Collaborated with: Asuka Li, Daryl Chan

Task 1 - Attribute Description

Person ID
Description: This is the ID column for the data table. Each instance in this attribute is a unique integer and increments by 1.
Type: Nominal

Age
Description: The attribute is a positive integer that reflects the patient/person's age
Type:
Ratio Data; has a meaningful 0 point (just born)

Gender
Description: The attribute reflects a person's gender (has values "M" for male and "F" for female)
Type: Nominal Binary Data

Chest Pain Type
Description: This dataset is on angina chest pains which are caused by the heart not receiving enough oxygen. There are four types of angina chest pains (labeled "Type 1", "Type 2", "Type 3", "Type 4"): Stable, Unstable, Microvascular, and Variant Angina.
Type: Nominal Data; there are 4 categories

Resting Blood Pressure
Description: The blood pressure when the person is at a resting state
Type:
Interval Data - each interval/unit is of equal size
Ratio - blood pressure can't be negative

Serum Cholesterol in mg/dl
Description: Amount of cholesterol content in the blood
Type:
Interval Data - each unit of cholesterol content increases by the same amount
Ratio - Can also have a meaningful zero point; cholesterol content can't be negative

Fasting blood sugar > 120
Description: This is the amount of glucose content in the bloodstream after an overnight fast
Type: Binary Data; two data values (yes or no)

Resting Electrocardiographic results
Description: The state of the heart at resting. There are three types in this data set.
"Hypertrophy of heart" - enlargement of the hearts main pumping chamber
"Myocardia infarction" - Heart attack
"Ischemia" - Inadequate supply of blood

Type: Nominal Data; there are three categories

Maximum Heart Rate achieved
Description: The maximum beats per minute that the heart achieved
Type: Ratio Data; number cannot be negative and can be 0

Exercise-induced angina
Description: Whether or not the patient's chest pain/angina is caused by physical activity
Type: Binary Data; two data values "yes" or "no"

Oldpeak
Description: ST depression induced by exercise relative to rest
Type: Ratio Data; the 0 state is provided by a neutral/no problem state

The slope of the peak exercise ST segment
Description: There are three types of slopes: upward sloping "1" flat "2" and downward sloping "3"
Type:
Nominal; three categories of data

Number of major vessels colored by fluoroscopy
Description: Fluoroscopy is a real time video of an X-ray of the body; this attribute tracks how many vessels are colored (integer from 0-3)
Type:
Ordinal Data; since the number of vessels 0, 1, 2, 3 are in order increasing
Numeric Interval - Can also argue the numeric side since there could be more vessels than just 3 (should probably use this for dissimilarity values)

Thal
Description: Thal is a blood disorder where the body has fewer oxygen-carrying protein; there are 3 types in this data set "3" - normal, "6" - fixed defect, "7" - reversible defect
Type: Nominal Data

Has heart disease
Description: Whether or not the patient has heart disease (Values: "yes" "no")
Type: Binary Data

Task 2 - Proximity Selection

Age
Resting Blood Pressure
Serum Cholesterol
Maximum Heart Rate

Oldpeak

Assuming that we are only looking at the numeric attributes separately without considering its effect on heart disease, we would use Euclidean distance. When asking the question: “Are two people similar by their age, blood pressure, cholesterol, heart rate, and oldpeak”, they are only similar if their actual numbers are similar. For example, when looking at age, two people are similar only if their ages are similar; we cannot make separate inferences about the two people otherwise. We also don’t know if one attribute should be weighted more than another, therefore, the Euclidean distance of two data instances will tell you the dissimilarity. Hence, we should have variance in scaling and translating the numerical values for our attributes.

Task 3 Dissimilarity

Part 1

What does the code look like?

- I used python to create a nested loop to create a combination of all of the pairs of data instances
- For each of the attributes (not including the ID column; due to the fact that this column doesn’t provide value) I calculated the dissimilarity between the two pair values
    - For numerical attributes, I took the absolute difference between the two values
    - For nominal attributes, if the values were equivalent, the dissimilarity is ‘0’ and if they were not, the dissimilarity is ‘1’
- The attributes I used absolute difference were the 5 attributes mentioned above along with “Number of Major vessels”, I used the ‘0’ and ‘1’ dissimilarity for the rest of the attributes

Part 2

What does the code look like?

