Course Project Final Report

Goals and Business Objective:

The goal is to make some marketing profiles through finding key factors that determine the individuals' income and figuring out the relations between different factors on the income.

Business objective: based on the above marketing profiles, UVW College can develop and implement new marketing campaigns, so it can improve their enrollment.

Assumptions:

The team has chosen a salary of $50K as a key element to determine the criteria for developing marketing campaigns. We assume that the key element is the most important factor in developing an effective marketing campaign.

The team is using the data set from 1994 United States Census Bureau. We assume that the data set is accurate and legitimate.

The team has acknowledged that many key variables that must be assessed for individuals making less than and more than $50K, including age, sex, education, marital status, occupation, etc.

Among the key variables that affect the individuals' income, we choose 8 features to further analysis. We assume that the chosen 8 features contribute the most to earning more than and less than $50K income.

User Stories:

User story #1: the marketing team would like to know the relationship between race and income. More specifically, the team wants to know the percentage of people whose income over 50k on each race.

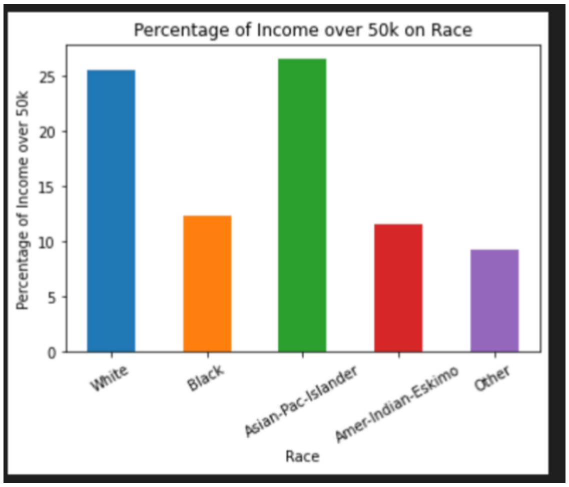
User story #2: the marketing team would like to explore the combination of age, education numbers and hours per week's impact on income.

User story #3: the marketing team wants to know the sex and marital status' impact on income. For the groups of people whose income is above 50K and below 50K respectively, how is the sex and marital status distribution?

User story #4: the marketing team wants to know the capital gain and age have any bearing on income.

User story #5: the marketing team would like to know the correlation between native country and income.

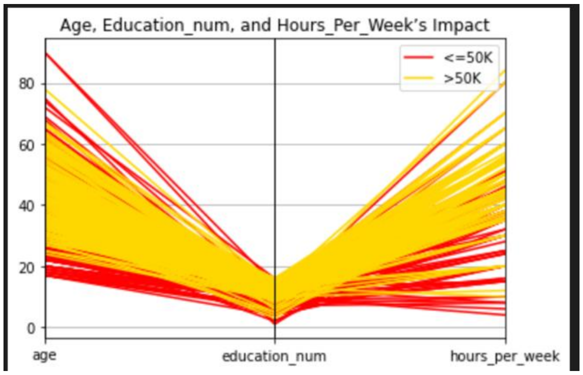
Visualization:



Graph1: for user story #1, I chose to use bar chart to visualize the data. The major reason is that bar chart is useful for showing segment information. Vertical bar charts are good for comparing different categorical data. Race is categorical data. Through the bar chart, we can compare the percentage of people whose income over 50k on each race.

I set the x axis represents the races and the y axis represents the percentage of income over 50K. Firstly, I created two dictionaries count1 and count2. Count1 is the sum of people on each race. Count2 is the sum of people whose income is over 50K on each race. The key to both dictionaries is race. Then, I created a dictionary to represent the percentage. Loop the count1 dictionary, initialize the percentage to 0, and calculating percentage on each race as count2[r] * 100 / count1[r]. In the loop, I also plot the bar chart.

