CSE 578: DATA VISUALIZATION

SYSTEM DOCUMENTATION REPORT

Roles and Responsibilities:

Team Members

1. Ruthwik Reddy Ratna
2. Siva Rama Pavan Kumar Buddi
3. Sandeep Kumar Reddy Onteddy
4. Shivani Devulapally
5. Nithish Kumar Reddy Nallamjompalli
6. Sudarshan Reddy Mallipalli

Product Owner:
XYZ Corporation

Stakeholder:
UVW College

Responsibilities of each team member is as follows:

1. Siva Rama Pavan Kumar Buddi worked on 4 user stories which determine an individual's salary.
2. Ruthwik, Sandeep worked on the remaining 4 user stories which helps in identifying an individual's salary.
3. Sudarshan, Nithish worked on a system documentation report.
4. Shivani worked on data cleaning/preprocessing.

Team Goals:
The goal of our team is to create an application which finds factors that determine the individual’s income. Developing user stories with different attributes so as to find relevant factors in determining an individual’s income.

Business Objective:

1. As a marketing team at XYZ Corporation, our task is to identify the attributes that determine an individual's income.
2. The goal our team as a marketing team is to improve admissions of UVW college where individual salary is main attribute

Assumptions:

- **Dataset Accuracy:** For any visualization dataset is the major thing. Without assuming the dataset as correct we can not achieve good results. Our team trusts that the data provided is accurate and latest.
- **Timeliness and Relevance:**We can safely assume that data is collected at right time.Dataset which is collected long back might not describe the current situation in a right way and may misinterpret the results.For example, if the salaries are taken at the time of recession or any other

situation which impacts the salary then the dataset will be biased and will not produce right results and might misinterpret the situation.

- **Selecting right attributes:** Selecting right attributes is the key while using visualizations. To display accurate results selecting attributes plays an important role. Our team believes that salary is the main attribute in selecting other parameters.

User Stories:

User Story #1: Determining whether **Education-num** is a relevant factor in identifying an individual's income.
User Story #2: Determining whether **sex** is a relevant factor in identifying an individual's income.
User Story #3: Determining whether **capital-loss, capital-gain, age** are relevant factors in identifying an individual's income.
User Story #4: Determining whether **education** is a relevant factor in identifying an individual's income.
User Story #5: Determining whether **workclass** is a relevant factor in identifying an individual's income.
User Story #6: Determining whether **Hours per week** is a relevant factor in identifying an individual's income.
User Story #7: Determining whether **Age** is a relevant factor in identifying an individual's income.
User Story #8: Determining whether **Race** is a relevant factor in identifying an individual's income.

Visualizations:

USER STORY #1: Education-num is a factor in determining an individual's income.

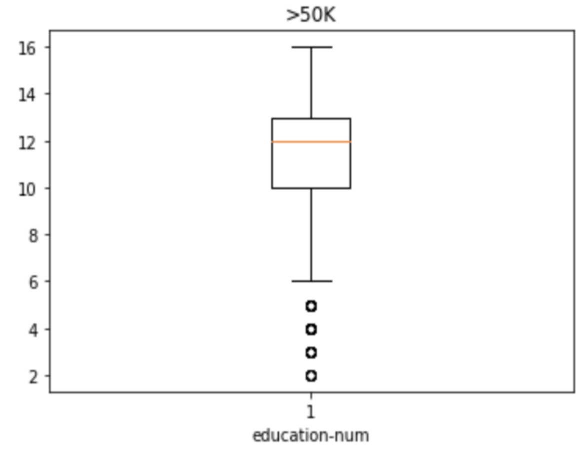


Figure 1(a). Box plot of education-num having salary >50k.

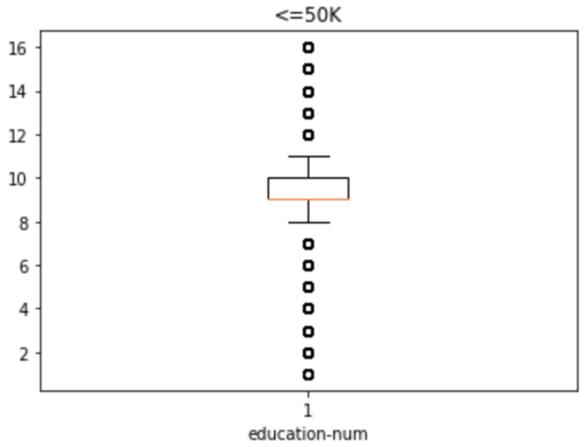


Figure 1(b). Box plot of education-num having salary <=50k.

The data is used to compare between groups, so using a box plot(whisker plot) for the visualization helped us understand how education-num level is a good factor to determine salary. For individuals making > 50k a year, a large number of them have at least a college degree or more. Whereas for the individuals making <= 50k a year, the majority have a high school education or less. Based on the charts, the marketing team should target individuals with an education-num less than or equal to high school.

USER STORY #2: sex is a factor in determining an individual's income.

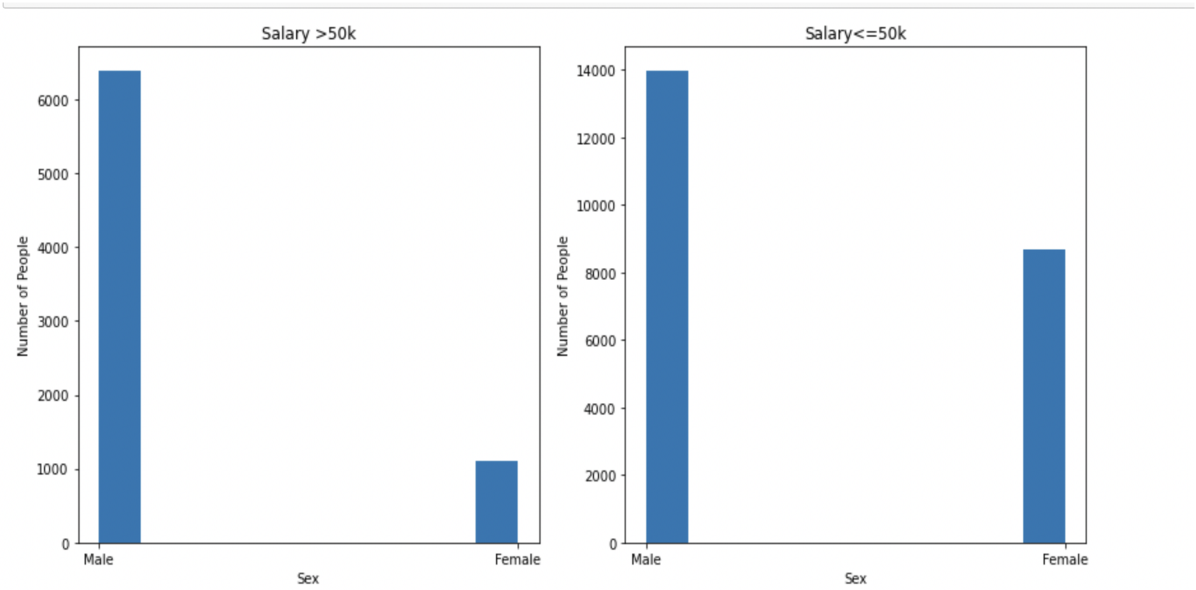


Figure 2(a). Histograms of employee sex having salary >50k. Figure 2(b). Histograms of employee sex having salary <=50k.