- Also used python; I looped through each data instance, and added all of the differences. At the end, I divided by the total number of attributes just like the formula
- This creates a column of dissimilarity values for each pair of data instances (~10k rows)

What does each dissimilarity tell you?

- The aggregate dissimilarity value tells us how different one data instance is from another. Since we normalized the data, each attribute is weighted equally (even if an attribute had values that were big, this doesn’t matter with attributes that have small values).
- Since we used the absolute distance between the numbers, the value tells us how far apart the actual numeric values are from each other

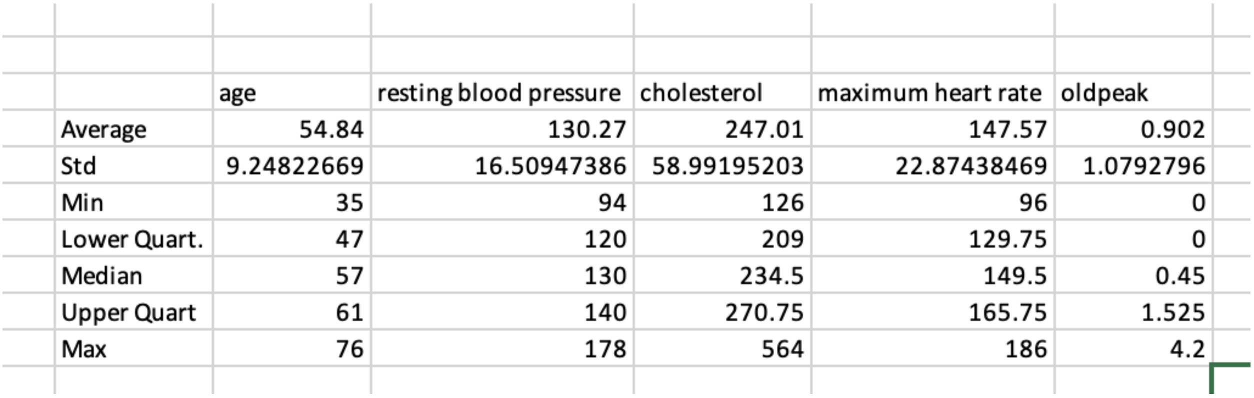
Largest Dissimilarity Pair

- 0.74953656

Smallest Dissimilarity Pair

- 0.03456872

Task 4 - Summary Statistics

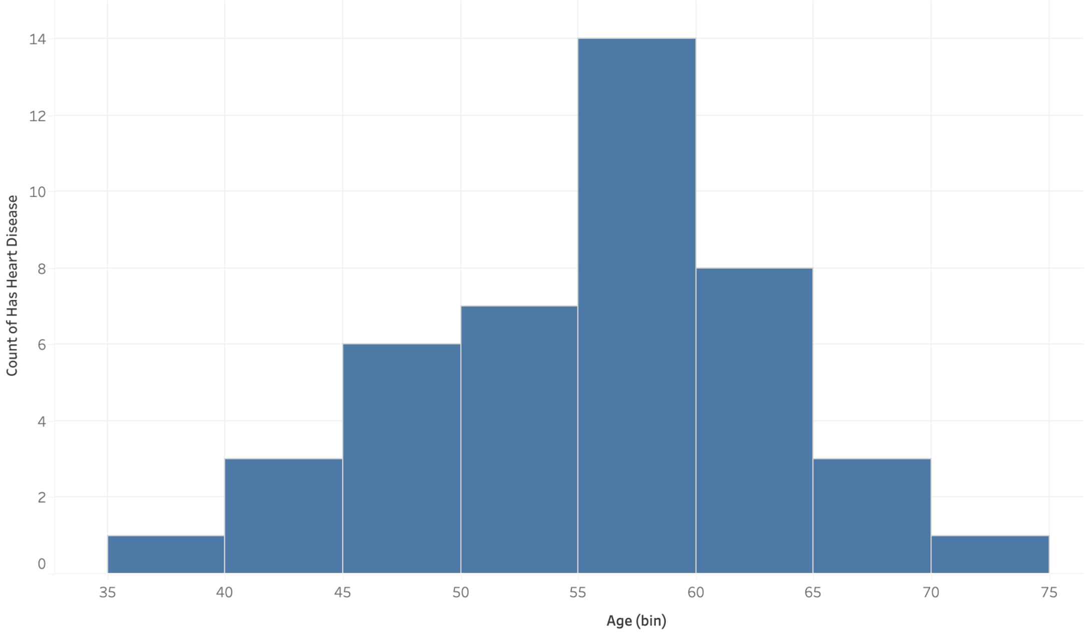


	age	resting blood pressure	cholesterol	maximum heart rate	oldpeak
Average	54.84	130.27	247.01	147.57	0.902
Std	9.24822669	16.50947386	58.99195203	22.87438469	1.0792796
Min	35	94	126	96	0
Lower Quart.	47	120	209	129.75	0
Median	57	130	234.5	149.5	0.45
Upper Quart	61	140	270.75	165.75	1.525
Max	76	178	564	186	4.2