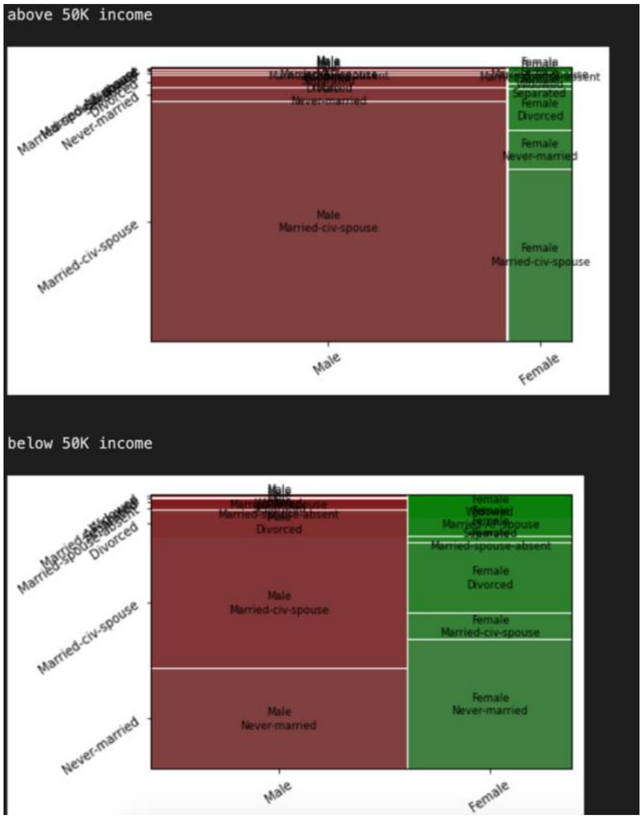
We can see in the bar chart that race White (25%) and Asian-Pac-Islander(a bit over 25%) have the similar percentage. This means that among White people, 25% White earn income over 50K, and among Asian-Pac-Islander, 26% Asian-Pac-Islander earn income over 50K. For Black and Amer-Indian-Eskimo, the percentage is between 10% and 15%. Other's percentage is near to 10%. Given that, White and Asian-Pac-Islander people have the highest percentage of people who earn more than 50K, comparing with other races. Race is a reliable indicator for predicting the income of individual. It can help the marketing team to develop different marketing strategies on different race.



Graph2: I chose to use parallel coordinate plot. The major reason is that parallel coordinate plot is ideal for comparing many variables together and identifying the trends from the charts. In user story #2, I would like to explore the factors of age, education numbers and hours per week's impact on income. I want to plot the pattern through using the three variables together. Parallel coordinate is a good choice.

First, I created a new column df['larger_section']. The 'larger_section' column shows 1 if the income above 50k and 0 if the income is below or equal to 50k. For the parallel coordinate frame (pc frame), df['larger_section'] = pc_frame['larger_section']. I divided it into two parts, pc_above_50K and pc_below_50K using the pc_frame['larger_section'] == 0(below 50K) and pc_frame['larger_section'] == 1(above 50K). Since the original data is too large, I chose 200 samples to better illustrate the pattern. I got the 200 samples through pc_frame[].sample(n = 200). Then, I concat the pc_above_50K and pc_below_50K together as pc_frame.

In the chart, yellow line represents the income above 50K, and red line shows the income below or equal to 50K. The pattern we can find is that most yellow line lie in the range that age between 25-60, education number is above 10 years and hours_per_week is around 40-60 hours. This pattern can help the marketing team in crafting strategies to bolster program enrollment.

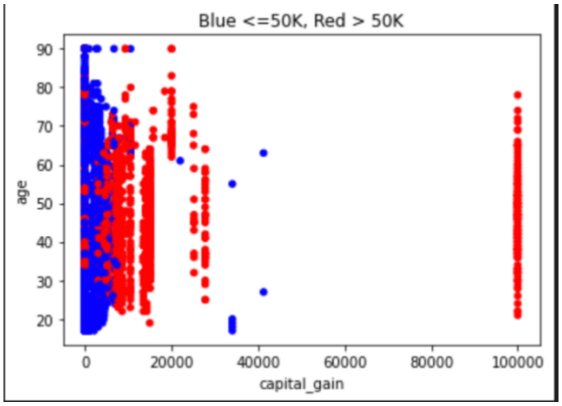

above 50K income

below 50K income

Graph3: I used mosaic plot to show the relationship of sex and marital status on income. Mosaic plot can show the correlations between distinct variables. Mosaic plot is good for nominal data's visualization. In this user story, sex and marital status are nominal data, so mosaic plot will be helpful for visualizing the relationship.

