CSE 578 Data Visualization

Apoorva Rastogi
1222362258
arasto17@asu.edu

Roles and responsibilities: Apoorva Rastogi is responsible for this project. This project uses data provided by XYZ Corporation to develop marketing profile for UVW College.

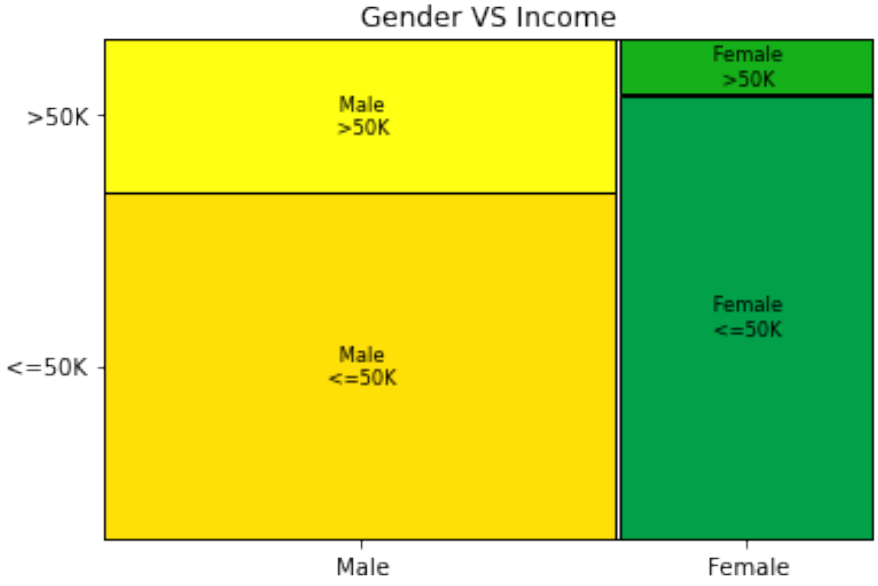
Goals and a business objective: XYZ Corporation uses data to develop marketing profiles on people. UVW College is a local college looking to bolster enrollment. Here we develop marketing profiles using data supplied by the United States Census Bureau, and we have focused on $50,000 as a key number for salary.Successfully identify the factors which affect income of an individual so that they can be used for the prediction tool.

Assumptions:

- UVW College has chosen salary of $50K as a key demographic to determine the criteria for marketing its programs.
- The college marketing department would like to create profiles based on factors in the dataset.
- The data source analyzed is a biased extract from the 1994 US Census database.
- The data is correct representation of the demographic present even its it does not cover entire demographic.

User Stories:

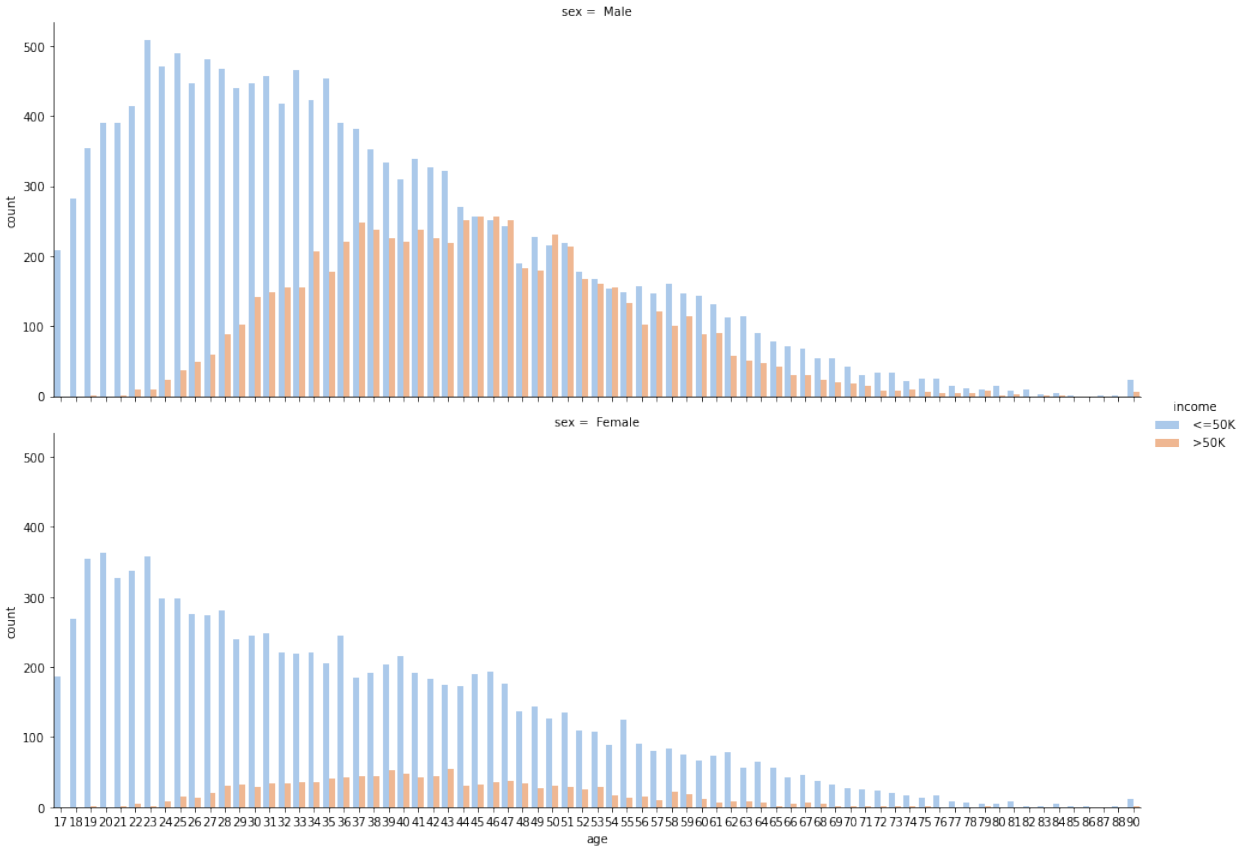
User Story #1: As a member of the UVW , I want to know if the gender and age of an individual is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into prediction tool.



As this is the first visualization I have created simple and integrated plot for both the factors to help understand the data as well as visualization better. As you can see in the above mosaic plot no of instances of male is much larger than female and among the males a much larger portion of the population is earning more than 50 k compared to female population.

In this second visualization we see the distribution of all the data we have with respect to the age and we can see that it is right skewed and also starts from the minimum age of 17. We see that number of working people are more in age of 20-50 and gradually decline.

