CSE 578: Data Visualization (2022 Summer - B)
Shih Chi Lin
1225180333

Systems Documentation Report

- **Roles and Responsibilities**

  I am a data analyst at XYZ Corporation. My company is working on a project with UVW College. I will analyze the data provided by the U.S. Census Bureau and find useful information for UVW College to market their degree programs to specific groups.

- **Goals and Objectives**
  - Goals: XYZ Corporation will analyze the data to find the key factors that affect an individual's income. Factors may be composed of multiple features.
  - Objectives: UVW College wants to recruit more students and to tailor and market to different segments.
- **Assumptions**

  The data used for the analysis came from the 1994 U.S. Census Bureau database. In order to establish criteria for marketing its degree programs, UVW College has made the income its key demographic. The college's marketing division wants to develop profiles based on variables including age, gender, education level, ethnicity, and income.

- **User Stories**

  8 attributes in use: income, workclass, occupation, native-country, education, marital-status, capital-gain and capital loss

  1. User Story #1

     As a member of the marketing team at UVW, I would like to know the relationship between income and workclass.

  2. User Story #2

     I would like to know if occupation affects income and the distribution of native country.

  3. User Story #3

     The worker wants to know the relationship between education and marital status.

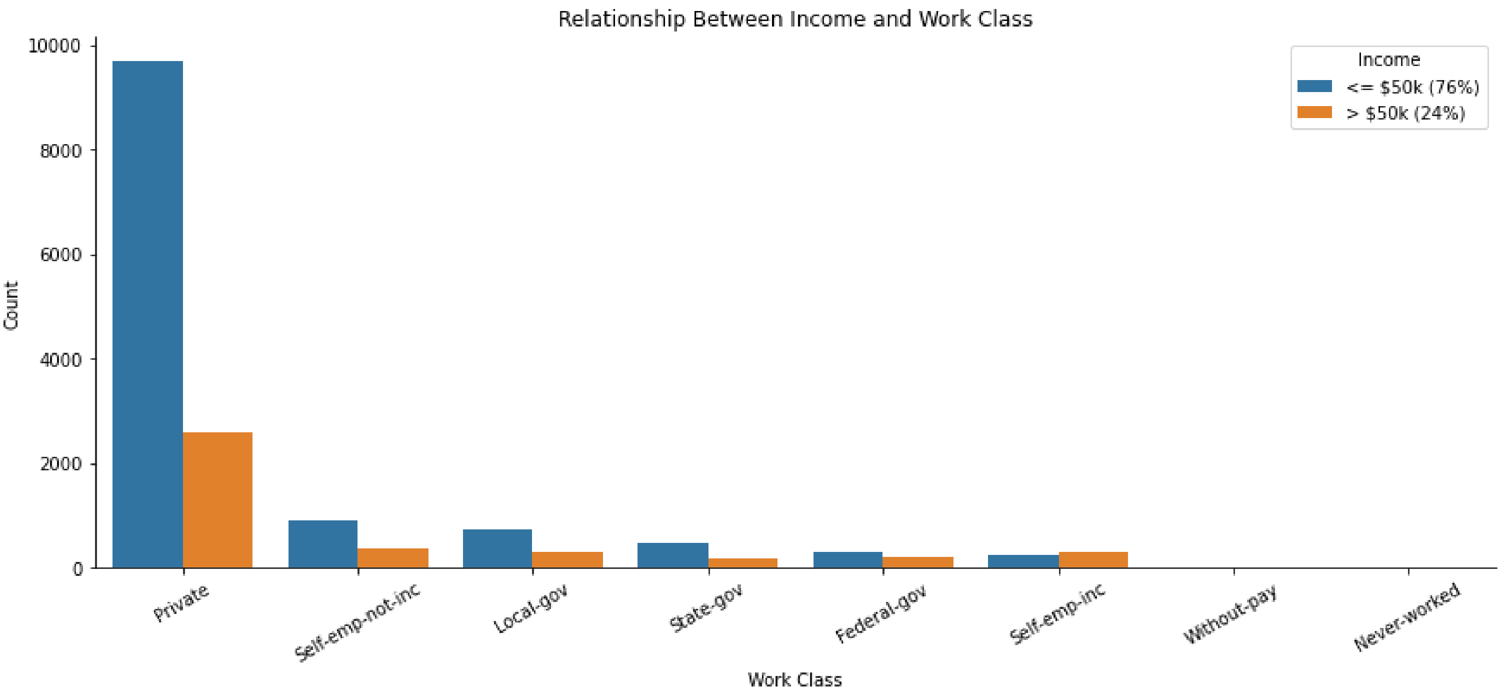
  4. User Story #4

     The Director of Marketing would like to know how capital gain and capital loss relate to income.

