Final project Report

Goals and objectives:

This project identifies and visualizes demographic factors that contribute to the likelihood of an individual's income. UVW college, the sponsor of this project, is looking to increase enrollment via targeting marketing using the outcome of this project. This project develops an application, and the users use this application to boost enrollment at the UVW college. The individual's income is the demographic for the marketing team of UVW for its degree programs. Whit this in mind, this project aims to use the data supplied by the USCB to create a marketing profile. Many variables come to play to determine an individual's income, including education, occupation, age, gender, marital status, and other features. The project provides the stakeholders with a summary report. The summary report details the characteristics that determine the individual's income. The summary report also explains the interrelation between different factors which drive the individual's income. Here, $50K is a crucial number this project uses as a distinguished individual's income earned.

Assumptions:

The project assumes that the provided data set is valid and complete. It also assumes that the dataset is a normal distribution of the characteristic of the average people. With that, the determination is that the dataset directly correlates with most of the population. Another assumption is that ethical or legal guidelines do not limit the dataset to restrict the result depending on the different features used in discrimination on certain population groups. This project is an academic project, and the project description was provided to guide the project along the way better. This project does not interview any UVW college marketing team members and makes assumptions about user stories. The project excludes Unknown values and features from the visualization.

User Stories:

User Story #1: Sex versus Salary:

The project would want to know if an individual's age determines their salary. If it is, then the project should include them in the tool.

User Story #2: Education versus Salary

The project would want to know if an individual's education determines their salary. If it is, then the project should include them in the tool.

User Story #3: Relationship and Salary

The project wants to see the relationship and the salary have the strong correlation or not. If it has, the project should include them in the tool.

User Story #4: Hours-per-week and Salary:

The project wants to understand a total work per week affects an individual's income in both men and women. If it affects, the project should include them in the tool.

User Story #5: Age versus Salary

The project would want to know if an individual's age determines their salary. If it is, then the project should include them in the prediction tool.

User Story #6: Marital Status versus Salary

The project also analyzes whether an individual's marital status determines their salary. If it is, then the project should include them in the prediction tool.

User Story #7: Race versus Salary

The project would want to examine the relationship between the salary and the native race. If it is, then the project should include them in the prediction tool.

User Story #8: Age and hours per week versus Salary

The project wants to understand how   an hours per week and age affects an individual's income. If it affects, the project should include them in the prediction tool.

From the stakeholders' perspective, the project analyzes the higher income probability among different occupations to decide whether to integrate into the prediction tool.

Visualizations:

Story #1: Sex and versus Salary

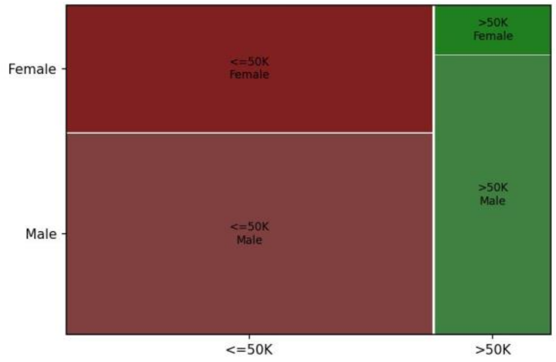


Fig:1 Mosaic plot of sex versus salary

Categorical variables sex and salary are displayed using the mosaic plot. This mosaic plot provides an overview of how the $50K income splits between males and females. The mosaic plot shows the relation in between sex and the salary of different individuals. From the chart, males earn more than females.

Visualization for Story #2: Education versus Salary

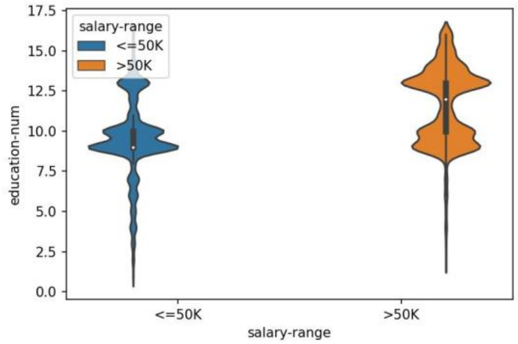


Fig:2 Violin Plots

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.