Task 5 - Charts

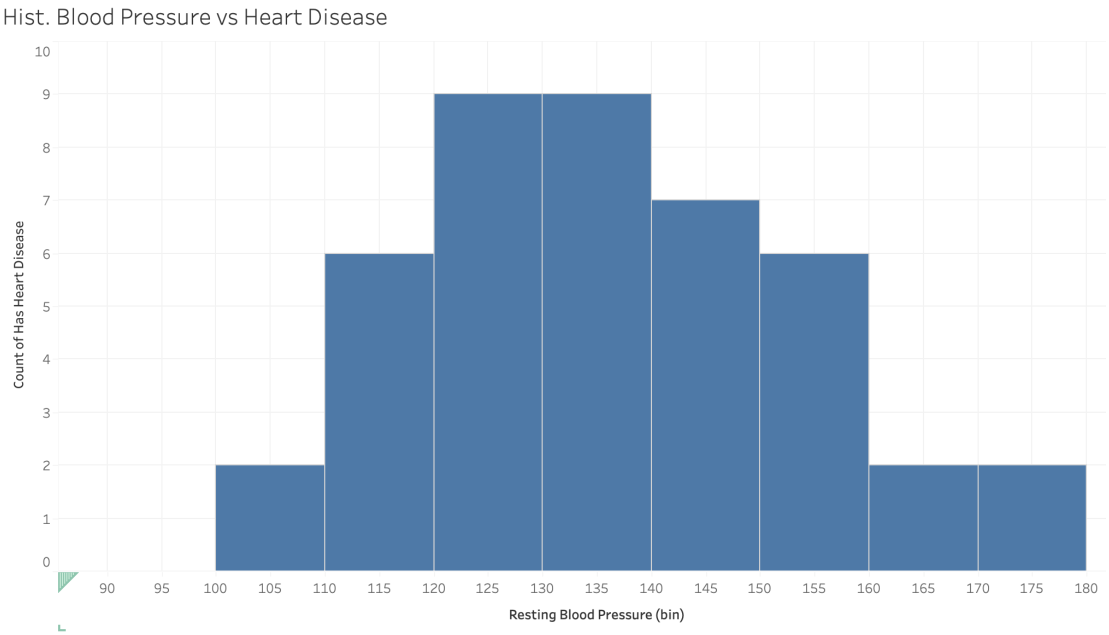
Chart 1

Hist. Age vs Heart Disease



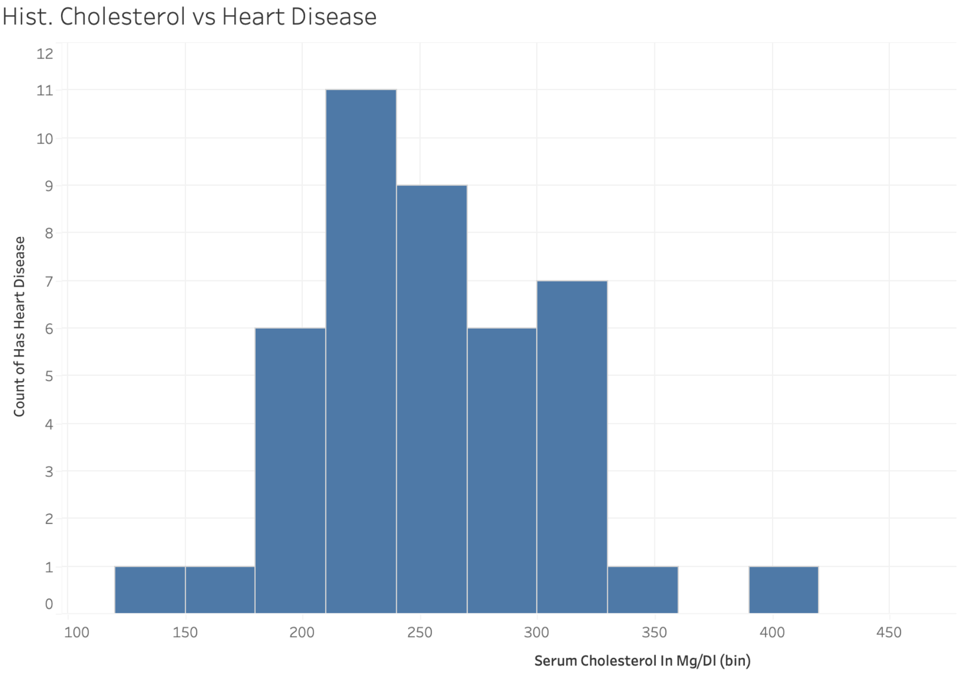
- Plots the amount of people who have heart disease at different age ranges
- To see whether or not old age may have a significance to heart disease
- Observation: Looks like there is a skew towards older age

