Homework 2

COSC 3337

Dr. Rizk

Problem Statement

Your task for this homework will be to perform exploratory data analysis and predict if a person is prone to a heart attack (0 or 1).

Answer the following:

Is this a classification or regression task and why?

About The Data

The dataset that we will be using for this homework contains the following information:

- age
- sex
- cp: chest pain type
- trtbps: resting blood pressure
- chol: cholesterol
- fbs: fasting blood sugar
- restecg: resting electrocardiographic results
- thalachh: maximum heart rate achieved
- exng: exercise induced angina
- oldpeak: previous peak
- slp: slope
- caa: number of major vessels
- thall: thal rate
- output: target/labels 0 or 1

Step 1

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

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Answer the following using Pandas:

How many observations are there in total?

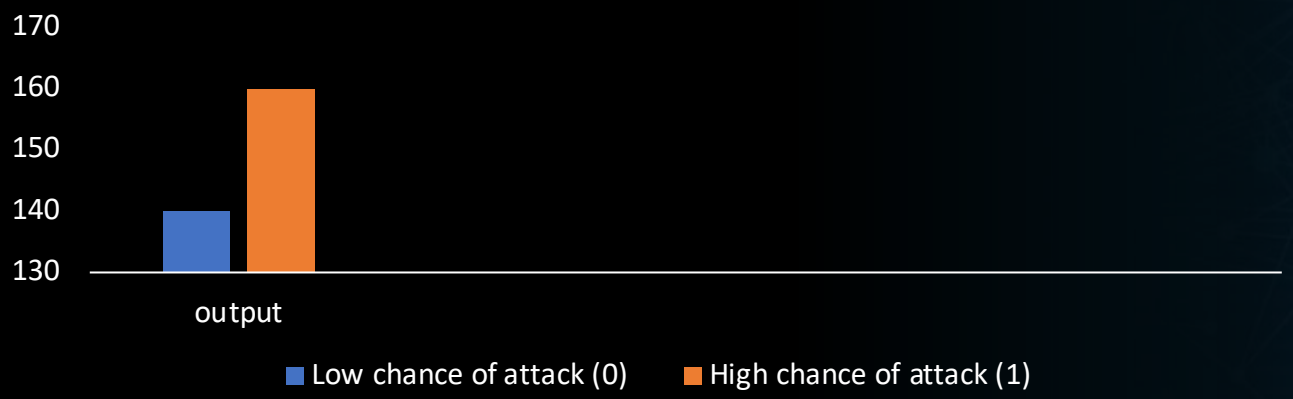
Are there any missing values?

How many unique values are in each column?

Which columns will you treat as categorical, which will you take as continuous, and why?

Step 2

Create a plot of your target variable on the x-axis and counts on the y-axis.



Answer the following:

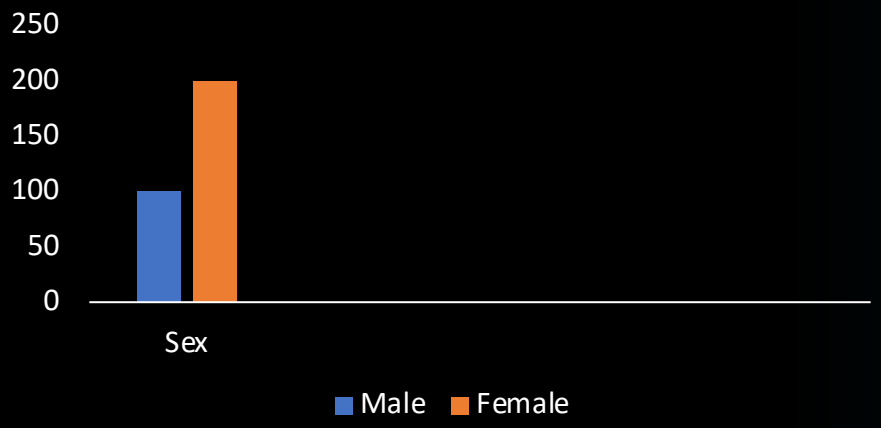
Is the dataset balanced? In other words, is there an equal representation of people prone to heart attacks and those who are not.

Is working with a balanced dataset important? Why or why not?

How can we deal with an imbalanced dataset?