Visualization for Story #3: Relationship versus Salary

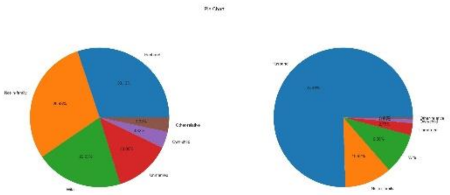


Fig:3 Pie Plot

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.

Story #4: Hours per week and Salary

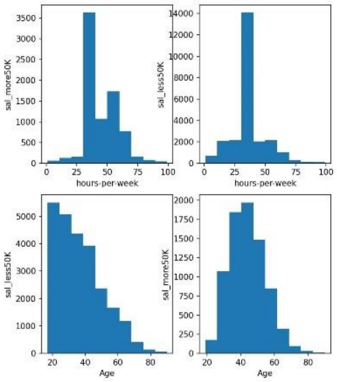


Fig. 4: Top Histogram plot

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 individual 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.

Story #5: Age versus Salary

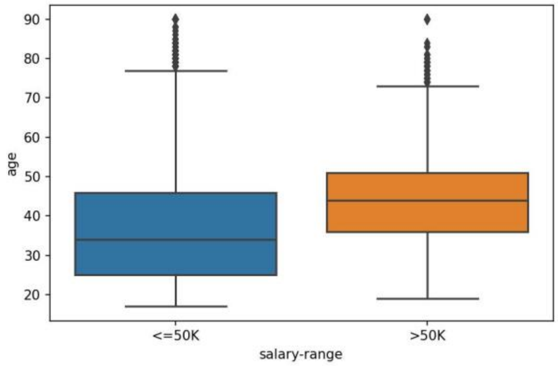


Fig: 5 Income Bracket split Box and Whisker Plot

This plot is used to compare datasets when the given dataset has a different sample size. Here, a box and whisker plot displays 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.

Visualization for Story #6: Marital status and Salary

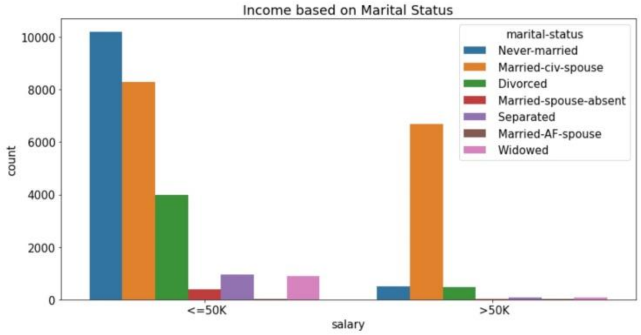


Fig:6 Marital status and salary

Based on the graph it can be said that whose income is 50k or below 50k they are either 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 50k .On the other hand, it can be also inferred that a significant chunk of data consists of people who are married -civ- spouse and have the higher chance of earning above 50k.

Visualization for Story #7: Race and Salary

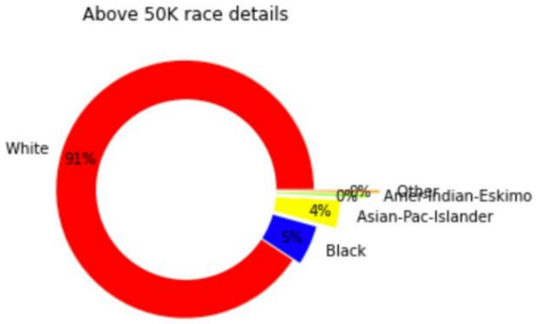


Fig:7(1) Salary above 50k

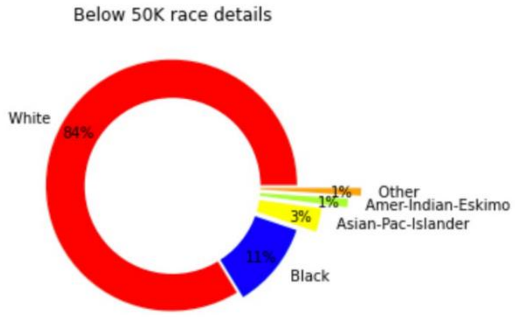


Fig:7(2) Salary below 50k

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.

