**Data Visualization Final Report**

**Sai Kumar Kanna**

**Heart Disease Diagnostic Analysis**

**Introduction**

Roughly 610,000 people a year die of heart Disease in the United Sates – that is, roughly one in every four deaths ca be attributed to heart disease. Heart disease is the leading cause of death in the U.S for both men and women, and every year, about 735,000 Americans suffer a heart attack. The importance, then, of understanding heart disease and its contributing factors, such as diabetes, high blood pressure, smoking, and the consumption of alcohol, cannot be overstated. Because heart disease is so prevalent, and because there are so many factors that can increase the likelihood that one will suffer from heart disease, heart disease perfectly illustrates the usefulness of data visualization.

Using data from the University of California Irvine’s machine learning repository, I am going to analyze the occurrence of heart disease and the factors that contribute to heart disease. I am going to analyze heart disease by types of chest pain, considering the maximum BPM experienced with different types of chest pain, the age and sex of those suffering chest pain, and the presence of existing conditions that would exacerbate their heart conditions and also find any correlation between the factors. I am also going to consider how blood pressure, and cholesterol variations with age impact heart disease, and consider how average BMI impacts heart disease.

**Research Questions**

1. What are the maximum heart rates associated with different types of chest pain?
2. What correlation is there between a person’s age, sex, and preexisting conditions and the types of chest pain they experience?
3. What role does a person’s age play on their average resting blood pressure?
4. What role does a person’s age play on their cholesterol levels?
5. What correlation is there between ST depression and age in people with heart disease?
6. What is the correlation between BMI and heart disease, and in which regions is this correlation most prevalent?
7. How does Thal type affect heart disease?
8. What correlation is there between a person’s cholesterol, fasting blood sugar, and average resting blood pressure and the presence of heart disease?
9. What role does the presence of exercise angina play on a person’s average maximum heart rate, and how is this impacted by their age?

**Motivation**

My motivation with this project is to understand what factors contribute to heart disease, and to understand regional differences in the prevalence of heart disease, with and without comorbidities like high blood pressure, high cholesterol, and an unhealthy BMI. Additionally, I factored age, sex, Thal type, and the presence or absence of angina into my data visualizations.

Thoroughly understanding an issue is the first necessary step to solving the issue, and I hope that research like I have done will eventually lead to decreased instances of heart disease.

**Methodology**

The data used for this project was collected from the University of California Irvine’s machine learning repository. The dataset has 918 rows with 16 columns, as follows:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. Chest Pain Type: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. Resting BP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. Fasting BS: fasting blood sugar [yes: if Fasting BS > 120 mg/dl, no: otherwise]
7. Resting ECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. Max HR: maximum heart rate achieved [Numeric value between 60 and 202]
9. Exercise Angina: exercise-induced angina [Y: Yes, N: No]
10. Old peak: old peak = ST [Numeric value measured in depression]
11. ST Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: down sloping]
12. ca: The number of major vessels (0–3)
13. thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously

Value 1: fixed defect (no blood flow in some part of the heart)

Value 2: normal blood flow

Value 3: reversible defect (a blood flow is observed but it is not normal)

1. Heart Disease: output class [1: heart disease, 0: Normal]

Using data from the UCI repository, I created a new data set, in which I collated the data. I created novel patient IDs for each individual in the original data set, and indicated in the columns the patient’s height, weight, activity level, BMI, smoking habits, and alcohol consumption which has been extracted from the original 76 attributes in the original parent repository. I did this to more clearly underscore the correlation between a person’s so-called vices and the prevalence of heart disease.

The five datasets used for its curation are:

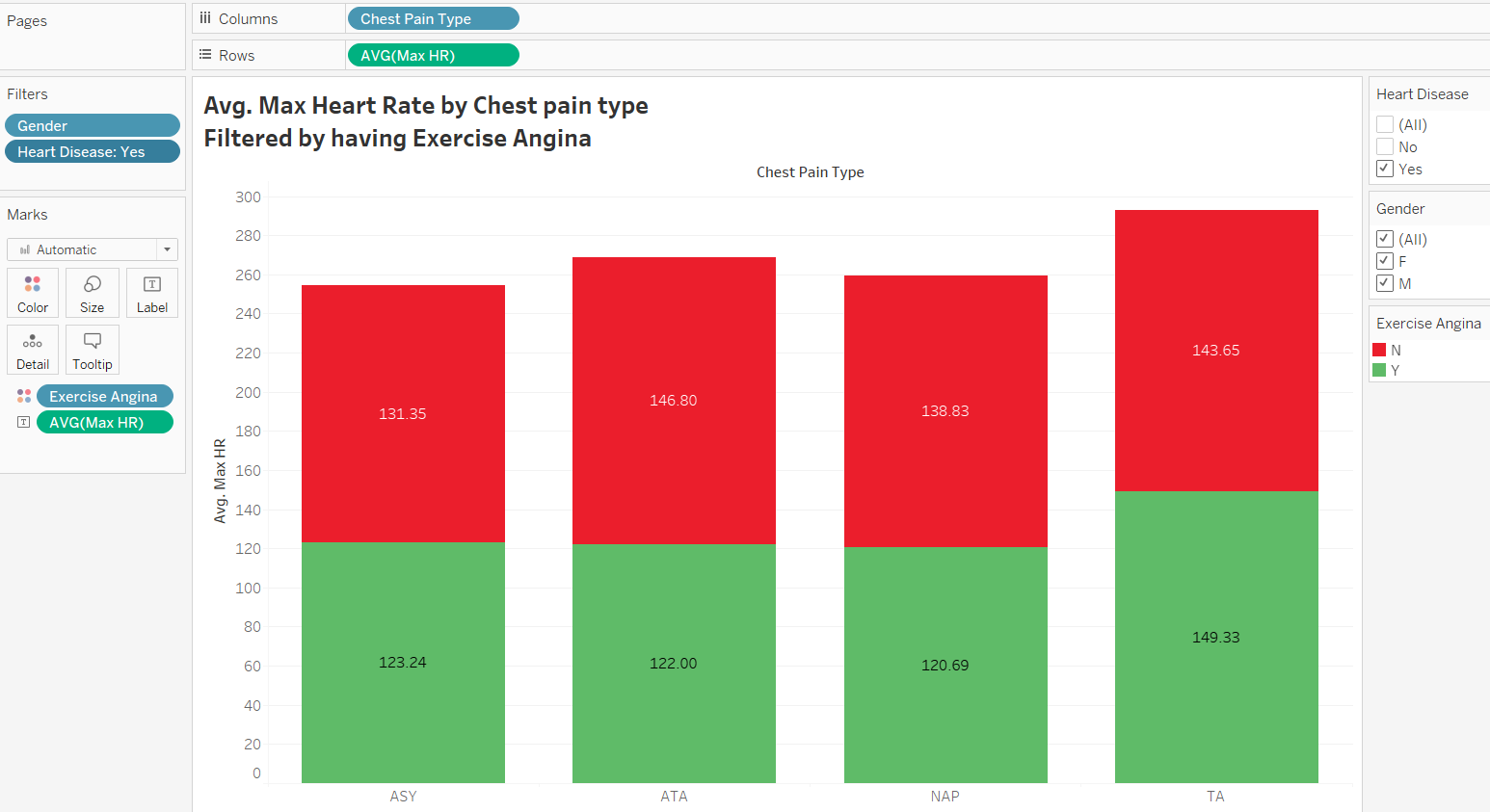
* Cleveland: 303 observations (under column location “U.S.A”).
* Hungarian: 294 observations
* Switzerland: 123 observations
* VA of Long Beach: 200 observations (under column location “U.S.A”).

I also cleaned the data set of any missing and null values.

**Analysis**

Below, I have included the Tableau sheets, dashboard, calculated fields and measures visualizing the aforementioned data.