Step 3

Create a count plot for ***each*** of your categorical variables. That is, the variable on the x-axis and counts on the y-axis. Here's an example of sex and slp:

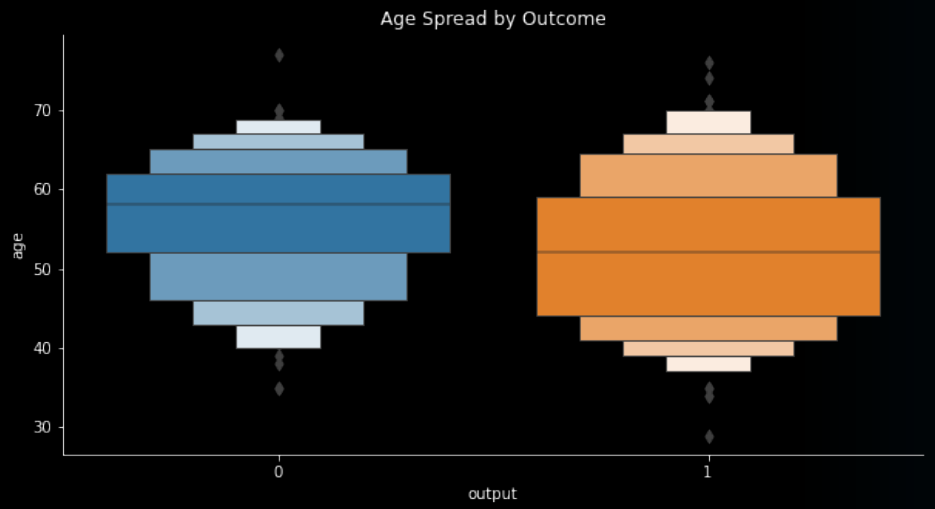


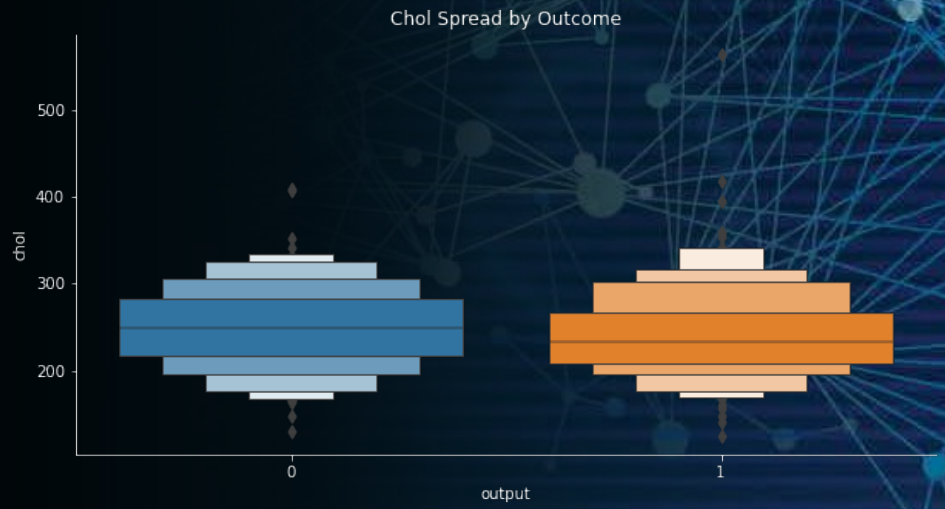
Answer the following:

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

Step 4

Create box plots by outcome for ***each*** of your continuous variables. Here's an example of age and chol:



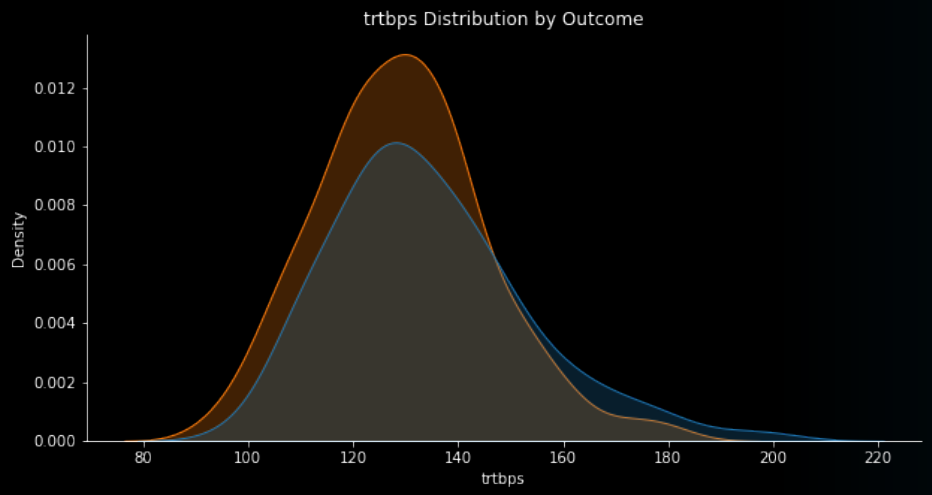


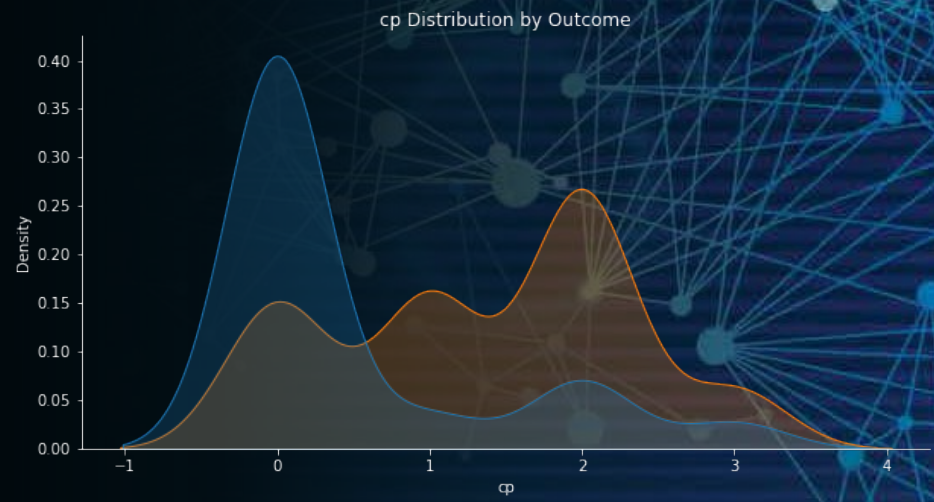
Answer the following:

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

Step 5

Create distribution plots by outcome for ***each*** of your continuous variables. Here's an example of trtbps and cp:



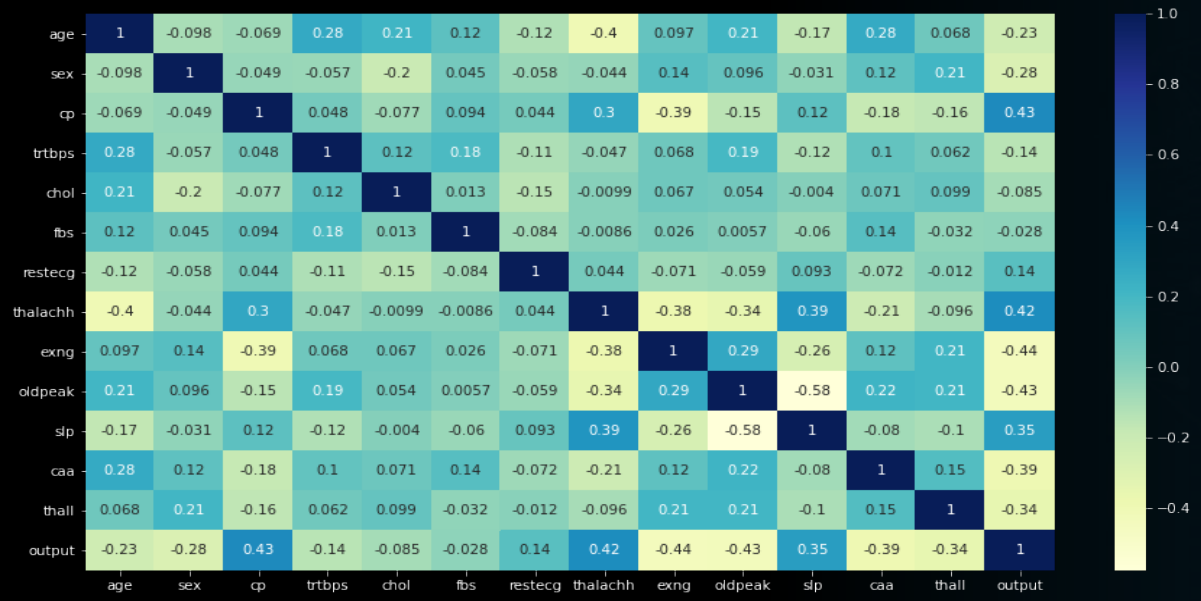


Answer the following:

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

Step 6

Create a heatmap of your data. Here's an example:

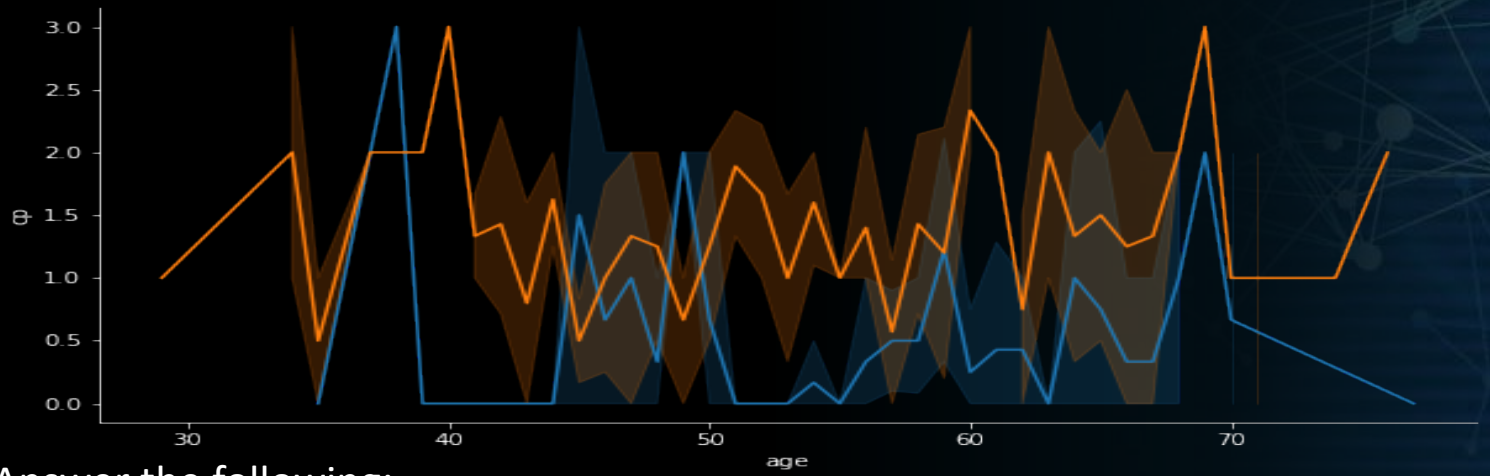


Answer the following:

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

Step 7

Explore the data a bit on your own and include a few additional graphs of your choice. For example, here's a line plot of cp vs. age:



Answer the following:

What can you conclude from the plots that you created? Did you find anything interesting?

Step 8

Answer the following:

Name two different models that you can use to solve the problem statement.

What is the difference between label encoding and one hot encoding, and when should you use one over the other?

What is multicollinearity, and why do we care about it when creating models? How can we check to see if there's significant multicollinearity in our data?

Why is scaling data important?

For the two different models that you named earlier, are they using a parametric or non-parametric learning algorithms? What's the difference?

Suppose that we had missing values in our dataset. What are different ways we could handle them?

Step 9

Choose 2 different models to solve the problem statement. Apply any necessary encoding, scaling, and train test splits to your data and construct the 2 models you selected. Provide a classification report and confusion matrix for both models.

Do the following:

Write a conclusion (~1 paragraph) detailing the main points you discovered while exploring the data. Also include things like: Did you scale your data? If so, which scaling method did you use and why? Is there a specific reason you selected these 2 models? How did the 2 models compare against each other?

Note: Be sure to use cross validation when comparing models. See sklearn's cross_val_score if you're stuck.

Examples of Confusion Matrix Results