Lucy Zhang

ID:1229576766

02/20/2023

Course Project Final Report

Goals and a business objective:

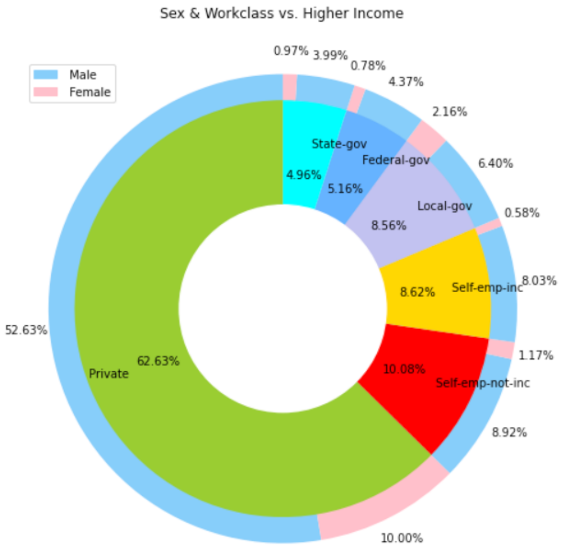
As a data analyst at XYZ company, I have been given a task to bolster enrollment at UVW College by using the United States Census Bureau data and focusing on a $50,000 salary as a key demographic to create data visualizations. In this project, I will review all the features in the data to determine the key factors for higher and lower income, and I will visualize them with graphs accordingly.

● Assumptions:

Based on UVW college's request to bolster enrollment, I assumed that making a profit from enrollment is not considered a priority. By choosing a $50,000 salary as a key demographic, I assumed that enrollment only targeted employed individuals. Also, UVW would like to develop an application to find the factors that determine the individual's income so that the application predicts the income of an individual, which can target a specific group of people for marketing purposes in the future. Lastly, I assumed that the United States Census Bureau data provided by the XYZ company is a biased extract from the 1994 US Census database.

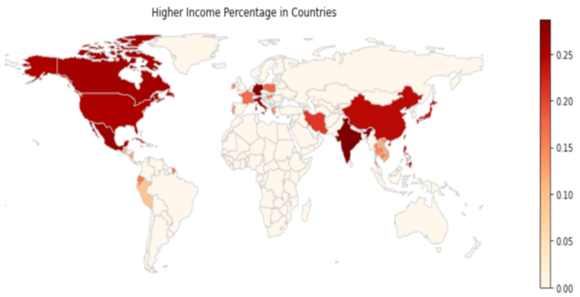
● User Stories:

1. Sex diversity in school and the workplace has always been an important topic for society today. From the visualization of Sex & Workclass vs. Salary observation, the school can market to females with different approaches, such as offering financial aid and scholarships for female students taking business or entrepreneur programs to support women-owned businesses.



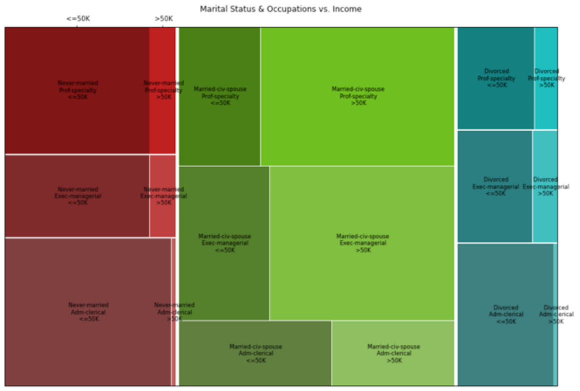
Visualization explains: Using a Nested Pie Chart, I observed that the percentage of women with the higher income is significantly low. Especially in the self-employed incorporated category, the proportion of females compared to males is less than 8%. I choose to create it because a Nested Pie Chart can show the proportion of individuals in each work class and the ratio of different sex in each work class that falls into higher income categories. It allows for easy visualization of comparing work class distribution on higher income levels and sex in different work classes. The steps of the design process differentiated higher and lower income and showed the proportion of each work class categories has taken in them. Since some categories have 0 percent, I have hidden the labels on the graph for better visualization. Each work class has labels and proportions in higher income demonstrated on the graph and a legend for color that corresponds to sexes. From the graph above, we can quickly tell that work class and sex are relatively reliable indicators for predicting an individual's income using a Nested Pie Chart as visualization.

2. According to the Country vs. Income graph observation, we can create a program that targets international students interested in obtaining a higher level of education in the United States and improving their chances of finding a higher-paying job. Also, the school can provide grand or some financial relief plan for students from Outlying US, Holland Netherlands, Philippines, and Honduras where the proportion of higher-income individuals are relatively low.