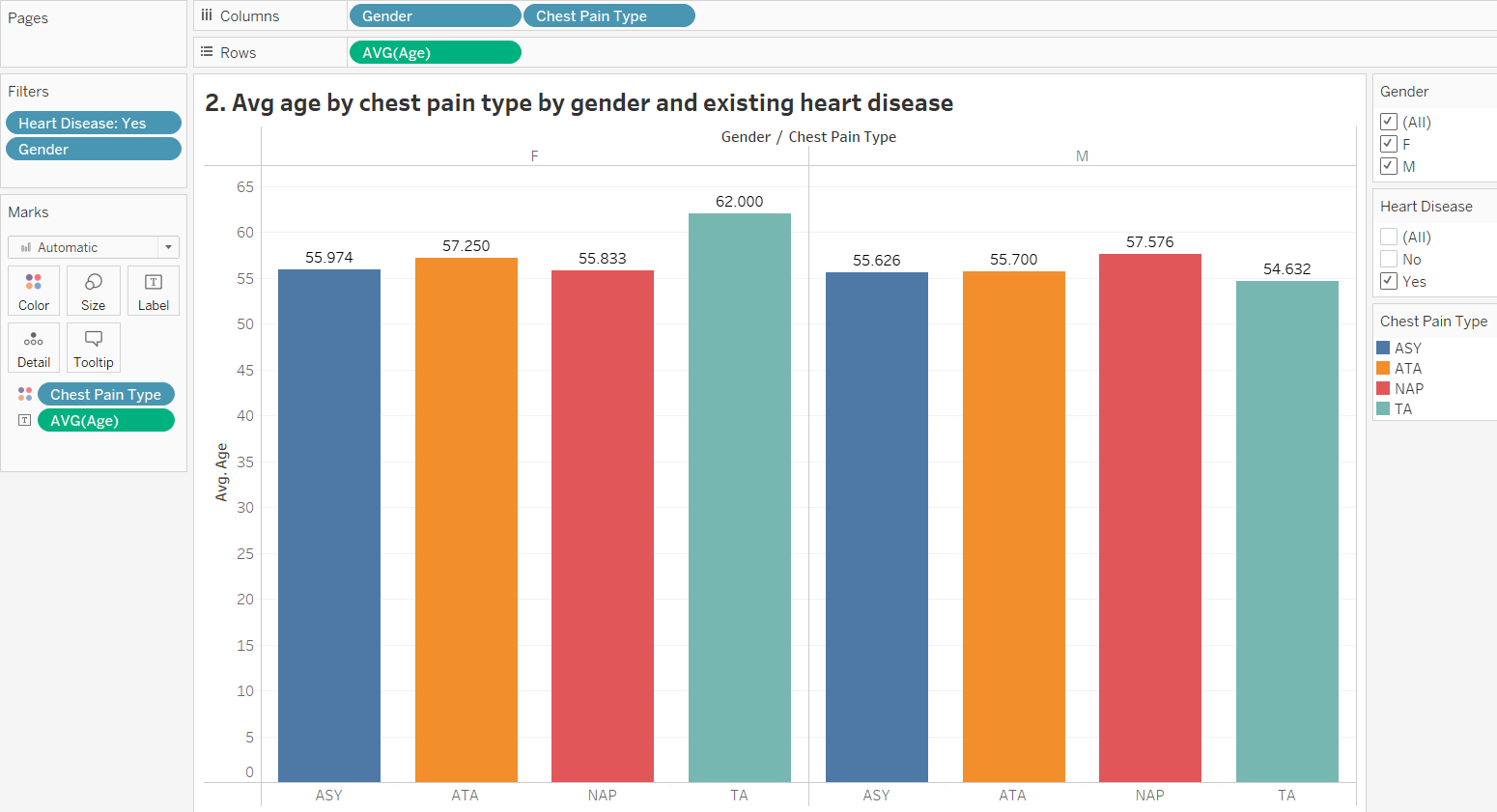
**Visualization 1**



*Fig.1 Average of max heart rate achieved by people filtered by Chest pain type, gender, presence of Exercise Anigna.*

The stacked bar graph above illustrates the average maximum heart rates associated with each type of chest pain, and filters them by sex and the presence or absence of exercise angina. From this graph, we can see that the average of maximum heart rate is lower for people with exercise angina than for people without exercise angina for people who have ATA chest pain type. Moreover, we can also see that there is a significant increase in the avg. of max heart rate for people with NAP chest pain type but are still diagnosed with heart disease. For people who are not diagnosed with heart disease, there is no significant difference in the max heart rate achieved irrespective of chest pain type.

