Stat 340 Group Project Final Report
Group Narwal
12/18/2021

1. Name/NetIDs:

Christine Ruan - yruan23
Charlotte Xu - xxu382
Isabel An - ban22
Young Yang - xyang532
Aaron Chen - kchen339

2. Description of the dataset:

Heart Failure Prediction Dataset from fedesoriano.
(https://www.kaggle.com/fedesoriano/heart-failure-prediction)

This dataset was created by combining different datasets already available independently but not combined before. The author collected and combined 5 datasets on heart diseases over 11 common features, and the included datasets are from Cleveland, Hungarian, Switzerland, Long Beach VA and Staglog heart dataset. Every dataset used here was collected from UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/. This leads to the final dataset with 918 observations and 11 common features on heart diseases.

This dataset combines five heart datasets with 11 shared characteristics to provide the biggest heart disease dataset available for research purposes to date. People with cardiovascular disease or who are at high cardiovascular risk (due to the existence of one or more risk factors such as hypertension, diabetes, hyperlipidemia, or previously existing illness) require early identification and care, which can be greatly aided by a machine learning model.

The main statistical question we are going to answer is what's the major factors that affect the heart which leads to heart disease.

3. A brief description of why your group found this data set, who gathered it and some background information (in particular to be of interest and why your reader should care).

According to the WHO, Cardiovascular diseases (CVDs) are the leading cause of death globally. About 17.9 million people died from Cardiovascular diseases in 2019 which form 32% of all global deaths. The high ratio of global deaths prompt us to think and understand why and what factors cause Cardiovascular diseases to happen. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol. Regular health examination is also a great way to early discover if a person has Cardiovascular diseases or not and can reduce the risk of death.

4. Variable description

a. Age: Age of the patient (years)
b. Sex: Sex of the patient (1: Male, 0: Female)
c. ChestPainType: chest pain type.
   (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
d. RestingBP(mm Hg): resting blood pressure.
e. Cholesterol(mm/dl): Serum cholesterol.
f. FastingBS: fasting blood sugar. The value will be 1 if FastingBS > 120 mg/dl; and the value will be 0 if otherwise.
g. RestingECG: resting electrocardiogram results. The data is expressed in description: 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).
h. MaxHR: maximum heart rate achieved. The data is a numeric value between 60 and 202.
i. ExerciseAngina: exercise-induced angina. The data is 1 if it's a Yes and 0 if it's a No.
j. Oldpeak: oldpeak = ST. The data is a numeric value measured in depression.
k. ST_Slope: the slope of the peak exercise ST segment. The data is expressed in slope description: Up(upsloping), Flat(flat), Down(downsloping).
l. HeartDisease: output class. The value will be 1 if the patient has heart disease and 0 if the patient is normal.

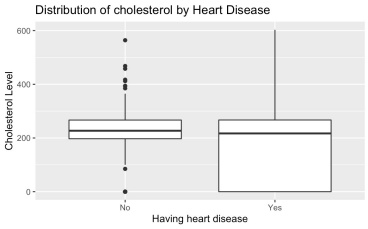
5. A block of code showing how to load the data into R (if your data set is too large to permit this please speak to us promptly)

```
heart <- read.csv('heart.csv')
```

6. Progress:

First, we use Logistic Regression to primarily analyze the dataset.

*Before we do that, we found out that the column 'Cholesterol' has many missing entries which are replaced by 0. After some simple preprocessing and analyzing, we had a box plot with 25% quantile at Cholesterol = 0, and 172 entries of data with cholesterol = 0 out of a total of 918 entries.