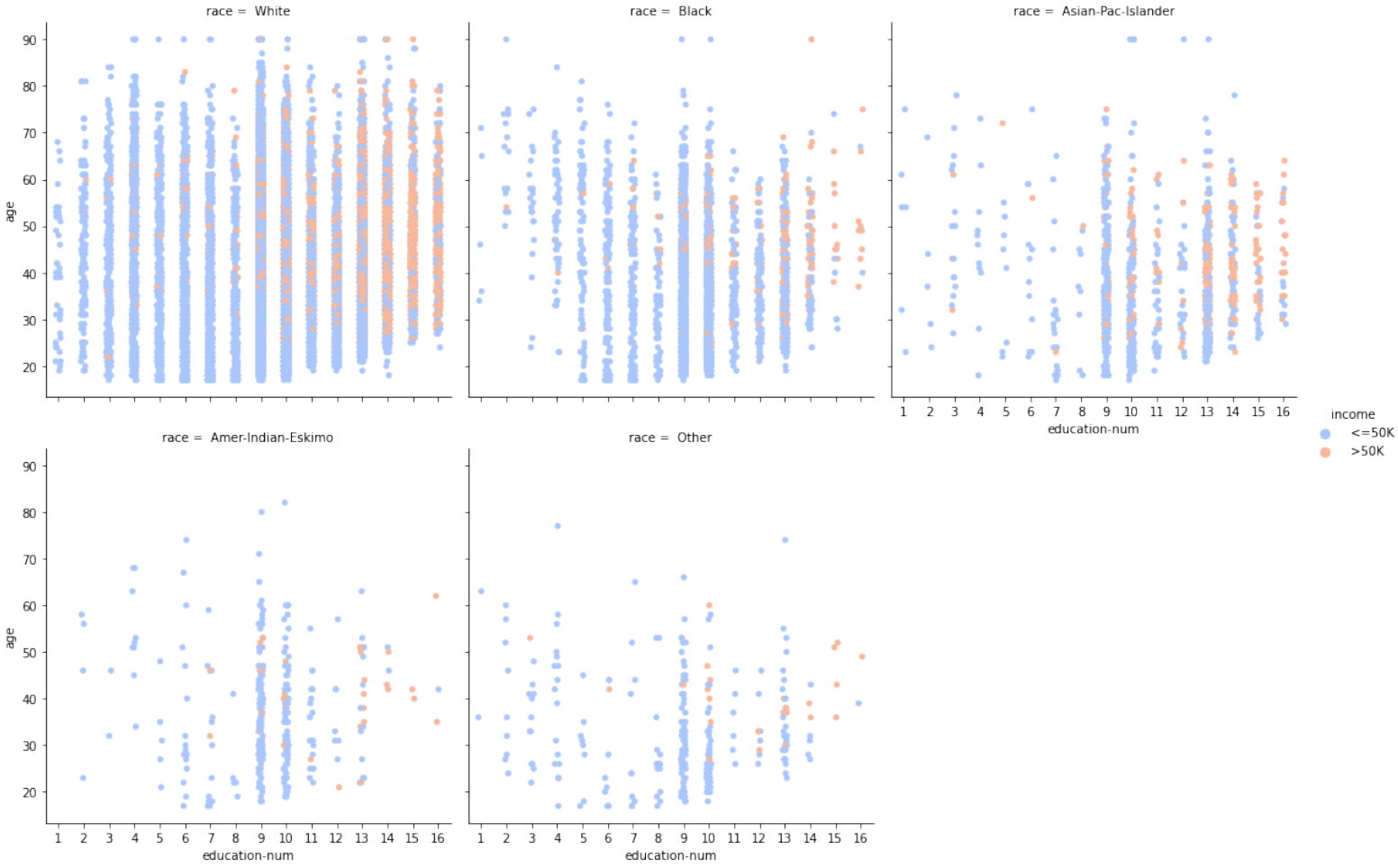
Gender VS Income VS age



Combining these two we can clearly see both of these in the graph above. Both the graphs are right skewed which means that for both the gender people majority of working people are young. Also we can clearly see no of males is higher than the females. Also, the orange distribution which shows the people earning more that 50 k is greater for men. IN men for ages 40-60 its almost the same height as < 50 k but for women it never reaches even close to <50k. In women greater than 50 k is always much less than less than 50 k.

User Story #2: As a member of the UVW , I want to know if the education and race an individual is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into prediction tool.