I define a method as mosaic_plot_class() with two variables col1 and col2. In this user story, when I call this method, I pass 'sex' and 'marital_status' into it. It produces mosaic plot presenting the relationship of sex and marital status. To better illustrate the relations, I created two mosaic plots. One shows the group of people whose income is above 50K, and another represents the group who earn below or equal to 50K. Therefore, I created two data frame: below_50K and above_50K, based on their income. When creating mosaic plot, pass corresponding data frame to the mosaic(). For example, for above 50K chart, we pass above_50K data into mosaic(). X axis is col1 and y axis is col2. In this case, col1 is sex and col2 is marital status.

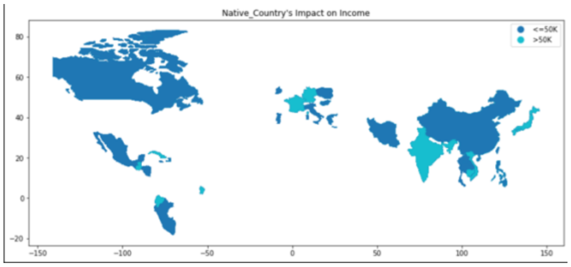
From the graph, we can tell that in the group above 50K, the largest amount of people is males who are married-civ-spouse, and the second largest is females who are married-civ-spouse. From gender's perspective, in the male group, married-civ-spouse people are far more than the other marital status people. However, in female group, married-civ-spouse people are the largest proportion, and never married and divorced have the similar proportion. This means that three types of marital status females earn more than 50K but only one type marital status males have such income. What's more, in the group below 50K, never married males and married males occupy most proportion of males who earn income below 50K. Never married and divorced females whose income are below 50K. In both gender group, never married status people who earn income below 50K. Both charts can provide useful information for the marketing team and help them initialize appropriate marketing campaigns.



Graph4, I chose scatter plot to represent the capital gain and age's impact on income. Scatter plot uses dots to show the values for two different numeric variables. In this user story, both capital gain and age are numeric variables, so scatter plot is a good way to visualize the relation.

I started with creating a scatter plot frame sp_frame with columns 'capital_gain', 'age' and 'income'. There is two different colors of dots in the chart and each color represent one income group. I used the df['larger_section'] I created before to make a color map(colors) through df['larger_section'].map(). In the map(), I set '0' as blue and '1' as red. This means, income below or equal to 50K is blue and income above 50K is red. Then I plot the scatter chart and set x axis is 'capital gain', y axis is 'age' and color is colors(the color map).

From the graph, we can see that most blue dots are in the range of capital gain far below 20k and around 10k-ish, and blue dots are in all age range. Red dots are distributed between 10k to 20k capital gain and 20-60 years old range. Some red dots are in the capital gain around 30k and on the 100K. This chart shows us that less capital gain indicates less income. In certain capital gain range(like 20k, 30k and 100k), more people earn income above 50K. This shall be useful for the marketing team.



Graph5, I used global map to represent the relation between native country and income. Since the variable is native country, a map is intuitively the best way to show the relation.

I imported geopandas to create global map. I first read data file ‘naturalearth_lowers’ as gdf to create a base map. This file includes all countries name and latitude and longitude information. Then, I get map_frame from data set df[] including columns ‘native_country’ and ‘income’. I merged the gdf and map_frame on the right(which is map_frame data) through common column (‘name’ on gdf and ‘native_country’ on map_frame). Lastly, I plot the merged map(merged_country) on the column ‘income’. Legend shows that dark blue is income below or equal to 50K and bright blue is income above 50K. From the graph, we can easily tell that which country whose income is above or below 50K. Like, China and USA are dark blue meaning that their income is below 50K. The marketing team can develop strategies to target people whose native country are from China and USA.

Questions:

1, Are all features contributing to determining individuals’ income?

Intuitively, we can come up with some factors related to the income, such as age, sex, education, hours per week. For example, we can imagine that long working hours per week shall result in more income. Some factors such as capital gain, marital status, may look like not directly relate to the income. Therefore, we can combine different factors and figure out their impact on the income. For example, I combine capital gain with age together and it tells us that different income groups lie in different capital gain and age range.