In consideration of the rest of the dataset, we decided to delete entries with ('Cholesterol' = 0). We have 746 rows left. The result of Logistic Regression is shown below:

```
Call:
glm(formula = HeartDisease ~ ., family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3733  -0.4320   0.1768   0.4906   2.4528

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.944187   1.291602   1.505 0.132259
Age               0.014656   0.012563   1.167 0.243385
Sex               0.645028   0.132669   4.862 1.16e-06 ***
ChestPainTypeATA -1.865462   0.307884  -6.059 1.37e-09 ***
ChestPainTypeNAP -1.574688   0.249041  -6.323 2.57e-10 ***
ChestPainTypeTA  -1.364108   0.412494  -3.307 0.000943 ***
RestingBP         0.004466   0.005728   0.780 0.435560
Cholesterol      -0.003327   0.001049  -3.170 0.001523 **
FastingBS         0.521679   0.131810   3.958 7.56e-05 ***
RestingECGNormal -0.119670   0.261378  -0.458 0.647065
RestingECGST     -0.270653   0.337633  -0.802 0.422773
MaxHR            -0.007653   0.004771  -1.604 0.108699
ExerciseAngina    0.515141   0.117060   4.401 1.08e-05 ***
Oldpeak           0.311933   0.116281   2.683 0.007305 **
ST_Slope         -1.630662   0.206719  -7.888 3.06e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1262.14  on 917  degrees of freedom
Residual deviance:  642.35  on 903  degrees of freedom
AIC: 672.35

Number of Fisher Scoring iterations: 5
```

According to the summary of the linear regression model, the p-value of "Sex", "ChestPainType", "FastingBS", "ExerciseAngina", "Oldpeak", "ST_Slope" are less than 0.05, meaning that they are statistically significant, then we can reject the null hypothesis of these variances.

Hence, from the logistic regression above, we have several factors to discuss:

1. Cholesterol

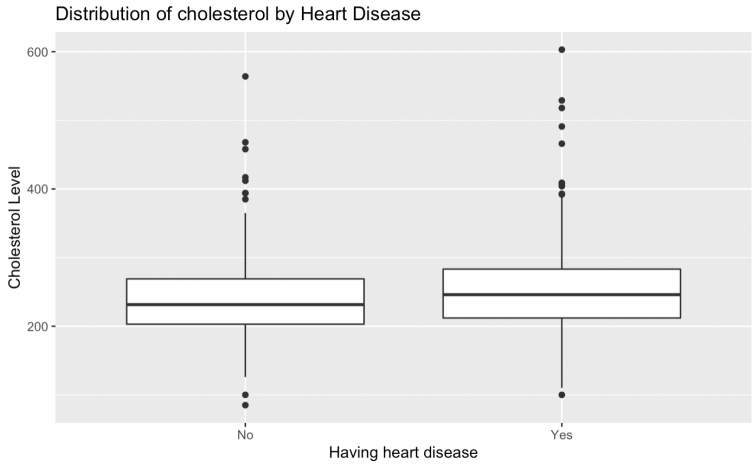
Why this factor:

According to the Logistic regression above, all three types of chest pain have a p-value smaller than 0.05, meaning that they are all statistically significant and we can reject all these three's null hypothesis.

Preprocessing:

We first divided the population into two groups and mutated a new column ‘HD’: “Yes” for those with heart disease, “No” for those without heart disease. Then, we make a boxplot to explore the distribution of cholesterol level by if those have heart disease.

Summarizing plot:



Plot 1.1 Cholesterol level comparison between diseased and normal population

From the boxplot, the graph on the right column(diseased) has a higher median and quantile values than the one on the left column. We could assume that **diseased patients have a higher cholesterol level than normal patients**.

To further prove the correlation between heart disease and Cholesterol, we go ahead to do logistic regression:

```
glm(formula = HeartDisease ~ Cholesterol, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.646  -1.124  -1.016   1.210   1.445

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.969940   0.321903  -3.013  0.00259 **
Cholesterol  0.003592   0.001282   2.802  0.00508 **
---
```

Applying the statistical model further proved our observation above. The p-value is smaller than the significant value, and the close-to-0 slope indicates that cholesterol level has a positive correlation, by **increasing the heart disease risk as lifting cholesterol level**.

Conclusion:

Cholesterol levels higher than normal can increase the rate of heart disease.