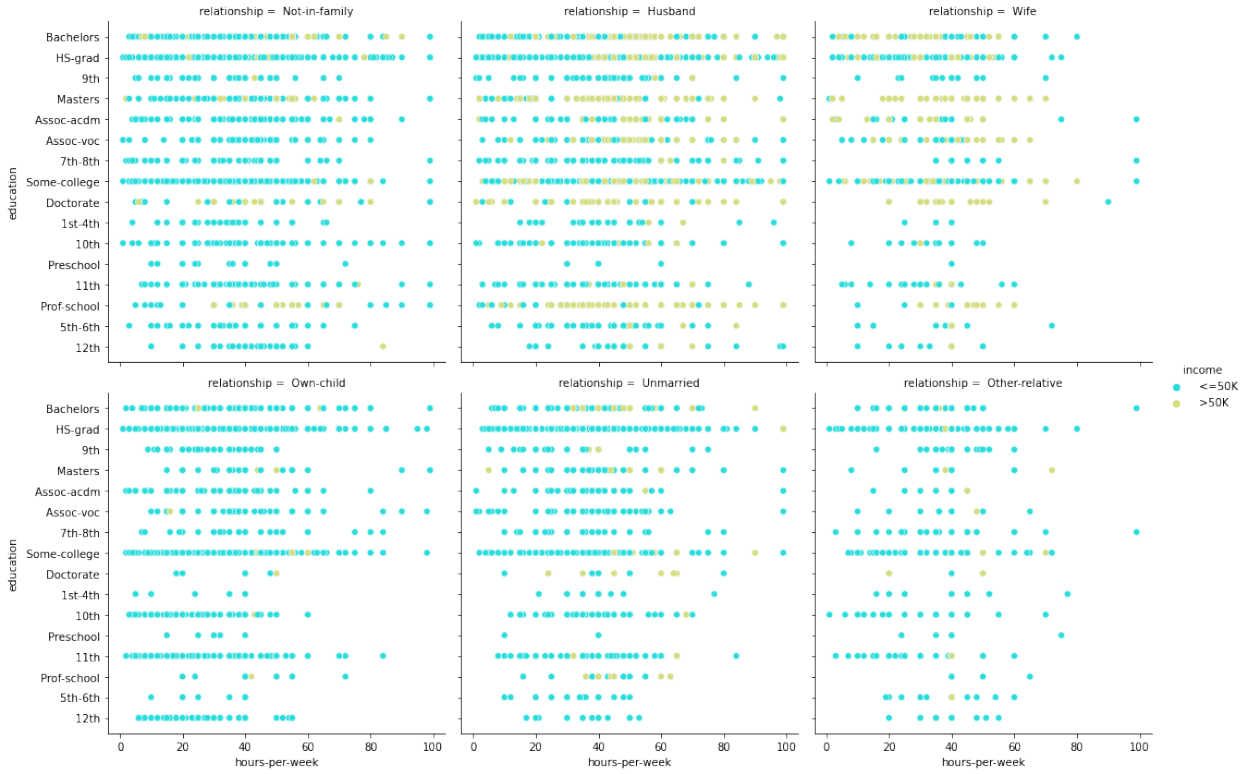
Race vs age vs Education-num



As we can see from the visualization above that people with higher no of education years are more likely to earn more which is true for any and all race. Although we have less data points for some races but we can see that higher education is helpful in achieving higher salary nevertheless. We can also see that the graph is kind of tapered from lower right which means we do not have people with young age having higher degree which is obvious as you need certain years to complete higher education. Also, although at every age people are earning > 50 k, majority seems to be between 30-60.

User Story #3: As a member of the UVW , I want to know if the education, relationship and hours-per-week an individual is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into prediction tool.

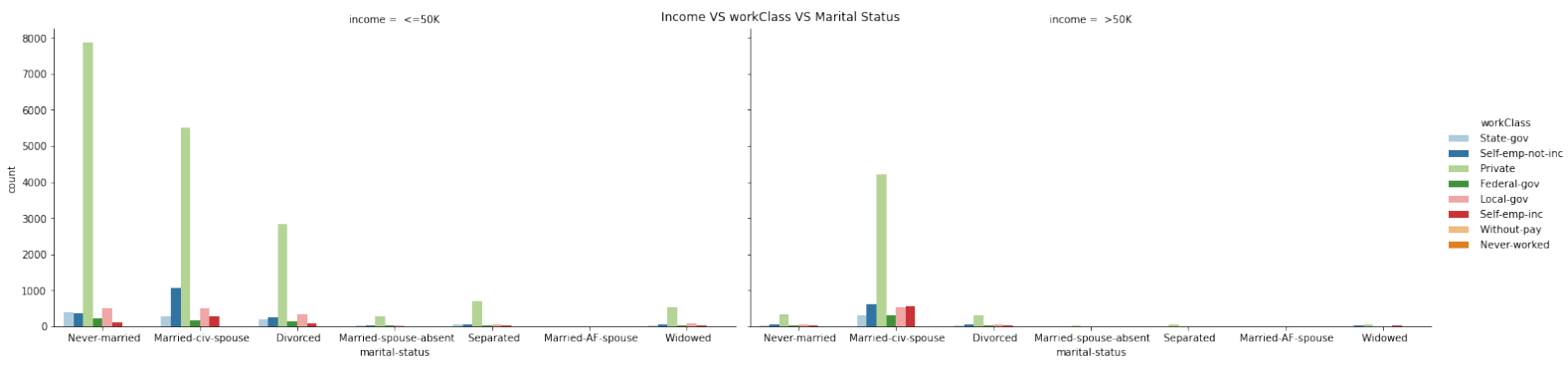
relationship VS education VS hours-per-week