Visualization for Story #8: Age and hours per week versus Salary

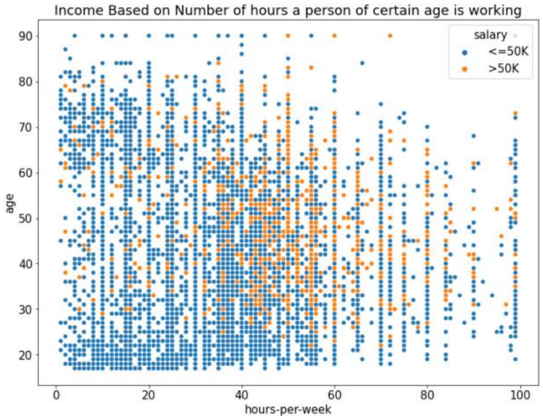


Fig:8 Hours per week and age versus Salary

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.

This graph does not show any correlation between age and working hours per week.

Question:

The initial work starts with the visualizations after making initial assumptions about the features necessary for this project. After that, the project determines which features are more important among the initially selected features. Although some of the features seem very relevant, they are excluded due to the lack of a spread-out dataset to predict the salary. Some wrong assumptions about some essential features lead to eliminating some of the visualizations. The project considers using some programming languages for visualization, such as

Since all the assignments are on Python in this class, the project uses them as visualization tools. Also, Python is more flexible and dynamic with having any library used for data visualization. To provide more flexibility and customization project uses Python for Matplot, seaborn, and state models. To discover the significance of different features of the dataset Pie chart is more useful than other charts for univariate analysis of categorical data. Mosaic plots are more convenient for multivariate analysis to compare different features. Histograms and box-whisker plots are suitable for continuous features of the dataset.

Not Doing:

The project plans to implement an interactive visualization as a future scope that displays more information about the dataset It would help the users understand the features' significance by analyzing the visualization only. The project should also research the performance of the new data set when new data samples add to the existing dataset. This is important because the given dataset has asymmetric instances in different income levels.

The United States datasets are more fulfilled than the datasets of the other countries. It also should concede that individuals in the United States earn more than individuals from the others due to their economic status.

The project should gather more information and feedback from the users while reviewing the executive report and add them to the next volume.

Appendix:

Story #1: Mosaic plot

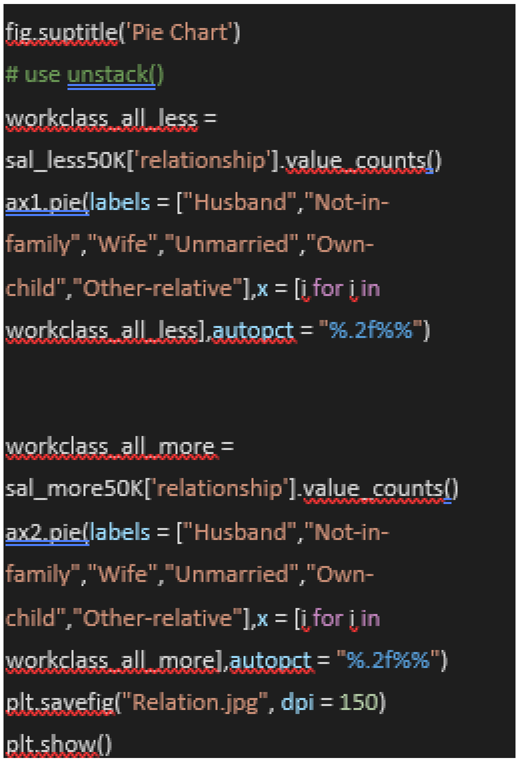
```
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
from statsmodels.graphics.mosaicplot import
mosaic

plt.rcParams["figure.figsize"] = [7.00, 3.50]
plt.rcParams["figure.autolayout"] = True

df = pd.read_csv("../input/adult-census-
income/adult.csv", sep=",")

mosaic(data=df, index=['salary-range', 'sex'])
plt.savefig("Mosaic.jpg", dpi=150)
plt.show()
```