2. Age

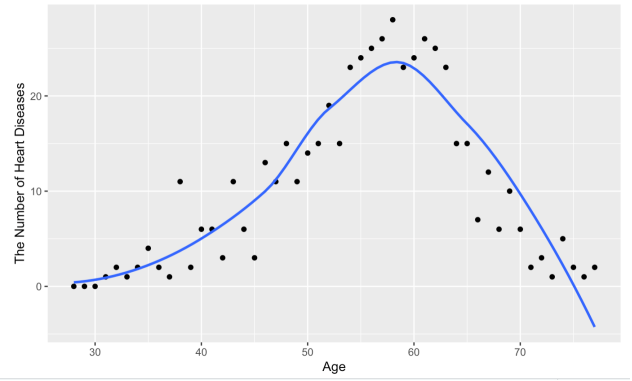
Why this factor:

From the Logistic Regression, 'Age' has a p-value that's larger than the significant value. According to the model, we are supposed to assume that Age does not have any significant influence on heart disease, while our common sense disagrees.

Preprocessing:

Instead of using the 'Age' value from every data entry, we group by 'Age' and summarize the data by n(), which gives us the count of the number of patients with heart disease under the same 'Age'. Therefore, we plotted the distribution here:

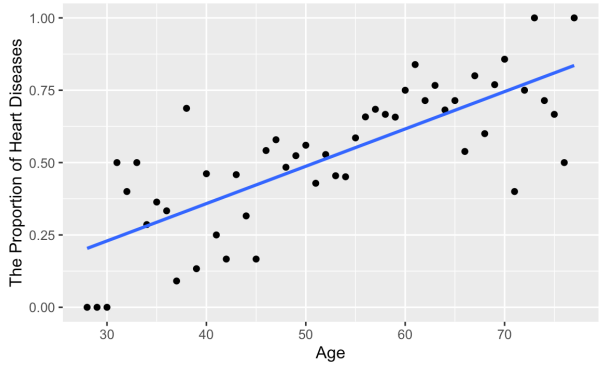
Summarizing plot:



Plot 2.1(Old) 'Age' with population of heart diseased patients

As we observe, the plot indicates that we have most patients in the dataset during 50-60, while the number of patients are much fewer in the older age group.

We don't think this plot could give us a helpful conclusion about 'Age' and heart disease; Instead, we count the percentage of diseased patients, which is the amount of diseased patients over the total population at that 'Age'. The new plot is here:



Plot 2.1(New) 'Age' with proportional population of heart diseased patients

In this plot, we clearly see that there's an increasing trend in the proportional value of the diseased patient with 'Age'. Therefore, we could conclude that **as the patient grows older, they are more likely to get a heart disease**.

To further prove the correlation between heart disease and age, we go ahead to do logistic regression:

```
Call:
glm(formula = HeartDisease ~ Age, family = "binomial", data = heart_with_Cholesterol)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.836  -1.081  -0.686   1.102   1.889

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.724016   0.471568  -7.897 2.85e-15 ***
Age          0.068484   0.008742   7.834 4.73e-15 ***
```

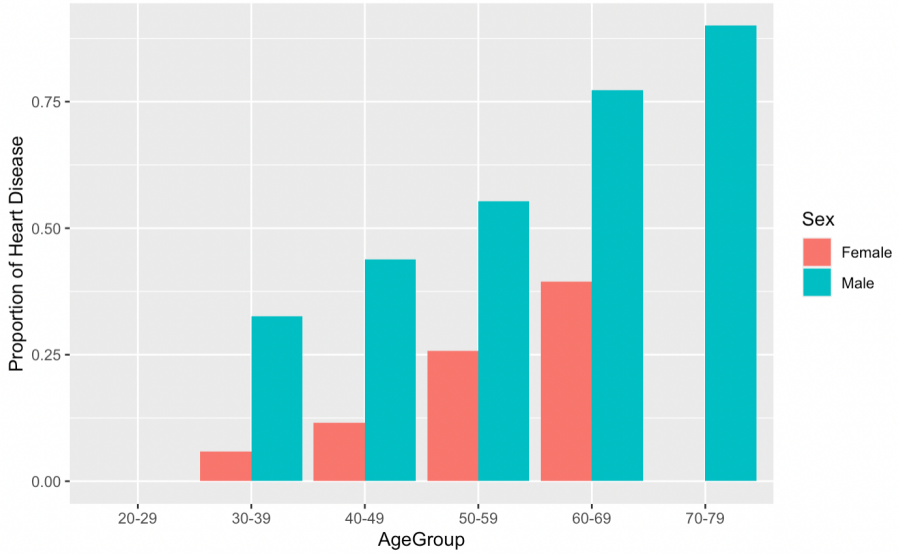
Applying the statistical model further proved our assumption. The p-value is less than the significant value, and therefore we can conclude that **the age has a positive correlation with the proportional population of heart diseased patients.**