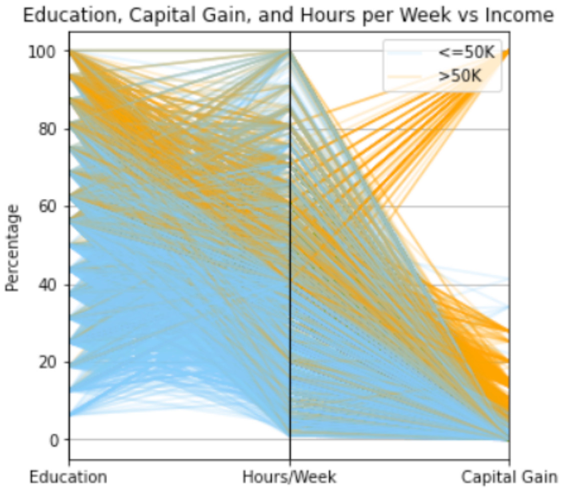
Visualization explains: By using Choropleth Maps, I observed that outlying areas of the U.S. and Holland Netherlands have very few or no people making more than $50,000. Also, Philippines and Honduras have less than 10% of people with higher income. However, other countries have a similar ratio of around 15% of people in the higher income group. I choose to create it because it allows us to see the spatial distribution of income across different regions of the world, compare the income levels of different countries, and provide a clear and concise representation of the data. The design process steps were identifying each country in the database and calculating the proportion of higher and lower income in each country. Countries not in the database show the lightest color since the value of higher-income individuals has been set to 0. From the graph above, we can easily tell that country is a relatively reliable indicator for predicting an individual's income by using Choropleth Maps as visualization.

3. Since visualization data from Marital Status & Occupations vs. Income showed that there are many never-married individuals have lower income compared to married individuals, I assumed the majority are under age 45, have a full-time job during the daytime, and live alone. Thus, we can create a night school program on skill development, such as Executive Management and Professional Specialty classes, because many higher-income individuals fall into those two occupation categories.



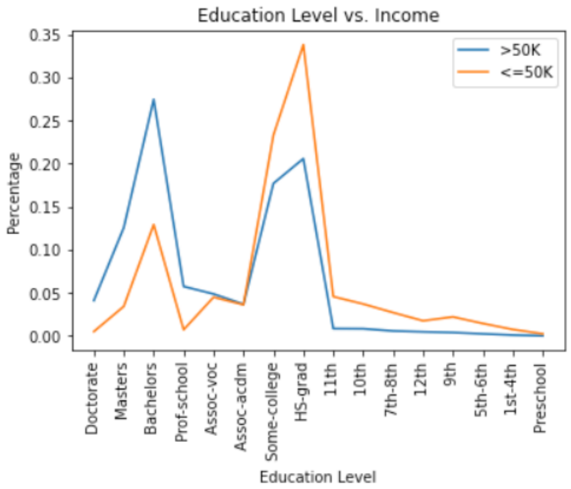
Visualization explains: By using Mosaic Plot, I observed that only a tiny portion of people who never married have higher incomes, and more than half of the married individuals in this data set have higher incomes. I choose to create it because a mosaic plot is a graphical display that allows us to visualize the relationship between three categorical variables and can help us better understand the relationship within occupations, marital status, and income. The steps of the design process were identifying each marital status and occupations with different income group. Since there were many variables in marital status and occupations and most of them are only a small proportion, I only showed top three values under marital status and occupation categories and hidden the labels on the axis for a better visualization. From the graph above, we can easily tell that marital status and occupations are a relatively reliable indicator for predicting an individual's income by using Mosaic Plot as visualization.

4. Given that individuals with lower educated individuals tend to work fewer hours per week, we can target lower-income individuals who are interested in continuing their education while working part-time. By offering flexible scheduling options and online courses, they can pursue higher education to increase their capital gain on their free time.

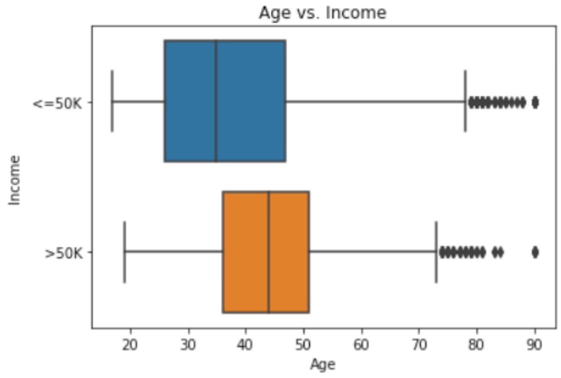


Visualization explains: Using Parallel Coordinates Plot, I observed that individuals with higher education tend to work more hours per week and have higher capital gain and higher income than those with lower education levels. I choose to create it because Parallel Coordinates Plot is a good choice for analyzing and visualizing multivariate data. In this case, it analyzed the relationship between each individual's education level, working hours, capital gains, and income. The steps of the design process were choosing features with numerical data and calculating them by percentage against each feature's max value for better visualization. From the graph above, we can easily tell that education number, hours per week, and capital gain are relatively reliable indicators for predicting an individual's income by using the Parelle Coordinates Plot as visualization.

