Homework 3

COSC 3337

Dr. Rizk

Problem Statement

A business provides you with the following data and wants you to find patterns in their customer's habits so that they can make better recommendations and target certain groups in the future.

Answer the following:

Would we approach this as a supervised or unsupervised learning task and why?

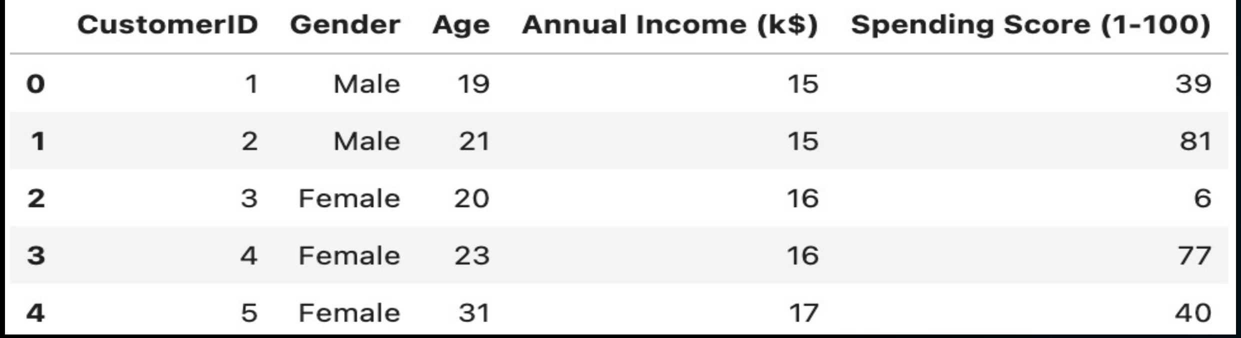
About The Data

The data we'll be using for this homework contains the following customer attributes:

- ID
- Gender
- Spending Score
- Age
- Income

Step 1

Begin by importing the data and displaying the first 5 observations.



	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Answer the following using Pandas:

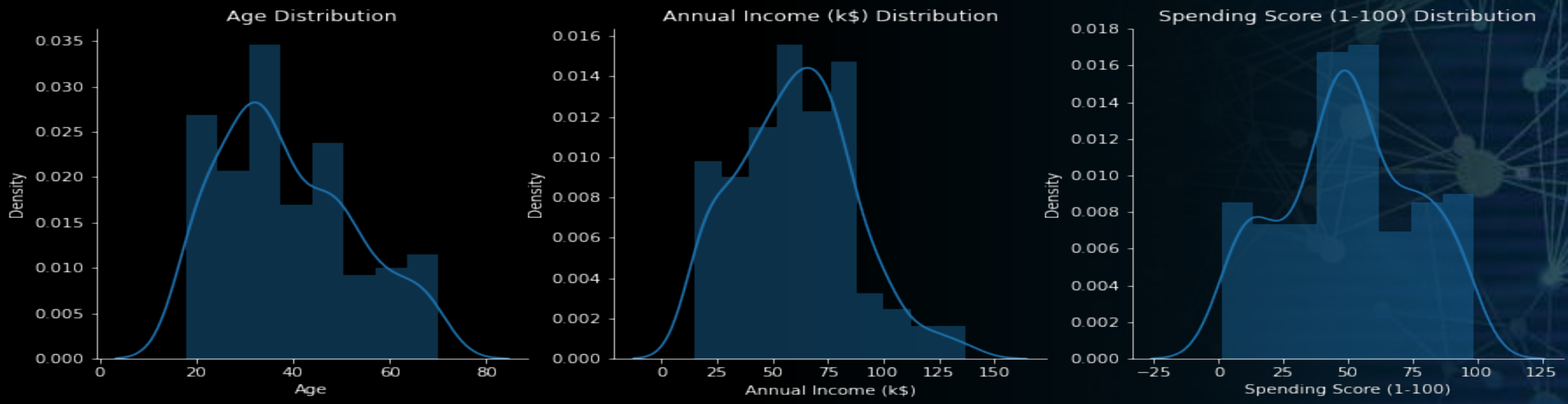
How many observations are there in total?

Are there any missing values?

How many unique values are in each column?

Step 2 (visualizations)

Create the following plots: A histogram of Age, Annual Income (k$), and Spending Score (1-100).

