Final Project

Systems Documentation Report

CSE 578: Data Visualization

Arizona State University – Summer 2022

Course Instructor: Samira Ghayekhloo

Submitted By:

Aishwarya Baalaji Rao

ASU ID: 1222423228

Date of Submission: 6th August 2022

1. Roles and Responsibilities

Stakeholders: UVW College
Product Owner: Aishwarya Baalaji Rao – Data Analyst, XYZ Corporation

As a member of the data analyst team at XYZ Corporation, the task is to create an application for UVW College that predicts income and boost their enrollment. Salary has been regarded as a crucial factor in determining marketing criteria for their degree programs. This forecast will be used to direct their marketing efforts toward certain people. The application will be built utilizing a dataset from the United States Census Bureau, with an emphasis on $50,000 as a critical figure for compensation.

The responsibilities include analyzing the data and relationships between the attributes to decide the best factors to combine and visualize the effects on income. After this analysis, the responsibilities include plotting the visualizations, narrating the story behind the visualization, and documenting the findings.

2. Goals and Business Objective

This project's goal is to detect patterns in the dataset by producing visuals to help identify the aspects that contribute to deciding an individual's income and to provide this information to UVW executives. The business objective is to boost the enrolment for UVW college programs. Additionally, the objective is to build an application that groups factors that contribute to the income earned so they can predict the income by varying the input parameters to effectively target their marketing efforts.

3. Assumptions

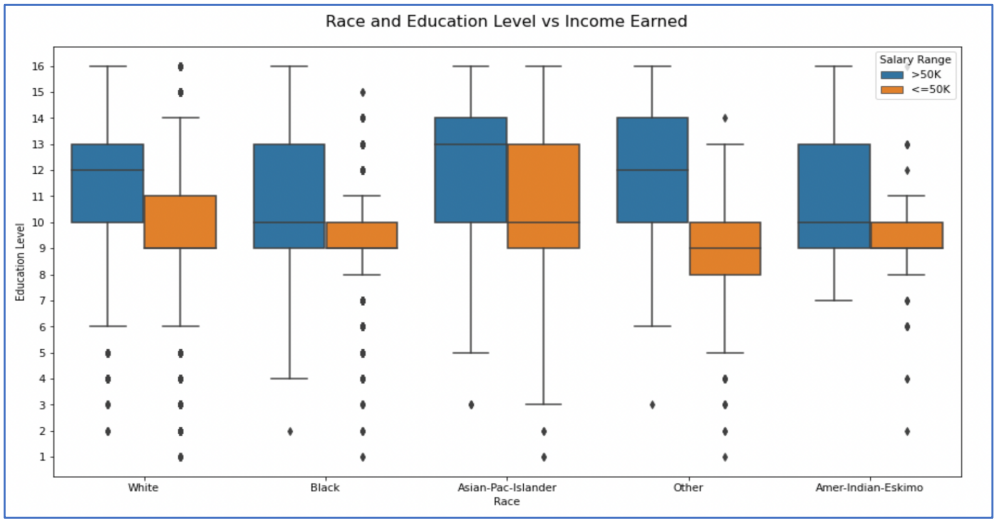
- **Data Accuracy and Quality**: It is assumed that the data is accurate and based on realistic situations so that the visualizations plotted convey a true story. It is also assumed that the quality of the underlying data is good and reliable.
- **Bias**: The data source analyzed is a biased extract from the 1994 US Census database.
- **Context**: It is assumed that the 14 attributes mentioned in the dataset gives enough context to come up with effective and targeted marketing campaigns and it is not missing any crucial information or attribute to our understanding of the data.
- **Threshold**: It is assumed that the 50K salary or income earned threshold is made intuitively and quantitatively to base all our visualizations on based on the pricing of the programs at UVW college.
- **Feature Selection**: It is assumed that the features that contribute the most to the income earned gives the most viable stories to report findings and hence most of the visualizations use these selected features among the 14 attributes.

4. User Stories

- **User Story 1 –** As a member of the UVW marketing team, we want to know if the ***race*** and ***education level*** have an impact on the income earned.
- **User Story 2 –** As part of the UVW marketing team, we want to know if the education level, ***capital gain***, ***age*** of people, and ***hours worked per week*** has an impact on the income earned.
- **User Story 3 –** As a member of the UVW marketing team, we want to know if the ***occupation*** of people, capital gain, and their ***work class*** has an impact on the income earned.
- **User Story 4 –** As a member of the UVW marketing team, we want to know if the age of people, the hours worked per week, and their ***gender*** has an impact on the income earned.