5. Individuals in the lower income group are more likely to have only a high school degree and are in their early 30s, which suggests that cost may be a barrier to higher education. A college can provide financial assistance, such as scholarships or grants, to help these individuals afford the cost of tuition and other expenses. On the other hand, people in the higher income group tend to have higher levels of education, which suggests that they may be more likely to pursue careers in fields that require advanced degrees. College can offer degree programs that provide them with the skills and knowledge they need to advance their careers, such as healthcare or technology classes.



Visualization explains: Using a Line Graph, I observed that many people in the higher income group have bachelor's degrees, and people in the lower income group have a large portion with only a high school degree. Only a few or none with higher income have an education level lower than a high school degree.



Visualization explains: By using Box and Whisker Plots, I observed around age 25 to 47 fall into the interquartile lower income range, and the median is around age 35. On the other hand, around age 35 to 51 falls into the interquartile range of higher pay, and the median is around age 45.

Questions:

The questions that arose during the project progression were mostly about which visualization graph is more suitable for each feature, such as whether the "race" feature should be using Pie Chart or Bar Chart and whether Choropleth Maps is a better option than Segmented Bar Chart for "country" feature.

My solution was to differentiate each feature's data type. For example, a bar chart or pie chart is more suitable if a feature has a nominal data type. If a feature has an interval data type, we can use Parallel Coordinates Plot.

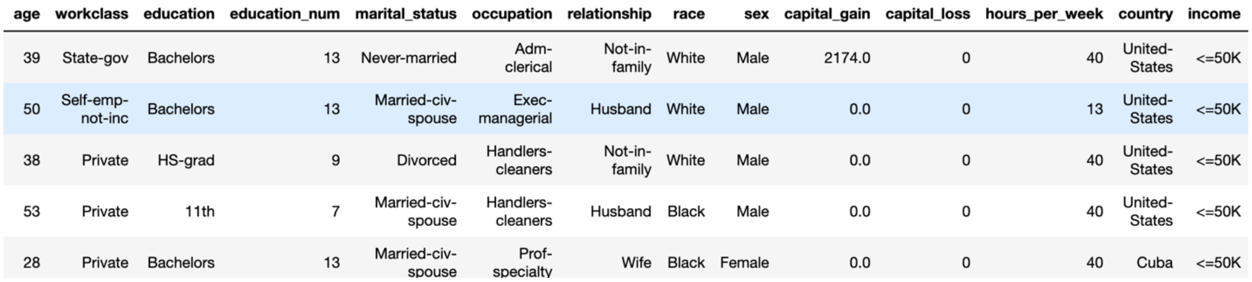
Then, I check for any outlier in the graph and adjust the graph accordingly to present the best visualization on the dataset.

Not doing:

The correlation matrix helps find the correlation between multiple variables. It can be used to identify strong positive or negative correlations between variables and can be a helpful tool in identifying patterns or trends in large datasets. In the future, we should first use a correlation matrix to find out which features strongly correlate with income.