2, How can I make up user stories?

After choosing the eight features (race, age, education numbers, capital gain, hours per week, sex, marital status, and native country), I am thinking the user stories. I came up with two user stories using single factors (race’s impact on income and native country’s impact on income), and three stories using combination of different factors. For the combinations, numerical data can be put together and categorical data can be combined. In the following question, I will answer what the numerical data is and what the categorical data is.

3, How to choose appropriate visualization tools to represent the relations?

The eight features have two data type: numerical data and categorical data. Numerical data includes age, education numbers, capital gain, and hours per week. We can use parallel coordinate chart, scatter plot to represent numerical data. Categorical data includes race, age, sex, marital status, and native country. Bar chart and mosaic plot(multivariant plot) are good for categorical data. I use map to visualize native country because only native country feature applies to map tool.

4, Are the user story validated through charts?

Yes. I carefully choose appropriate visualization tools to show the data relations, so all user stories have been examined and validated. The conclusions can be helpful for the marketing team to develop marketing campaigns.

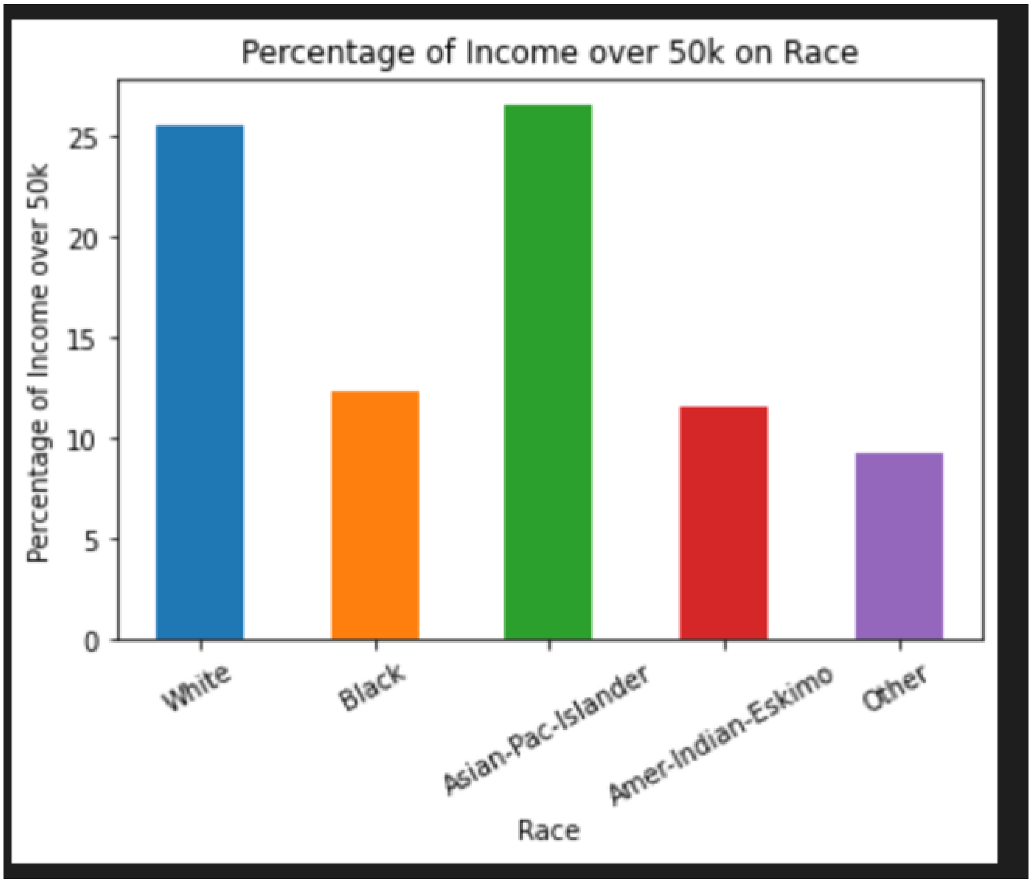
Not Doing:

I have only examined eight features’ impact on the income. We can explore all the 14 features in the future, so as to figure out all the factors’ relations and effects on the income. This can be helpful for targeting the specific audience for the program and bolster the enrollment.