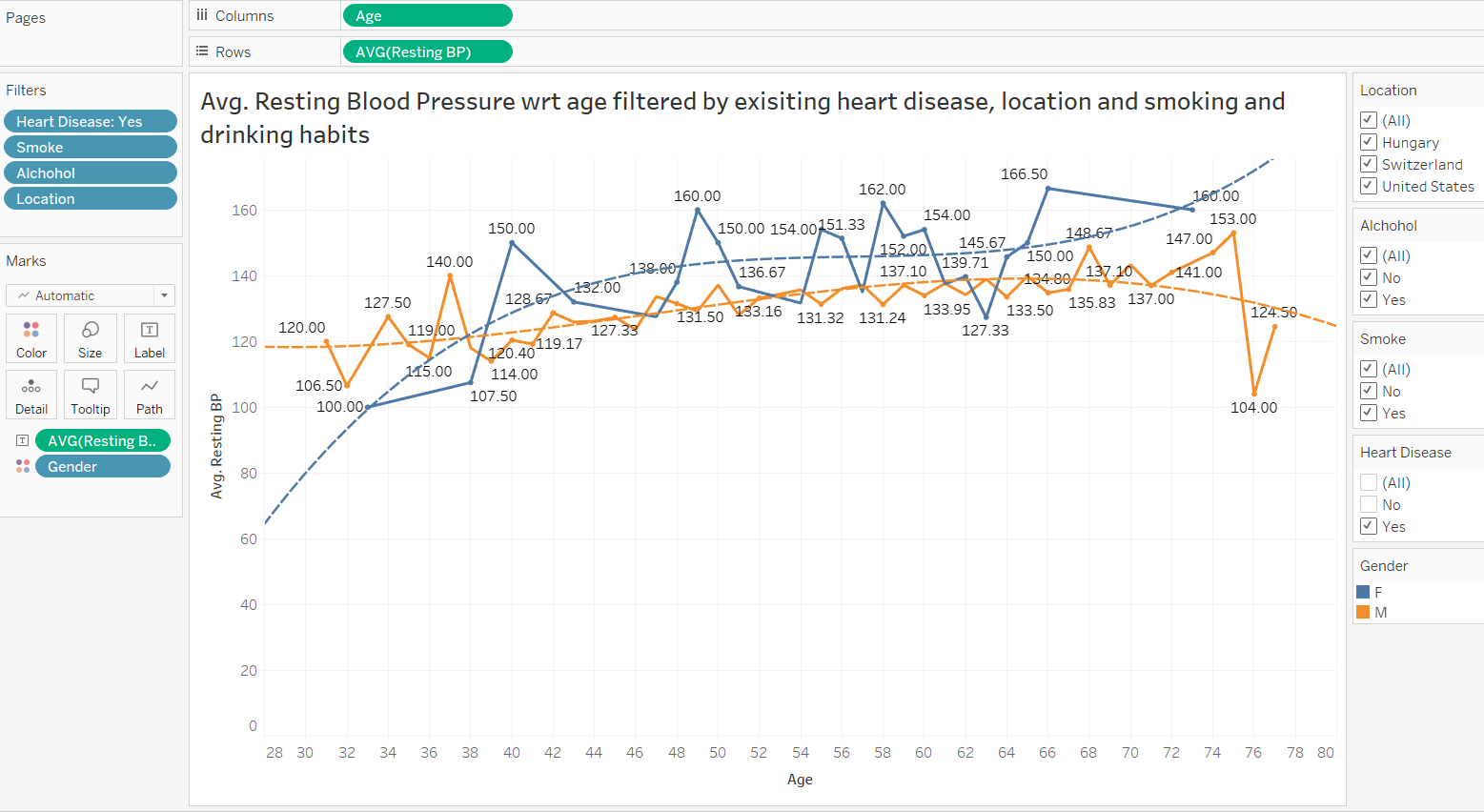
**Visualization 2**



*Fig.2. Average age by chest pain type by gender and existing heart disease*

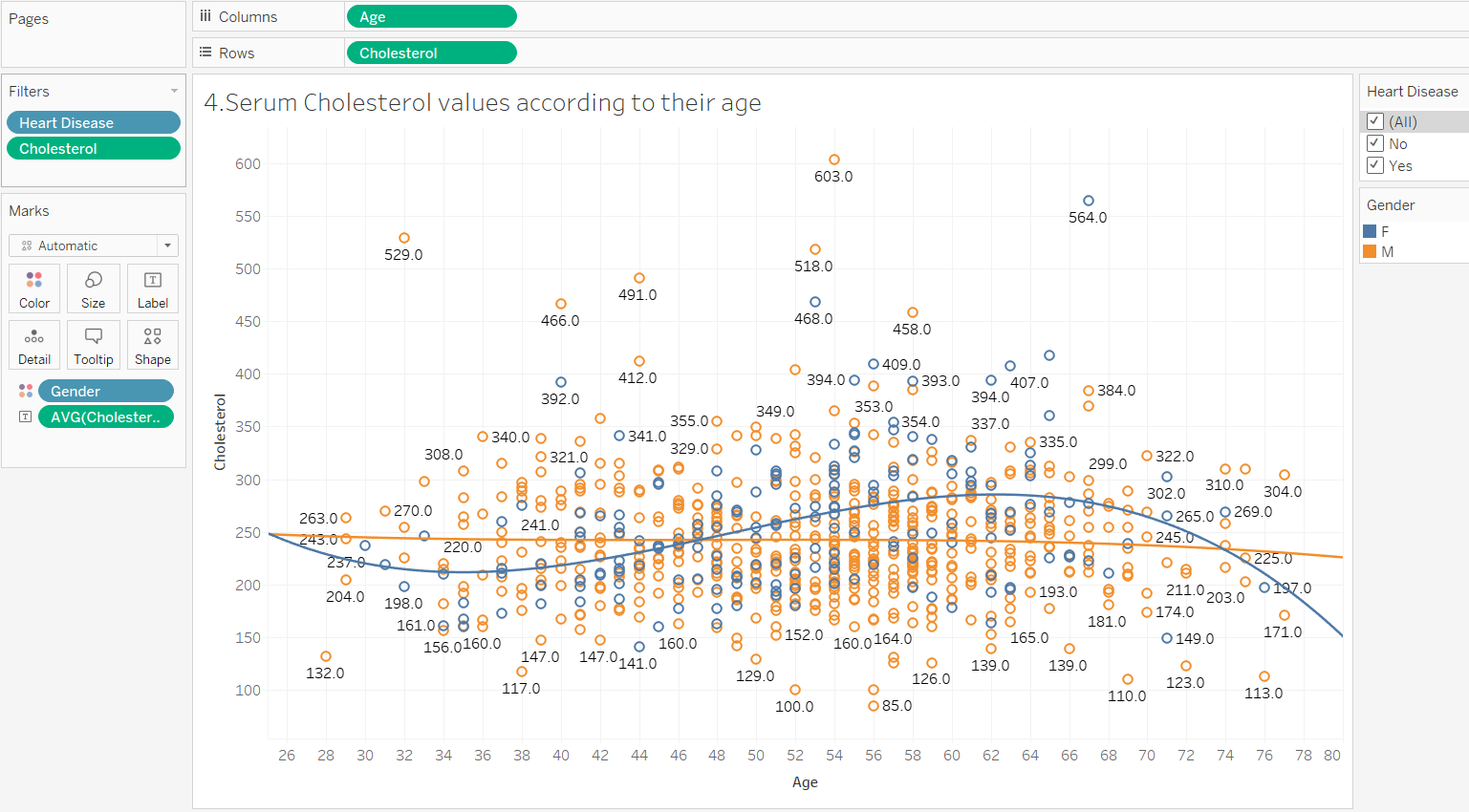
The graph above illustrates the average age at which people experience chest pain, and filters this information by the type of chest pain they experience, their sex, and the presence or absence of pre-existing heart disease. We can see that female, with existing heart disease detected and Typical Angina Chest type pain, have average age significantly higher than the remaining chest pain types. But, in males, we can see that the average age of the person irrespective of chest pain type doesn’t show any peculiar traits.

**Visualization 3**

****

*Fig.3. Average Resting BP by age filtered by gender, existing heart disease , smoking and drinking habits*

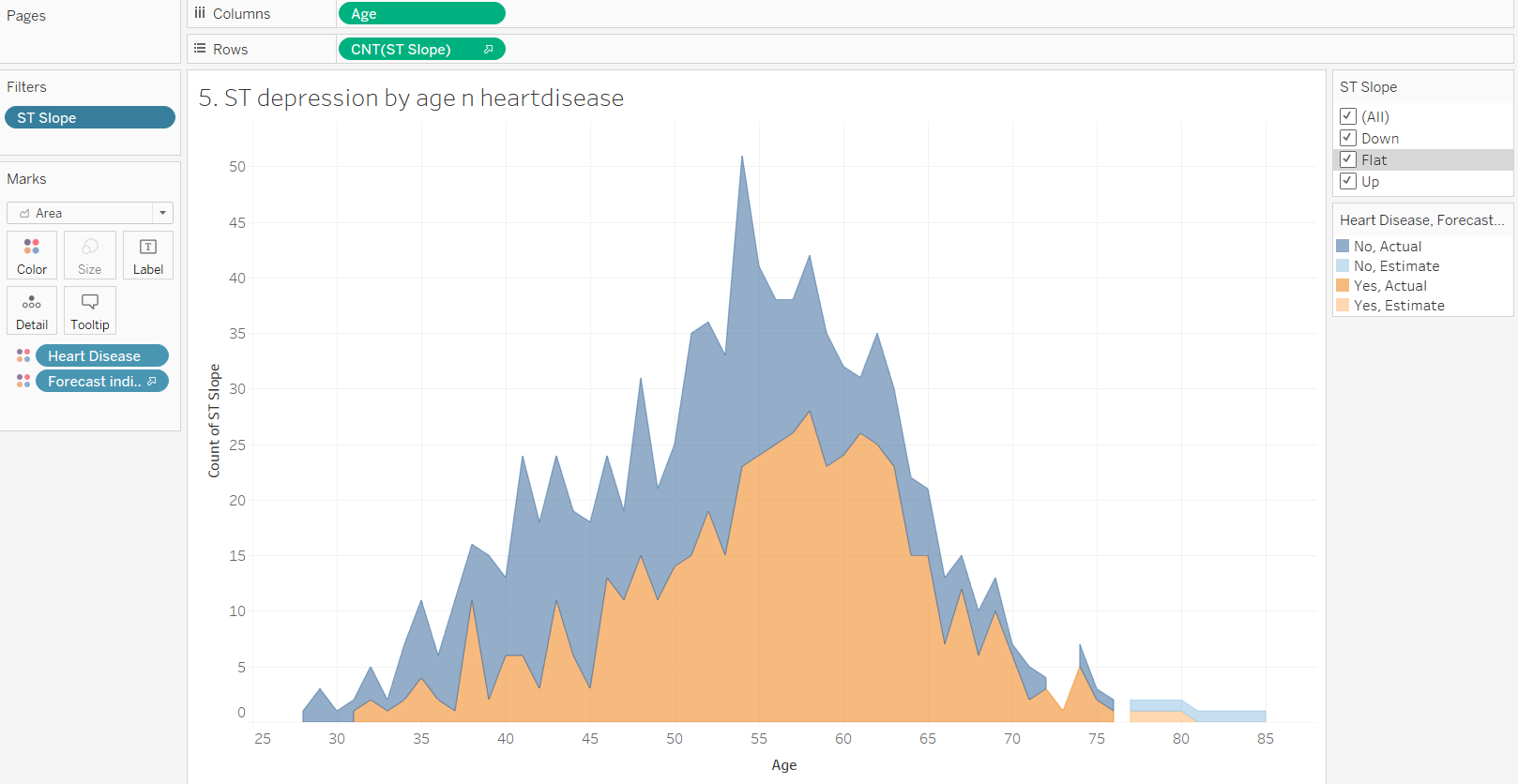
According to the graph above, we are measuring the average resting blood pressure with respect to age, filtered by location, alcohol, smoking habits, and presence of existing heart disease. We can see that for females, the average blood pressure is significantly higher than males with existing heart disease. We also are using trend lines to project the average resting blood pressure of men and women based on their location and their smoking and drinking habits. We can see that females have significantly higher resting blood pressure than males with existing heart disease.