The above visualizations are the histograms which are used to determine the salary of an individual. From Figure 1(a), more number of male employees have a salary >50K than female employees. There is a significant increase in male and female employees where the salary is <=50K. It can be clearly deduced that more have salaries <=50k. Of the people making a salary >50K are male employees. In terms of recruitment to bolster admissions of UVW college, they need to focus more on recruiting female employees where the majority of them are having salaries <=50K.

USER STORY #3: capital-gain, capital-loss, age are factors in determining an individual's income.Figue

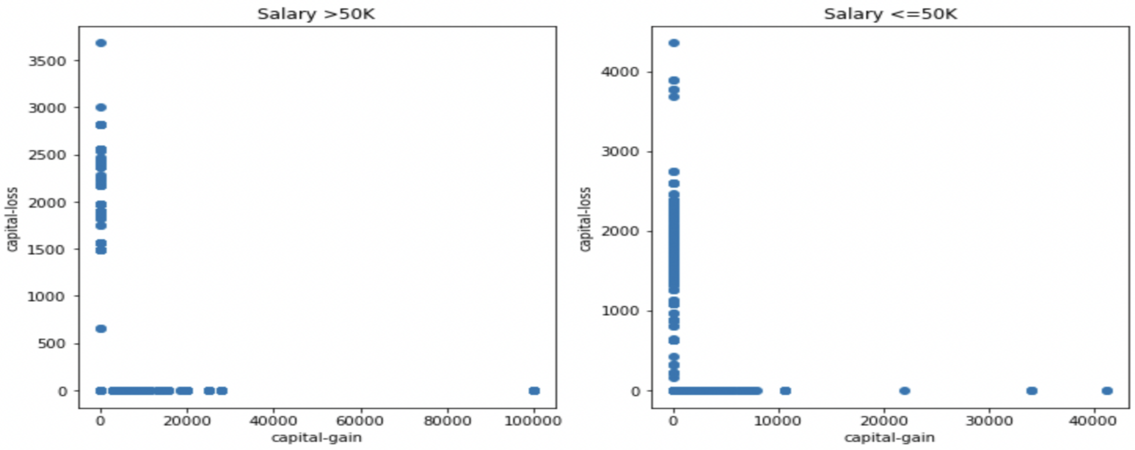


Figure 3(a): scatterplot for capital-gain and capital-loss having salary >50K

Figure 3(b): scatterplot for capital-gain and capital-loss having salary <=50K

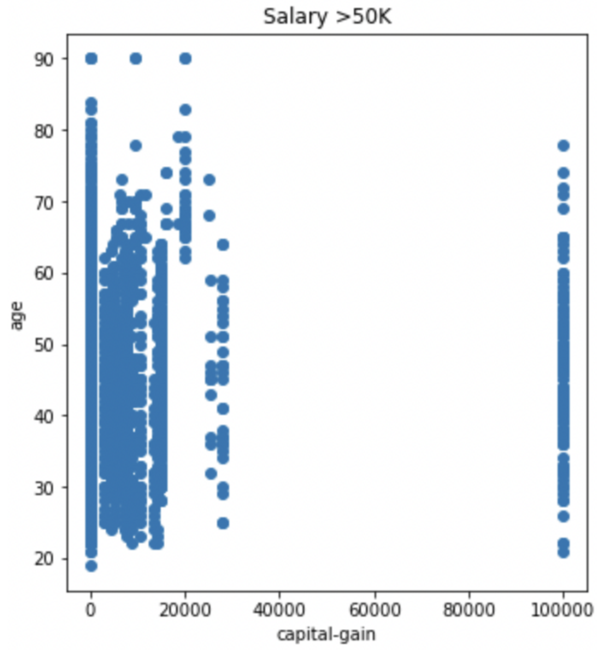


Figure 3(c): scatterplot for capital-gain and age having salary >50K

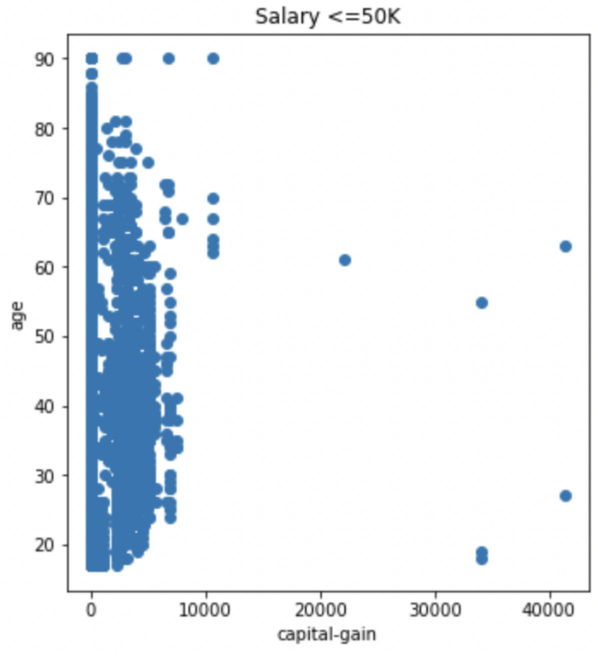


Figure 3(d): scatterplot for capital-gain and age having salary <=50K

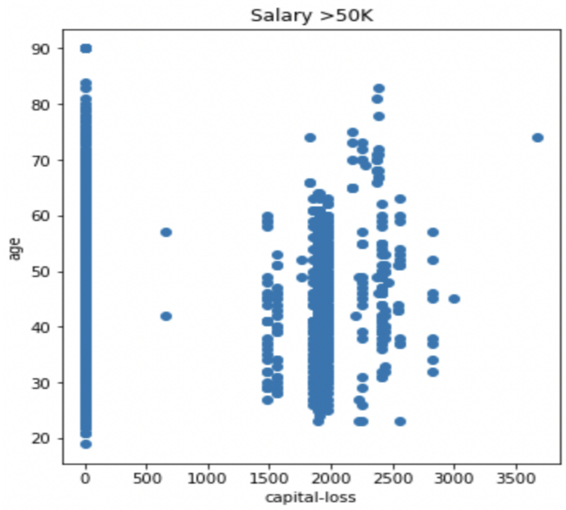


Figure 3(e): scatterplot for capital-loss and age having salary >50K

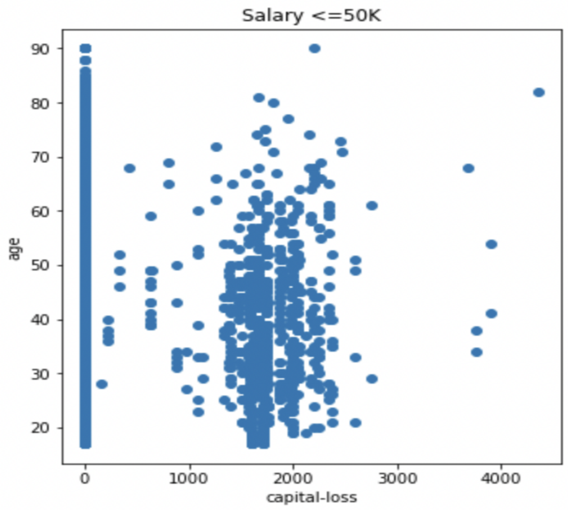


Figure 3(f): scatterplot for capital-loss and age having salary <=50K

Above visualizations are scatter-plots to determine an individual’s income with respect to capital-loss, capital-gain, age. Figure 3(a), Figure 3(b) are the scatter plots of capital-gain and capital-loss with salary >50K, <=50K. Figure 3(c), Figure 3(d) are the scatter plots of capital-gain and age with salary >50K, <=50K. It can be inferred that most people with a salary >50K of all ages have capital-gain ranging from 0 to 20000. On the other hand, employees with a salary <=50k have capital-gain ranging from 0 to 10000. Individuals with high capital-gain are likely to have salaries >50K. Figure 3(e), Figure 3(f) are the scatter plots of capital-loss and age with >50K, <=50K. It can be inferred that most of the people are having capital-loss ranging from 1000 to 3000.