5. Visualizations

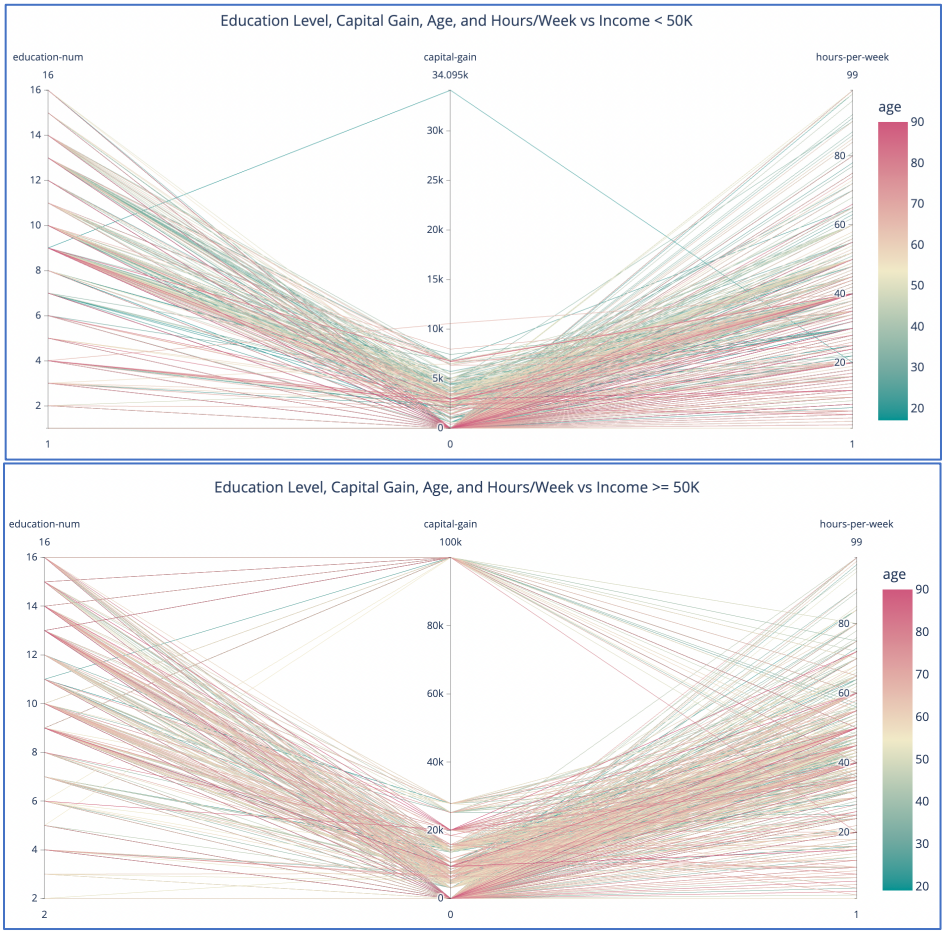
User Story 1 – Marketing wants to know the relationship between income earned and race, education level



9	HS-grad
14	Masters
13	Bachelors
10	Some-college
16	Doctorate
15	Prof-school
11	Assoc-voc
12	Assoc-acdm
4	7th-8th
8	12th
6	10th
7	11th
5	9th
3	5th-6th
2	1st-4th
1	Preschool

Visualization 1: The above box plot is a simple multi-variate analysis with 3 variables – Race, Education Level and Income Earned. The probability of a person's salary range is associated with their education level and race. We can see that people of all races have an income greater than 50K with their education level between 9-14 (HS-Grad, Some College, Assoc-voc, Assoc-acdm, Bachelors, Masters), indicating the importance of college education. In Others race category, we can observe that people are gaining <=50K income with education levels being below 10. In Black race for income <=50K, we observe a lot of outliers in the data. In the Asian-Pac-Islander race we can see that people with education level up to 13 (Bachelors), are earning <=50K but with an extra push of education level 14 (Masters), they can push income to >50K. Therefore, we can conclude that the race and education level can predict the income of an individual.