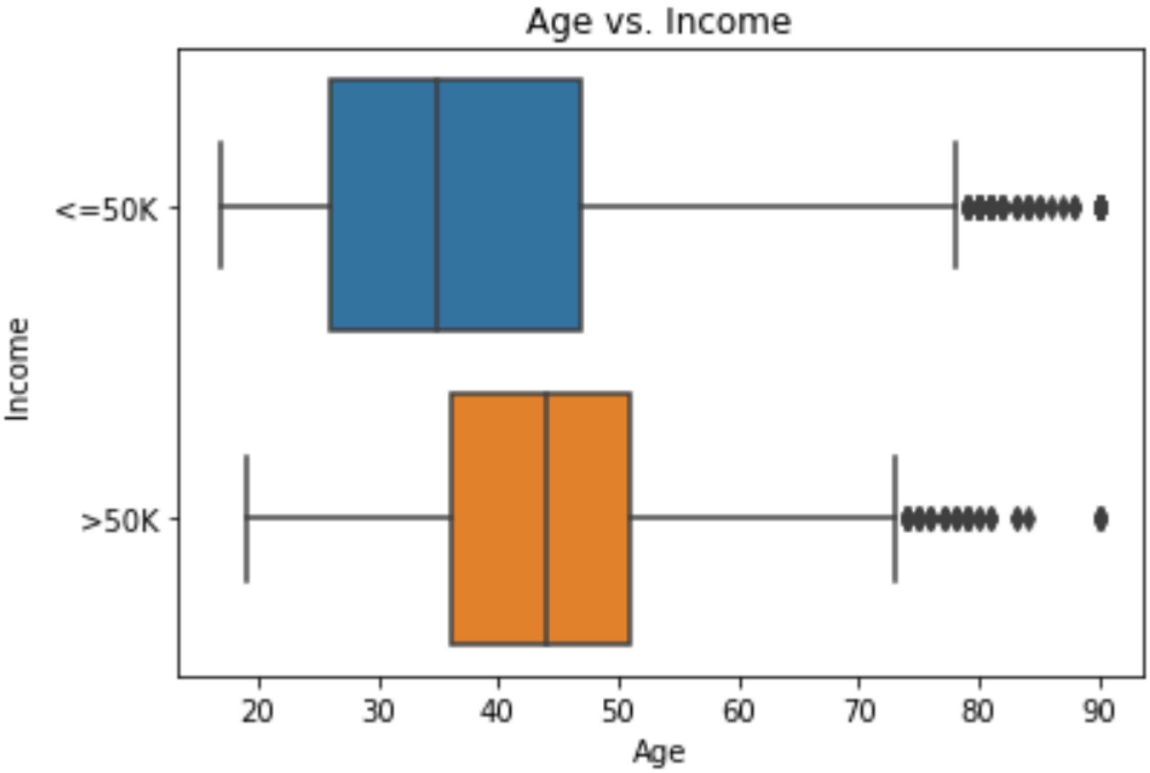
Appendix:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
from collections import OrderedDict
from statsmodels.graphics.mosaicplot import mosaic
import geopandas as gpd

data = pd.read_csv('adult.data',names=data_types,index_col=None,na_values=" ?" ,comment='|')
data = data.drop('fnlwgt', axis=1)
df = pd.DataFrame(data).drop_duplicates()

df.head()
```



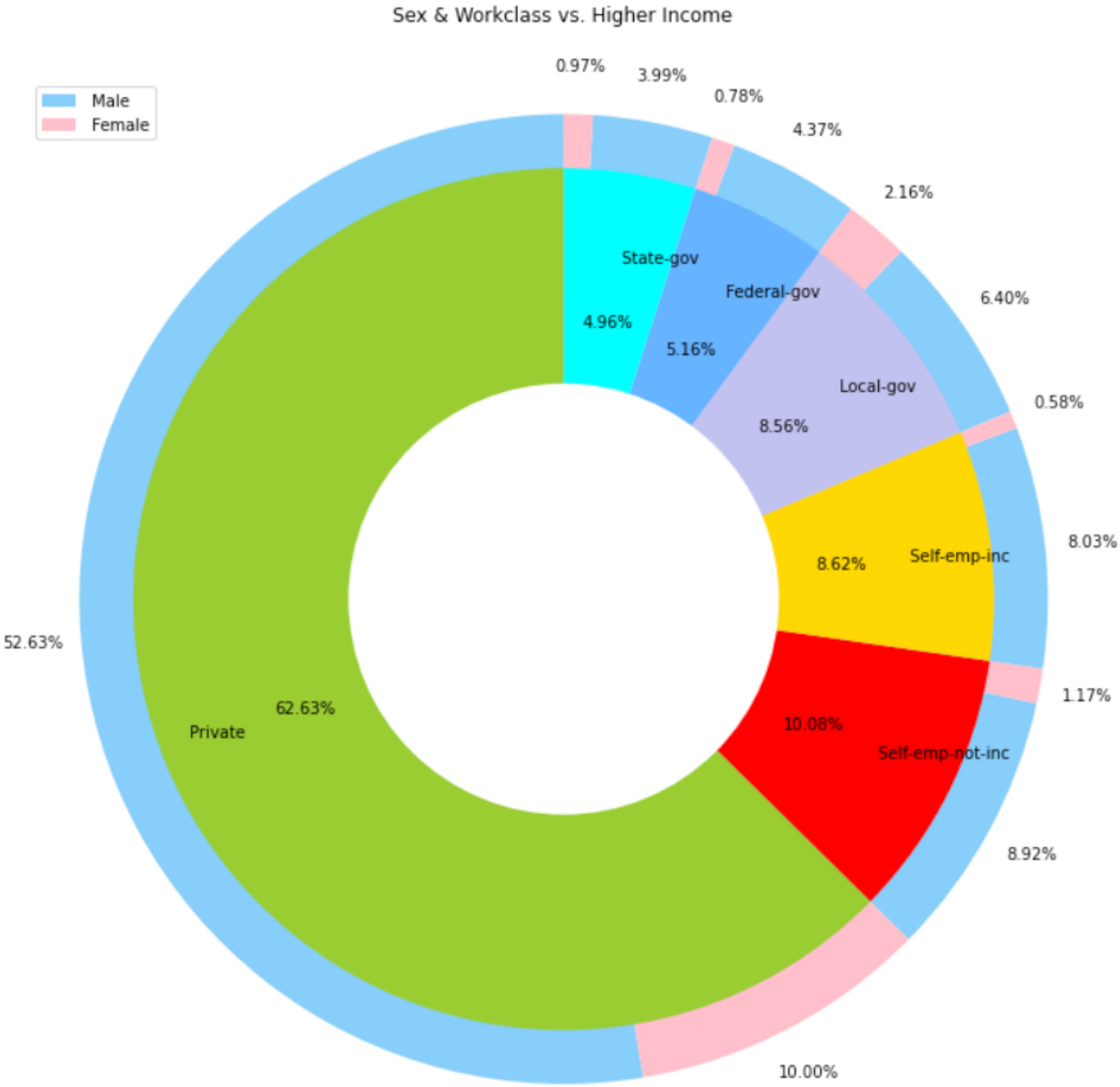
age	workclass	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	country	income
39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0	40	United-States	<=50K
50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0	13	United-States	<=50K
38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0	40	United-States	<=50K
53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0	40	United-States	<=50K
28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0	40	Cuba	<=50K

