Craft-repair
Prof-specialty
Exec-managerial
occupation
age


pink <=50K, purple >50K
capital-gain
occupation

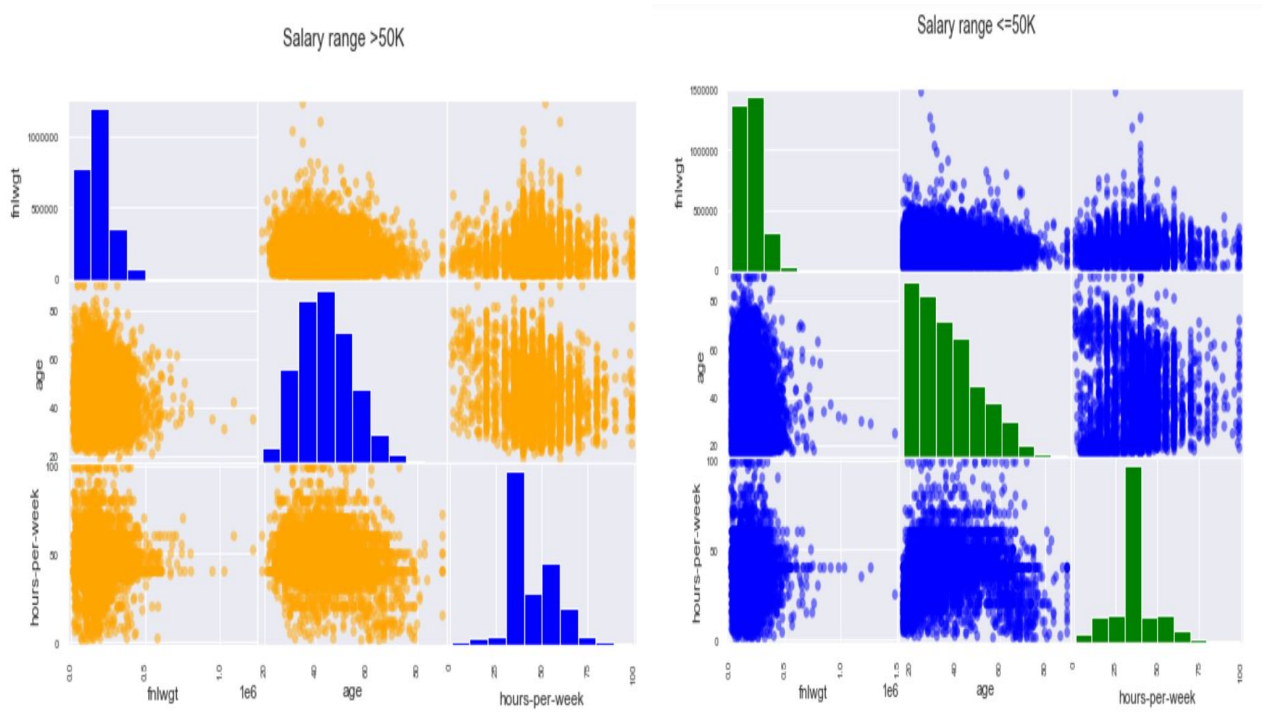
Observation:

- The category “age” ranging from 15 to 60, “capital gain” in the range of 1000 to 3000 earning **less** than 50k are the targeted audience.
- The category “age” ranging from 15 to 60, “capital gain” in the range of 1000 to 3000 earning **more** than 50k are the targeted audience.

5. SCATTER MATRIX(less and greater than 50k)

- **Attributes:** Fnlwgt, Age, Hours-per-week, Salary-Range
- **Values:** Range of Fnlwgt, Range of age, Range of hours-per-week

Graph:



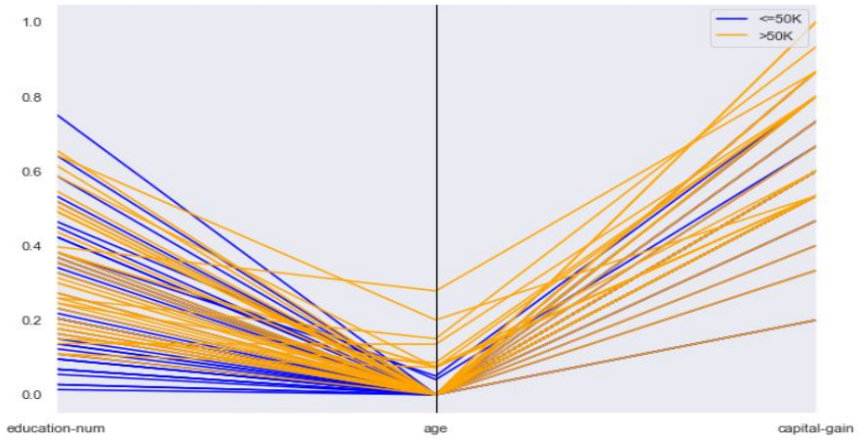
Observation:

- The rest of the data appears to be distinct, with the exception of a few outliers.
- This visualization demonstrates the combined effects of age, hours worked, and fnlwgt on pay increases and decreases.
- People with large fnlwgt are more likely to make above $50,000 annually.

6. PARALLEL COORDINATE PLOT

- **Attributes:** Capital-gain, Age, Hours-per-week, Salary-Range
- **Values:** Divorced, Married-AF-Spouse, Married-CIV-Spouse, Married-spouse-absent, Never Married, Separated, Windowed, Education range from 10th - Masters

Graph:



Observation:

- The parallel coordinate plot reveals that this set of traits significantly distinguishes between the two classes. Orange and blue are distinct from one another.
- An individual with more education is likely to earn a salary of at least $50,000.
- Younger people are anticipated to earn less than elderly people.

Tools Used

Python, Jupyter Notebook, Scatter Plot, 3D- Scatter Plot, Parallel Co-ordinate Plot, Tree Map, Histogram.

Future Works

- In the future, i intend to apply machine learning methods to create machine learning models that can more accurately identify the elements affecting salary. In order to detect the features, we can also create visual recommenders using query builders.
- Can design a user interface where the user inputs the variable they wish to visualize, and the software then offers all of the possible visualization possibilities.
- Include seasonality, price, promotions, and product lifecycles together with local events.

Appendix+Code

Link: https://github.com/SherinSeba/Data_Visualization.git