**Visualization 4**

*Fig.4. Serum cholesterol values according to age filtered by existing heart disease gender with trend lines*

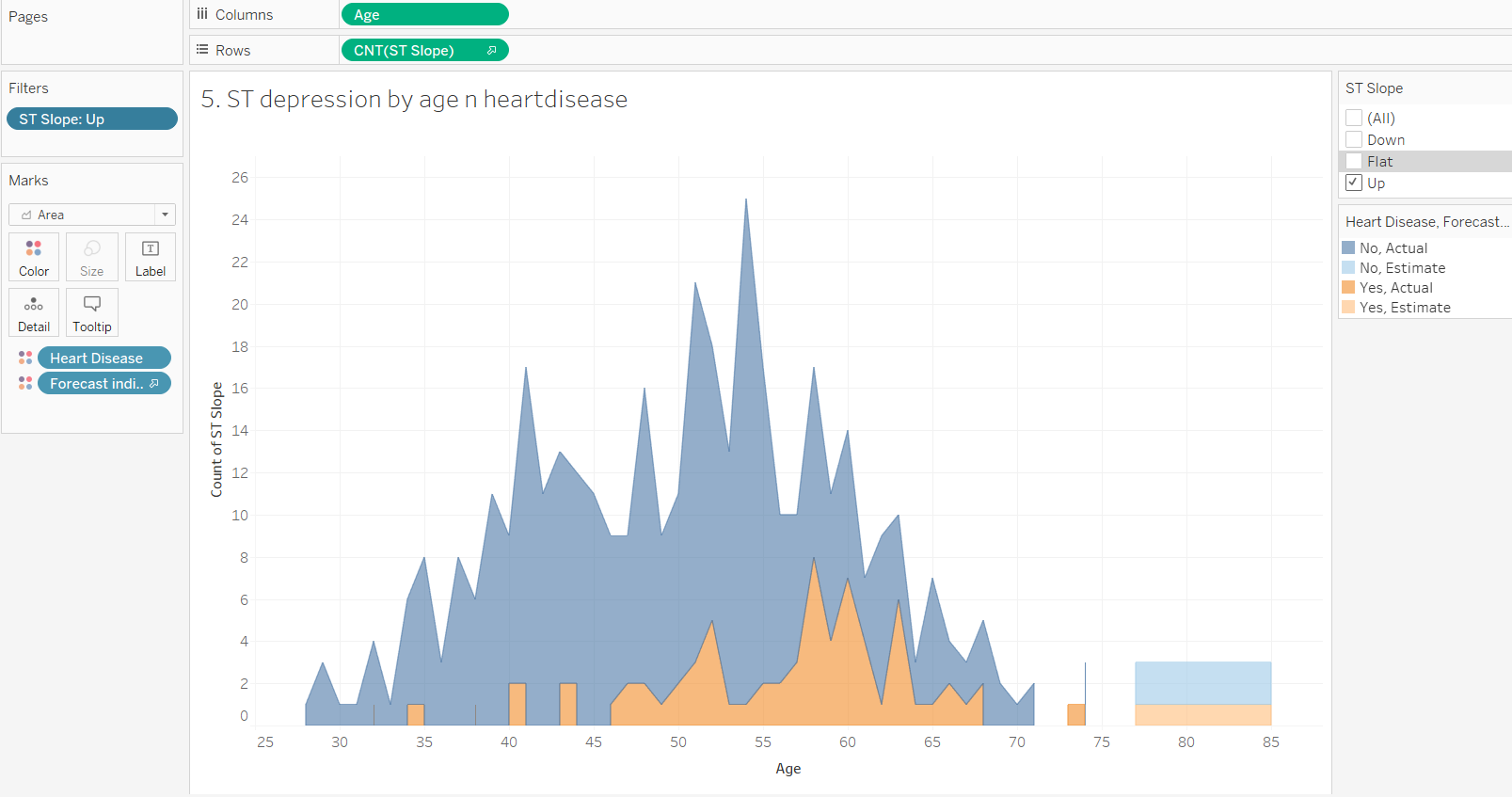
The above scatter plot shows the serum cholesterol levels of individuals based on their age and sex. From the graph we can see that males tend to have relatively stable cholesterol levels irrespective of age and the presence or absence of pre-existing heart disease, while females tend to have lower cholesterol levels when they are below 46-48, and higher cholesterol levels when they are 48 and older.

**Visualization 5**



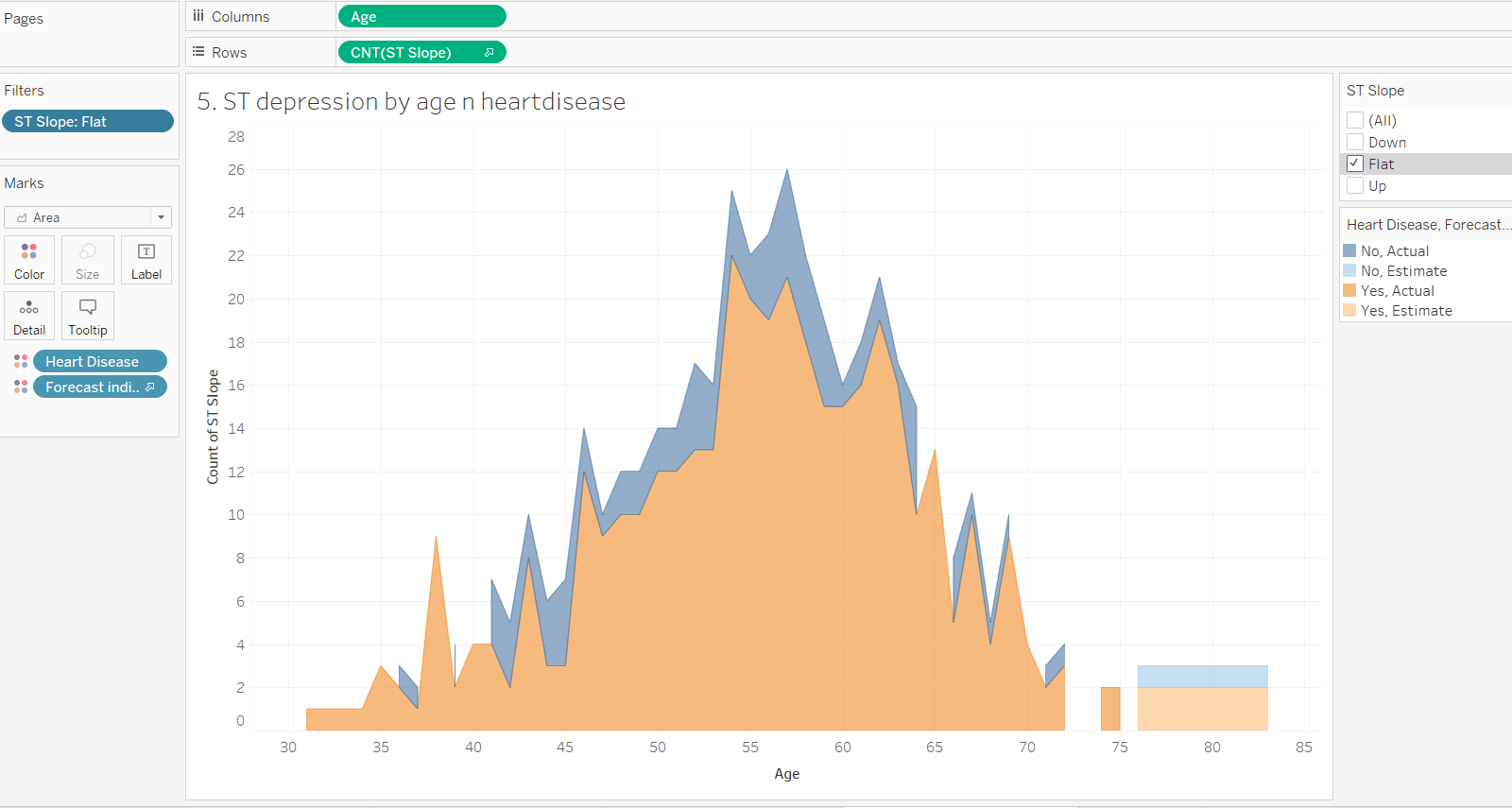
*Fig.5. Count of people with ST segment depression with presence and absence of heart disease by age*

The graph above shows the correlation between ST depression and age filtered by the presence of heart disease and the type of ST segment depression. We also forecasted the ST depression levels after age 75 using Tableau. We can see the ST depression levels are lower in people with heart disease than in people without heart disease.