User Story 2 – Marketing wants to know the relationship between income earned and education level, capital gain, hours/week, and age



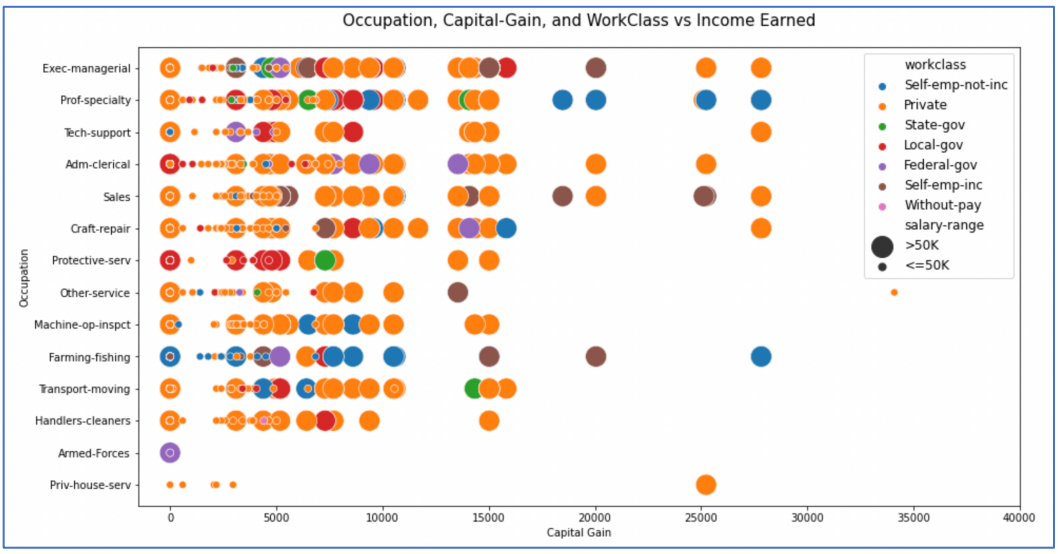
9	HS-grad
14	Masters
13	Bachelors
10	Some-college
16	Doctorate
15	Prof-school
11	Assoc-voc
12	Assoc-acdm
4	7th-8th
8	12th
6	10th
7	11th
5	9th
3	5th-6th
2	1st-4th
1	Preschool

Visualization 2: The above parallel co-ordinate plot is a multi-variate analysis with 5 variables – Education Level, Capital Gain, Age, Hours/Week and Income Earned. The data is split in 2 classes with income >=50k and <50K before feeding it to the plot.

1. For people with salary <50K: We can observe that overall, there is a green tinge going from education level 7-16, earning a capital gain of about 0-8K, and working in the range of 20-60 hours/ week indicating that most of the demographic in this range are of younger age between 20-45 years and earning less than 50K. Conversely, the red tinge indicating older people of age 65-90 years who are earning <50K hold an education level varying from 4-16, have a capital earning between 0-5k and are working less than 40 hours/week.
2. For people with salary >=50K: We can observe that the graph is dominated by a red tinge indicating that older people of age 60-90 years old are the majority earning over 50K, holding education level between 4-16, capital gain of 0-30K (with outliers at 100K) and working between 20-80 hours/week. Younger people are also found in this category for the same range of variables but is majorly dominated by the older people.