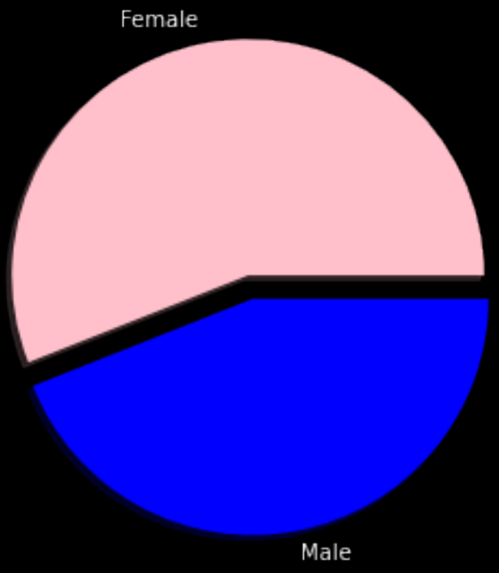


Answer the following:

What can you conclude from the plots you created? Are there any interesting findings?

Visualizations Continued…

Create a pie chart showing the proportions of male to female in our data.

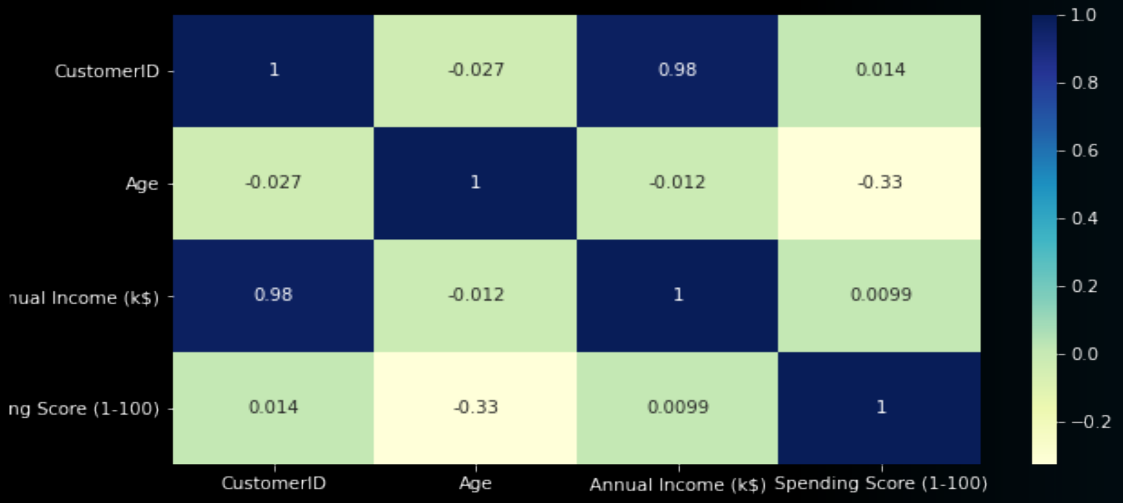


Answer the following:

What can you conclude from the plot you created?

Visualizations Continued…

Create a heatmap of the data.



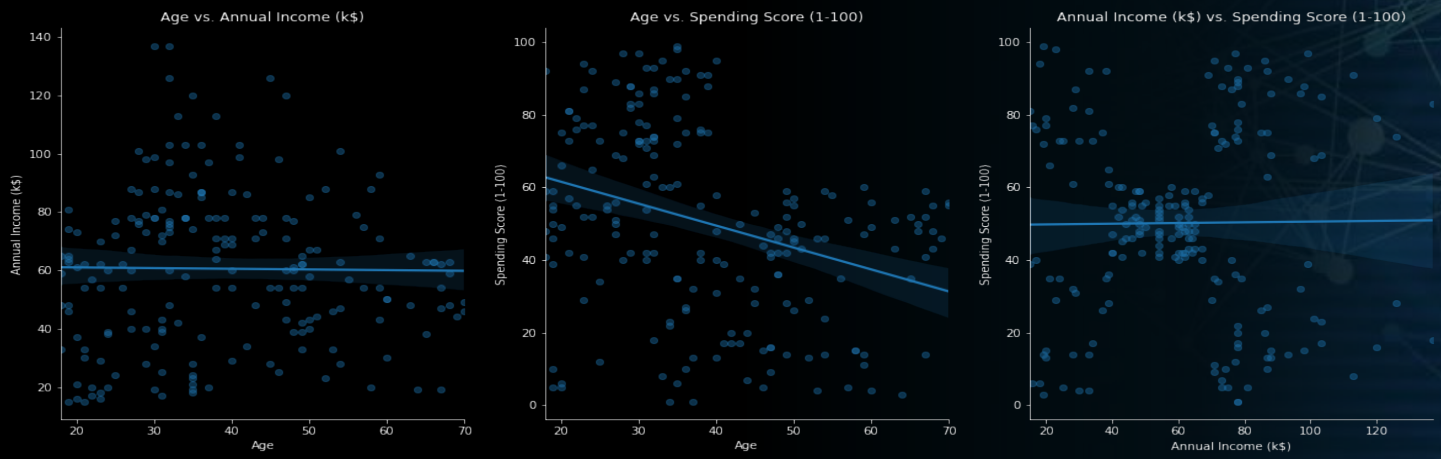
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
CustomerID	1	-0.027	0.98	0.014
Age	-0.027	1	-0.012	-0.33
nual Income (k$)	0.98	-0.012	1	0.0099
ng Score (1-100)	0.014	-0.33	0.0099	1