Story#2: -Violin Plots

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import warnings
warnings.filterwarnings("ignore")

df = pd.read_csv("../input/adult-census-
income/adult.csv", sep=",")

sns.violinplot(df["salary-range"],
y=df["education-num"], hue = df["salary-
range"],
box=True,
points='all'
)

plt.savefig("violiPlot.jpg", dpi=150)

plt.show()
```

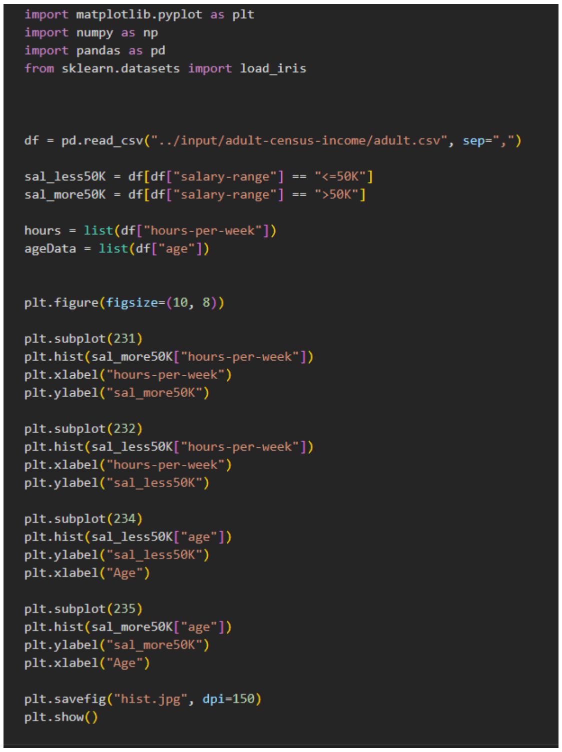
Story #3-Pie plot

```
import warnings
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

warnings.filterwarnings("ignore")
plt.style.use('seaborn-whitegrid')
%matplotlib inline




df = pd.read_csv("../input/adult-census-
income/adult.csv", sep=",")
sal_less50K = df[df["salary-range"] == "<=50K"]
sal_more50K = df[df["salary-range"] == ">50K"]

labels = list(sal_more50K.workclass.unique())
# plot data
fig, (ax1, ax2) = plt.subplots(nrows=1, ncols = 2,
figsize = (20,8))
```



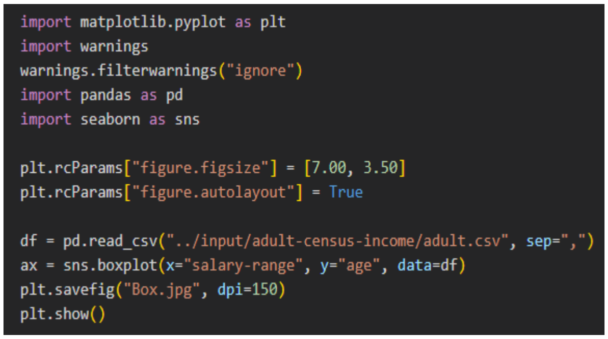
```
fig.suptitle('Pie Chart')
# use unstack()
workclass_all_less =
sal_less50K['relationship'].value_counts()
ax1.pie(labels = ["Husband","Not-in-
family","Wife","Unmarried","Own-
child","Other-relative"],x = [i for i in
workclass_all_less],autopct = "%.2f%%")