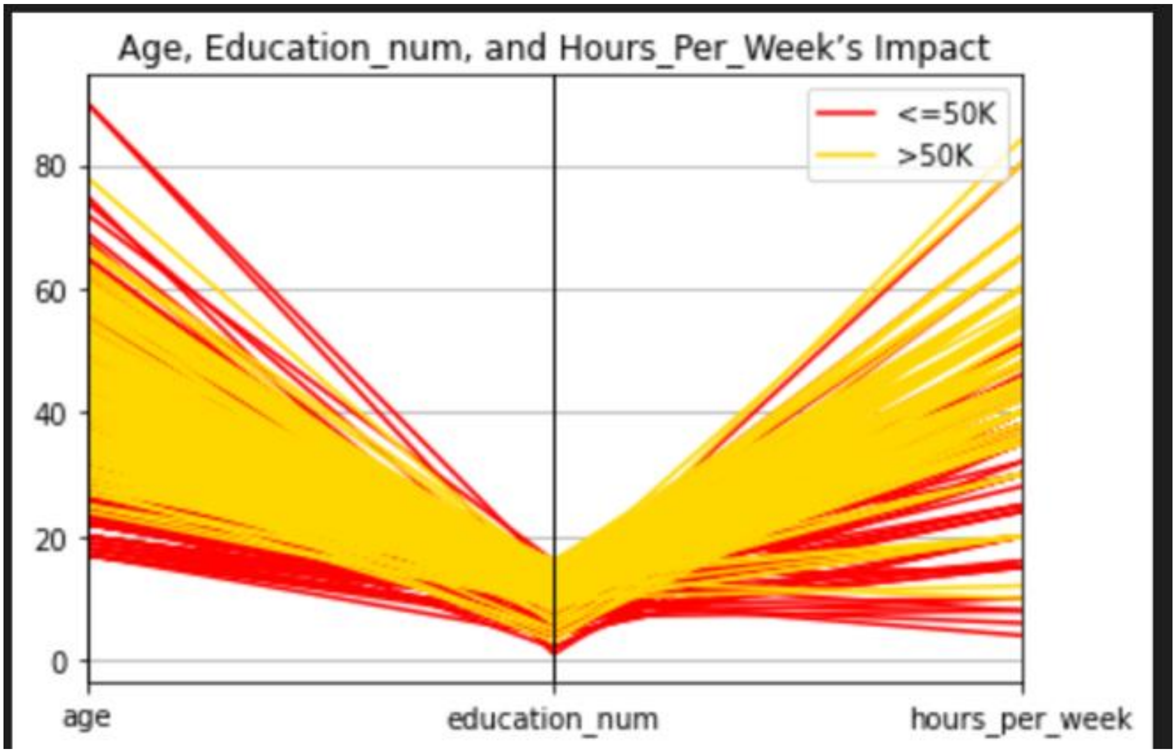
I have made up profiles that show the 8 factors impact on the income, but I didn't develop models to predict the income. In the future, we can use the patterns we have found out to train the data and develop models for predicting.

