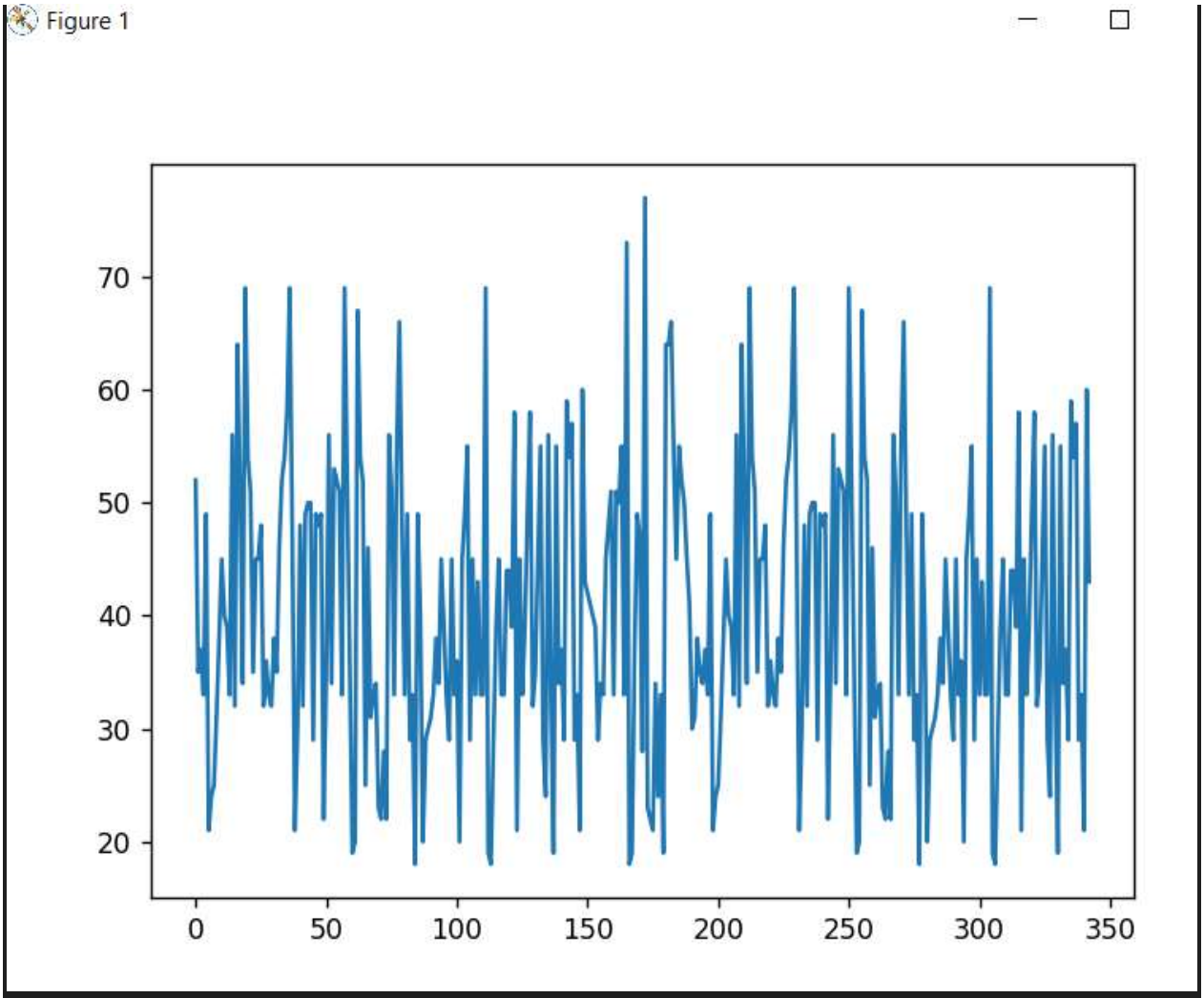
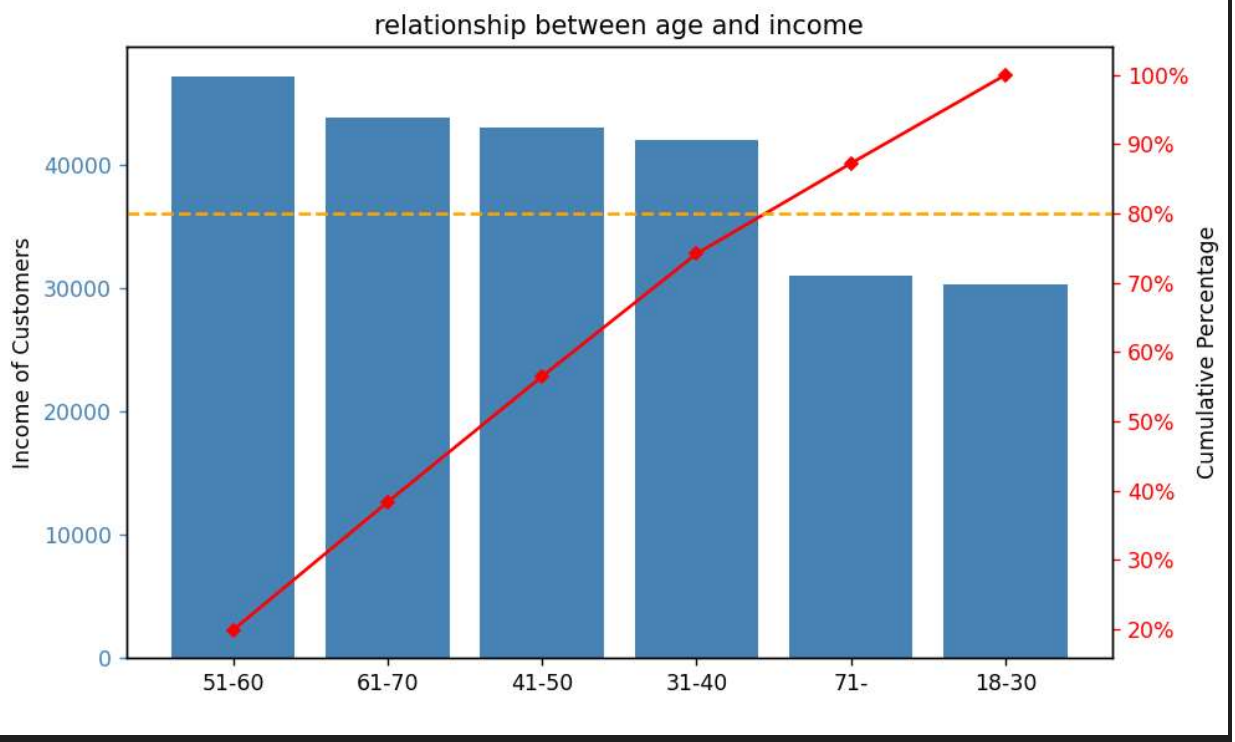
```
Name: Shi Yu Yeo(Le Roy)
Course# BAN 5733 online 63320


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

df = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')
res = 0

np.mean(df['female'])
print("%.2f" % np.mean(df['female']))

age = np.mean(df['age'])
print("%.0f" % age)

income = np.mean(df['income'])
print("%.2f" % income)

plt.plot(df['age'])
plt.show()
```




```
#creating a pareto chart, taking the average income from a age group in 10s,
example 30-40,40-50,
#grouping into buckets for the age
#age 18 to 30
new_df = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')

new_df = new_df[new_df.age >=18]

new_df2 = new_df[new_df.age <=30]
print(new_df2)

income18to30 = np.mean(new_df2['income'])
print("%.2f" % income18to30)

#age20to30 = np.mean(new_df2['age'])
#print("%.2f" % age20to30)
```

```
#age 31 to 40
new_df3140 = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')

new_df = new_df3140[new_df3140.age >=31]
new_df3 = new_df[new_df.age <=40]
income31to40 = np.mean(new_df3['income'])
print("%.2f" % income31to40)

#age 41 to 50
new_df = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')

new_df = new_df[new_df.age >=41]
new_df4 = new_df[new_df.age <=50]
income41to50 = np.mean(new_df4['income'])
print("%.2f" % income41to50)

#age 51 to 60
new_df = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')

new_df = new_df[new_df.age >=51]
new_df5 = new_df[new_df.age <=60]
income51to60 = np.mean(new_df5['income'])
print("%.2f" % income51to60)

#age 61 to 70
new_df = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')

new_df = new_df[new_df.age >=61]
new_df6 = new_df[new_df.age <=70]
income61to70 = np.mean(new_df6['income'])