workclass_all_more =
sal_more50K['relationship'].value_counts()
ax2.pie(labels = ["Husband","Not-in-
family","Wife","Unmarried","Own-
child","Other-relative"],x = [i for i in
workclass_all_more],autopct = "%.2f%%")
plt.savefig("Relation.jpg", dpi = 150)
plt.show()
```

Story#4: -Histogram



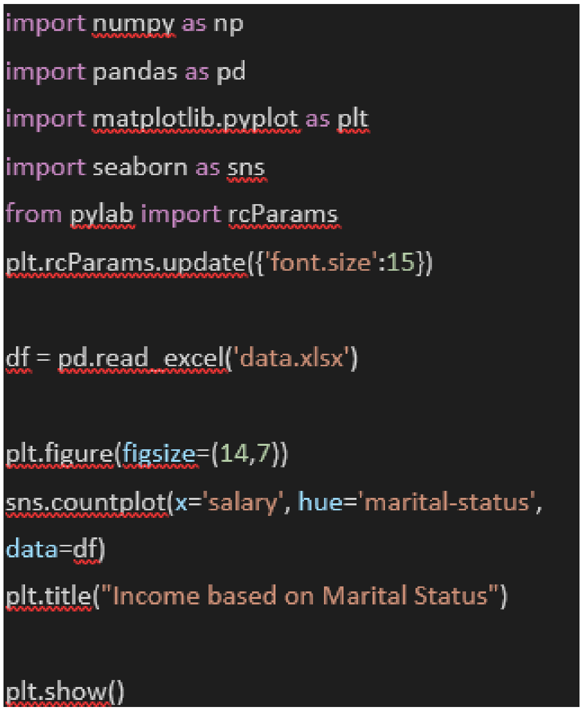
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.datasets import load_iris



df = pd.read_csv("../input/adult-census-income/adult.csv", sep=",")

sal_less50K = df[df["salary-range"] == "<=50K"]
sal_more50K = df[df["salary-range"] == ">50K"]

hours = list(df["hours-per-week"])
ageData = list(df["age"])


plt.figure(figsize=(10, 8))

plt.subplot(231)
plt.hist(sal_more50K["hours-per-week"])
plt.xlabel("hours-per-week")
plt.ylabel("sal_more50K")

plt.subplot(232)
plt.hist(sal_less50K["hours-per-week"])
plt.xlabel("hours-per-week")
plt.ylabel("sal_less50K")

plt.subplot(234)
plt.hist(sal_less50K["age"])
plt.ylabel("sal_less50K")
plt.xlabel("Age")

plt.subplot(235)
plt.hist(sal_more50K["age"])
plt.ylabel("sal_more50K")
plt.xlabel("Age")

plt.savefig("hist.jpg", dpi=150)
plt.show()
```

Story #5 -Box and Whisker Plot.