- **Visualizations**

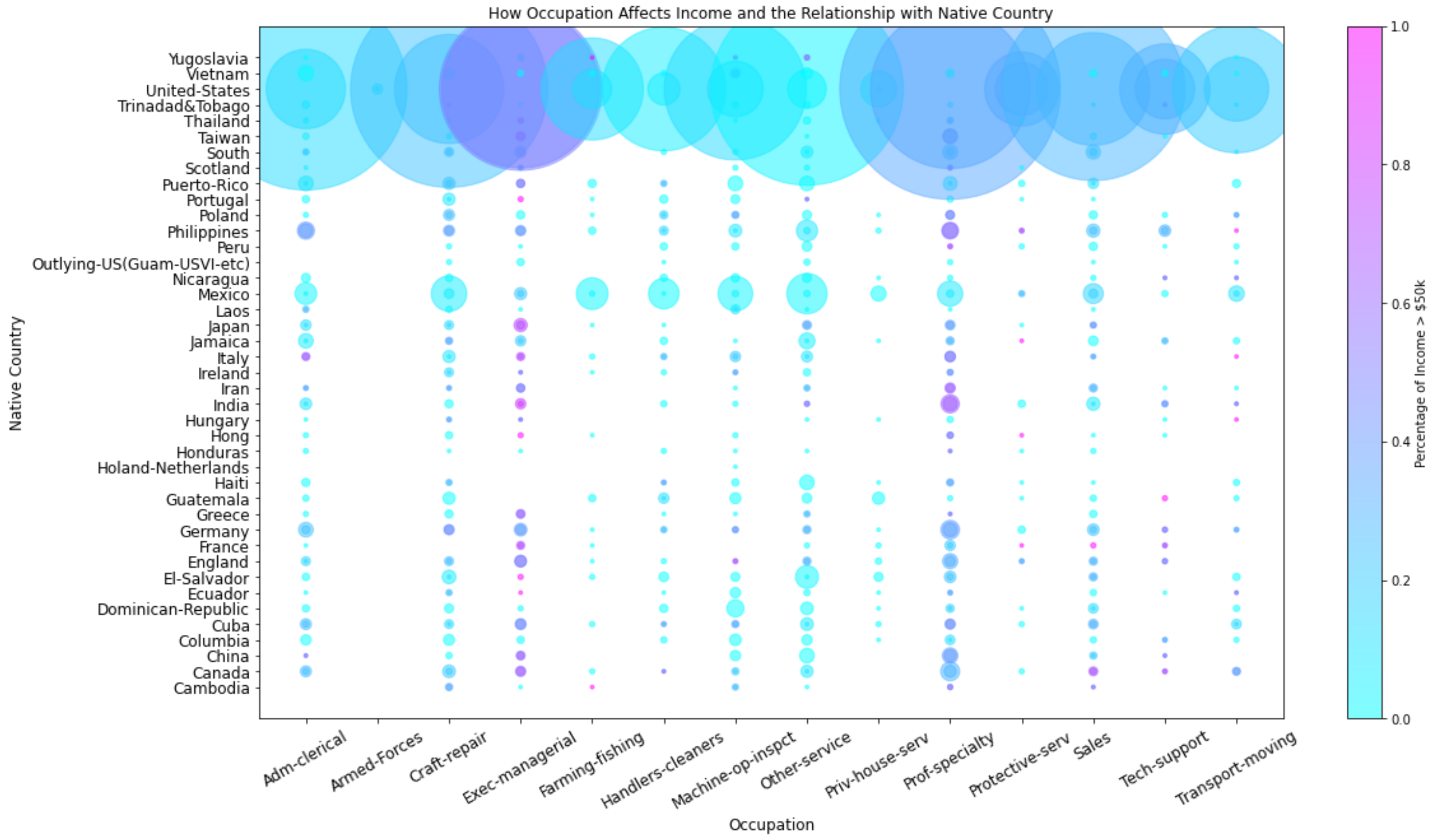
1. User Story #1: As a member of the marketing team at UVW, I would like to know the relationship between income and workclass.