```
sns.boxplot(df['age'],df['income'])
plt.ylabel('Income')
plt.xlabel('Age')
plt.title('Age vs. Income')
plt.show()
```



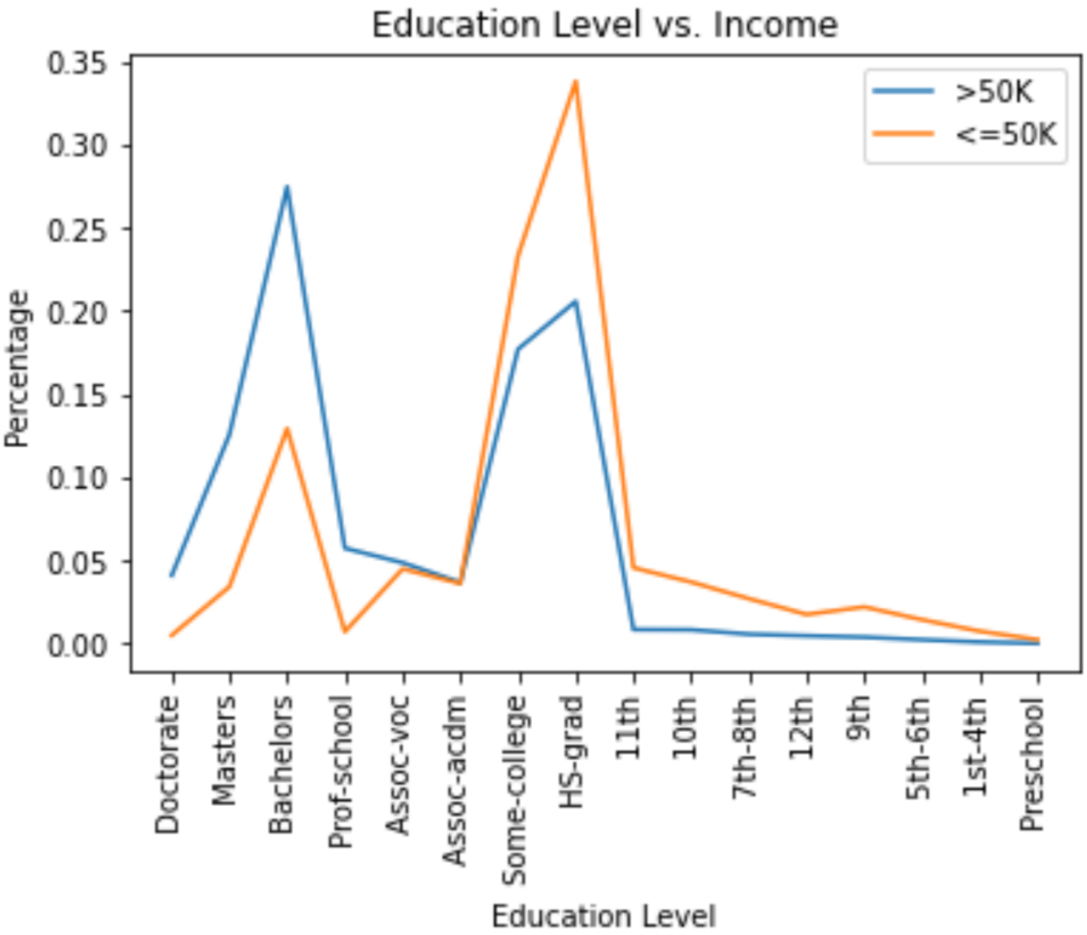
```
workclass_hi = df.loc[(df.income == " >50K"),["workclass"]].value_counts()
workclass_hi = workclass_hi[workclass_hi != 0]
labels = []
for i in workclass_hi.keys():
   labels.append(i[0])
sex_ = [' Male',' Female']
sizes = []
for i in labels:
   values = df.loc[(df.income == " >50K") & (df.workclass == i),['sex']].value_counts()
   sizes.append(values[0])
   sizes.append(values[1])