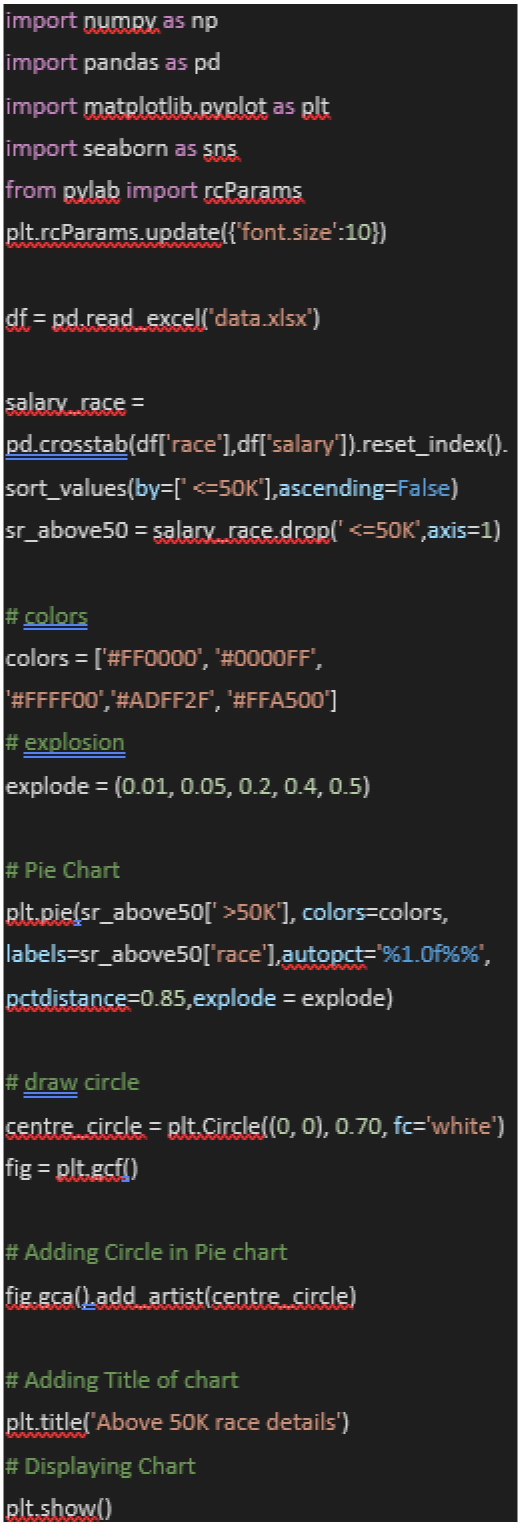
```
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import seaborn as sns

plt.rcParams["figure.figsize"] = [7.00, 3.50]
plt.rcParams["figure.autolayout"] = True

df = pd.read_csv("../input/adult-census-income/adult.csv", sep=",")
ax = sns.boxplot(x="salary-range", y="age", data=df)
plt.savefig("Box.jpg", dpi=150)
plt.show()
```

Story #6: -Bar graph.

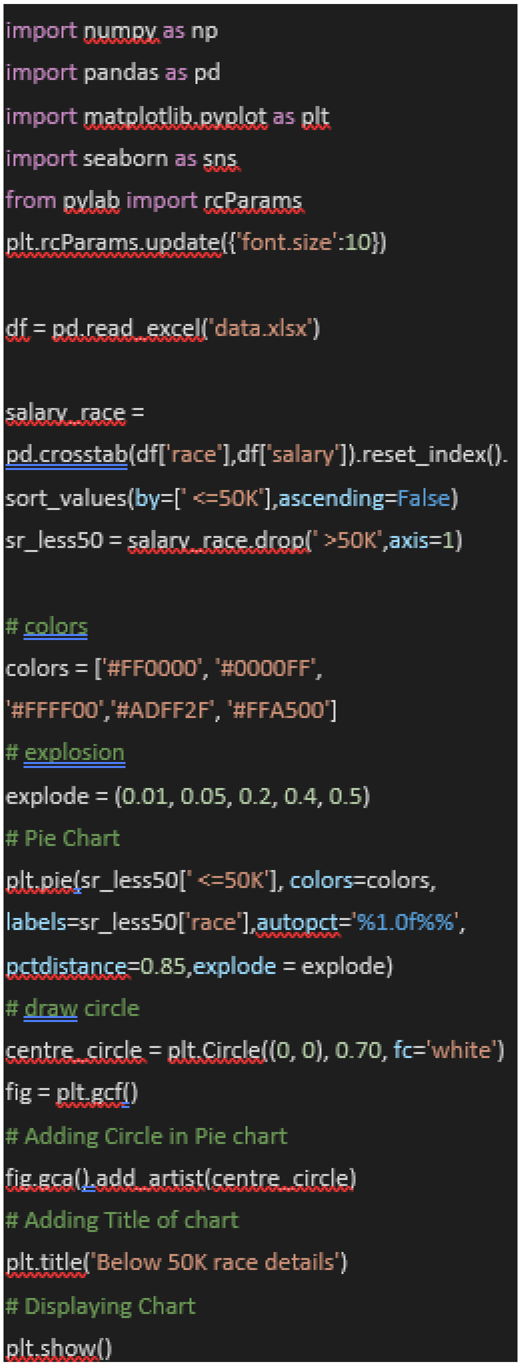


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pylab import rcParams
plt.rcParams.update({'font.size':15})