   The following chart shows the bar chart of the work class. The analysis shows that 76% of the people in this data earn less than or equal to $50k and the workclass of majority of them is private regardless of their income level. Except for incorporated self-employment, the number of people earning more than 50k in the rest of the job class is less than the number of people earning less than 50k.



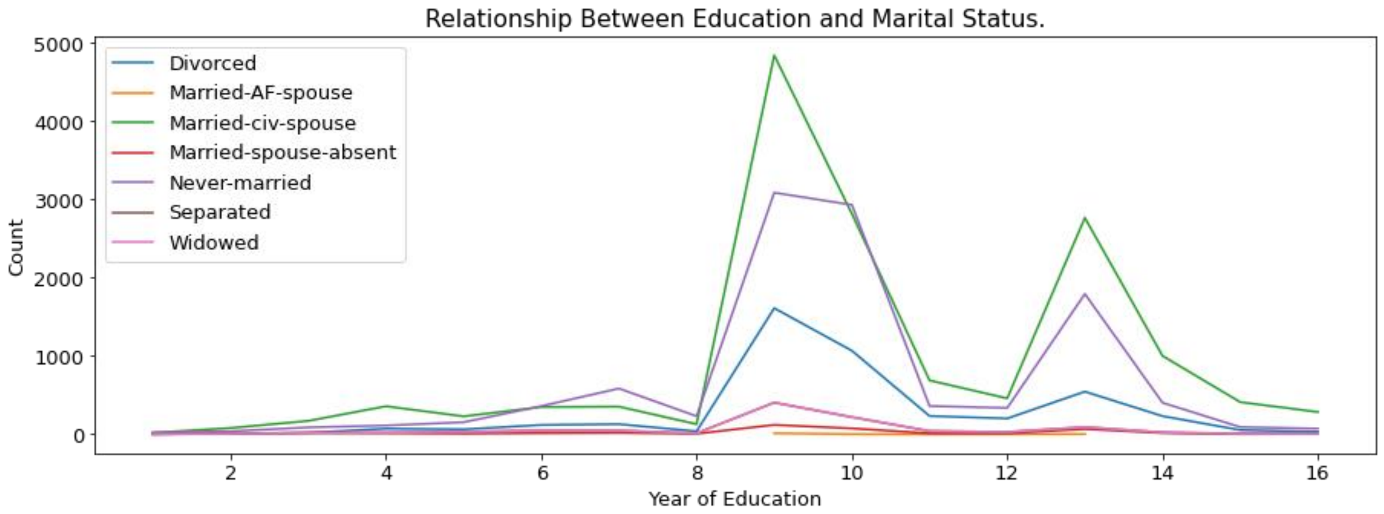
2. User Story #2: I would like to know if occupation affects income and the distribution of their native country.

   In the figure below, the larger the circle, the more people belong to the native country, therefore, we can know that there are bias of the population in this data. In addition, the pinker the color is, the higher the percentage of people with income over 50k in the circle. From this graph, we can learn that people who are not from the United States have a relatively high percentage of income above 50k, and their occupations are concentrated in exec-managerial and prof-specialty.