USER STORY #4: education is a factor in determining an individual's income.

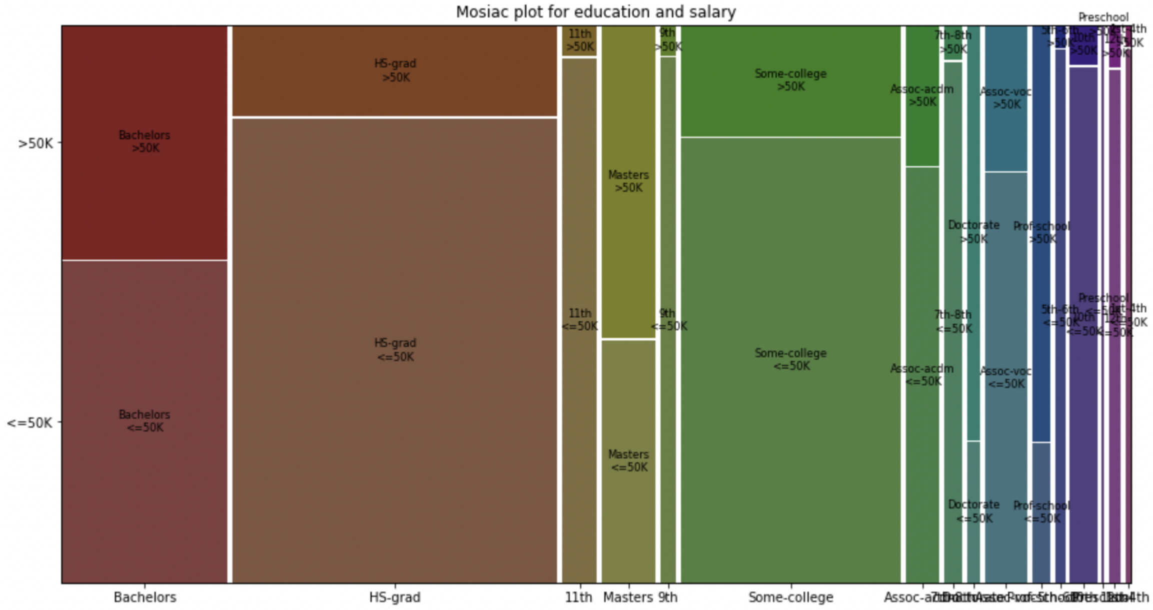
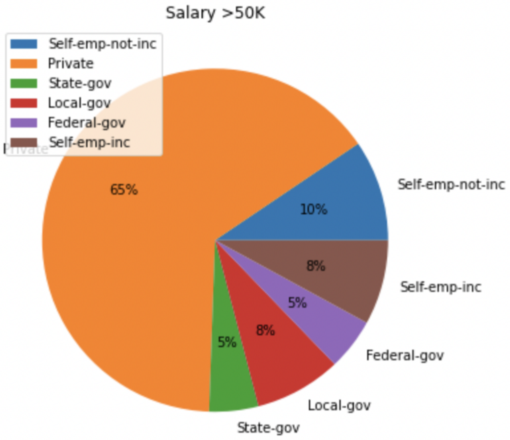


Figure 4: Mosaic plot for education in determining salary of an individual

Mosaic plots give a good visualization when trying to find the relationship between data with two or more categories using the visual attributes of 'Area' and 'Length'. From Figure 4 we can deduce that the majority of the portion is covered by high school grads making a salary <=50K. Half of the people are from bachelors and high school grads having salaries >50K and <=50K. Furthermore, the majority of individuals belonging to the lower education levels to the right of Figure 4 also make less than or equal to 50k a year. So, UVW college must focus on marketing for students with less education levels.

USER STORY #5: Workclass is a relevant factor in determining an individual's income.



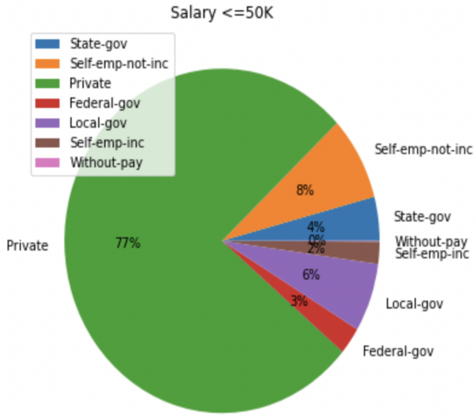


Figure 5(a) Pie chart showing of workclass of employees having salary >50K

Figure 5(b) Pie chart of workclass of employees having salary <=50K

There are seven types of workclass of employees working in the marketing team. For categorical data, pie charts are the best for visualization to deduce the results. From figure 5(a) it can be inferred that most of the employees from private workclass are having salaries >50K. But people having salaries <=50K(Figure 5(b)) are from private workclass. But in both the charts the majority of the employees are having workclass as private. 77% of the employees having salaries <=50K are from private workclass. So UVW could focus on marketing to the people from other workclass.

USER STORY #6: Hours per week is a factor in determining an individual's income.

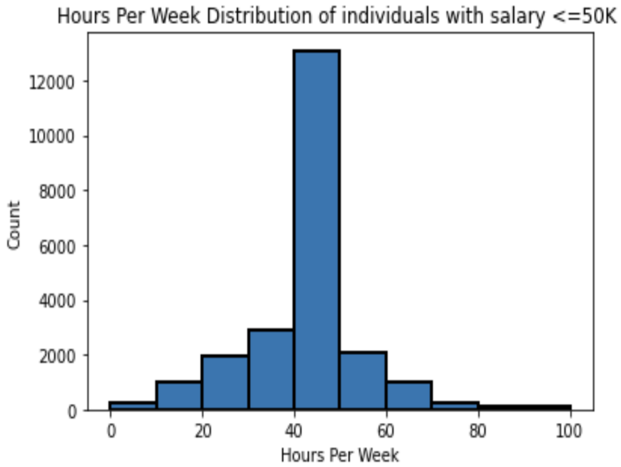


Figure 6(a). Histogram of Hours per Week having salary <=50k.

