INDICATORS OF HEART DISEASE (FUNDAMENTALS OF DATA VISUALIZATION - FINAL PROJECT)

INTRODUCTION AND BACKGROUND

The data used in this project derives from the dataset titled "Heart Failure Prediction Dataset", which can be accessed from the following source: https://www.kaggle.com/fedesoriano/heart-failure-prediction. The dataset is a combination of healthcare data, provided by the following locations: Cleveland (Ohio, USA), Hungary, Switzerland, Long Beach (California, USA), and UCI (University of California Irvine). This dataset contains patient information and provides details on their physical characteristics and health status, including whether or not if they have heart disease.

TASKS AND OBJECTIVES

The main task is to assess common attributes among patients who have and do not have heart disease, with the concluding objective of establishing a pattern of common indicators among the given patients as well as among different age groups. For the development process, a relative reference frame is used to compare selected data to other data from the dataset. The task influences the visualizations by including comparative analysis, which will result in the visualizations having data respresentation for patients who have and do not have heart disease. By dividing the dataset into these two categories, the visualizations will reveal significant patterns and commonalities from indicators of heart disease, such as cholesterol levels.

This task was conducted with the use of Python (specifically with the Python package Altair), Jupyter Notebook, and Binder to generate the relevant visualizations. To gain full access to these visualizations, access the following source:

https://mybinder.org/v2/gh/IsraelsLibrary/Fundamentals_of_Data_Visualization/2c6e36bfc4c9ba1f9866744d69e31da1bffb7bd4?urlpath=lab%2Ftree%2FFundamentals_of_Data_Visualization_Final_Project.ipynb

Before running the file "Fundamentals_of_Data_Visualization_Final_Project.ipynb", make sure that the following Python libraries are installed: pandas, altair, and numpy. If they are not installed, open a terminal window in the Binder version of Jupyter Notebook and execute the following command: pip install pandas altair numpy

DESIGN APPROACH

The dataset contains the following attributes: Age, Gender, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting ECG Levels, Maximum Heart Rate, Exercise-Induced Angina, Old Peak, ST Slope, and Heart Disease Status. For the visualizations, certain attributes from this dataset will be emphasized, which include the following: Age, Gender, Chest Pain Type, Cholesterol, Maximum Heart Rate, Resting Blood Pressure, Exercise-Induced Angina, and Heart Disease Status. I included resting blood pressure as an additional attribute later in the project after further *research confirmed how resting blood pressure is a great indicator for patients at risk of heart attacks (and this will serve as my justification for this additional element). As for the gender attribute, I discovered that there is not a full representation, because non-binary and other genders were not included in the original dataset. The dataset is only limited to two genders that were provided: male and female.