Chart 2



- Plots Number of Heart Disease for different Blood Pressure ranges
- See whether blood pressure is significant to heart disease
- Observation: Seems like the 135-140 range is the mean/median for heart disease

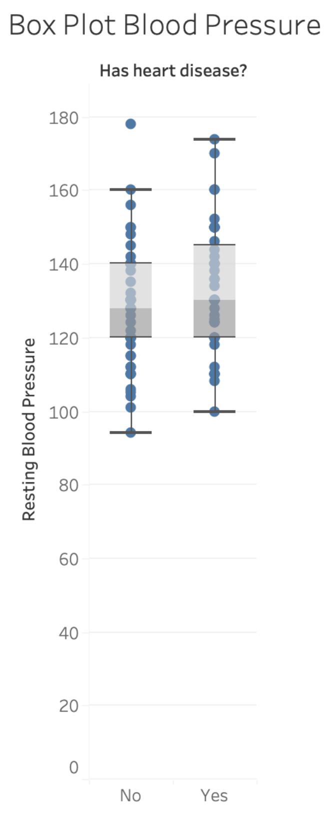
Chart 3



- Plots number of people with heart disease with amount of Cholesterol
- See whether cholesterol is significant with heart disease

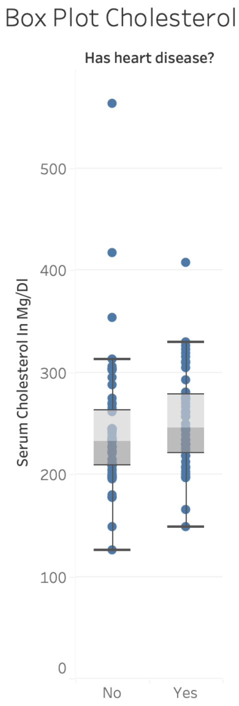
- Observation: Higher cholesterol seems like it has a significance to heart disease

Chart 4



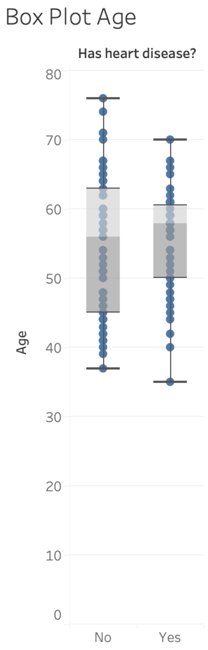
- With box plots we can see each attribute among people with and without heart disease for better comparison
- Whether blood pressure has an effect on heart disease
- Observation: Looks like the blood pressure is higher among those with heart disease

Chart 5



- We would like to see if cholesterol has any effect on heart disease
- Observation: heart disease does seem to be implicated by higher cholesterol levels