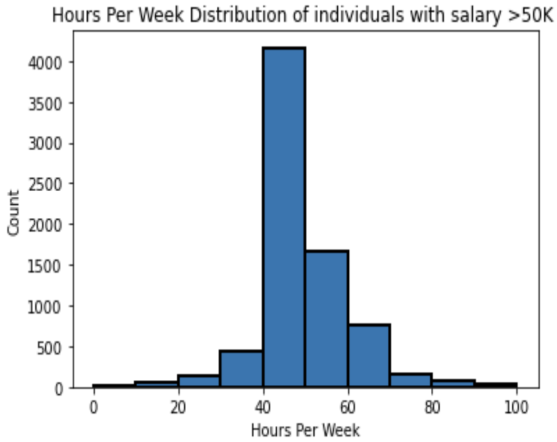


Figure 6(b). Histogram of Hours per Week having salary >50k.

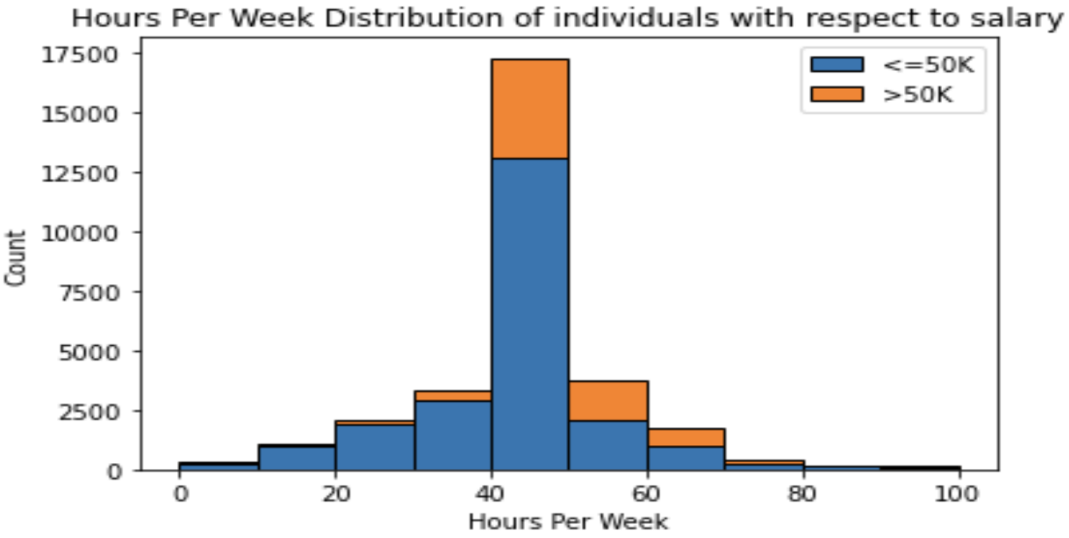


Figure 6(c). Histogram of Hours per Week of individuals with respect to salary.

This analysis of hours per week to salary is shown in the above visualization, it is shown that the hours per week representation to salary as <=50K and the hours per week representation with respect to salary as >50K.In the context of UVW's recruitment approach, if they're looking for potential students who make less than $50,000 per year, they might want to look for those who work 20 to 40 hours per week or fewer. This is because they are the people who are most likely to have time for a university program and are most likely working a part-time job rather than a full-time job.

USER STORY #7: Age is a factor in determining an individual’s income.

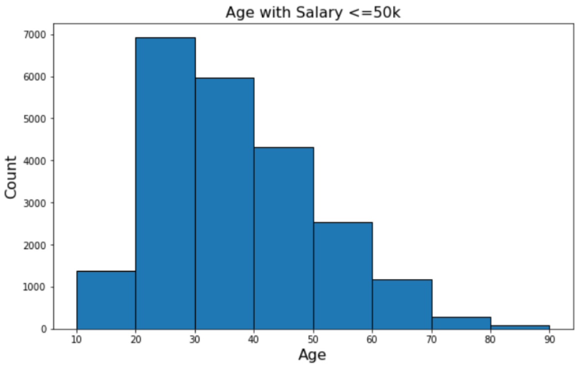


Figure 7(a). Histograms of individuals age having a salary >50k.

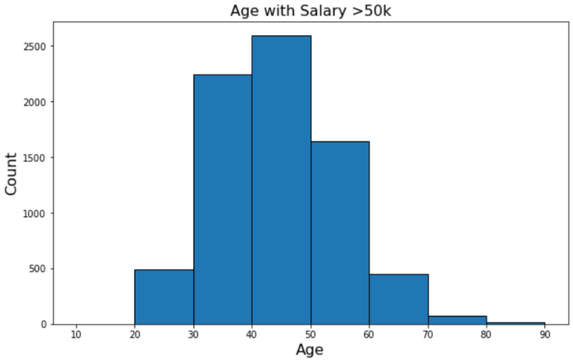


Figure 7(b). Histograms of individuals age having salary <=50k.

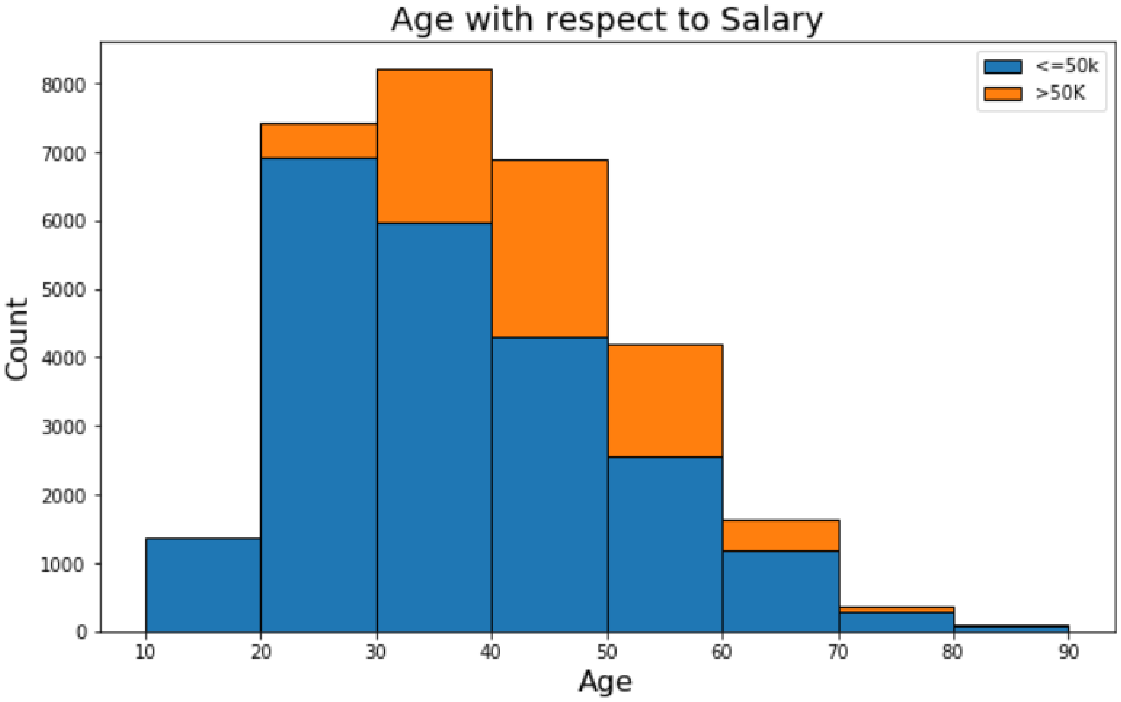


Figure 7(c). Histograms of individuals age having a salary >50k.