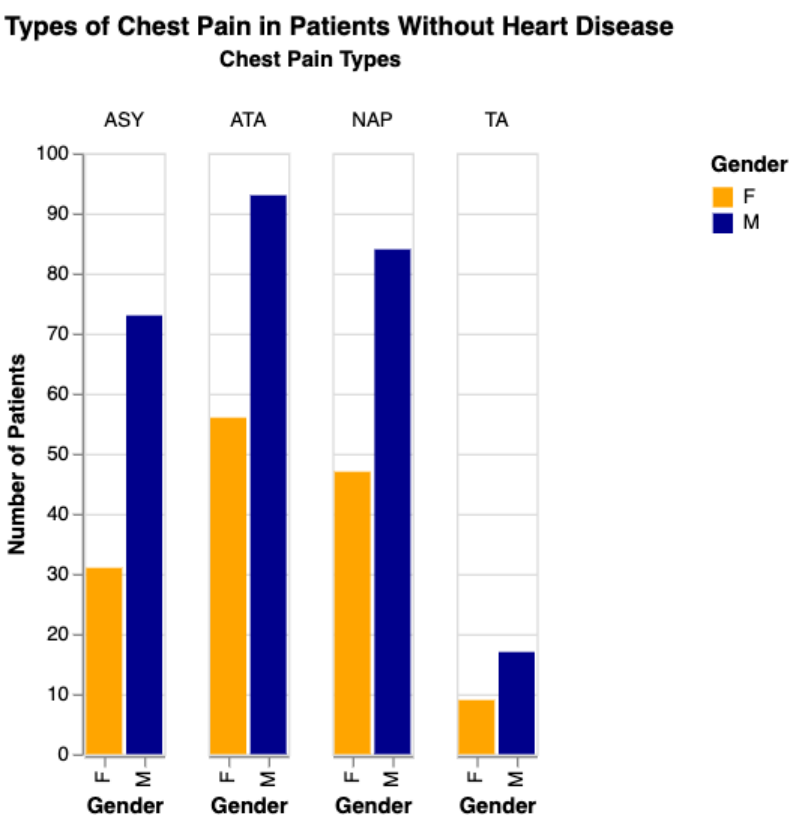
As a result, the following visualizations and analysis reflect this detail for data related to the two given genders. The justification for selecting these other attributes includes the goal to see connections between the selected attributes and how they impact each other. They include some of the best indicators of heart disease. According to the **Mayo Clinic, chest pain and chest discomfort (angina) are common symptoms found in patients who have heart disease. In addition to these symptoms, the Mayo Clinic also mentions that one of the main risk factors that lead to a development of heart disease is high cholesterol levels. With these selected attributes, users will gain a better understanding of the patterns formed by these attributes once revealed through the visualizations. To generate the desired visualizations, the first step would be to split the dataset into two subsets of data: patients without heart disease and patients with heart disease. Development will include mapping the data for both categories of patients and revealing their attributes. Multiple visualizations will cover the selected attributes from the data. Once the visualizations are made, they will reveal key patterns for both categories of patients. Low-fidelity prototypes and other related resources for this project can be accessed at the following GitHub source: https://github.com/IsraelsLibrary/Fundamentals_of_Data_Visualization

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

VISUALIZATIONS

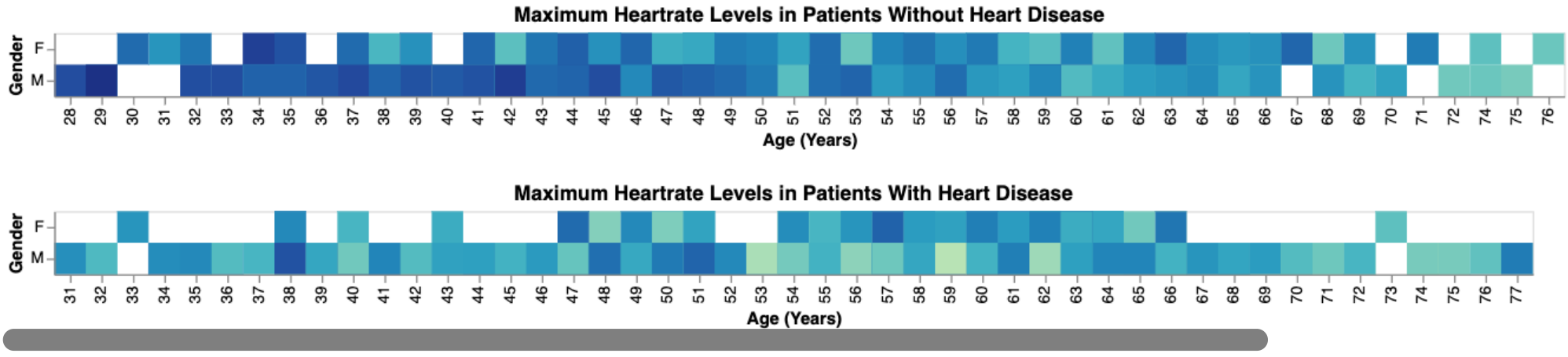
For the first set of visualizations, these include two grouped bar charts that reveal the number of patients that suffer from different types of chest pain. The chest pain categories are as follows: Asymptomatic (ASY), Atypical Angina (ATA), Non-anginal Pain (NAP), and Typical Angina (TA). Interactive features of the charts will allow users to hover over any one of the bars from either barplot, and they will receive information on the number of patients that suffered a specific type of chest pain as well as the gender that group is affiliated with. These visualizations can be seen from below.