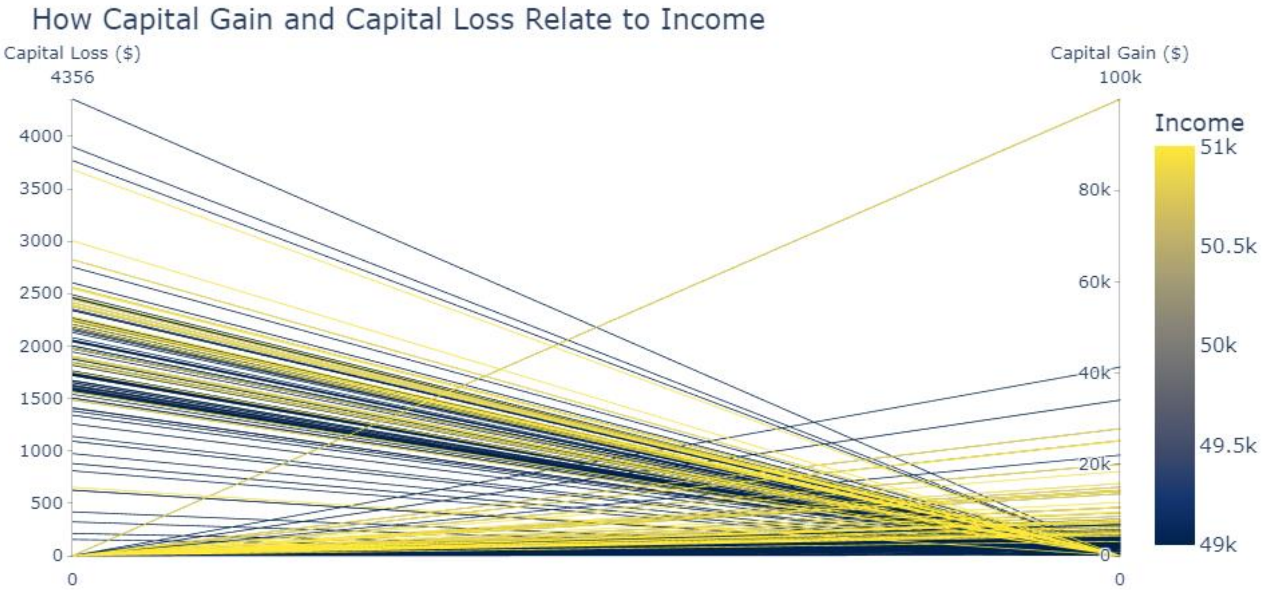


3. User Story #3: The worker wants to know the relationship between education and marital status.

   This graph shows the year of education of people in various marital statuses. It can be seen that the majority of people in the three marital statuses had 9 to 10 and 13 years of education, and