Appendix

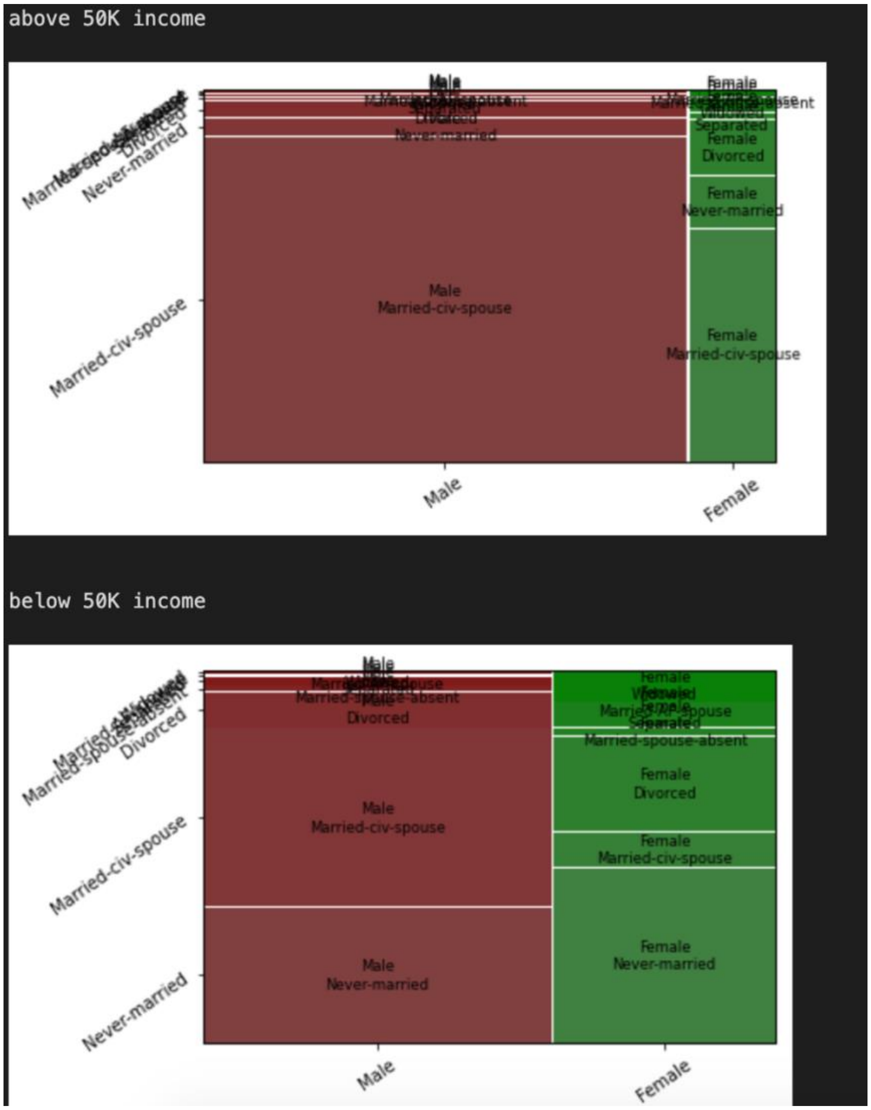
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from statsmodels.graphics.mosaicplot import mosaic
%matplotlib inline
from pandas.plotting import parallel_coordinates
import geopandas as gpd

col_names = ['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 = col_names, header = None)

#remove question mark
for col in col_names:
    if not isinstance(df[col][0],np.int64):
        df[col] = df[col].apply(lambda x : x.strip() if x.strip() != '?' else None)


# above 50k income is larger_section
df['larger_section'] = (df['income'] == '>50K') * 1

# scenario1: race impact on income
races = df['race'].values
sum = len(races)

count1 = {} # count the sum of each race
count2 = {} # count the num of over 50k on each race
for i in range(sum):
    if not races[i] in count1:
        count1[races[i]] = 0
        count2[races[i]] = 0

    count1[races[i]] = count1[races[i]] + 1

    if df['income'][i] == '>50K':
        count2[races[i]] = count2[races[i]] + 1


percentage = {}
for r in count1.keys():
    if not r in percentage:
        percentage[r] = 0
    percentage[r] = count2[r] * 100 / count1[r]
    # print(percentage[r])
    plt.bar(r, percentage[r], width= 0.5)

plt.xticks(rotation = 30)
plt.xlabel('Race')
plt.ylabel('Percentage of Income over 50k')
plt.title('Percentage of Income over 50k on Race')

plt.show()
```



```
#scenario2: combination of age, education_num, and hours_per_week’s impact on income
plt.close()

pc_frame = df[['age', 'education_num', 'hours_per_week']].copy()

pc_frame['income'] = df['income']
pc_frame['larger_section'] = df['larger_section']
# print(pc_frame['larger_section'])
pc_frame.columns = ['age', 'education_num', 'hours_per_week', 'income', 'larger_section']

pc_below_50K = pc_frame[pc_frame['larger_section'] == 0].sample(n = 200)
pc_above_50K = pc_frame[pc_frame['larger_section'] == 1].sample(n = 200)
pc_frame = pd.concat([pc_below_50K, pc_above_50K])