3. Sex

Why this factor:
According to the Logistic Regression which we computed above, "Sex" has a p-value smaller than 0.05, indicating this variable will influence a person who will get heart disease or not.

Preprocessing:
Based on the 'Age' analysis, we go on by using groupby(AgeGroup, sex) to get two groups of data for both males and females in different age groups. We divide every 10 years into an age group and consider the difference between sex in different age groups. The proportion is calculated by dividing the number of diseased patients by the total population in that group. The plot is shown below:

Summarizing plot:



Plot 3.1 'Sex' and 'Age' Group with proportional population of heart diseased patients

The blue bars stand for males and the red bars stand for females. We can observe from the plot that as the age increases, **both male and female populations have an increasing trend on heart disease rate**. We also noticed that there's no female sample from Age Group 70-79. We assume that female sample size seems to be a lot smaller than male sample size, so we counted from the dataset and we get:

Sex <chr>	n <int>
Female	182
Male	564

2 rows

Among all the data entries, only 182 of them show data from female patients while 564 are from male patients. **Nevertheless, we can easily see a pattern from the plot - with an increase in age, male patients have a higher probability of getting heart disease than female patients.**

To further prove the correlation between heart disease and Sex, we go ahead to do logistic regression:

```
Call:
glm(formula = HeartDisease ~ Sex + Age, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9931  -1.0170  -0.4606   0.9839   2.3773

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.342254   0.504490  -8.607  < 2e-16 ***
Sex          0.793146   0.103721   7.647 2.06e-14 ***
Age          0.071838   0.009157   7.845 4.33e-15 ***
---
```

Applying the statistical model further proved our assumption. **The p-value is smaller than the significant value, and the slope of sex is close to 1, which confirms the above conclusion that men (Sex = 1) are more likely to get heart disease.**

4. Fasting BS and ExerciseAngina

Why this factor:

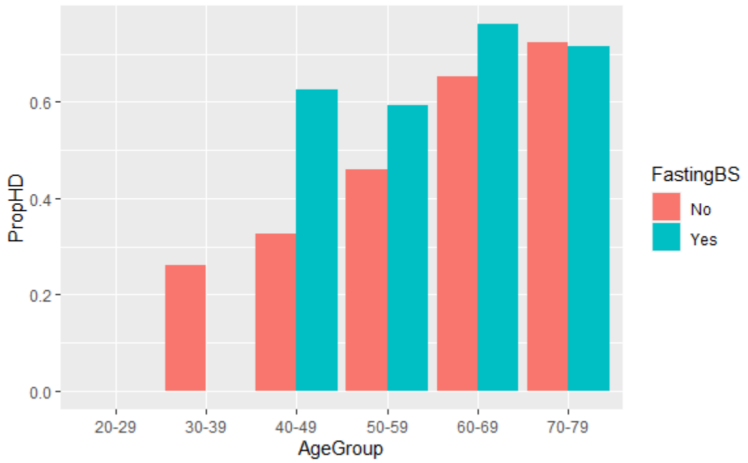
According to the Logistic Regression which we computed above, both “Fasting BS” and “ExerciseAngina” also have a p-value smaller than 0.05, indicating these variables will influence a person who will get heart disease or not.