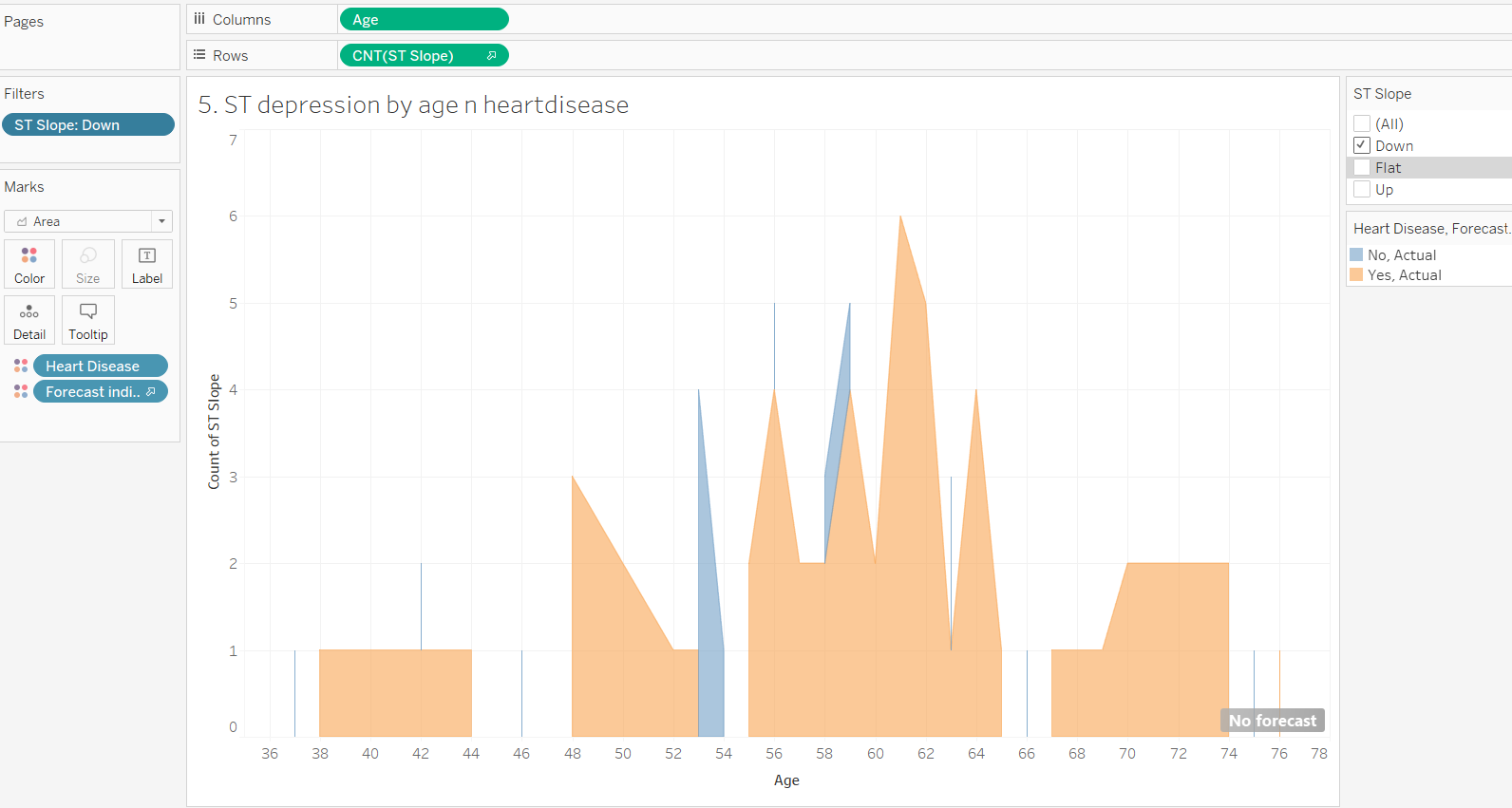
*Fig.5a. Count of people with ST segment (UP sloping) depression with or without heart disease by age*

As the graph illustrates, there are fewer people with upsloping ST depression than there are with up sloping ST depression.



*Fig.5b. Count of people with ST segment (Flat) depression with or without heart disease by age*

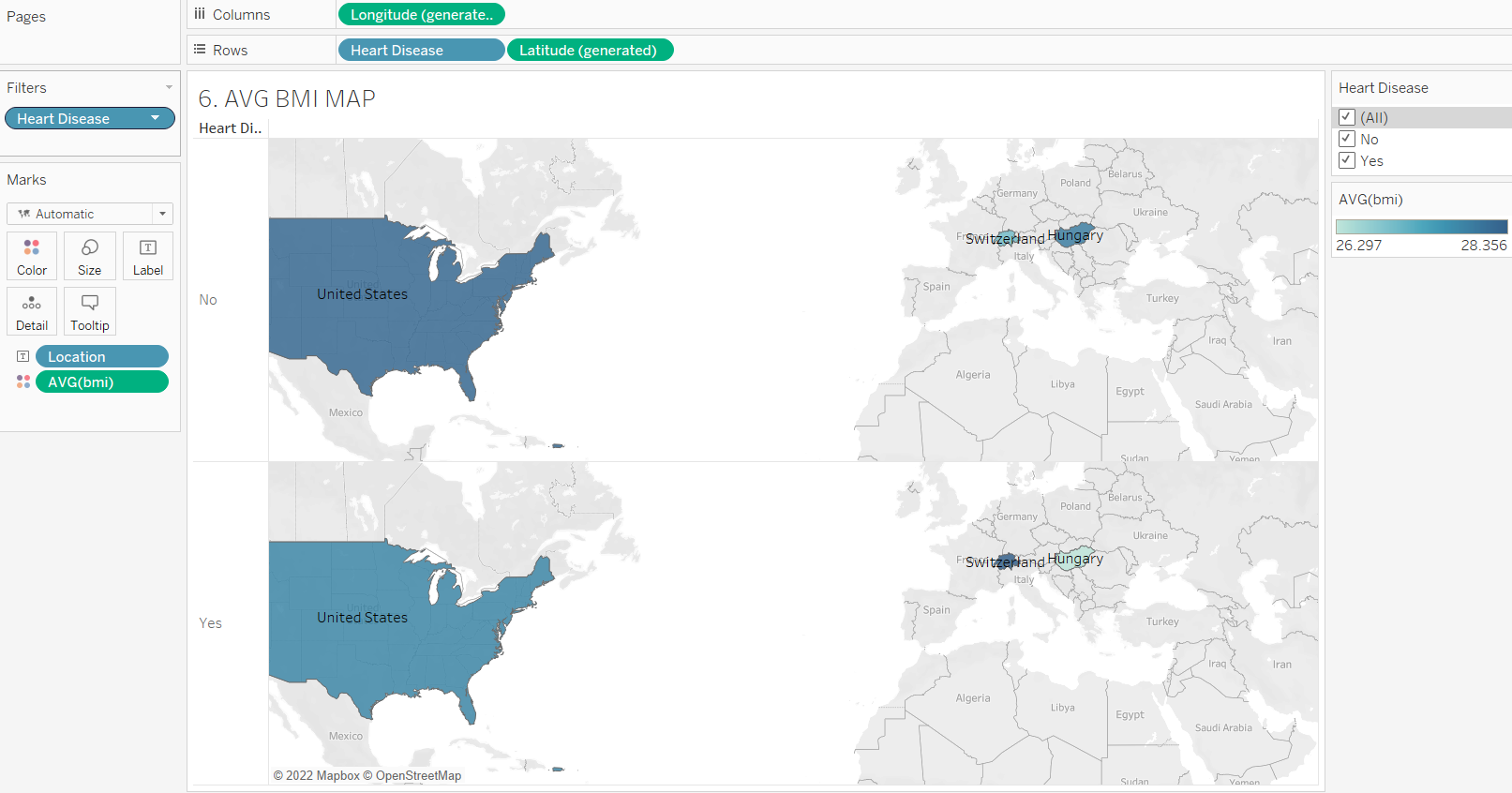
We can also see that when the ST depression is flat, the presence or absence of pre-existing heart disease does not have much impact on the ST depression.