From these visualizations, we can draw the following conclusions: 1.) in both categories of patients with heart disease and patients without heart disease, more male patients have suffered from various types of chest pain than female patients, 2.) atypical angina is the most prominent type of chest pain among patients without heart disease, and 3.) the highest number of patients with heart disease have suffered from asymptomatic chest pain

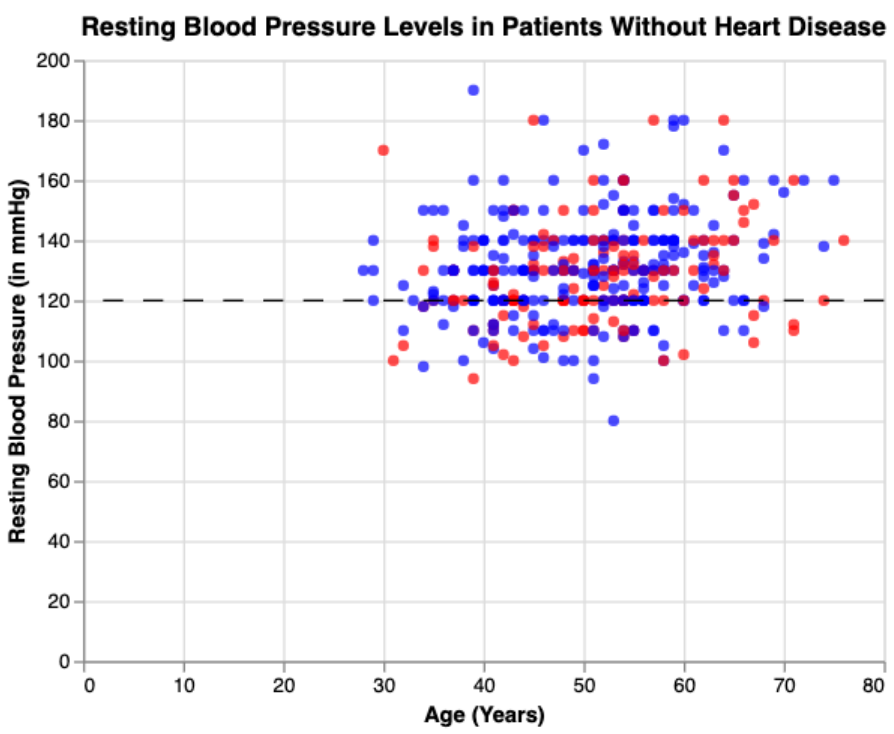


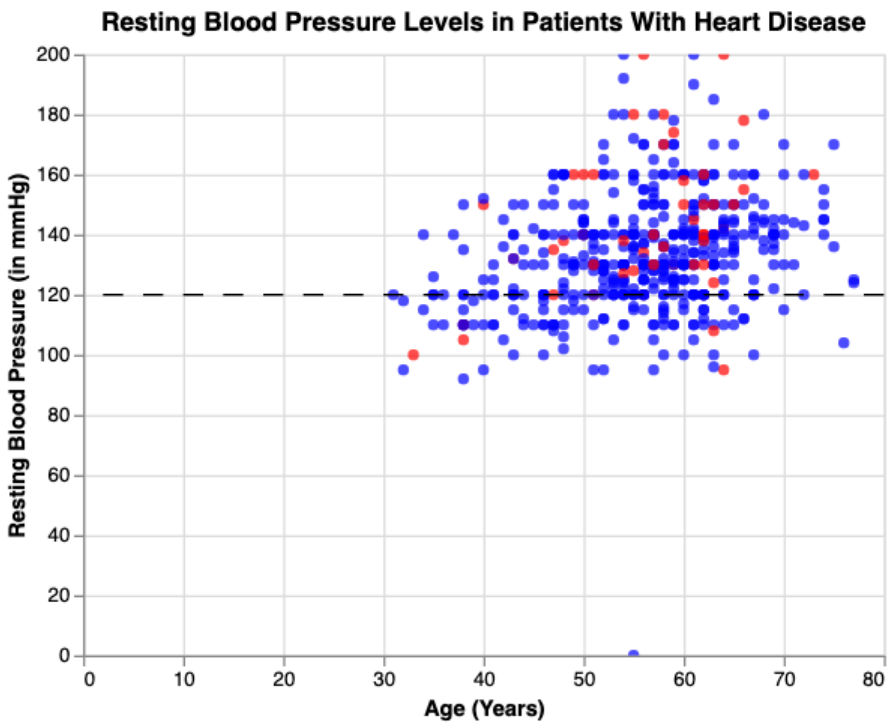
As part of the second set of visualizations below, these included two heatmaps, with one representing patients without heart disease and the other heatmap representing patients with heart disease. This set of heatmaps utilizes a divergent color scale to not only show maximum heartrates for each patient, but to also provide an overview and high density of common heartrate levels for patients of a given gender and age group. Entries in both heatmaps that appear to be white indicate that there was no recorded entry for a patient of that given gender and age. While there has been much debate on whether or not maximum heartrate is a reliable indicator of heart disease, extremely high or low maximum heartrates have still been proven dangerous. According to an **article released by Harvard Health Publishing of Harvard Medical School, people who have a low maximum heartrate are at a higher risk of experiencing a heart attack or even death.