Preprocessing:

Fasting BS: Based on the ‘Age’ analysis, we go on by using group_by(AgeGroup, FastingBS) to get two groups of data. The plot is shown below (3.1).

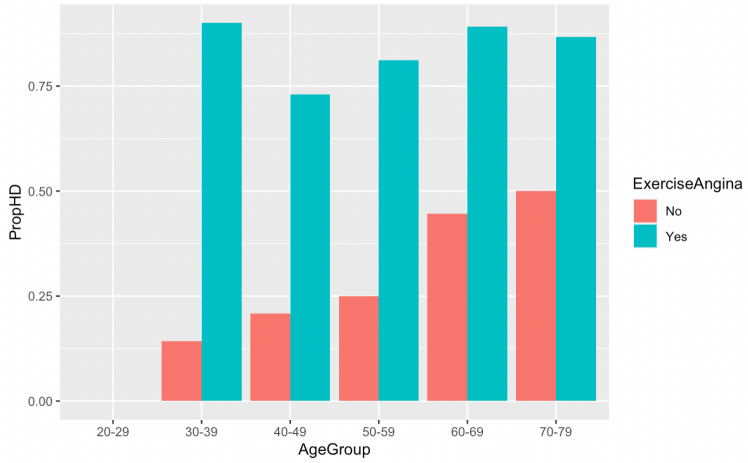
ExerciseAngina: Based on the ‘Age’ analysis, we go on by using group_by(AgeGroup, ExerciseAngina) to get two groups of data. The plot is shown below (3.2):

Summarizing plot:



Plot 4.1 FastingBS with proportional population of heart diseased patients

The red columns stand for FastingBS < 120 mg/dl, and blue columns stand for FastingBS > 120 mg/dl. **For patients whose fasting blood sugar is less than 120 mg/dl**, we can see a trend that as the age grows, people's probabilities of getting heart disease will increase. **The heart disease risk is mainly influenced by age**.
On the other hand, **for patients with > 120 mg/dl blood sugar**, the proportional population remained a comparative high throughout all age groups. Analyzing these two trends, we can assume that **having a FastingBS > 120 mg/dl will increase the risk of heart disease. Moreover, for ages between 60-80, no matter what the value of FastingBS is, the probability of heart disease is very high.**



Plot 4.2 ExerciseAngina with proportional population of heart diseased patients

For **patients without ExerciseAgina**(red columns), the proportional population showed a similar pattern with that in the Age graph. For **patients with ExerciseAgina** (green columns), the proportional population remained a comparative high throughout all age groups. Analyzing these two trends, we can assume that **having an Exercise Angina will increase the chance of heart disease, no matter what age group is.**

To further prove the correlation between heart disease and with FastingBS, and ExerciseAngina, we go ahead to do logistic regression:

```
Call:
glm(formula = HeartDisease ~ Age + FastingBS, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9617  -1.0741  -0.6845   1.0944   1.8831

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.289691   0.499079  -6.592 4.35e-11 ***
Age          0.063762   0.008918   7.150 8.70e-13 ***
FastingBS    0.273918   0.107201   2.555   0.0106 *
```

Applying the statistical model further proved our assumption. The p-value is smaller than the significant value. For FastingBS, the positive and smaller-than-0.5 slope indicates that there's a **slight positive correlation between having FastingBS and heart disease**. And the p-value of 0.0106, which is 0.05 level, explains why the proportion of heart disease is the same in the age between 60-80 no matter what FastingBS is. That is because there is no obvious correlation between FastingBS and the probability of heart disease in this age group.

```
Call:
glm(formula = HeartDisease ~ Age + ExerciseAngina, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3321  -0.7896  -0.5107   0.6587   2.1185

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.60957    0.54282  -4.807 1.53e-06 ***
Age             0.05363    0.01002   5.352 8.70e-08 ***
ExerciseAngina  1.23855    0.09606  12.894  < 2e-16 ***
```

For ExerciseAgina, the p-value is smaller than the significant value. The positive and over 1 slope indicates that **having ExerciseAgina**(ExerciseAgina = 1) **will lead to a higher risk of heart disease.**

5. Chest Pain

Why this factor:
According to the Logistic regression above, all four types of chest pain have a p-value smaller than 0.05, meaning that we can reject all four's null hypothesis.