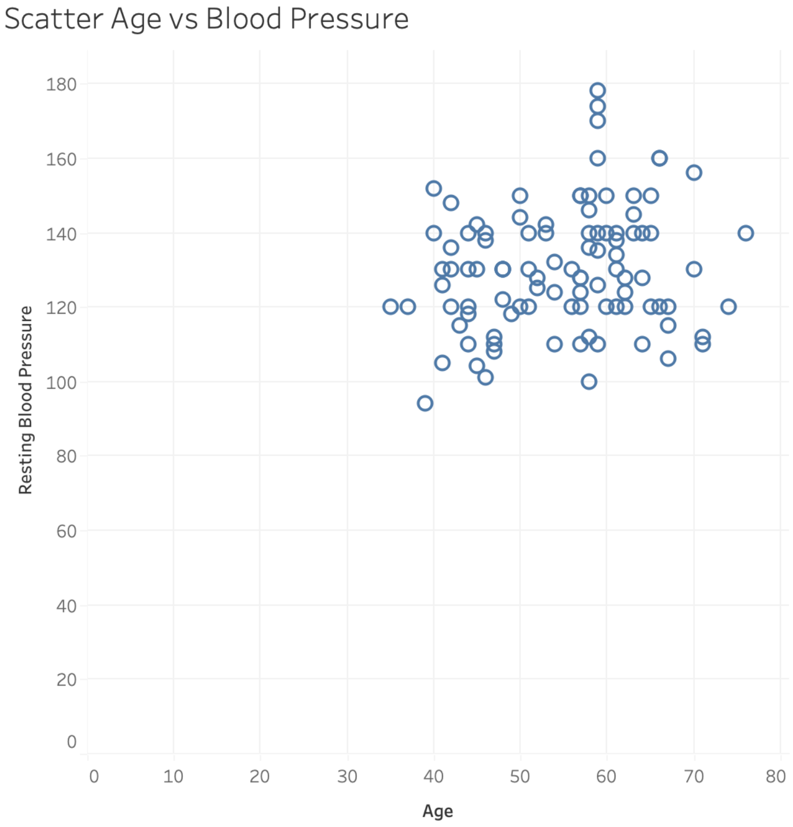
Chart 6



- Would like to see if age has anything to do with heart disease

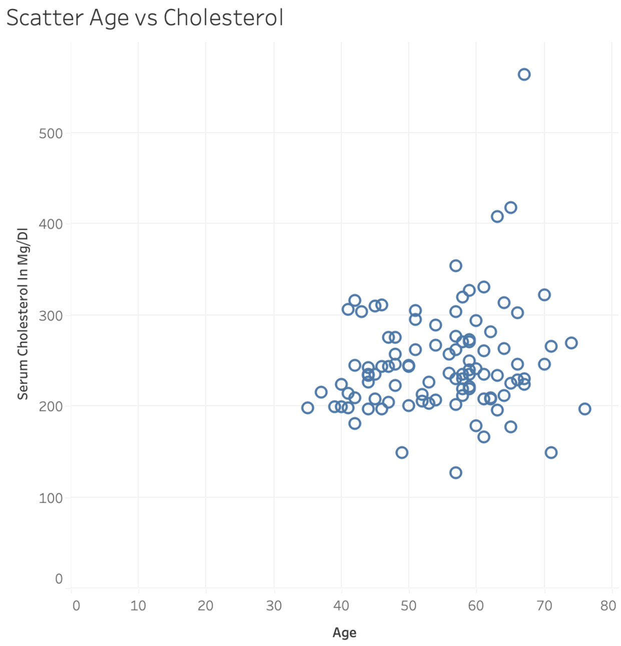
- Observation: It does not seem so, although the median for heart disease has a higher age, overall, the age instances are quite similar

Chart 7



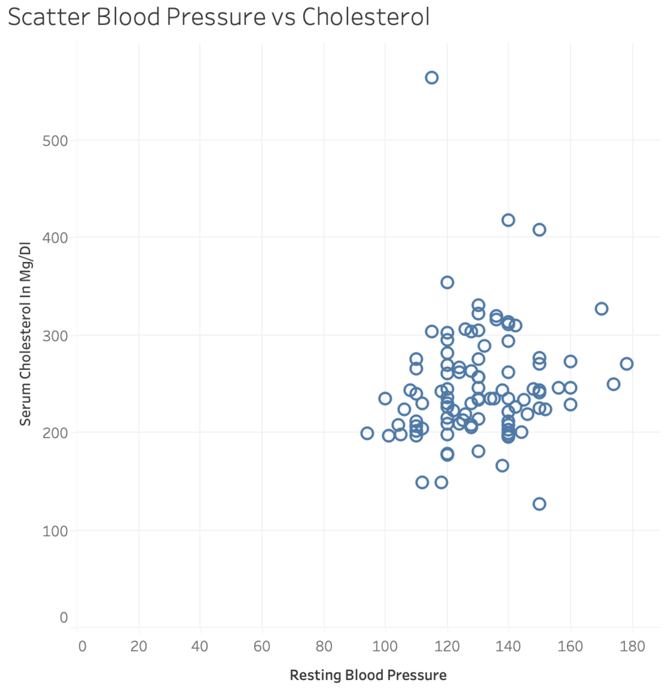
- Displaying a scatter plot that shows the age against blood pressure to see the correlation between the two numeric attributes
- Observation: Seems to have a positive correlation (higher age → higher blood pressure)