The highest maximum heartrate on average is considered to by 200 bpm (beats per minute). That knowledge lead to further analysis, which revealed that only one patient had a maximum heartrate above 200 bpm, who was a 29-year-old male patient without heart disease. The heatmap visualizations led to the following conclusions (despite how insightful these conclusions were, they did not show any relevant patterns related to heart disease nor confirmed if maximum heartrate levels are a reliable indicator of heart disease): 1.) higher maximum heartrate levels were found in patients without heart disease, 2.) for patients without heart disease, the greatest density of high maximum heartrate levels appear to be for patients between ages of 28 and 56, and 3.) only one patient (male patient of age 29, and without heart disease) was found to have a maximum heartrate above 200 bpm. However, the actual maximum heartrate (202 bpm) was relatively close and not significantly higher.



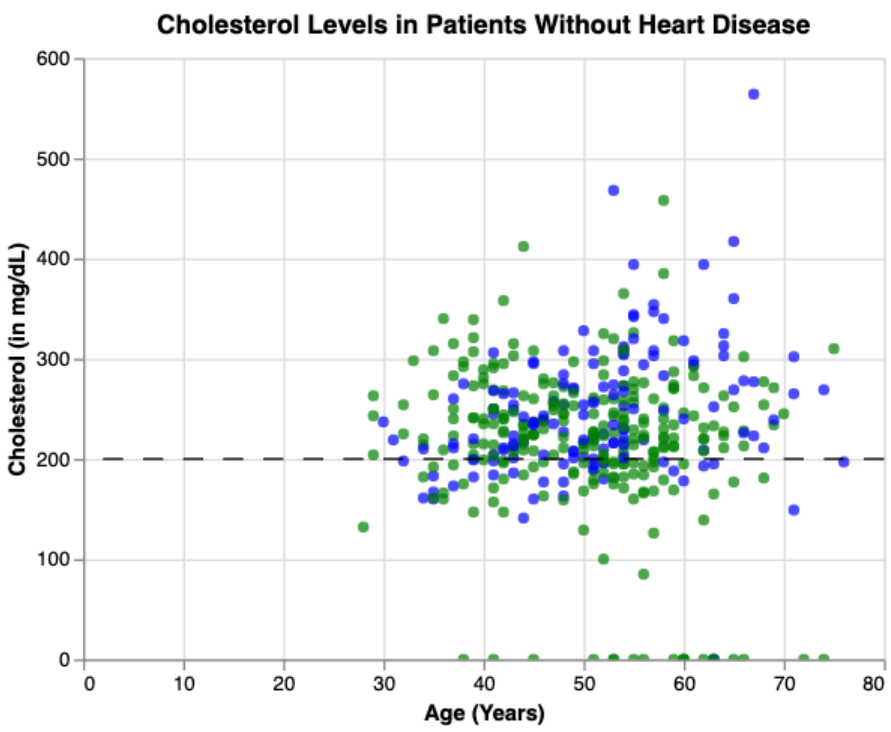
For the third set of visualizations shown below, the plots focus on resting blood pressure levels for patients among different ages. Two scatterplots are provided (one representing patients with heart disease and the other that represents patients without heart diseases) along with thresholds where entries above the thresholds (120 mmHg or millimeters of mercury) are considered to be in the category of resting high blood pressure. The third set of visualizations and further analysis led to the following conclusions: 1.) overall, more male patients were found to have high levels of resting blood pressure than female patients, and 2.) the age range of patients with high levels of resting blood pressure appear to be similar between patients with heart disease and patients without heart disease. This age range appeared to be for patients from age 30 up to age 77 (with the exception of four patients who appear to be in their late 20s.)