*Fig.5. Count of people with ST segment (down sloping) depression with or without heart disease by age*

Finally, when the ST depression is down sloping, there is a much greater correlation between pre-existing heart disease and ST segment depression.

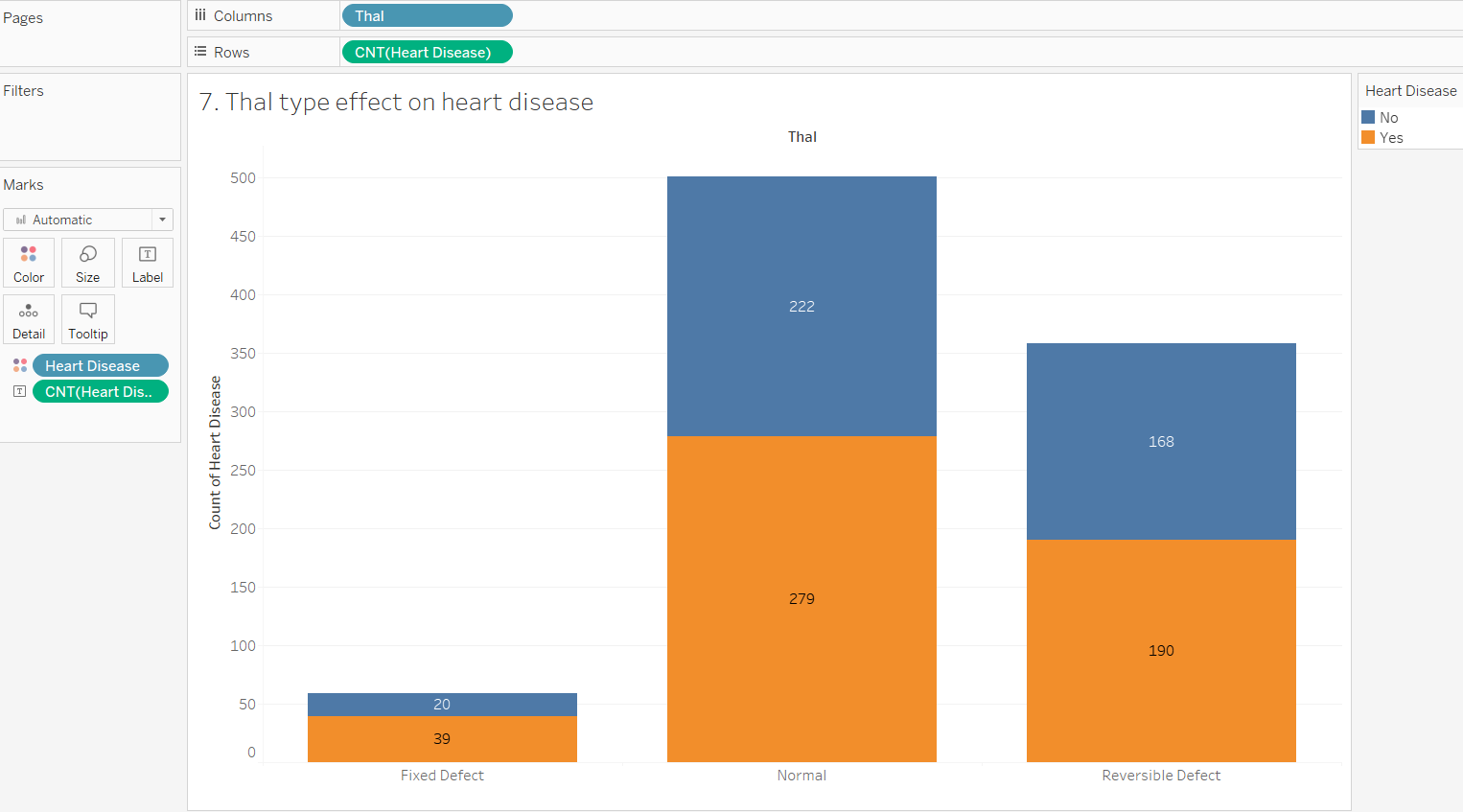
**Visualization 6**



*Fig.6. Average BMI index by location*

For this visualization I used the population’s height and weight to create a calculation field, BMI, calculated by the following formula: BMI = weight / height2. Unsurprisingly, there is a correlation between a high BMI and the presence of heart disease. Individuals with a BMI of 25 – 39.9, are considered overweight, and individuals with a BMI > 40 are considered obese. Membership in both of these categories is associated with a greater risk of heart disease. As the map shows, when we consider the population of people with heart disease, the population in the U.S and Switzerland tend to have higher BMIs than the population of people in Hungary. When we consider the population of people without heart disease, this is reversed, with people in Switzerland having lower BMIs than people in Hungary. In both cases, both in the presence of heart disease and in the absence of heart disease, the population of the U.S tends to have a higher BMI.