Preprocessing:
Based on the original dataset, we group-by(AgeGroup, ChestPainType). We then summarize the count of the population, the sum of diseased patients, and the percentage of diseased populations.

Summarizing plot:

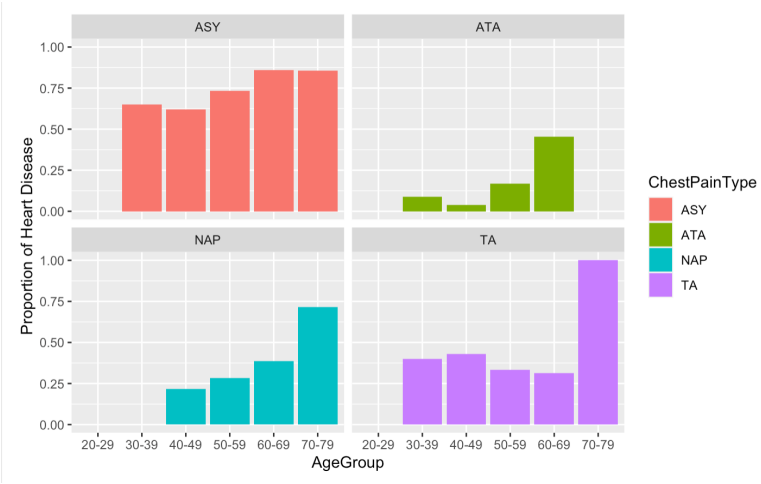
To understand each type of chest pain better, we counted them by different categories:

A tibble: 4 × 2 | Groups: ChestPainType [4]

ChestPainType <chr>	n <int>
ASY	370
ATA	166
NAP	169
TA	41

4 rows

From the count above, ASY (asymptomatic chest pain) is the major symptom for most patients. To further discover the pattern inside this data, we drew the proportional diseased populations with each type of chest pain based on the previous age group:



Plot 5.2 Proportion of Heart Disease with Age Group in Chest pain categories

From the plot above, we can conclude that:

1) **ASY:** ASY chest pain enlarge heart disease risk **throughout all age groups.**
2) **ATA & NAP:** these two chest pains **increase** the risk of heart disease **when age increases.**
3) **TA:** TA chest pain has a **subtle influence** on heart disease rate, but **increases the risk** of heart diseases **for patients whose age is between 70-79.**

To further prove the correlation between heart disease and ChestPainType, we go ahead to do logistic regression:

```
Call:
glm(formula = HeartDisease ~ Age + ChestPainType, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1948  -0.7272  -0.3549   0.7674   2.4116

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.12287    0.54459  -3.898 9.69e-05 ***
Age               0.05916    0.01008   5.867 4.44e-09 ***
ChestPainTypeATA -2.79948    0.26651 -10.504  < 2e-16 ***
ChestPainTypeNAP -2.03529    0.21693  -9.382  < 2e-16 ***
ChestPainTypeTA  -1.73914    0.36204  -4.804 1.56e-06 ***
---
```

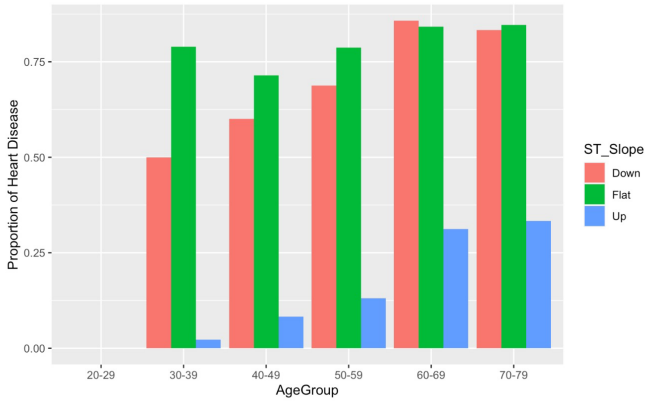
Applying the statistical model further proved our assumption. All p-values are smaller than the significant value, and the model indicates that **both Age and ChestPainTypes are significan**t by showing the importance of three stars.