From figure 7(a) we can see that the histogram is right skewed which shows that the majority for people whose salary is less than 50k are people who are young or less age. From figure 7(b) it is evident that the majority of people earning a salary >50k are between 40-50. So in order to target people earning <=50k salary focus on people with age 20-40 and similarly for people earning >50k focus on the age group 40-50.

USER STORY #8: Race is a factor in determining an individual's income.

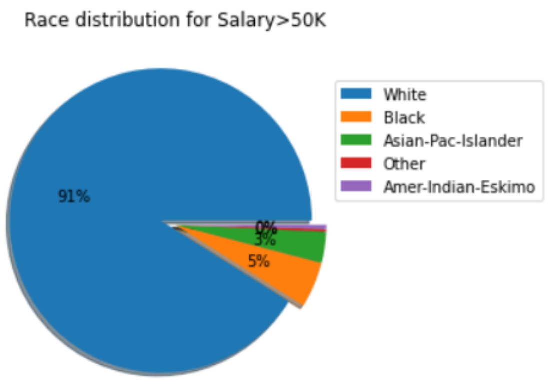


Figure 8(a). Histograms of individuals Race having salary >50k.

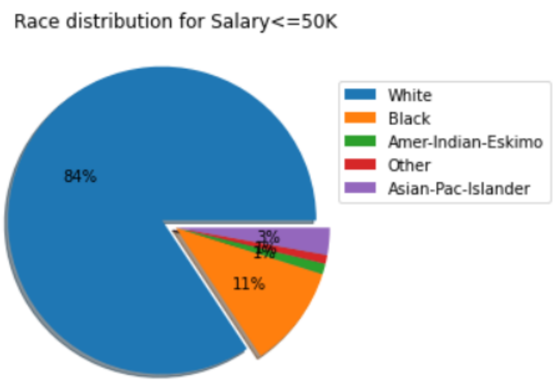


Figure 8(b). Histograms of individuals Race having salary <=50k.

The Race attribute has 5 categorical values which have their contributions to two classes (salary <=50k and >50k) as above. From the above figures it is evident that the majority of both the classes have a high white race percentage. So we will not be able to judge the impact of the other categorical values on the salaries as the data is skewed more towards White race.

Questions:

- **What features to use and how to assign weightage towards each feature in the dataset?**
  This data set in total contains 14 features out of which 6 are continuous variables and rest are categorical variables.All the features do not contribute to the prediction of class.Therefore we analyzed each of the features using visualization tools and drawn relationships between them for further analysis.Later, we have ranked the features based on their importance and selected the most relevant features.

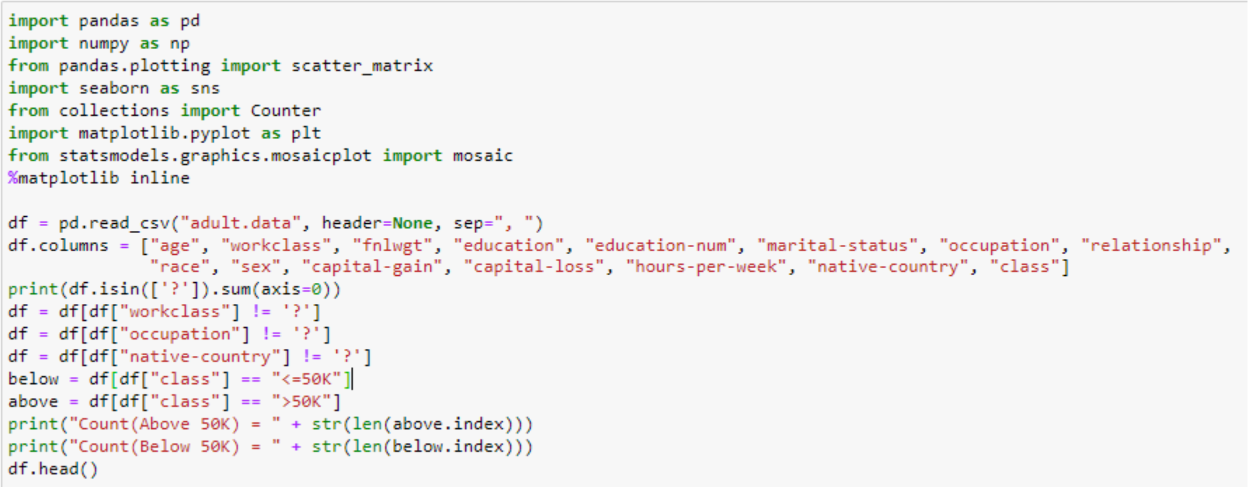
- **Should we ignore the data that contains null values in one of their attributes**?
  We have studied the dataset thoroughly and noticed that null values only occurred in three attributes namely workclass, occupation and native country.Therefore we cannot completely remove the entire columns which contain these null values.Therefore we have analyzed the data carefully and removed those variables which had null values in them particularly.

Not doing:

- Currently we are not predicting the class labels using the range set data.This means predicting salary for the group of people in the age range of 35-50 and working for 20-30 hours per week.
- Our project does not contain a predictive model which will classify given the key variable as parameter. This means that our project cannot classify the individual making >=50000$ or <=50000$ per year.

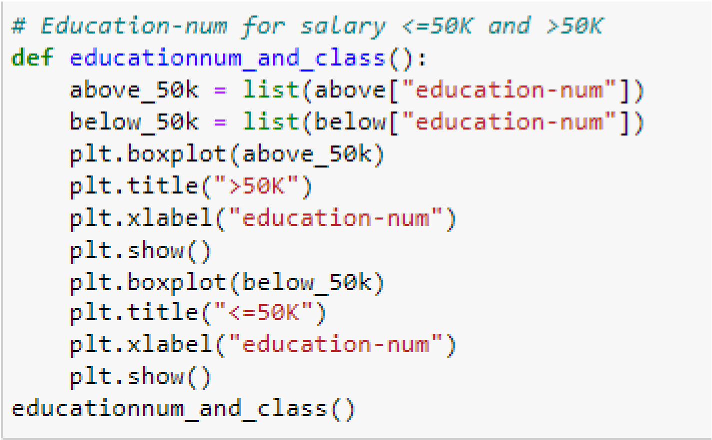
Appendix:

Data Preprocessing:



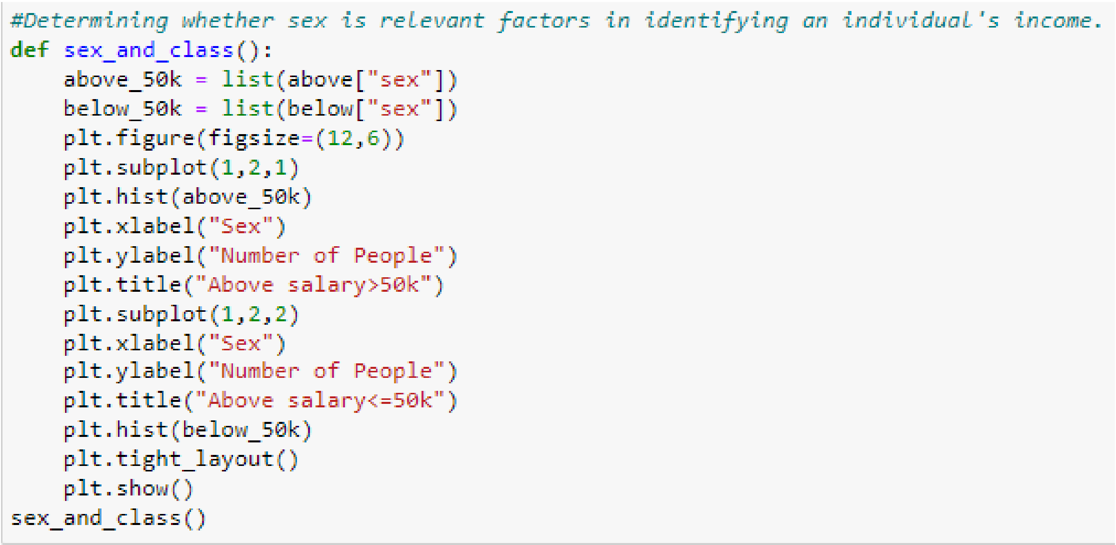
```
import pandas as pd
import numpy as np
from pandas.plotting import scatter_matrix
import seaborn as sns
from collections import Counter
import matplotlib.pyplot as plt
from statsmodels.graphics.mosaicplot import mosaic
%matplotlib inline

df = pd.read_csv("adult.data", header=None, sep=", ")
df.columns = ["age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation", "relationship",
              "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "class"]
print(df.isin(['?']).sum(axis=0))
df = df[df["workclass"] != '?']
df = df[df["occupation"] != '?']
df = df[df["native-country"] != '?']
below = df[df["class"] == "<=50K"]
above = df[df["class"] == ">50K"]
print("Count(Above 50K) = " + str(len(above.index)))
print("Count(Below 50K) = " + str(len(below.index)))
df.head()
```

User Story #1:



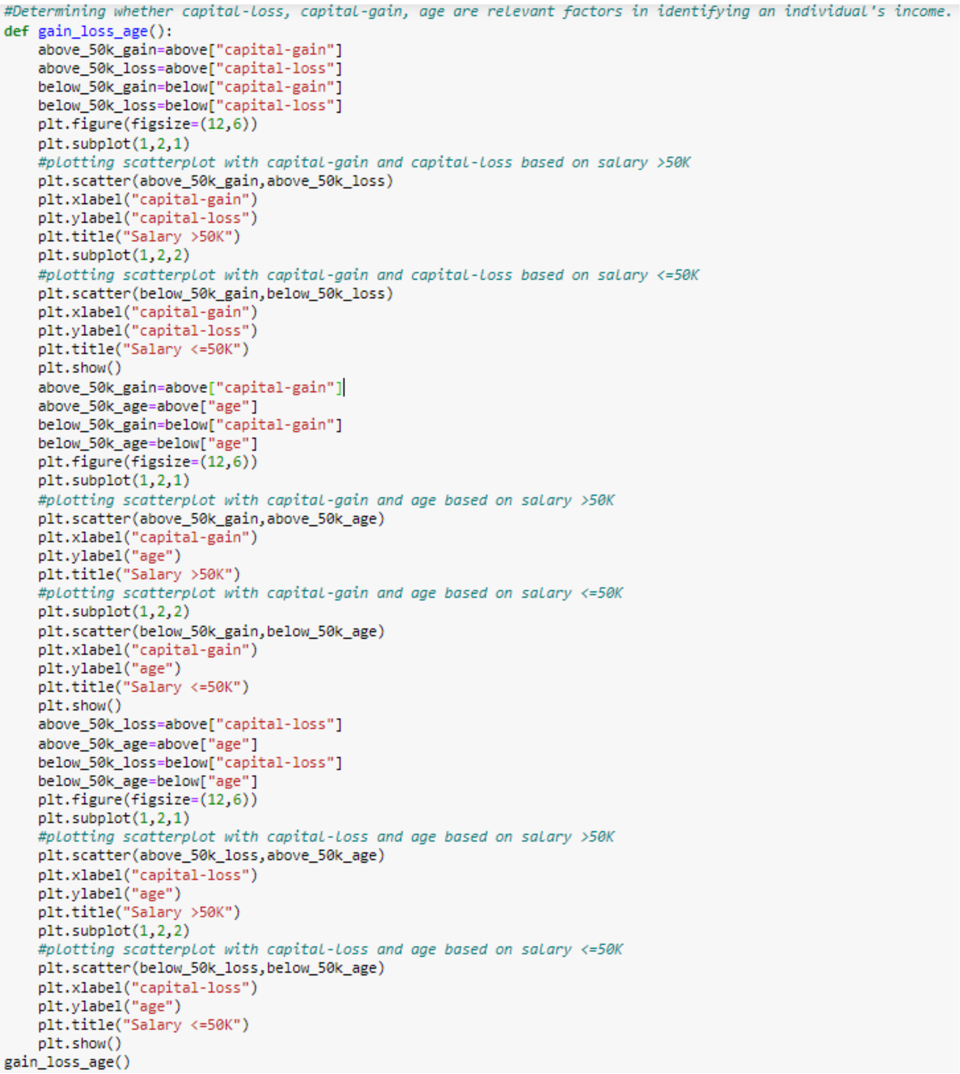
```
# Education-num for salary <=50K and >50K
def educationnum_and_class():
    above_50k = list(above["education-num"])
    below_50k = list(below["education-num"])
    plt.boxplot(above_50k)
    plt.title(">50K")
    plt.xlabel("education-num")
    plt.show()
    plt.boxplot(below_50k)
    plt.title("<=50K")
    plt.xlabel("education-num")
    plt.show()
educationnum_and_class()
```

User Story #2:



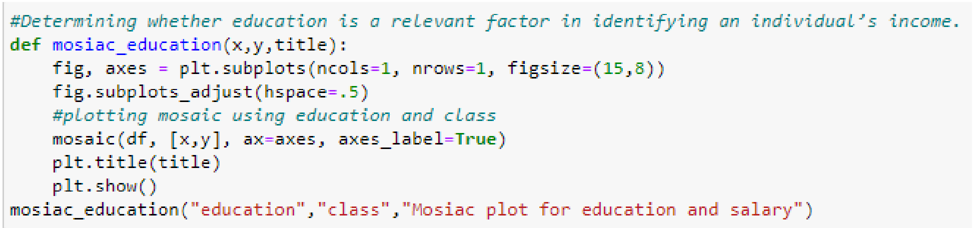
```
#Determining whether sex is relevant factors in identifying an individual's income.
def sex_and_class():
    above_50k = list(above["sex"])
    below_50k = list(below["sex"])
    plt.figure(figsize=(12,6))
    plt.subplot(1,2,1)
    plt.hist(above_50k)
    plt.xlabel("Sex")
    plt.ylabel("Number of People")
    plt.title("Above salary>50k")
    plt.subplot(1,2,2)
    plt.xlabel("Sex")
    plt.ylabel("Number of People")
    plt.title("Above salary<=50k")
    plt.hist(below_50k)
    plt.tight_layout()
    plt.show()
sex_and_class()
```

User Story #3:



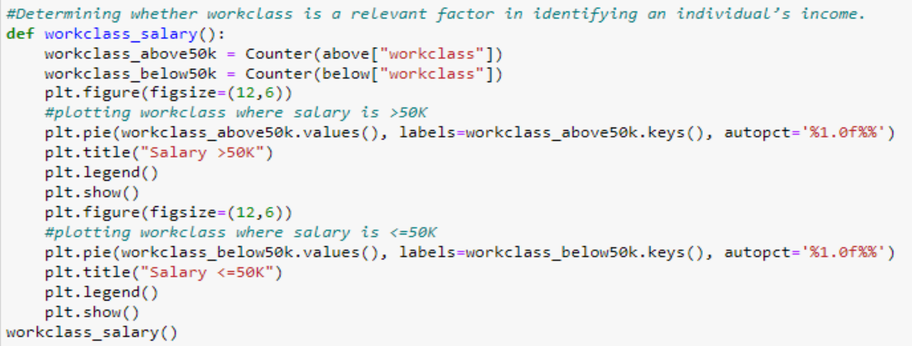
```
#Determining whether capital-loss, capital-gain, age are relevant factors in identifying an individual's income.
def gain_loss_age():
    above_50k_gain=above["capital-gain"]
    above_50k_loss=above["capital-loss"]
    below_50k_gain=below["capital-gain"]
    below_50k_loss=below["capital-loss"]
    plt.figure(figsize=(12,6))
    plt.subplot(1,2,1)
    #plotting scatterplot with capital-gain and capital-loss based on salary >50K
    plt.scatter(above_50k_gain,above_50k_loss)
    plt.xlabel("capital-gain")
    plt.ylabel("capital-loss")
    plt.title("Salary >50K")
    plt.subplot(1,2,2)
    #plotting scatterplot with capital-gain and capital-loss based on salary <=50K
    plt.scatter(below_50k_gain,below_50k_loss)
    plt.xlabel("capital-gain")
    plt.ylabel("capital-loss")
    plt.title("Salary <=50K")
    plt.show()
    above_50k_gain=above["capital-gain"]
    above_50k_age=above["age"]
    below_50k_gain=below["capital-gain"]
    below_50k_age=below["age"]
    plt.figure(figsize=(12,6))
    plt.subplot(1,2,1)
    #plotting scatterplot with capital-gain and age based on salary >50K
    plt.scatter(above_50k_gain,above_50k_age)
    plt.xlabel("capital-gain")
    plt.ylabel("age")
    plt.title("Salary >50K")
    #plotting scatterplot with capital-gain and age based on salary <=50K
    plt.subplot(1,2,2)
    plt.scatter(below_50k_gain,below_50k_age)
    plt.xlabel("capital-gain")
    plt.ylabel("age")
    plt.title("Salary <=50K")
    plt.show()
    above_50k_loss=above["capital-loss"]
    above_50k_age=above["age"]
    below_50k_loss=below["capital-loss"]
    below_50k_age=below["age"]
    plt.figure(figsize=(12,6))
    plt.subplot(1,2,1)
    #plotting scatterplot with capital-loss and age based on salary >50K
    plt.scatter(above_50k_loss,above_50k_age)
    plt.xlabel("capital-loss")
    plt.ylabel("age")
    plt.title("Salary >50K")
    plt.subplot(1,2,2)
    #plotting scatterplot with capital-loss and age based on salary <=50K
    plt.scatter(below_50k_loss,below_50k_age)
    plt.xlabel("capital-loss")
    plt.ylabel("age")
    plt.title("Salary <=50K")
    plt.show()
gain_loss_age()
```

User Story #4:



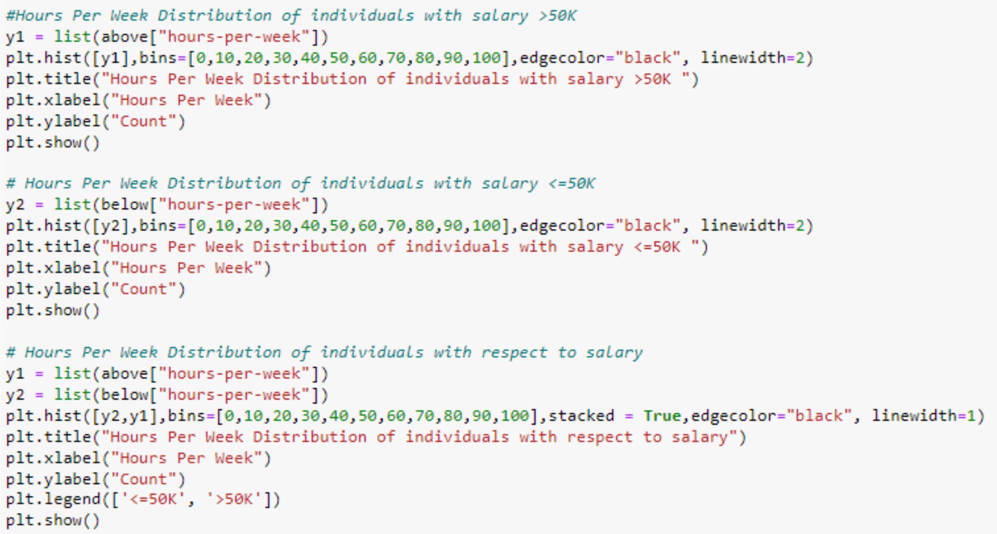
```
#Determining whether education is a relevant factor in identifying an individual’s income.
def mosiac_education(x,y,title):
    fig, axes = plt.subplots(ncols=1, nrows=1, figsize=(15,8))
    fig.subplots_adjust(hspace=.5)
    #plotting mosaic using education and class
    mosaic(df, [x,y], ax=axes, axes_label=True)
    plt.title(title)
    plt.show()
mosiac_education("education","class","Mosiac plot for education and salary")
```

User Story #5:



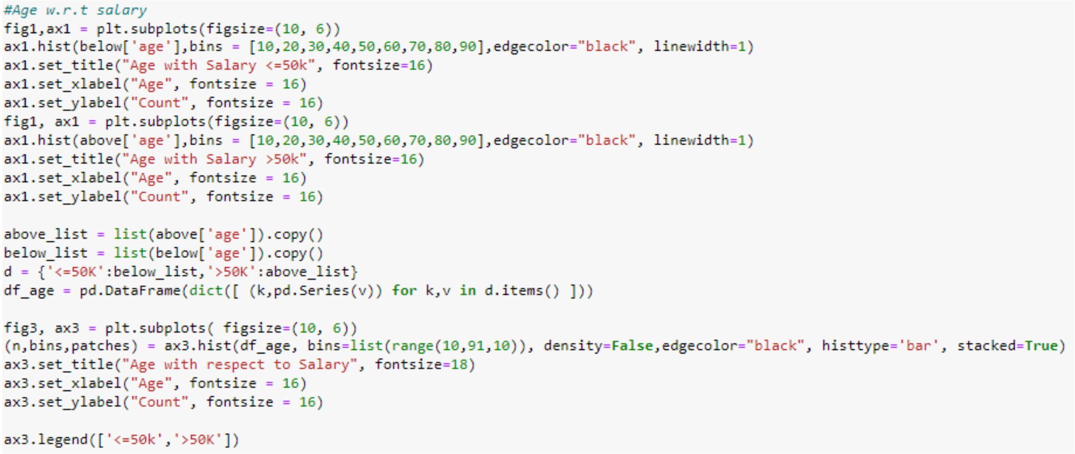
```
#Determining whether workclass is a relevant factor in identifying an individual’s income.
def workclass_salary():
    workclass_above50k = Counter(above["workclass"])
    workclass_below50k = Counter(below["workclass"])
    plt.figure(figsize=(12,6))
    #plotting workclass where salary is >50K
    plt.pie(workclass_above50k.values(), labels=workclass_above50k.keys(), autopct='%1.0f%%')
    plt.title("Salary >50K")
    plt.legend()
    plt.show()
    plt.figure(figsize=(12,6))
    #plotting workclass where salary is <=50K
    plt.pie(workclass_below50k.values(), labels=workclass_below50k.keys(), autopct='%1.0f%%')
    plt.title("Salary <=50K")
    plt.legend()
    plt.show()
workclass_salary()
```

User Story #6:



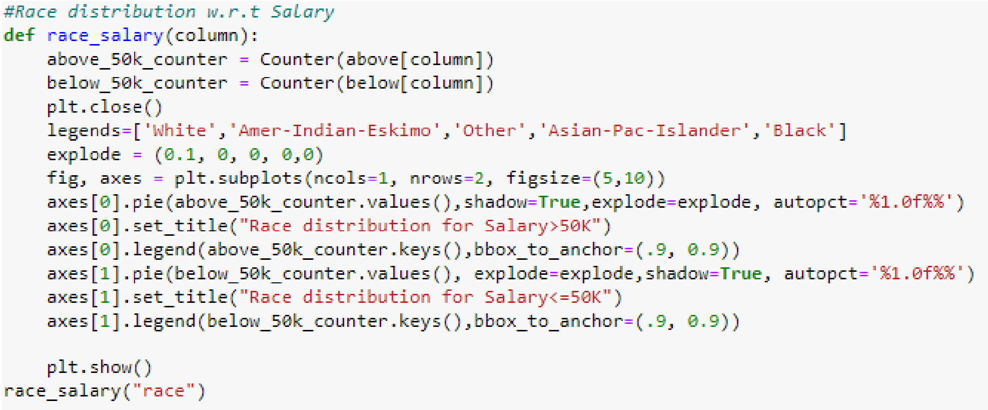
```
#Hours Per Week Distribution of individuals with salary >50K
y1 = list(above["hours-per-week"])
plt.hist([y1],bins=[0,10,20,30,40,50,60,70,80,90,100],edgecolor="black", linewidth=2)
plt.title("Hours Per Week Distribution of individuals with salary >50K ")
plt.xlabel("Hours Per Week")
plt.ylabel("Count")
plt.show()

# Hours Per Week Distribution of individuals with salary <=50K
y2 = list(below["hours-per-week"])
plt.hist([y2],bins=[0,10,20,30,40,50,60,70,80,90,100],edgecolor="black", linewidth=2)
plt.title("Hours Per Week Distribution of individuals with salary <=50K ")
plt.xlabel("Hours Per Week")
plt.ylabel("Count")
plt.show()

# Hours Per Week Distribution of individuals with respect to salary
y1 = list(above["hours-per-week"])
y2 = list(below["hours-per-week"])
plt.hist([y2,y1],bins=[0,10,20,30,40,50,60,70,80,90,100],stacked = True,edgecolor="black", linewidth=1)
plt.title("Hours Per Week Distribution of individuals with respect to salary")
plt.xlabel("Hours Per Week")
plt.ylabel("Count")
plt.legend(['<=50K', '>50K'])
plt.show()
```

User Story #7:



```
#Age w.r.t salary
fig1,ax1 = plt.subplots(figsize=(10, 6))
ax1.hist(below['age'],bins = [10,20,30,40,50,60,70,80,90],edgecolor="black", linewidth=1)
ax1.set_title("Age with Salary <=50k", fontsize=16)
ax1.set_xlabel("Age", fontsize = 16)
ax1.set_ylabel("Count", fontsize = 16)
fig1, ax1 = plt.subplots(figsize=(10, 6))
ax1.hist(above['age'],bins = [10,20,30,40,50,60,70,80,90],edgecolor="black", linewidth=1)
ax1.set_title("Age with Salary >50k", fontsize=16)
ax1.set_xlabel("Age", fontsize = 16)
ax1.set_ylabel("Count", fontsize = 16)

above_list = list(above['age']).copy()
below_list = list(below['age']).copy()
d = {'<=50K':below_list,'>50K':above_list}
df_age = pd.DataFrame(dict([ (k,pd.Series(v)) for k,v in d.items() ]))

fig3, ax3 = plt.subplots( figsize=(10, 6))
(n,bins,patches) = ax3.hist(df_age, bins=list(range(10,91,10)), density=False,edgecolor="black", histtype='bar', stacked=True)
ax3.set_title("Age with respect to Salary", fontsize=18)
ax3.set_xlabel("Age", fontsize = 16)
ax3.set_ylabel("Count", fontsize = 16)

ax3.legend(['<=50k','>50K'])
```

User Story #8:



```
#Race distribution w.r.t Salary
def race_salary(column):
    above_50k_counter = Counter(above[column])
    below_50k_counter = Counter(below[column])
    plt.close()
    legends=['White','Amer-Indian-Eskimo','Other','Asian-Pac-Islander','Black']
    explode = (0.1, 0, 0, 0,0)
    fig, axes = plt.subplots(ncols=1, nrows=2, figsize=(5,10))
    axes[0].pie(above_50k_counter.values(),shadow=True,explode=explode, autopct='%1.0f%%')
    axes[0].set_title("Race distribution for Salary>50K")
    axes[0].legend(above_50k_counter.keys(),bbox_to_anchor=(.9, 0.9))
    axes[1].pie(below_50k_counter.values(), explode=explode,shadow=True, autopct='%1.0f%%')
    axes[1].set_title("Race distribution for Salary<=50K")
    axes[1].legend(below_50k_counter.keys(),bbox_to_anchor=(.9, 0.9))

    plt.show()
race_salary("race")
```