parallel_coordinates(pc_frame, class_column = 'income', cols = ['age', 'education_num', 'hours_per_week'],
color=('#FF0000','#FFD700'))
plt.title("Age, Education_num, and Hours_Per_Week’s Impact")
plt.show()
```

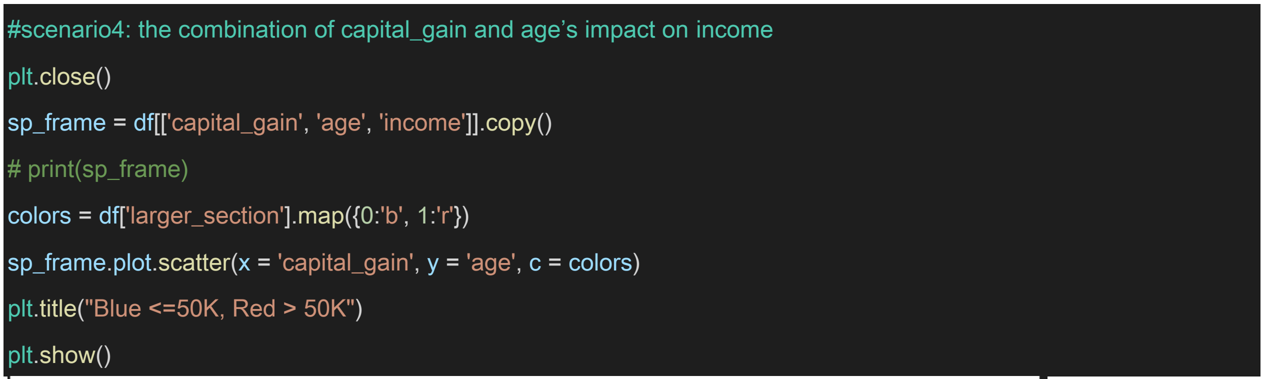


```
#scenario3: the combination of sex and marital_status's impact on income
def mosaic_plot_class(col1, col2):
    plt.close()
    below_50K = df[df['income'] == '<=50K']
    above_50K = df[df['income'] == '>50K']

    print("above 50K income")
    mosaic(above_50K, [col1, col2], label_rotation = 35)
    plt.show()

    print("below 50K income")
    mosaic(below_50K, [col1, col2], label_rotation = 35)
    plt.show()

mosaic_plot_class('sex', 'marital_status')
```

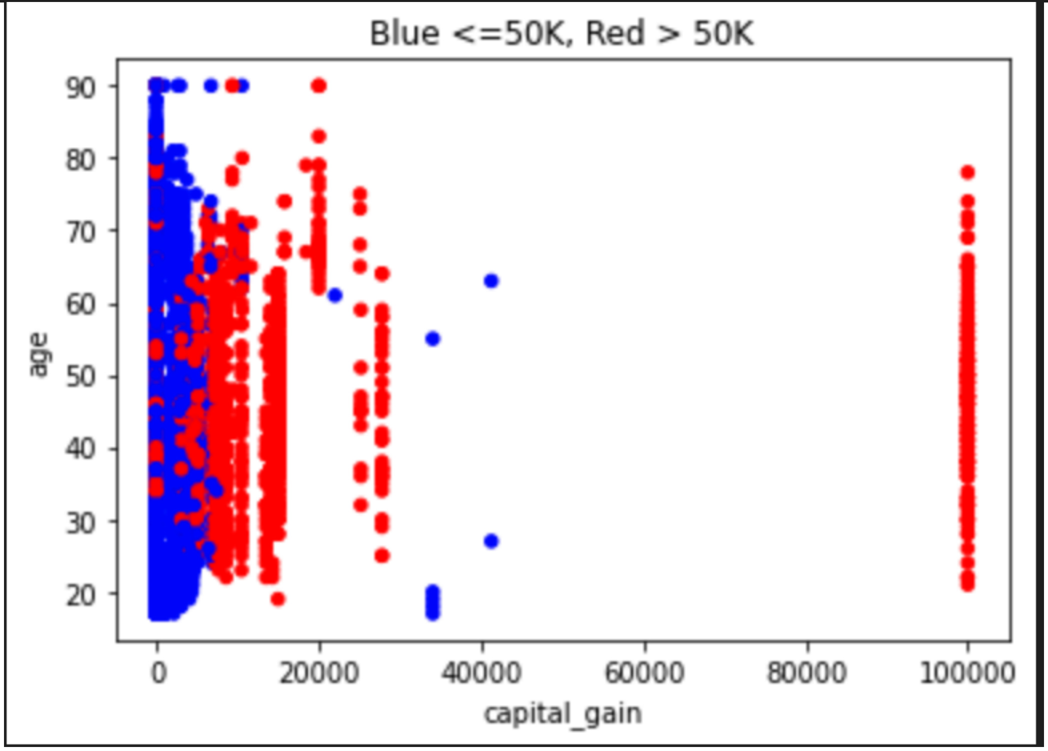