their marital statuses were divorced, never-married, and married-civ-spouse. It is clear that there is no significant correlation between marital status and education.



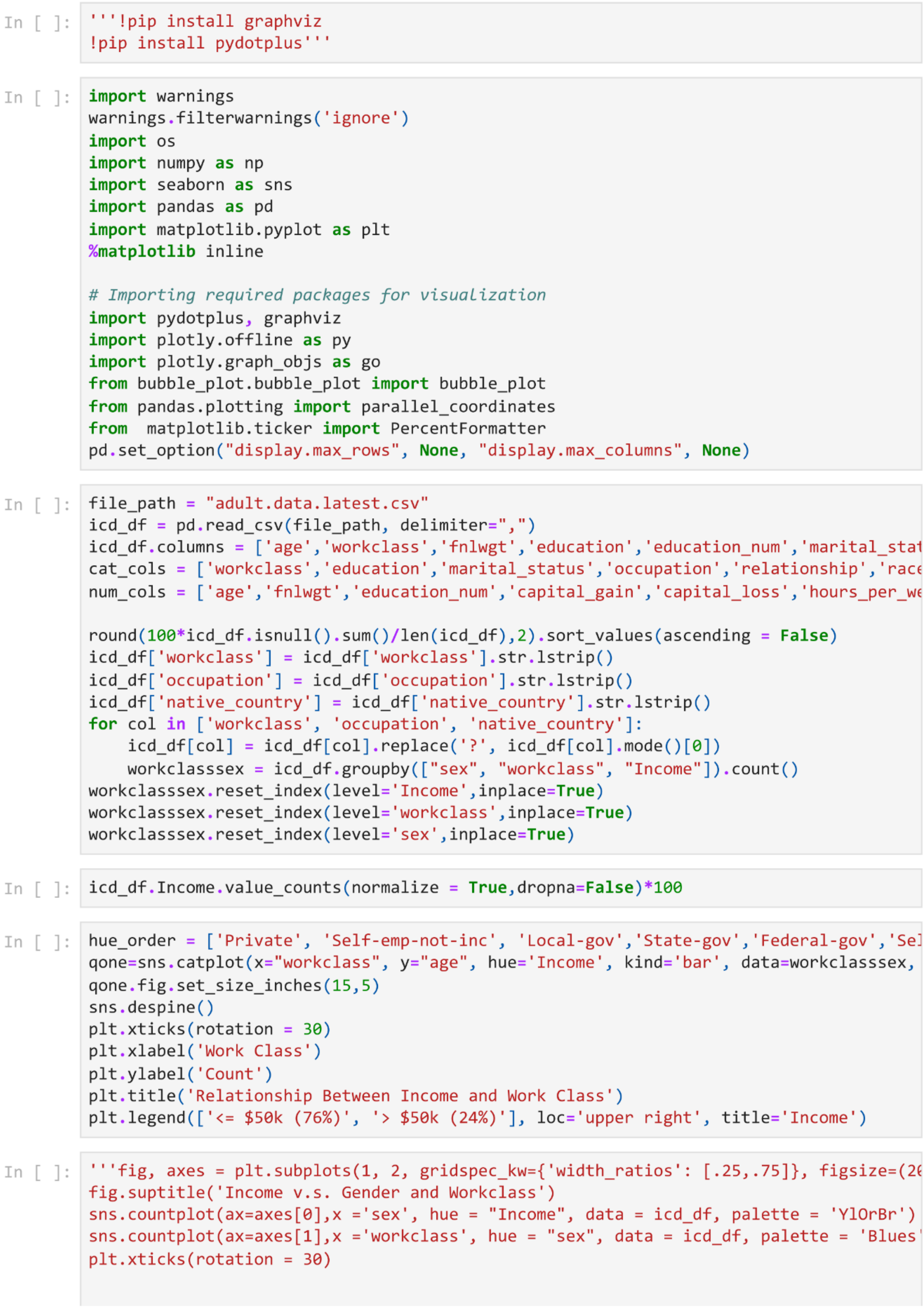
4. User Story #4: The Director of Marketing would like to know how capital gain and capital loss relate to income.