colors = ['yellowgreen','red','gold','#c2c2f0','#66b3ff','cyan','grey','purple',]
colors_g = ['lightskyblue','pink']
colors_gender = colors_g*6
def my_autopct(pct):
   return '{:.2f}%'.format(pct) if pct > 0 else "
fig, ax = plt.subplots(figsize=(10, 10))
plt.pie(sizes,colors=colors_gender,startangle=90, radius=2.25,autopct=my_autopct,
pctdistance=1.1)
plt.pie(workclass_hi, labels=labels, labeldistance = 0.8, colors=colors,
startangle=90,frame=True, radius=2, autopct=my_autopct, pctdistance=0.65)
ax.legend(labels, loc='center right')
```

```
legend_patches = [mpatches.Patch(color=c, label=l) for c, l in zip(colors_g, sex_)]
ax.legend(handles=legend_patches, loc='upper left')
centre_circle = plt.Circle((0,0),1,color='black', fc='white',linewidth=0)
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
ax.set_title('Sex & Workclass vs. Higher Income\n\n')
plt.axis('equal')
plt.tight_layout()
plt.show()
```

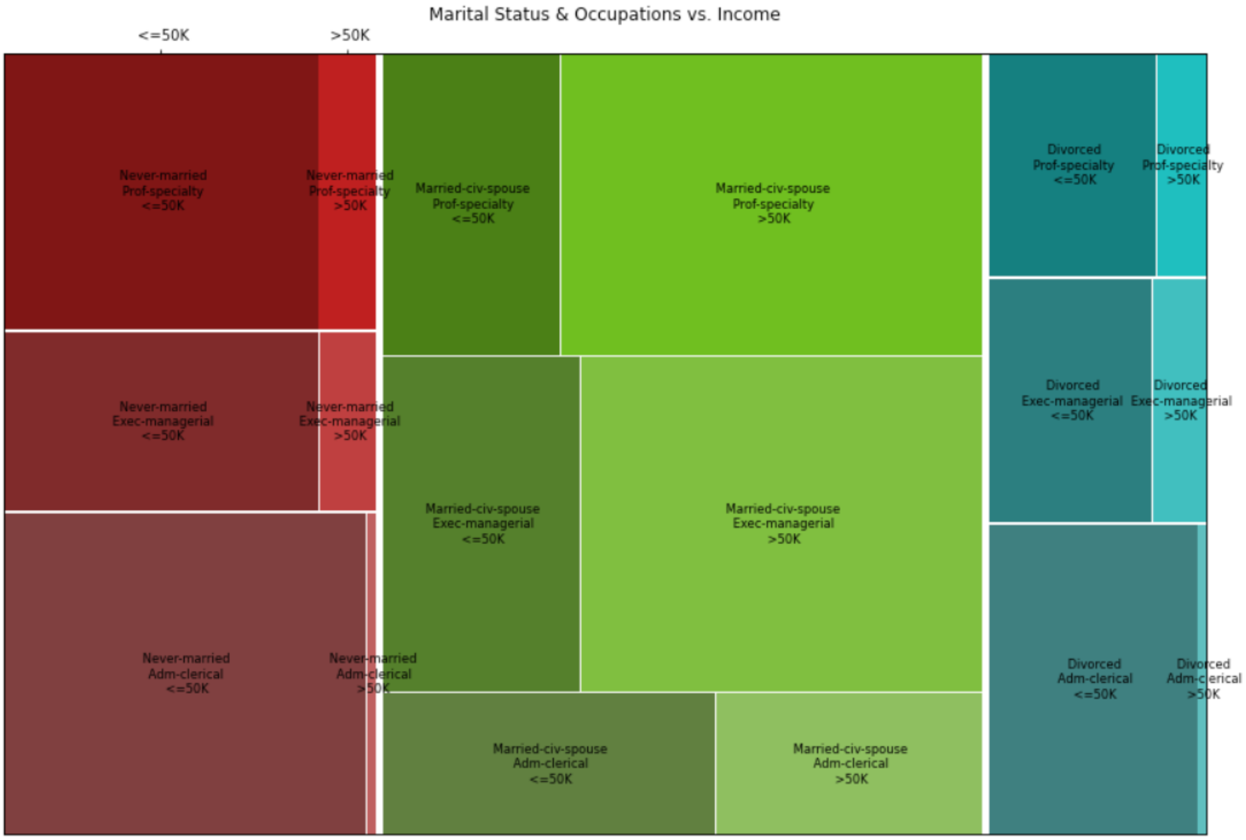