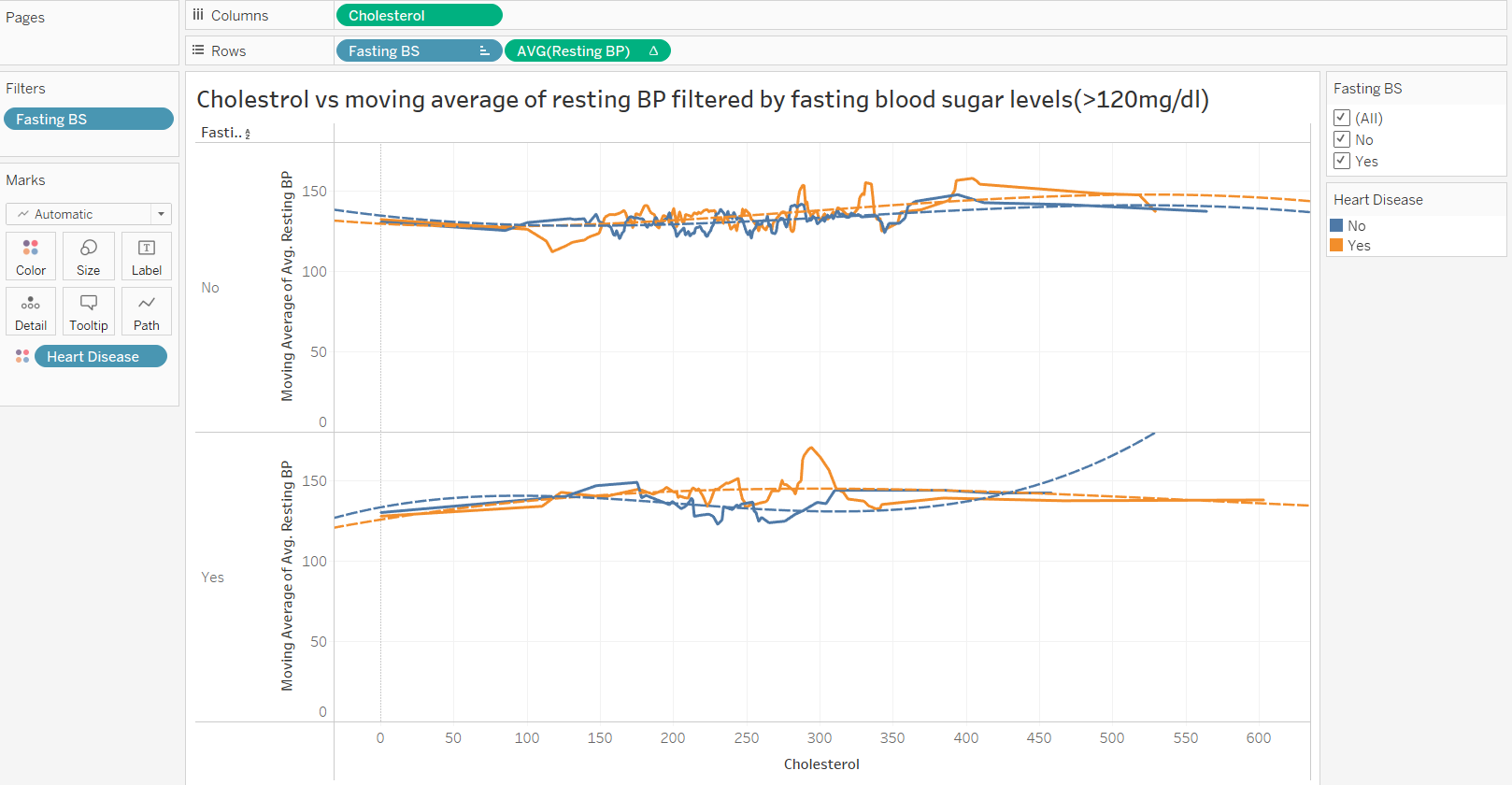
**Visualization 7**



*Fig.7. Thal type effect on heart disease*

The graph above illustrates that in the presence of thalassemia, a blood disease characterized by a deficiency of hemoglobin, the occurrence of fixed-effect heart disease is doubled. When there is no presence of thalassemia, or when there is a reversible-defect thalassemia, there is no significant effect on the prevalence of heart disease.

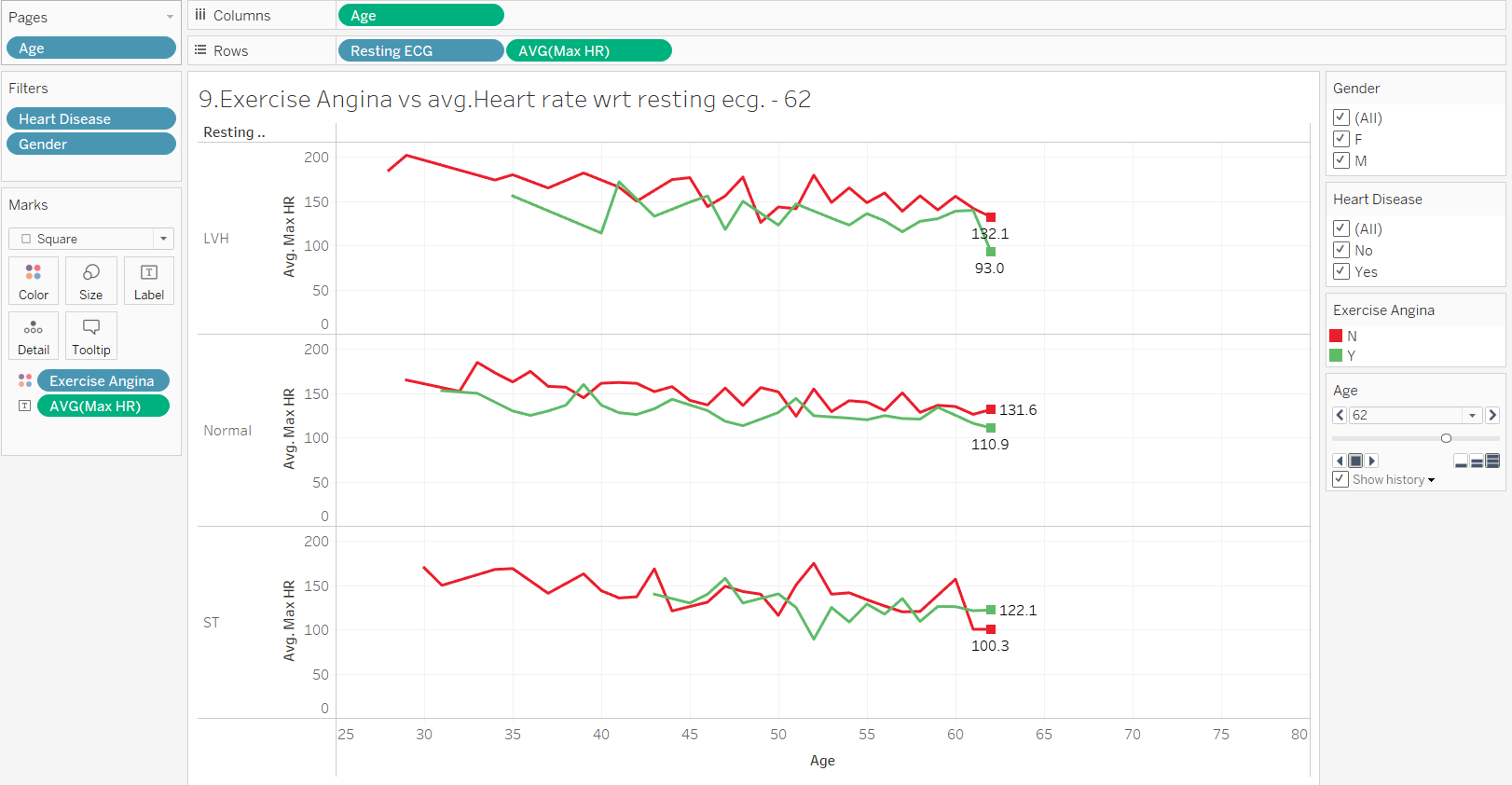
**Visualization 8**



*Fig.8. Cholesterol vs moving average of resting BP filtered by fasting blood sugar levels(>120mg/dl)*

The above graph shows the correlation between moving average of average resting BP and cholesterol, filtered by whether or not the fasting blood sugar levels are greater than 120 mg / dl. When the fasting blood sugar levels are less than 120 mg/dl, the trend lines used do a much better job of predicting the average BP since theR2 value is much less than 0.05. When the fasting blood sugar levels are greater than 120 mg/dl, the trend lines do not do an accurate job of predicting the average BP when there is no heart disease present. Overall, there doesn’t seem to be an accurate correlation between cholesterol and blood pressure, either in the presence of or in the absence of heart disease.

**Visualization 9**



*Fig.9 Average of max heart rate by age filtered by resting ECG and the presence or absence of exercise angina.*

The above graph shows the average max heart rate by age, as shown on an ECG, filtered by the presence and absence of LVH - left ventricular hypertrophy – and ST – T-wave abnormality, and exercise angina. In all three cases we can see that people with exercise angina, have lower heart rates compared to people who don’t. When we filter the data exclusively to show only the patients with pre-existing heart disease, we see that there is comparatively greater difference between Max Heart rates for abnormal Resting electrocardiographic measurements (LVH and ST).