df = pd.read_excel('data.xlsx')

plt.figure(figsize=(14,7))
sns.countplot(x='salary', hue='marital-status',
data=df)
plt.title("Income based on Marital Status")

plt.show()
```

Story #7 -Donut chart (Above 50k and Below 50k)

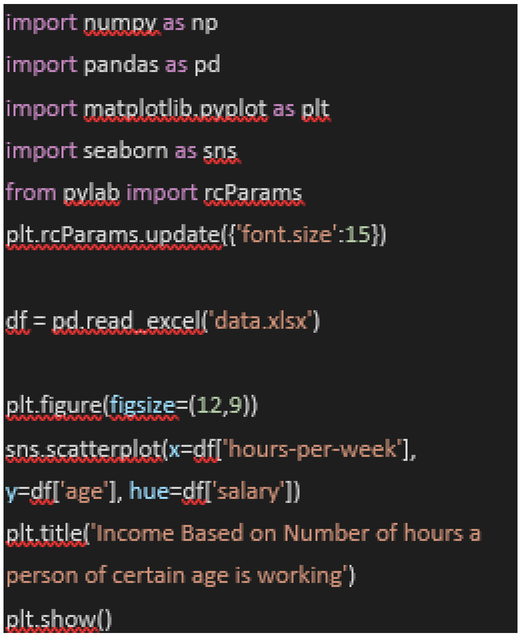


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pylab import rcParams
plt.rcParams.update({'font.size':10})

df = pd.read_excel('data.xlsx')

salary_race =
pd.crosstab(df['race'],df['salary']).reset_index().
sort_values(by=[' <=50K'],ascending=False)
sr_above50 = salary_race.drop(' <=50K',axis=1)

# colors
colors = ['#FF0000', '#0000FF',
'#FFFF00','#ADFF2F', '#FFA500']
# explosion
explode = (0.01, 0.05, 0.2, 0.4, 0.5)

# Pie Chart
plt.pie(sr_above50[' >50K'], colors=colors,
labels=sr_above50['race'],autopct='%1.0f%%',
pctdistance=0.85,explode = explode)

# draw circle
centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()

# Adding Circle in Pie chart
fig.gca().add_artist(centre_circle)

# Adding Title of chart
plt.title('Above 50K race details')
# Displaying Chart
plt.show()
```



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pylab import rcParams
plt.rcParams.update({'font.size':10})

df = pd.read_excel('data.xlsx')

salary_race =
pd.crosstab(df['race'],df['salary']).reset_index().
sort_values(by=[' <=50K'],ascending=False)
sr_less50 = salary_race.drop(' >50K',axis=1)

# colors
colors = ['#FF0000', '#0000FF',
'#FFFF00','#ADFF2F', '#FFA500']
# explosion
explode = (0.01, 0.05, 0.2, 0.4, 0.5)
# Pie Chart
plt.pie(sr_less50[' <=50K'], colors=colors,
labels=sr_less50['race'],autopct='%1.0f%%',
pctdistance=0.85,explode = explode)
# draw circle
centre_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()
# Adding Circle in Pie chart
fig.gca().add_artist(centre_circle)
# Adding Title of chart
plt.title('Below 50K race details')
# Displaying Chart
plt.show()
```

Story #8: -Scatter Plot.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from pylab import rcParams
plt.rcParams.update({'font.size':15})

df = pd.read_excel('data.xlsx')

plt.figure(figsize=(12,9))
sns.scatterplot(x=df['hours-per-week'],
y=df['age'], hue=df['salary'])
plt.title('Income Based on Number of hours a
person of certain age is working')
plt.show()
```