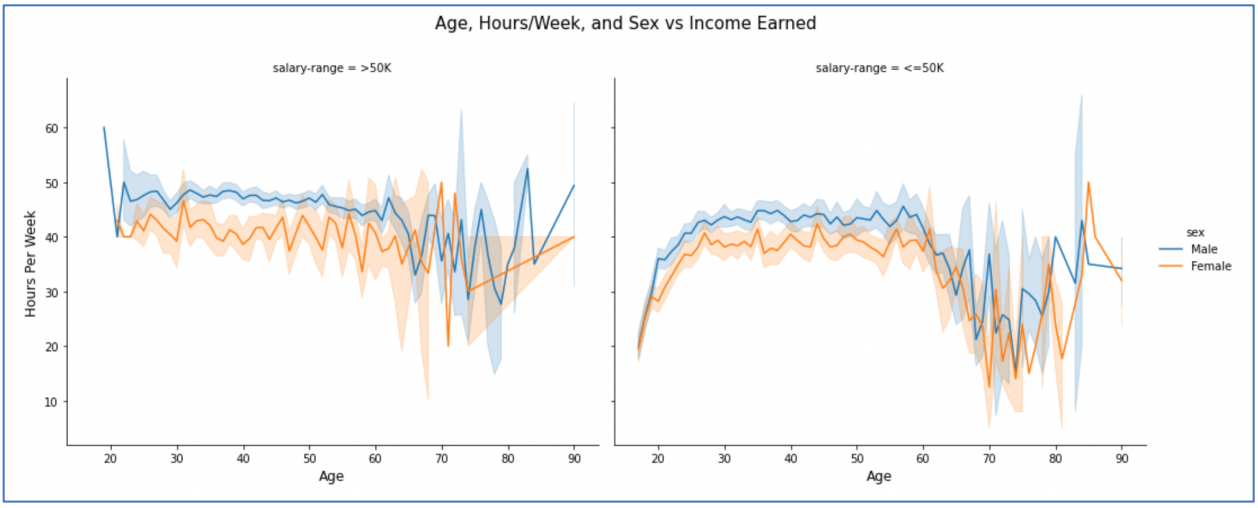
This concludes that the income of an individual can be effectively predicted in the application by using the education level, capital gain, age, and hours/week variables.

User Story 3 – Marketing wants to know the relationship between income earned and education level, capital gain, hours/week, and age



Visualization 3: The above scatter/bubble plot is a multi-variate analysis with 4 variables – Occupation, Capital Gain, Work Class, and Income Earned. *For efficient inference, the outliers above 40K were removed in the plot*. The probability of a person's income is seen to be related to their occupation, capital gain and work class. From the above visual, we can see that most of the people earning more than 50K salary belong to the private sector of work class in every occupation. Capital gain is mostly concentrated in the range of 0-15K for all the work class. This indicates that most of the marketing efforts can be catered towards people working in the private sector to boost enrollment for UVW college and can effectively predict the income.

User Story 4 – Marketing wants to know the relationship between income earned and age, hours/week, and gender



Visualization 3: The above line plot is a multi-variate analysis with 4 variables – Age, Hours/Week, Sex, and Income Earned. The probability of a person's income is related to the factors chosen for this plot. The above

visual is split into two columns – the left one showing the data for people who earn a salary of >=50K and the right one shows for people earning a salary of <50K. We can immediately see a salary gap between male and female from both the columns, suggesting that the pay is not equal for women working the same number of hours as men. For people earning >= 50K we see the trend that both male and females in the age of 20-65 years old are working between 40-60 hours a week. And people above 65 years old have varying hours of work between 20-55 per week. For people earning <50K, we see that both males and females in the age between 20-60 years old are working anywhere between 20-45 hours/week. After 65 years old, there is a high variety of hours put in work ranging between <20 to 50 hours/week. This concludes that the variables considered here can predict the income of an individual effectively.

Visuals Chosen and the Design Process

The visuals chosen for this project's objective are box plot, parallel coordinate plot, scatter plot and line plot. First the data and its columns were thoroughly analyzed to select important features to visualize. After that, research was conducted to find out what kind of plots represent numerical and categorical data for a multivariate analysis. Since color could potentially relate to a category value, box plot was the appropriate option for representing the data in the first visualization, which included two categorical variables. For the second visualization, a parallel coordinate plot was the optimal option for comparing four distinct forms of numerical data, which was then divided into two plots to illustrate the salary comparison. For the third visual, scatter plot could be used and converted to a bubble plot to incorporate 4-dimensional data by assigning two of them to size and color attributes. For the last visual, line plot was considered and made 4 dimensional by using the column split and the color attribute, which concluded the 4 required visualizations with 8 unique attributes from the data.

6. Not Doing

As part of future work, the visualizations could be plotted as a dashboard to create an application for the marketing executives to come up with effective and targeted campaigns. It would be helpful to get more diverse and reliable data to have better inferences from the visualizations being plotted. More helpful visualizations can also be plotted to boost enrollment, but we were limited to four for the scope of this project. A prediction model could also be built to effectively clean and process data to suggest effective marketing decisions.