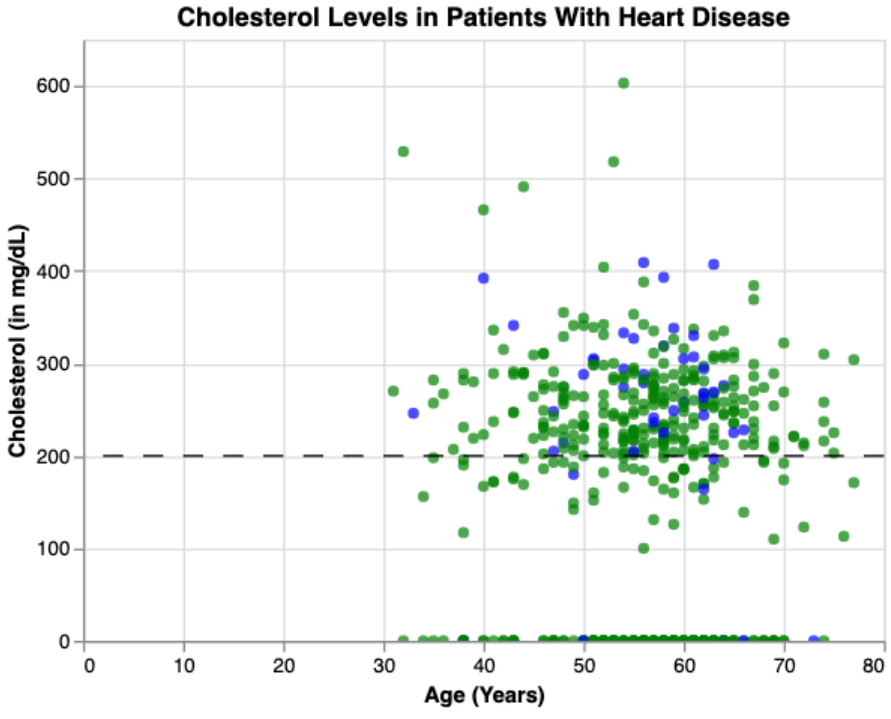




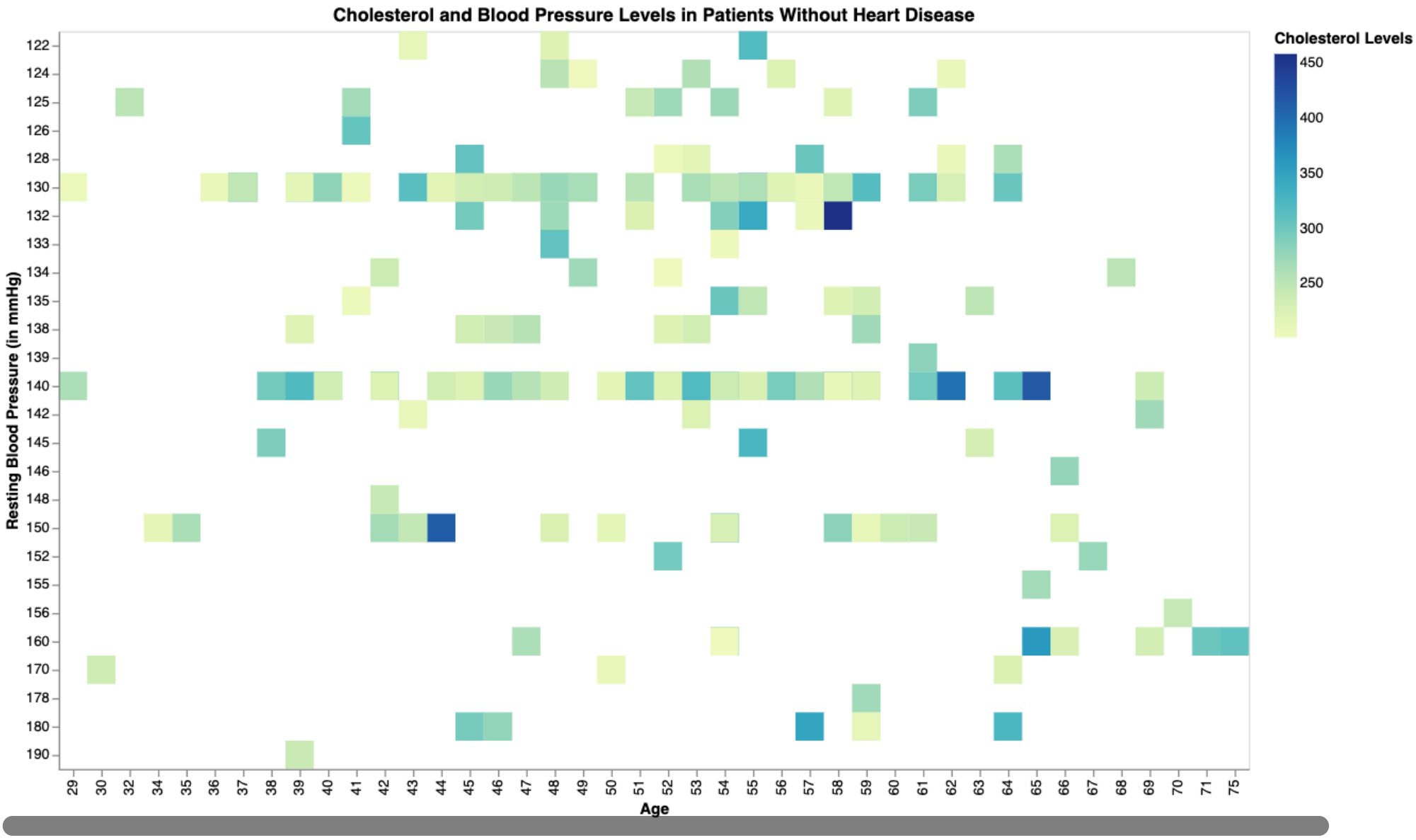
According to ***Johns Hopkins Medicine of Johns Hopkins University, the biggest indicator of heart disease is high cholesterol levels. This attribute will be the main focus of the final set of visualizations for this project. In the following visualizations, the last two sets of visualizations include the following: two scatterplots that display the cholesterol levels (in mg/dl or milligrams per deciliter) for patients and two heatmaps that show the relationship between high blood pressure and cholesterol levels. In addition to the plots of different patients, the scatterplots also include a threshold (200 mg/dl). Any plots above the thresholds indicate high levels of cholesterol for a given patient. However, there was a critical discovery after visualizations were made, particularly with the scatterplots. Many of the data entries revealed to have a cholesterol value of zero, which indicates that there was an error with the input data when the source dataset was originally formed. Unfortunately, in this situation, this prevents any further quantitative analysis, and we will have to conduct qualitative analysis to make general comparisons and find commonalities.