Chart 8



- Plotting age against cholesterol so that we can see the correlation between the two numeric attributes
- Observation: Has a positive correlation but is not as high as age and blood pressure

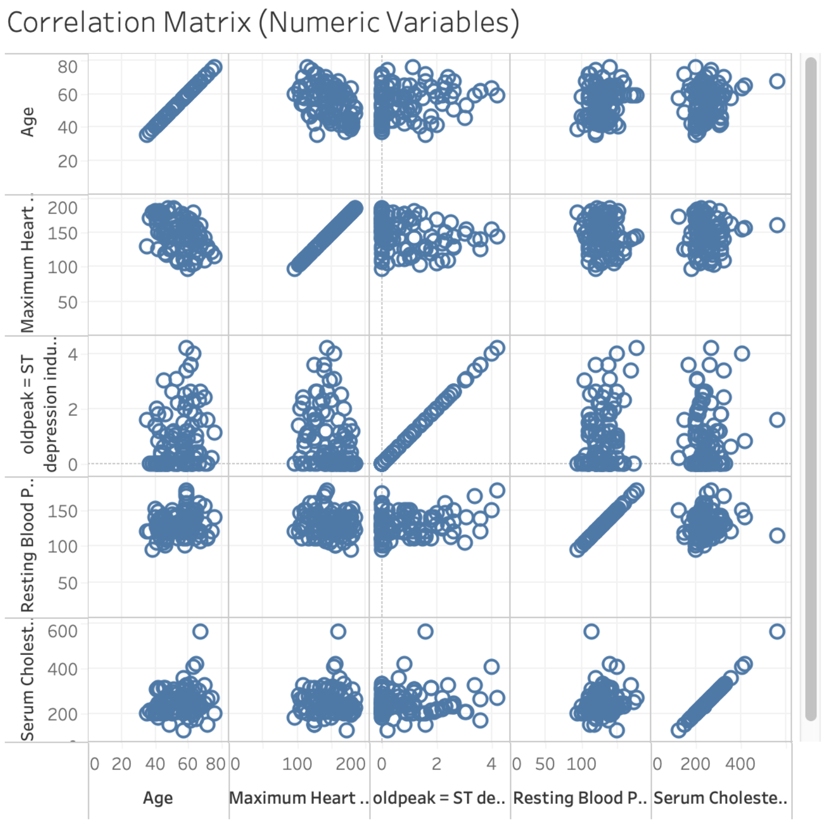
Chart 9



- Blood pressure against Cholesterol
- We would like to know if blood pressure and cholesterol have any correlation

- Observation: there is a slight positive correlation between the two attributes

Chart 10



- Although I already have 3 scatter plots, I decided to include this as my last graph since it gives an overall summary of the correlations between the numeric attributes
- Most of these correlations have a positive correlation with each other and it seems like old peak is the only attribute that has a neutral (0 correlation) with the other attributes

Task 6 - Tools and Languages

Dissimilarity

- I used Python to compute the dissimilarity between each pair of data instances
- The code was quite long; I didn't expect it at first, but I realized we had to normalize the data, compute the absolute difference, as well as compute the sum and division of total attributes.
- I initially tried to use Excel (I thought it would be simpler), however, I could not find a way to loop through all of the data instances and generate pairs, which is why, in the end, I decided on Python.

Summary Statistics

- I used Excel to compute to each number for the summary statistic
- Process was very simple: Create a data table with only the numeric attributes and compute each number using the average, stdev, and quartile formulas built into Excel

- I would say since there are formulas in Excel for these stats, it would be more efficient to use Excel since there is minimal coding

Charts

- I used Tableau to create the histograms, scatter plots, box plots, and correlation matrix
- Tableau is very logical to use - put the attributes in the x and y axis as you wish and you can pick the graph type you want.
- Compared to Python and R, Tableau requires minimal code, therefore, I would say it was the most efficient software to use. Compared to Excel, I think Tableau creates better looking graphs/charts and is a bit more customizable.