print("%.2f" % income61to70)

#age 71 to max
new_df = pd.read_excel(r'E:/OSU/Fall 2023/Module 1/ex 1/Demographics.xlsx')

new_df7 = new_df[new_df.age >=71]
income71 = np.mean(new_df7['income'])
print("%.2f" % income71)



df_graph = pd.DataFrame({'count': [income18to30, income31to40, income41to50,
income51to60, income61to70, income71]})
df_graph.index = ['18-30', '31-40', '41-50', '51-60', '61-70', '71-']
```

```
#sort DataFrame by count descending
df_graph = df_graph.sort_values(by='count', ascending=False)

#add column to display cumulative percentage
df_graph['cumperc'] = df_graph['count'].cumsum()/df_graph['count'].sum()*100

#view DataFrame
print(df_graph)

import matplotlib.pyplot as plt
from matplotlib.ticker import PercentFormatter

#define aesthetics for plot
color1 = 'steelblue'
color2 = 'red'
line_size = 4

#create basic bar plot
fig, ax = plt.subplots()
ax.bar(df_graph.index, df_graph['count'], color=color1)
ax.set_ylabel("Income of Customers")

#add cumulative percentage line to plot
ax2 = ax.twinx()
ax2.plot(df_graph.index, df_graph['cumperc'], color=color2, marker="D",
ms=line_size)
ax2.axhline(80, color="orange", linestyle="dashed")
ax2.yaxis.set_major_formatter(PercentFormatter())
ax2.set_ylabel("Cumulative Percentage")

#specify axis colors
ax.tick_params(axis='y', colors=color1)
ax2.tick_params(axis='y', colors=color2)

#display Pareto chart
plt.xlabel("Age of Customers")
plt.title("relationship between age and income")
plt.show()
```



The x-axis displays the different ages age income from highest to lowest frequency.

From the Pareto Chart above, it can be seen that ~80% of the average income earners are ages between 51-60, 61-70, 41-50, 31-40. So, age above 71 and below 30 are the least possible chances of getting a high income.