6. ST_slope

Why this factor:
According to the Logistic Regression which we computed above, “ST_slope” has a p-value smaller than 0.05, indicating this variable will influence a person who will get heart disease or not.

Preprocessing:
First, we use as.factor to make the ST_Slope variable numeric. We mutate a new column ‘ST_Slope’ which contains three types of ST_Slope : Down, Flat and Up and then group by ‘AgeGroup’ and ‘ST_Slope’ to get the summary of the proportion to get heart disease by different ‘AgeGroup’ and ‘ST_Slope’. Finally, we make a column plot that shows the proportion of heart disease by different ST-Slope and AgeGroup.

Summarizing plot:



Plot 6.1 ST-Slope types with the proportion of heart disease rate

From the plot above, we may conclude that

1. Patients with Down ST-slope have **a much larger possibility** of having heart disease, and this influence also **increases as the age increases**.
2. A Flat ST-slope would **enlarge the risk** for heart disease **throughout all age groups**. Patients with this type of ST-slope have a **higher risk** for heart disease, even though they are considered as **low-risk according to their age-group**.
3. Patients with Up ST-slope share a **larger risk** for heart disease **as they age**, but the influence is much l**ess significant** than the other two ST-slope types.

To further prove the correlation between heart disease and ST-Slope types, we go ahead to do logistic regression:

```
Call:
glm(formula = HeartDisease ~ Age + ST_Slope, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0196  -0.6176  -0.4062   0.7599   2.2521

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.53889    0.57416  -2.680  0.00736 **
Age          0.04702    0.01053   4.465 8.01e-06 ***
ST_Slope    -2.56049    0.18589 -13.774  < 2e-16 ***
---
```

Applying the statistical model further proved our assumption. The p-value is obviously smaller than the significant value, and the model indicates that both Age and ST_Slope are **significant** by showing the importance of three stars.

7. Oldpeak

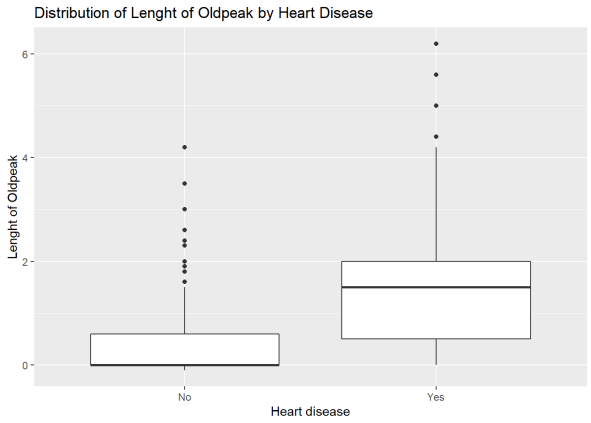
Why this factor:

According to the Logistic Regression which we computed above, “Oldpeak” has a p-value smaller than 0.05, indicating this variable will influence a person who will get heart disease or not.

Preprocessing:

We take the data with no zero value in the Cholesterol variable since the zero value may affect our final result. We also change the 1 and 0 values of heart disease to yes and no. We then use ggplot’s boxplot to present what is the length of oldpeak for most people who don’t have heart disease and who do have heart disease.

Summarizing plot:



Plot 7.1 Distribution of Length of Oldpeak by Heart Disease

The box plot above indicates that with a longer length of oldpeak, the risk of having heart disease significantly increases compared to people who don't have heart disease. The median length of oldpeak for people who **don't have** heart disease is approximately zero and 75% of people are below 1 mm. On the other hand, the median length of oldpeak for people who **have** heart disease is approximately 1.75 mm and 75% of people are below 2 mm.