```
education_hi = df.loc[(df.income == " >50K"),["education"]].value_counts()
education_li = df.loc[(df.income == " <=50K"),["education"]].value_counts()
labels = [' Doctorate',' Masters', ' Bachelors', ' Prof-school', ' Assoc-voc',  ' Assoc-acdm', ' Some-college',' HS-grad', ' 11th', ' 10th', ' 7th-8th', ' 12th', ' 9th', ' 5th-6th', ' 1st-4th', ' Preschool']
edu_hi_percent = [0] * len(labels)
```

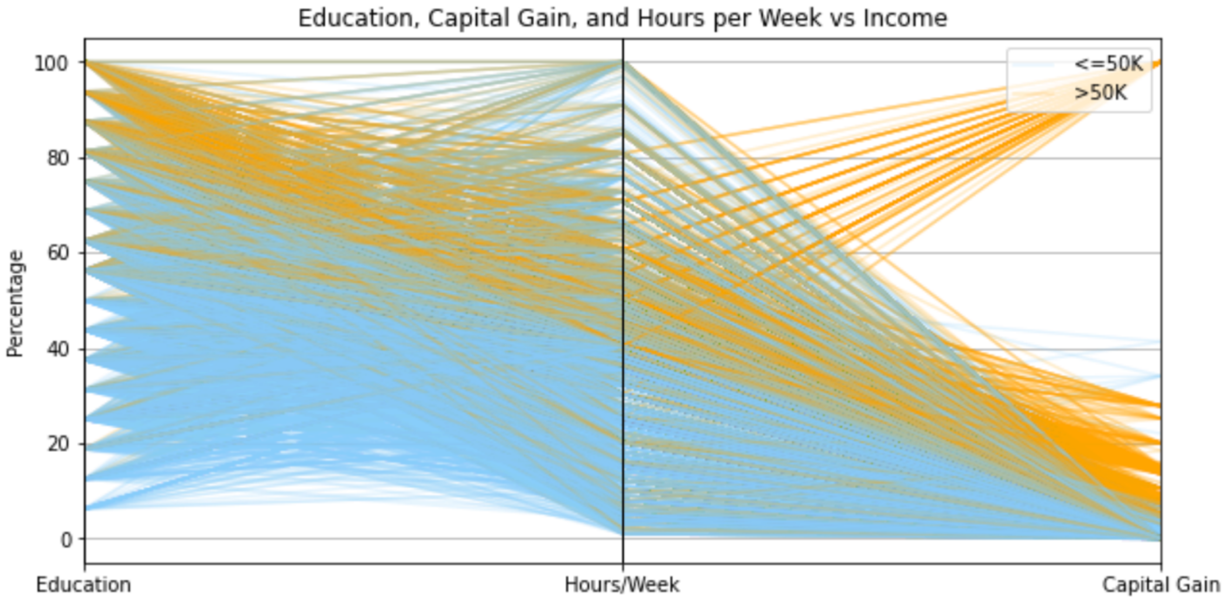
```
edu_li_percent = [0] * len(labels)
for i in range(len(labels)):
    if labels[i] in education_hi:
        edu_hi_percent[i] = education_hi[labels[i]]/education_hi.sum()
    if labels[i] in education_li:
        edu_li_percent[i] = education_li[labels[i]]/education_li.sum()