This graph shows the capital loss and capital gain situation. The yellow lines represent those who earned more than $50k; the dark blue ones represent those who earned less than $50k. Interestingly, no one had both loss and gain, so we see that all people with capital loss point to 0 capital gain, and vice versa. In addition, the number of capital gain is much larger than the number of capital loss, which can be seen on the left and right y-axis.

CSE 578: Data Visualization (2022 Summer - B)
Shih Chi Lin
1225180333

- **Appendix**



In [ ]:
```
'''!pip install graphviz
!pip install pydotplus'''
```

In [ ]:
```
import warnings
warnings.filterwarnings('ignore')
import os
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

# Importing required packages for visualization
import pydotplus, graphviz
import plotly.offline as py
import plotly.graph_objs as go
from bubble_plot.bubble_plot import bubble_plot
from pandas.plotting import parallel_coordinates
from  matplotlib.ticker import PercentFormatter
pd.set_option("display.max_rows", None, "display.max_columns", None)
```

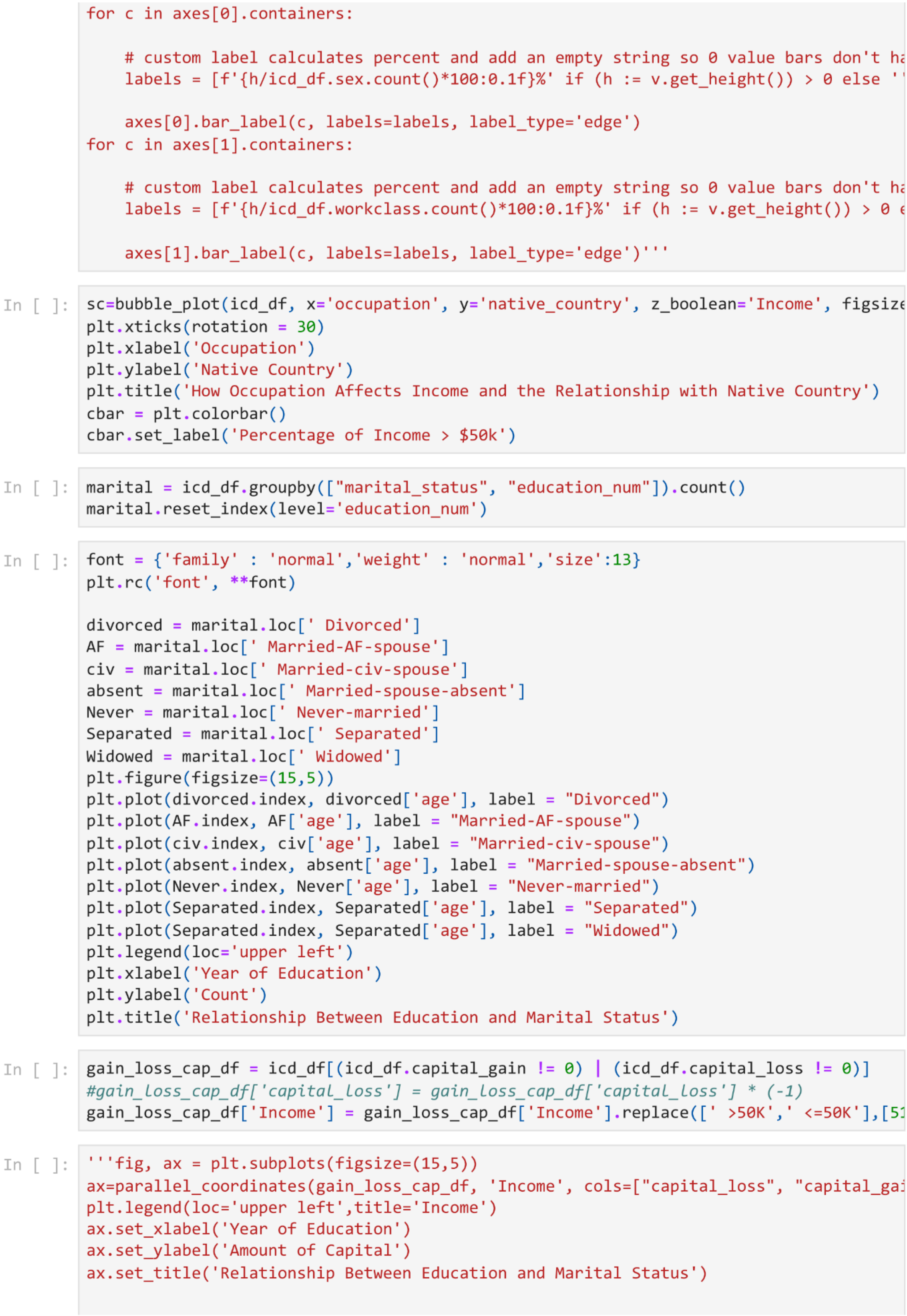
In [ ]:
```
file_path = "adult.data.latest.csv"
icd_df = pd.read_csv(file_path, delimiter=",")
icd_df.columns = ['age','workclass','fnlwgt','education','education_num','marital_stat
cat_cols = ['workclass','education','marital_status','occupation','relationship','race
num_cols = ['age','fnlwgt','education_num','capital_gain','capital_loss','hours_per_we

round(100*icd_df.isnull().sum()/len(icd_df),2).sort_values(ascending = False)
icd_df['workclass'] = icd_df['workclass'].str.lstrip()
icd_df['occupation'] = icd_df['occupation'].str.lstrip()
icd_df['native_country'] = icd_df['native_country'].str.lstrip()
for col in ['workclass', 'occupation', 'native_country']:
    icd_df[col] = icd_df[col].replace('?', icd_df[col].mode()[0])
    workclasssex = icd_df.groupby(["sex", "workclass", "Income"]).count()
workclasssex.reset_index(level='Income',inplace=True)
workclasssex.reset_index(level='workclass',inplace=True)
workclasssex.reset_index(level='sex',inplace=True)
```

In [ ]:
```
icd_df.Income.value_counts(normalize = True,dropna=False)*100
```

In [ ]:
```
hue_order = ['Private', 'Self-emp-not-inc', 'Local-gov','State-gov','Federal-gov','Sel
qone=sns.catplot(x="workclass", y="age", hue='Income', kind='bar', data=workclasssex,
qone.fig.set_size_inches(15,5)
sns.despine()
plt.xticks(rotation = 30)
plt.xlabel('Work Class')
plt.ylabel('Count')
plt.title('Relationship Between Income and Work Class')
plt.legend(['<= $50k (76%)', '> $50k (24%)'], loc='upper right', title='Income')
```