Answer the following:

After creating your heatmap, is there anything interesting? For example, any significant multicollinearity?

Visualizations Continued…

Create the following plots: A scatter plot of Age vs. Annual Income (k$), Age vs. Spending Score (1-100), and Annual Income (k$) vs. Spending Score (1-100).

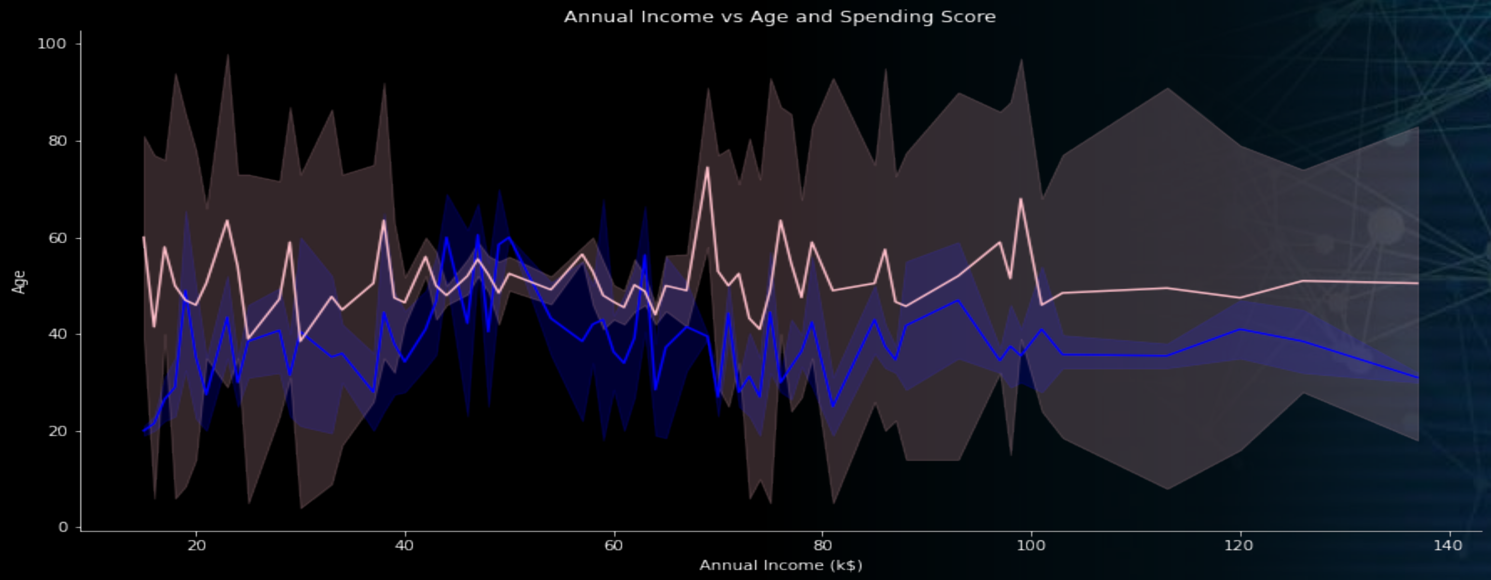


Answer the following:

What can you conclude from the plots you created? Are there any interesting findings?

Visualizations Continued…

Create a line graph of Annual Income (k$) vs. Spending Score (1-100) for both genders.