From the above visualization we can see that people with prof-school, doctorate, masters have majority of people earning greater than 50 k and after those are people with Bachelors , some college , HS-grad with good no of people earning > 50k. Data points in relation wife, own-child, unmarried and other-relative are more right skewed which means they typically work less no of hours. Relation Own Child has majorly 10th, 11th, 12th, Bachelors, HS-grad and some-college they are good to pursue higher education to get higher salaries. Mostly not in family is earning less than husband and wife relation for same education level and hours of work.

User Story #4: As a member of the UVW , I want to know if workClass, Marital Status of an individual is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into prediction tool.

We can see that majority of earning is from private jobs for any income or marital-status. Married-civ-spouse has almost similar if not same no of people in greater as well as less than 50 k but for others different is much larger. Majorly earning people are from Never-Married,Married-civ-spouse, divorced people out of which only few from never-married and divorced earn greater then 50 k.



Not-Doing:

1. How does age relate to marital status or relation to determine Income
2. How does capital gain and loss affect Income.
3. What age people like to work in what work Class.
4. How does occupation related to Income and Education.

Appendix:

Attributes Selected:
1. Age
2. Gender
3. Race
4. Education-No
5. Hours-Per-Week
6. Relation
7. Education-Degree
8. Work Class
9. Marital-Status

Name - Apoorva Rastogi
#1222362258
#arasto17@asu.edu

Code:

```
import sqlite3
import pandas as pd
import itertools
import matplotlib.pyplot as plt
import numpy as np
from statsmodels.graphics.mosaicplot import mosaic
import seaborn as sns
columns = ["age", "workClass", "fnlwgt", "education", "education-
num","marital-status", "occupation", "relationship",
        "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-
country", "income"]
df = pd.read_csv('adult.data', names=columns,sep=',', na_values=' ?')
```

```
df = df.drop('fnlwgt',axis=1)
```

UserStory1:

```
props={}
props[(" Male"," <=50K")]={'facecolor':'xkcd:dandelion','edgecolor':'black'}
props[(" Male"," >50K")]={'facecolor':'xkcd:yellow','edgecolor':'black'}
props[(" Female"," <=50K")]={'facecolor':'xkcd:emerald','edgecolor':'black'}
props[(" Female"," >50K")]= {'facecolor':'xkcd:green','edgecolor':'black'}
mosaic(df, ['sex', 'income'],properties=props)
plt.title("Gender VS Income")
plt.show()

ax =
sns.catplot(x="age",hue='income',col='sex',data=df,aspect=2.7,col_wrap=1,
kind="count",palette='pastel')

ax1 = sns.catplot(x="age",data=df,aspect=2.7, kind="count",color='skyblue')
ax1.fig.suptitle('age VS No of Earning People')
```

UserStory2:

```
 sns.catplot(
    data=df,
    #x='sex',
    x="education-num",
    y='age',
    palette='coolwarm', col="race",
    hue="income",col_wrap=3
)
```

UserStory3:

```
plot = sns.relplot(x='hours-per-week',data=df,y='education',
palette='rainbow',hue='income',col="relationship",col_wrap=3)
```

UserStory4:

```
ax = sns.catplot(x="marital-status",hue='workClass', col="income",data=df,
kind="count",aspect =2,palette="Paired")
ax.fig.suptitle('Income VS workClass VS Marital Status')
```