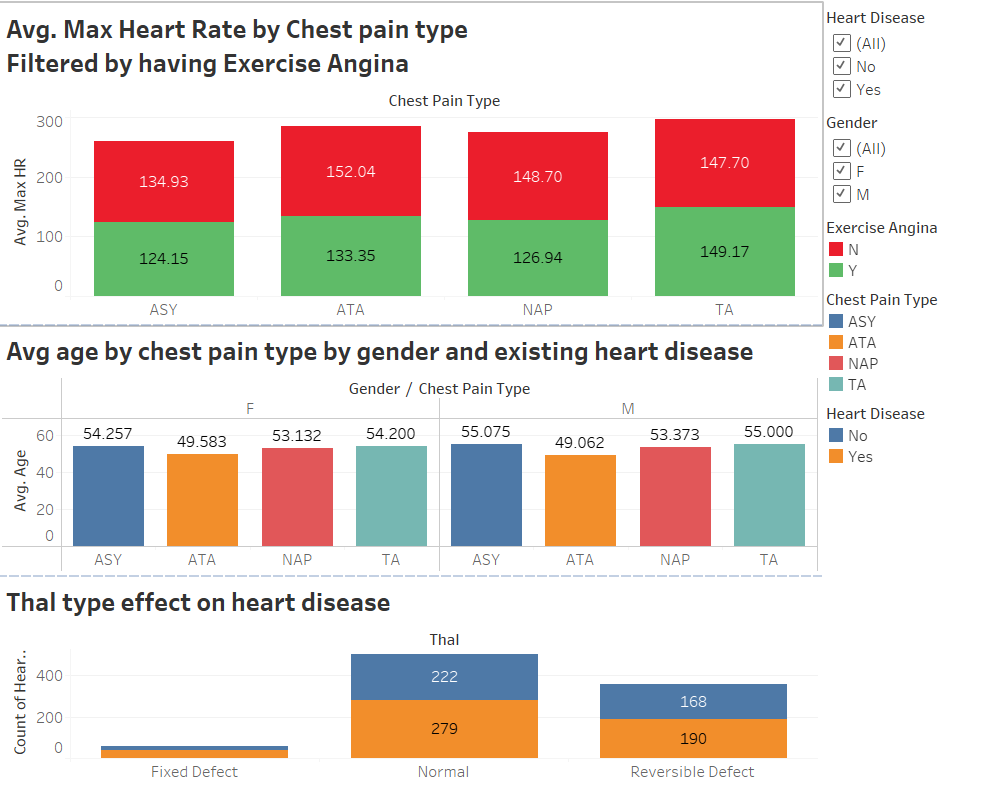
**Conclusion**

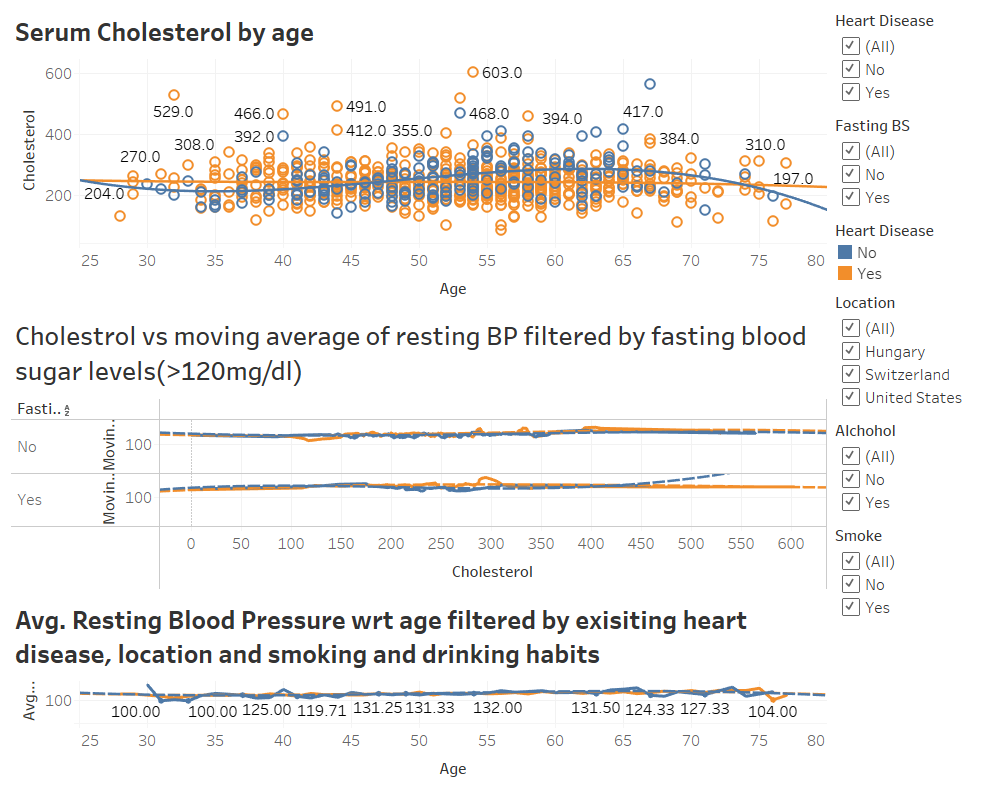
There are many factors that contribute to the presence of heart disease, and understanding all of these multifaced factors will help us prevent the occurrence of heart disease. A person’s age, sex, BMI, activity level, cholesterol levels, blood pressure levels, and their smoking and drinking habits all play a role in heart disease. Some of these factors result in a greater instance of heart disease than others, and, as we have seen, the presence of pre-existing conditions, like thalassemia, also exacerbate heart disease.

Age and sex are the most important risk factors in developing heart diseases. With each decade of life, there is a tripling of risk that one will develop heart disease. This is not surprising. As people age, they also become less active, and, consequently, may gain weight. As we have seen, BMI and activity level also play a role in the instances of heart disease. This, coupled with the fact that comorbidities, like high blood pressure and cholesterol levels, exacerbate heart problems, make it unsurprising that age is the most important risk factor in developing heart disease. 82% of the people that die of heart disease are 65 or older.

Sex is the second most important risk factor in developing heart diseases. While the risk of developing heart diseases increases with age for both males and females, males tend to develop issues with heart disease when they are 55 or over, and women tend to develop issues with heart disease when they are 50 or over. Males also tend to have more frequent instances of asymptomatic chest pain, which is associated with a higher chance of heart disease, than females.

The following are the dashboards for the visualizations made:





**Applications and Areas of Further Exploration**

This project has presented other areas of exploration in terms of the factors that cause heart disease, and in terms of how this data can be compiled to assess a person’s overall risk factor.

1. It has been shown that high blood pressure and smoking and drinking habits both increase the risk of heart disease. It seems fairly intuitive that these factors influence each other as well, as smoking is a vasoconstrictor and the consumption of alcohol acts as a vasodilator. Both smoking and drinking are used to relieve feelings of stress, and feelings of stress may be higher among people who already have high blood pressure. Thus, the presence of stressors and anxiety and the “treatments” for these stressors may both contribute to heightened blood pressure, and, consequently, the risk of heart disease. Further research into this topic could shed light on the correlations between smoking, drinking, high blood pressure, and heart disease.
2. The risk factors that cause heart disease are numerous, and many of these risk factors are interrelated, as is the case with blood pressure and the consumption of alcohol or tobacco. Thus, it would be beneficial to create a prediction system into which one could input all of their relevant data to assess their overall risk of heart disease. A tool that would factor in the patient’s age, sex, BMI, activity level, comorbidities, family history of heart disease, and alcohol and tobacco consumption would take time to develop, but it could have life-saving implications. As we have seen through some of our data visualizations, one factor alone does not appear very daunting. However, when multiple factors are presented in the same visualization, as was the case with visualization 9, the difference between the typical and atypical is alarming.

**References**

University of California. (n.d.). *Heart Disease Data Set.* UCI Machine Learning Repository.

<https://archive.ics.uci.edu/ml/datasets/heart+disease>

Deshmukh, H. (2020, June 18). *Heart Disease UCI-Diagnosis & Prediction.* Medium.

<https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7>

Rawat, S. (2019, august 10). *Heart Disease Prediction.* Medium.

<https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>

Fedesoriano. (n.d.). *Heart Failure Prediction Dataset.* Kaggle.

<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>

Detrano, R., Janosi, A., Steinbrunn, W., & Pfisterer, M. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease*. American Journal of Cardiology, 64,304--310.

<http://rexa.info/paper/b884ce2f4aff7ed95ce7bfa7adabaef46b88c60c>