Answer the following:

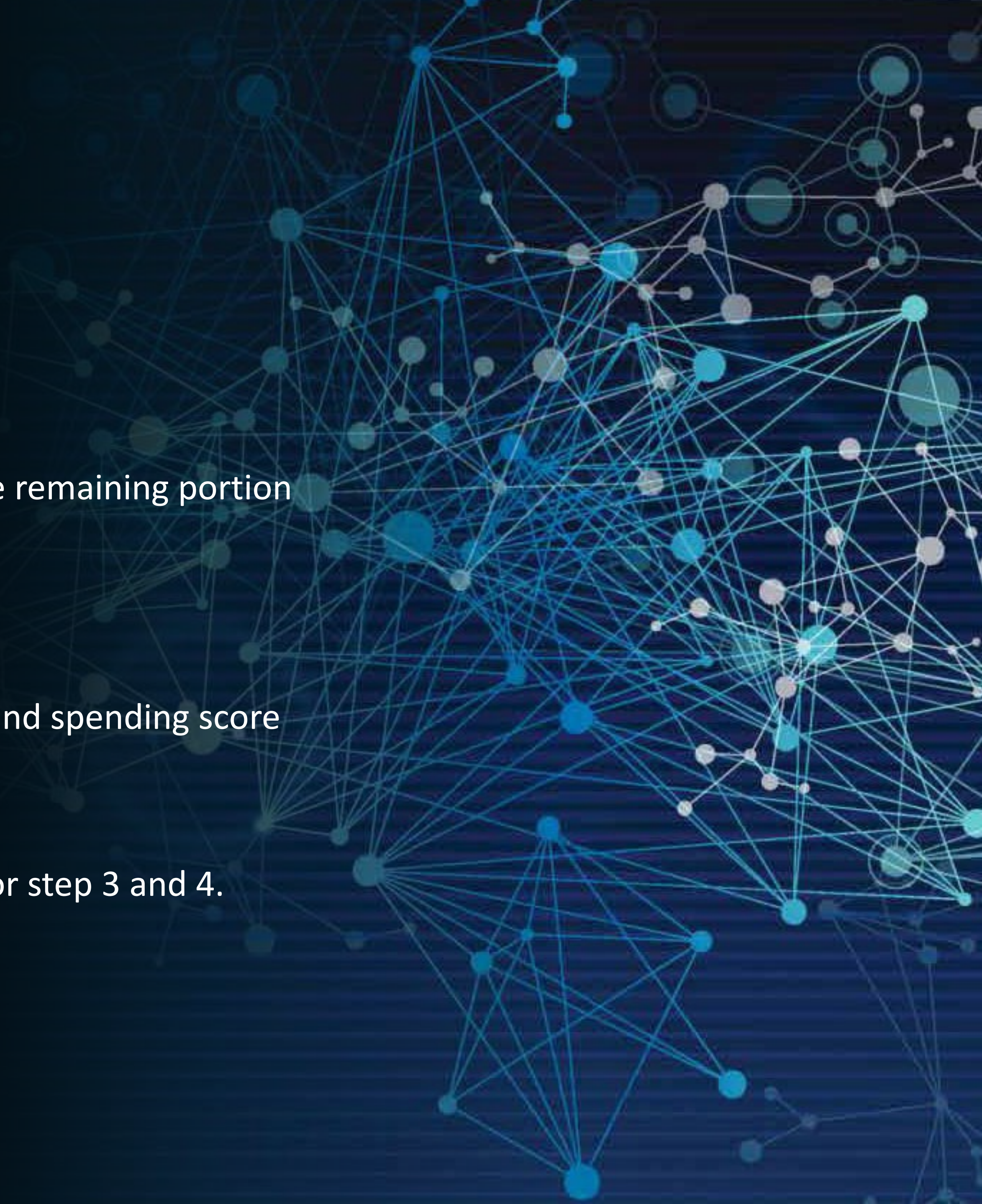
What can you say about the plot you created?

Step 3 (k-means clustering)

For visualization purposes, only use spending score and income for the remaining portion of this homework. For this step, perform the following:

- Use the elbow method to find the optimal number of clusters.
- Create a K-Means model using your optimal number of clusters
- Visualize the clusters by plotting them (annual income on one axis and spending score on the other)

Note: refer back to labs if you’re having trouble creating the clusters for step 3 and 4.



Step 4 (Hierarchical Clustering)

Perform the following:

- Create a Dendrogram to find the optimal number of clusters. Use method = 'ward'.
  Hint: Look at the old lab on hierarchical clustering or scipy.cluster.hierarchy
- Visualizing the Clusters of hierarchical clustering

Step 5 (conclusion)

Write a brief conclusion summarizing your results and how K-Means and Hierarchical clustering differed.

Answer the following:

- Briefly explain how both k-means and hierarchical clustering work along with their advantages/disadvantages if any.
- Give an example of where you could use hierarchical clustering or k-means.