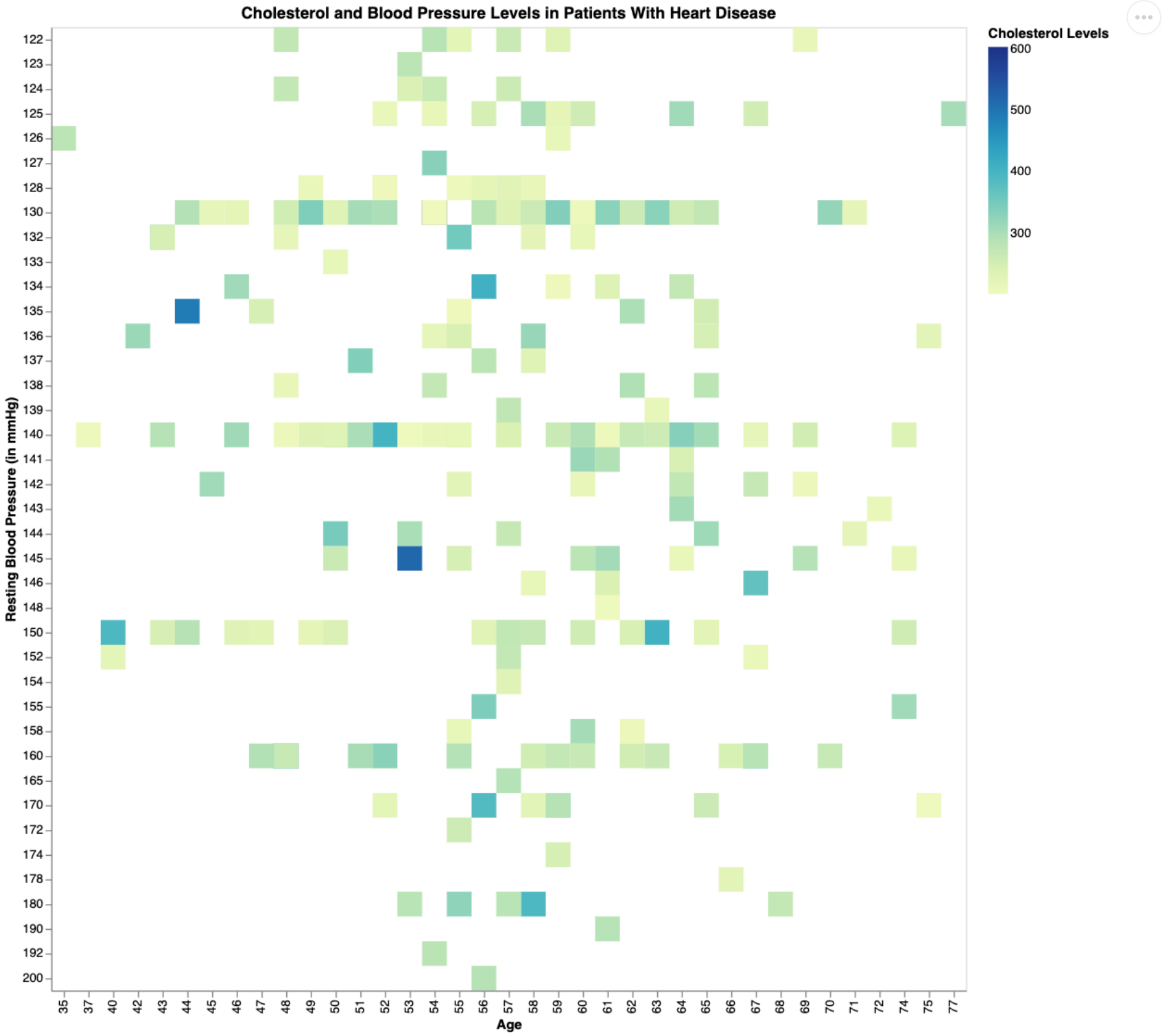
The advanced features of the charts will help users to get a better understanding of how high cholesterol levels are distributed among patients. Users can zoom in on the scatterplots and hover over different plots to gain more information on the age, gender, and cholesterol level of each patient.





After focusing on the results from the previous visualizations, the decision was made to transform the data further to focus only on patients who have high cholesterol levels and high blood pressure. The heatmaps shown below reveal some interesting commonalities of both attributes, when compared between the results for patients with heart disease and patients without heart disease. From both sets of visualizations related to cholesterol as well as extended analysis, the following conclusions were made: 1.) male patients are more likely to have high cholesterol and high blood pressure than female patients, 2.) high cholesterol is mostly to be found in patients who have a blood pressure of 130 mmHg or 140 mmHg (in both heart disease patients and patients without heart disease), and 3.) the highest concentration of patients (with heart disease and without heart disease) are found to be between the ages of 30 and 77 (with outliers being in the late 20s for patients without heart disease).


Cholesterol and Blood Pressure Levels in Patients Without Heart Disease
Resting Blood Pressure (in mmHg)
Age
Cholesterol Levels
450
400
350
300
250


Cholesterol and Blood Pressure Levels in Patients With Heart Disease
Resting Blood Pressure (in mmHg)
Age
Cholesterol Levels
600
500
400
300

FINAL EVALUATION

The evaluation process includes a journaling study approach for qualitative evaluation. Due to limited availability of certain experts, I was unable to recruit desired experts for this evaluation. However, I was able to recruit the following experts of certain fields: an artist/graphic designer, a healthcare marketing specialist, and a clinical data analyst.