In [ ]:
```
'''fig, axes = plt.subplots(1, 2, gridspec_kw={'width_ratios': [.25,.75]}, figsize=(2
fig.suptitle('Income v.s. Gender and Workclass')
sns.countplot(ax=axes[0],x ='sex', hue = "Income", data = icd_df, palette = 'YlOrBr')
sns.countplot(ax=axes[1],x ='workclass', hue = "sex", data = icd_df, palette = 'Blues'
plt.xticks(rotation = 30)
```

CSE 578: Data Visualization (2022 Summer - B)
Shih Chi Lin
1225180333



```
for c in axes[0].containers:

    # custom label calculates percent and add an empty string so 0 value bars don't ha
    labels = [f'{h/icd_df.sex.count()*100:0.1f}%' if (h := v.get_height()) > 0 else ''

    axes[0].bar_label(c, labels=labels, label_type='edge')
for c in axes[1].containers:

    # custom label calculates percent and add an empty string so 0 value bars don't ha
    labels = [f'{h/icd_df.workclass.count()*100:0.1f}%' if (h := v.get_height()) > 0 e

    axes[1].bar_label(c, labels=labels, label_type='edge')'''
```

```
In [ ]: sc=bubble_plot(icd_df, x='occupation', y='native_country', z_boolean='Income', figsize
        plt.xticks(rotation = 30)
        plt.xlabel('Occupation')
        plt.ylabel('Native Country')
        plt.title('How Occupation Affects Income and the Relationship with Native Country')
        cbar = plt.colorbar()
        cbar.set_label('Percentage of Income > $50k')
```

```
In [ ]: marital = icd_df.groupby(["marital_status", "education_num"]).count()
        marital.reset_index(level='education_num')
```

```
In [ ]: font = {'family' : 'normal','weight' : 'normal','size':13}
        plt.rc('font', **font)

        divorced = marital.loc[' Divorced']
        AF = marital.loc[' Married-AF-spouse']
        civ = marital.loc[' Married-civ-spouse']
        absent = marital.loc[' Married-spouse-absent']
        Never = marital.loc[' Never-married']
        Separated = marital.loc[' Separated']
        Widowed = marital.loc[' Widowed']
        plt.figure(figsize=(15,5))
        plt.plot(divorced.index, divorced['age'], label = "Divorced")
        plt.plot(AF.index, AF['age'], label = "Married-AF-spouse")
        plt.plot(civ.index, civ['age'], label = "Married-civ-spouse")
        plt.plot(absent.index, absent['age'], label = "Married-spouse-absent")
        plt.plot(Never.index, Never['age'], label = "Never-married")
        plt.plot(Separated.index, Separated['age'], label = "Separated")
        plt.plot(Separated.index, Separated['age'], label = "Widowed")
        plt.legend(loc='upper left')
        plt.xlabel('Year of Education')
        plt.ylabel('Count')
        plt.title('Relationship Between Education and Marital Status')
```

```
In [ ]: gain_loss_cap_df = icd_df[(icd_df.capital_gain != 0) | (icd_df.capital_loss != 0)]
        #gain_loss_cap_df['capital_loss'] = gain_loss_cap_df['capital_loss'] * (-1)
        gain_loss_cap_df['Income'] = gain_loss_cap_df['Income'].replace([' >50K',' <=50K'],[51
```

```
In [ ]: '''fig, ax = plt.subplots(figsize=(15,5))
        ax=parallel_coordinates(gain_loss_cap_df, 'Income', cols=["capital_loss", "capital_gai
        plt.legend(loc='upper left',title='Income')
        ax.set_xlabel('Year of Education')
        ax.set_ylabel('Amount of Capital')
        ax.set_title('Relationship Between Education and Marital Status')
ax.set_yticks([gain_loss_cap_df['capital_loss'].min(), gain_loss_cap_df['capital_loss'
'''
```

In [ ]:

```
import pandas as pd
import numpy as np
import plotly.express as px

fig = px.parallel_coordinates(
    gain_loss_cap_df,
    dimensions=["capital_loss", "capital_gain"],
    title = 'How Capital Gain and Capital Loss Relate to Income',
    color ='Income',
    labels={"Income": "Income","capital_loss": "Capital Loss ($)","capital_gain":"Capi
    color_continuous_scale=px.colors.sequential.Cividis
    #width=1600, height=900
)
fig.update_layout(
    font=dict(
        size=15
    ))
fig.show()
```