plt.plot(labels, edu_hi_percent, label='>50K')
plt.plot(labels, edu_li_percent, label='<=50K')
plt.xlabel('Education Level')
plt.ylabel('Percentage')
plt.title('Education Level vs. Income')
plt.xticks(rotation=90)
plt.legend()
plt.show()
```



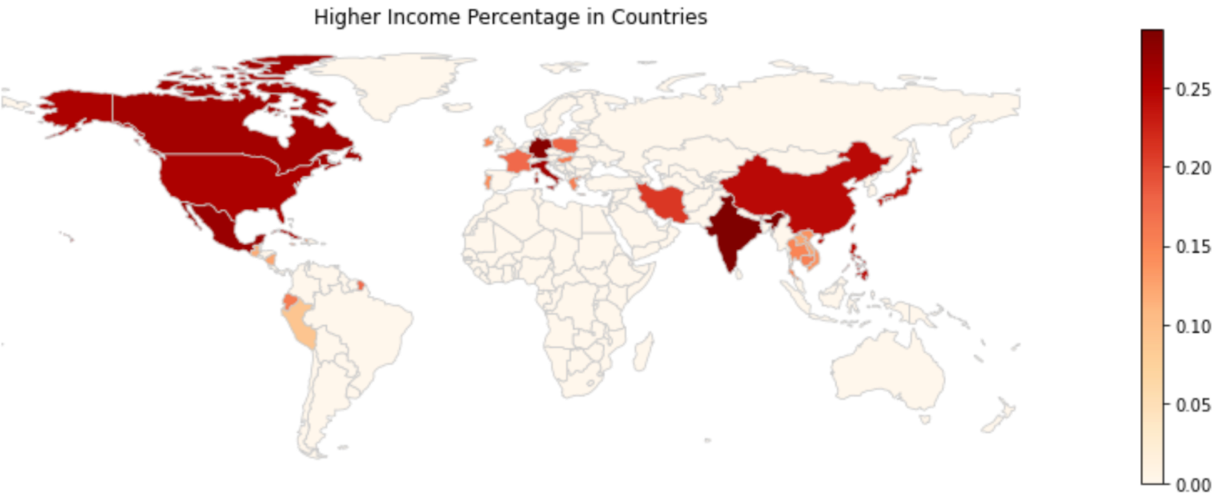
```
ms_top = df['marital_status'].value_counts()
occ_top = df['occupation'].value_counts()
ms_lis =[]
occ_lis =[]
for i in range(3):
    ms_lis.append(ms_top.keys()[i])
    occ_lis.append(occ_top.keys()[i])

ms_df = df[df['occupation'].isin(occ_lis) & df['marital_status'].isin(ms_lis)]

fig, ax = plt.subplots(figsize=(15, 10))
mosaic(ms_df, ['marital_status', 'occupation','income'],ax=ax)
plt.title('Marital Status & Occupations vs. Income')
ax.set_xticks([])
ax.set_yticks([])
```



```
df['education_num_percentage'] = df['education_num'] / df['education_num'].max() * 100
df['hours_per_week_percentage'] = df['hours_per_week'] / df['hours_per_week'].max() * 100
df['capital_gain_percentage'] = df['capital_gain'] / df['capital_gain'].max() * 100

coord = df[['education_num_percentage', 'hours_per_week_percentage',
'capital_gain_percentage','income']]
plt.figure(figsize=(10, 5))
ax= pd.plotting.parallel_coordinates(coord, 'income', alpha=0.2, color=['lightskyblue','orange'])
ax.set_xticklabels(['Education', 'Hours/Week', 'Capital Gain'])

plt.ylabel('Percentage')
plt.title('Education, Capital Gain, and Hours per Week vs Income')
plt.show()
```



```
country_hi = df.loc[(df.income == " >50K"),["country"]].value_counts()
country_li = df.loc[(df.income == " <=50K"),["country"]].value_counts()
country_labels = country_hi.keys()
labels = []
for i in country_labels:
    if i[0] == " United-States":
        labels.append('United States of America')
    labels.append(i[0].strip())

country_hi_v = []
for i in range(len(country_hi)):
    country_sum = country_hi[i] + country_li[i]
    country_hi_v.append(country_hi[i]/country_sum)

geodf = pd.DataFrame(list(zip(labels, country_hi_v)),columns=['country','percent'])

world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
world = world[(world.pop_est>0) & (world.name!="Antarctica")]
merged_data = world.merge(geodf, how='left', left_on='name', right_on='country')
merged_data['percent'] = merged_data['percent'].apply(lambda x: x if not pd.isna(x) else 0)
fig, ax = plt.subplots(figsize=(15, 5))
merged_data.plot(column='percent',legend=True, cmap='OrRd', linewidth=0.8, edgecolor='0.8',
ax=ax)
plt.axis("off")
plt.title('Higher Income Percentage in Countries')
plt.show()
```


Higher Income Percentage in Countries
0.25
0.20
0.15
0.10
0.05
0.00