For the artist, he found the visualizations to be well organized. He did recommend to modify the positions of the legends on certain graphs. Based on previous knowledge and information that he learned on the subject matter in the past, the data appeared to be accurate. He also noticed that the gender attribute did not have a full repesentation, and lacked the inclusion of nonbinary and other genders. Since this data did not include other genders, this could potentially impact the accuracy of overall data and its representation of patient records.

The healthcare marketing specialist found the visualizations to be informative in her presentation of the data. She did bring forth a suggestion of additional data which can be included in future datasets: whether or not if patients were taking medication at the time that the data was being recorded. This bit of data could provide more insight and change the data patterns that are displayed through the visualizations (i.e. medication that could impact heartrate levels or cholesterol levels). This would be key information to look for in similar datasets when future visualizations are made.

From the perspective of the clinical data analyst, he found the data to be very clarifying and that the visualizations have a great layout. He stated that most of the chosen attributes for the visualizations were great selections, and that they help to form data patterns that users can see for key indicators of heart disease. He did provide suggestions for future visualizations that have more advanced capability, such as nomograms, which provide visualizations of mathematical computations. In medicine, nomograms rely on biological and clinical data to generate a probability of a certain clinical event, such as cancer recurrence or other forms of prognosis.

To conclude this evaluation, I was able to answer the target question of this project: are there characteristics or common attributes from the data that can serve as good indicators of heart disease? Although not all of the attributes were considered to be great indicators, the following attributes were found to be great indicators: cholesterol, blood pressure, age, and gender (to a degree). I was able to meet all criteria that was established beforehand, which includes the following: revealing patient information for patient attributes that exceed a certain threshold, revealing the differences between patients with heart disease and patients without heart disease, and reveal which age groups are at high risk of heart disease.

SYNTHESIS AND CONCLUSION

In conclusion, the data visualizations resulted in the following discoveries: male patients are more likely to suffer from indicators that lead to heart disease, asymptomatic chest pain occurs in a high number of both heart disease patients and patients without heart disease, and the combination of high blood pressure and high cholesterol would mostly occur in patients between ages 30 and 77. Cholesterol, age, gender, resting blood pressure, and types of chest pain revealed to be good indicators to find in patients who are at risk of heart disease. While these elements worked well in the visualizations, certain aspects of these elements as well as other elements will need be refined for future iterations.

One of the more notable elements that needs refining is the use of maximum heartrate as a selected attribute. Based on the related visualizations and background research, no significant patterns nor connections could be made between maximum heartrate levels and heart disease. Although maximum heartrate levels cannot be confirmed as a reliable indicator for heart disease, they still serve as a major indicator for general heart attacks. Further analysis with this attribute will be needed for future iterations.

In order to conduct further quantitative analysis, corrections will have to be made for data inputs, especially related to cholesterol. These errors can impact the accuracy of any potential analysis and would not prove to be helpful for other data analysts or data scientists who would need this data for further use.

Another element that would need refining for later iterations would be the data representation. As mentioned before, the gender attribute was limited to male and female, and did not include patient records for patients who identify as a different gender. As the format of patient records continue to ****change, datasets and data visualizations will need to include all demographics of certain attributes in order to avoid potential bias and to improve overall accuracy of visualization results.

SOURCES AND CITATIONS

* Fuchs, Flavio D. Whelton, Paul K. "High Blood Pressure and Cardiovascular Disease", https://www.ahajournals.org/doi/full/10.1161/HYPERTENSIONAHA.119.14240

** Mayo Clinic, "Heart Disease", https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

*** Harvard Health Publishing, "What Your Heart Rate Is Telling You", https://www.health.harvard.edu/heart-health/what-your-heart-rate-is-telling-you#:~:text=However%2C%20an%20unusually%20high%20resting,can%20help%20down%20the%20road.

**** Johns Hopkins Medicine, "ABCs of Knowing Your Heart Risk", https://www.hopkinsmedicine.org/health/wellness-and-prevention/abcs-of-knowing-your-heart-risk#:~:text=Cholesterol%20levels,your%20risk%20of%20heart%20disease.

***** Burgess, Claire PhD. Kauth, Michael R.,PhD. Klemt, Caroline PsyD. Shanawani, Hasan MD,MPH. Shipherd, Jillian C.PhD. "Evolving Sex and Gender in Electronic Health Records", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6590954/