7. Appendix

Jupyter-Notebook Code:

Cell 1:

```
1.  import pandas as pd
2.  import matplotlib.pyplot as plt
```

```
import numpy as np
from statsmodels.graphics.mosaicplot import mosaic
import seaborn as sns
%matplotlib inline
import plotly.express as px

import warnings
warnings.filterwarnings("ignore")
```

Cell 2:

```
data = pd.read_csv('./adult.data', sep=", ", engine='python', header=None)
data.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation", "relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "salary-range"]
data = data[data["workclass"] != '?']
data = data[data["education"] != '?']
data = data[data["marital-status"] != '?']
data = data[data["occupation"] != '?']
data = data[data["relationship"] != '?']
data = data[data["race"] != '?']
data = data[data["sex"] != '?']
data = data[data["native-country"] != '?']
below50K = data[data["salary-range"] == "<=50K"].sample(n=7841)
above50K = data[data["salary-range"] == ">50K"]

data = pd.concat([above50K, below50K])
```

Cell 3:

```
# Visualization 1: The UVW marketing team wants to find out if Race and Education level
# have an impact on the Income Earned.

plt.figure(figsize=(15,8))
sns.boxplot(x='race',y='education-num', data=data, hue="salary-range")

plt.yticks(np.arange(1, 17, 1.0))
plt.title('Race and Education Level vs Income Earned', pad=20).set_fontsize(15)
plt.xlabel("Race")
plt.ylabel("Education Level")
plt.legend(loc='upper right', bbox_to_anchor=(1, 1), title="Salary Range")
plt.show()

educationNumArr = data['education-num'].unique();
educationArr = data['education'].unique();
concArr = np.stack((educationNumArr, educationArr), axis=1)
customLegend = pd.DataFrame(concArr)
print(customLegend)
```

Cell 4:

```
# Visualization 2: The UVW marketing team wants to find out if Education Level, Capital
# Gain, Age, and Hours Worked per Week have an impact on the income earned.

viz2a = px.parallel_coordinates(below50K, color="age",
```

```
                                dimensions = ['education-num', 'capital-gain', 'hours-per-week'],
                                color_continuous_scale=px.colors.diverging.Tealrose)
viz2a.update_layout(barmode='group',
                        title=dict(
                        text="Education Level, Capital Gain, Age, and Hours/Week vs Income < 50K",
                        x=0.5,
                        y=0.95,
                        xanchor='center',
                        yanchor='top',
                        font=dict(size=15)
                    ))


viz2a.update_layout(margin=dict(t=100))
viz2a.show()


viz2b = px.parallel_coordinates(above50K, color="age",
                                dimensions = ['education-num', 'capital-gain', 'hours-per-week'],
                                color_continuous_scale=px.colors.diverging.Tealrose)

viz2b.update_layout(barmode='group',
                        title=dict(
                        text="Education Level, Capital Gain, Age, and Hours/Week vs Income >= 50K",
                        x=0.5,
                        y=0.95,
                        xanchor='center',
                        yanchor='top',
                        font=dict(size=15)
                    ))


viz2b.update_layout(margin=dict(t=100))
viz2b.show()

educationNumArr = data['education-num'].unique();
educationArr = data['education'].unique();
concArr = np.stack((educationNumArr, educationArr), axis=1)
customLegend = pd.DataFrame(concArr)
print(customLegend)
```

Cell 5:

```
# Visualization 3: The marketing team at UVW wants to find out the relation between
# income earned vs the Occupation of people, Capital Gain and their Work Class.

plt.figure(figsize=(15,8))
viz3 = sns.scatterplot(x='capital-gain', y='occupation', size='salary-range', hue="workclass",
data=data, sizes=(50, 400))

plt.title('Occupation, Capital-Gain, and WorkClass vs Income Earned', pad=20).set_fontsize(15)
plt.xlim(-1500, 40000)
plt.xlabel("Capital Gain")
plt.ylabel("Occupation")
plt.legend(prop={'size': 12})
```

```
plt.show()
```

Cell 6:

```
# Visualization 4: The marketing team at UVW wants to find out the relation between
income earned
# vs the Age of people, the Hours they Work per Week and their Gender.

viz4 = sns.relplot(
    x = "age",
    y = "hours-per-week",
    data = data,
    height = 6,
    kind = 'line',
    hue = 'sex',
    col = 'salary-range', aspect=8/6.8);

viz4.fig.subplots_adjust(top=.85)
viz4.set_axis_labels("Age", "Hours Per Week", fontsize=12)
viz4.fig.suptitle('Age, Hours/Week, and Sex vs Income Earned', fontsize=15);
```