To further prove the correlation between heart disease and ST-Slope types, we go ahead to do logistic regression:

```
Call:
glm(formula = HeartDisease ~ Oldpeak, family = "binomial", data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9167  -0.7419  -0.7419   0.8705   1.6880

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1494     0.1147  -10.02   <2e-16 ***
Oldpeak       1.2830     0.1063   12.07   <2e-16 ***
```

Applying the statistical model further proved our assumption. The p-value is obviously smaller than the significant value. The positive and over 1 slope indicates that having a longer length of oldpeak will lead to a higher risk of heart disease. The model indicates that old is significant by showing the importance of three stars.

8. Conclusion

From the analysis above, we found that:

- The heart disease rate is influenced by the factors of **'Cholesterol', 'Age', 'Sex', 'Fasting BS', 'ExerciseAngina', 'ChestPain', 'ST_slope' and 'Oldpeak'**.
- There is a positive relationship between **cholesterol** level and rate of getting heart disease.
- **Age** has a positive correlation with the rate of getting heart disease, and every unit increment in age leads to a 6.8% possibility of heart disease increment.
- For **Sex**, Both male and female have a higher risk of getting the disease as their age increases; moreover, male patients have a higher probability of getting heart disease than female patients when age increases.
- For **'FastingBS'**, a person with FastingBS > 120 mg/dl will have a greater risk of getting heart disease; on the other hand, for people with FastingBS < 120 mg/dl, age is the only factor that can influence the risk of getting heart disease. An exception is for ages between 60-80, 'FastingBS' has a smaller impact and the risk of heart disease mainly results from other factors.

- For **'ExerciseAngina'**, the rate of getting heart disease will increase if a person has 'ExerciseAngina', even his age shows a low rate of getting heart disease. Moreover, people with an older age still have a high risk of getting the disease.
- Among the four types of **chest pain**, all of them (ASY, ATA, NAP and TA) increase the risk of getting heart disease to some extent. Furthermore, patients with chest pain of ASY should especially pay more attention, and patients whose ages are between 70-79 shall pay attention if there are risks of getting the disease.
- For **ST_Slope**, all three types of ST-Slope (Down, Flat and Up) increase the risk of getting heart disease.
- People with a longer length of **oldpeak** have a significantly higher risk of getting heart disease.

We rebuilt our model by the analyzed factors from above:

```
Call:
glm(formula = HeartDisease ~ Age + Sex + ChestPainType + Cholesterol +
    FastingBS + ExerciseAngina + Oldpeak + ST_Slope, family = "binomial",
    data = heart_with_Cholesterol)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4372  -0.4263  -0.1419   0.5107   2.6188

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.957788   0.917039  -2.135 0.032769 *
Age               0.036441   0.012950   2.814 0.004893 **
Sex               0.832413   0.147193   5.655 1.56e-08 ***
ChestPainTypeATA -1.693311   0.327192  -5.175 2.28e-07 ***
ChestPainTypeNAP -1.538702   0.280947  -5.477 4.33e-08 ***
ChestPainTypeTA  -1.565575   0.461686  -3.391 0.000696 ***
Cholesterol       0.003553   0.001904   1.866 0.062043 .
FastingBS         0.156939   0.154830   1.014 0.310764
ExerciseAngina    0.553528   0.123792   4.471 7.77e-06 ***
Oldpeak           0.360983   0.136854   2.638 0.008346 **
ST_Slope         -1.762942   0.236444  -7.456 8.91e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1032.6  on 745  degrees of freedom
Residual deviance:  521.9  on 735  degrees of freedom
AIC: 543.9

Number of Fisher Scoring iterations: 5
```

Lastly, according to the regression model of all factors, the p-value of FastingBS is greater than 0.05, meaning that the extent to which FastingBS affects heart disease is very small.

Finding:

Based on our conclusion, we suggest that old people keep lower cholesterol levels and the FastingBS level should keep lower than 120 mg/dl. Also, they should take extra care when they have an Exercise Angina, long oldpeak and their ST_slopes are not 'Up' type.

Future Questions:

- Which major factor is most significant among all factors?
- What are some possible factors that are not in the dataset but also could contribute to the model?