```
above 50K income
```

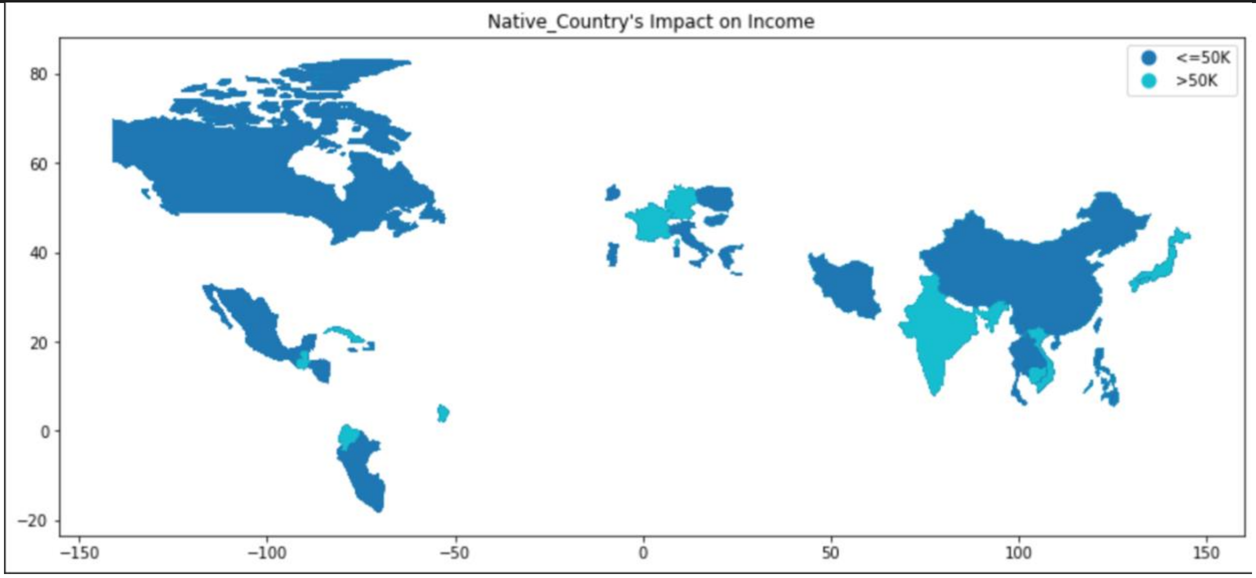
```
below 50K income
```



```
#scenario4: the combination of capital_gain and age’s impact on income
plt.close()
sp_frame = df[['capital_gain', 'age', 'income']].copy()
# print(sp_frame)
colors = df['larger_section'].map({0:'b', 1:'r'})
sp_frame.plot.scatter(x = 'capital_gain', y = 'age', c = colors)
plt.title("Blue <=50K, Red > 50K")
plt.show()
```



```
#Scenario 5: the native_country's impact on income
plt.close()
gdf = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
gdf.head()
# print(gdf)
map_frame = df[['native_country', 'income']].copy()
# print(country)
merged_country = gdf.merge(map_frame, left_on = 'name', right_on = 'native_country', how = 'right')
# print(merged_country)
plt.rcParams["figure.figsize"] = (25,6)
merged_country.plot(column = 'income', legend = True)
plt.title("Native_Country's Impact on Income")
plt.show()
```


Native_Country's Impact on Income
<=50K
>50K
80
60
40
20
0
−20
−150
−100
−50
0
50
100
150