There are 3 steps to this project, with the final deliverable being a formal paper (completed in R Markdown) that outlines the problem, shows the analysis done with the data, and concludes with your recommendation for next steps. Each step provides detailed information that you must include in each phase of the project.

Identify a topic or a problem that you want to research. Provide an introduction that explains the problem statement or topic you are addressing. Why would someone be interested in this? How is it a data science problem?

I would like to research heart disease. Heart disease is one of the leading causes of death. It's important to find out what factors might increase risks for heart disease so that we know who's at a high risk. By looking at mass data of those with heart disease, we can see what they have in common and what to watch out for, for people who haven't developed it yet.

At this point you should have framed your problem/topic, described the data, and how you plan to solve the problem. Now you need to move on to the next step of analyzing and preparing the data. Adding on to the draft you started in Step 1: Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.

```
library(knitr)
library(ggplot2)
library(purrr)
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## v dplyr   1.0.10
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stats)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(dplyr)
```

```r
#data importing: not sure why my code is cutting off???

data_1 = read.csv("C:/Users/brean/OneDrive/Desktop/final project/parkerDSC520final/heart_failure_predic

data_2 = read.csv("C:/Users/brean/OneDrive/Desktop/final project/parkerDSC520final/heart_disease_kaggle_

data_3 = read.csv("C:/Users/brean/OneDrive/Desktop/final project/parkerDSC520final/heart_data_kaggle_be

head(data_1)
```

```
##   Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
## 1  40   M           ATA       140         289         0     Normal   172
## 2  49   F           NAP       160         180         0     Normal   156
## 3  37   M           ATA       130         283         0         ST    98
## 4  48   F           ASY       138         214         0     Normal   108
## 5  54   M           NAP       150         195         0     Normal   122
## 6  39   M           NAP       120         339         0     Normal   170
##   ExerciseAngina Oldpeak ST_Slope HeartDisease
## 1              N     0.0       Up            0
## 2              N     1.0     Flat            1
## 3              N     0.0       Up            0
## 4              Y     1.5     Flat            1
## 5              N     0.0       Up            0
## 6              N     0.0       Up            0
```

```r
head(data_2)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  52   1  0      125  212   0       1     168     0     1.0     2  2    3
## 2  53   1  0      140  203   1       0     155     1     3.1     0  0    3
## 3  70   1  0      145  174   0       1     125     1     2.6     0  0    3
## 4  61   1  0      148  203   0       1     161     0     0.0     2  1    3
## 5  62   0  0      138  294   1       1     106     0     1.9     1  3    2
## 6  58   0  0      100  248   0       0     122     0     1.0     1  0    2
##   target
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
## 6      1
```

```r
head(data_3)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63   1  3      145  233   1       0     150     0     2.3     0  0    1
## 2  37   1  2      130  250   0       1     187     0     3.5     0  0    2
## 3  41   0  1      130  204   0       0     172     0     1.4     2  0    2
## 4  56   1  1      120  236   0       1     178     0     0.8     2  0    2
## 5  57   0  0      120  354   0       1     163     1     0.6     2  0    2
## 6  57   1  0      140  192   0       1     148     0     0.4     1  0    1
##   target
```

```
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```
#data cleaning: changing all names to be consistent

data_1 <- data_1 %>% rename(age = Age)
colnames(data_1)[1]
```

```
## [1] "age"
```

```
data_1 <- data_1 %>% rename(sex = Sex)
colnames(data_1)[2]
```

```
## [1] "sex"
```

```
data_1 <- data_1 %>% rename(chest_pain = ChestPainType)
colnames(data_1)[3]
```

```
## [1] "chest_pain"
```

```
data_2 <- data_2 %>% rename(chest_pain = cp)
colnames(data_2)[3]
```

```
## [1] "chest_pain"
```

```
data_3 <- data_3 %>% rename(chest_pain = cp)
colnames(data_3)[3]
```

```
## [1] "chest_pain"
```

```
data_1 <- data_1 %>% rename(rest_bp = RestingBP)
colnames(data_1)[4]
```

```
## [1] "rest_bp"
```

```
data_2 <- data_2 %>% rename(rest_bp = trestbps)
colnames(data_2)[4]
```

```
## [1] "rest_bp"
```

```r
data_3 <- data_3 %>% rename(rest_bp = trestbps)
colnames(data_3)[4]
```

```
## [1] "rest_bp"
```

```r
data_1 <- data_1 %>% rename(chol = Cholesterol)
colnames(data_1)[5]
```

```
## [1] "chol"
```

```r
data_1 <- data_1 %>% rename(fasting_bs = FastingBS)
colnames(data_1)[6]
```

```
## [1] "fasting_bs"
```

```r
data_2 <- data_2 %>% rename(fasting_bs = fbs)
colnames(data_2)[6]
```

```
## [1] "fasting_bs"
```

```r
data_3 <- data_3 %>% rename(fasting_bs = fbs)
colnames(data_3)[6]
```

```
## [1] "fasting_bs"
```

```r
data_1 <- data_1 %>% rename(max_heart_rate = MaxHR)
colnames(data_1)[8]
```

```
## [1] "max_heart_rate"
```

```r
data_2 <- data_2 %>% rename(max_heart_rate = thalach)
colnames(data_2)[8]
```

```
## [1] "max_heart_rate"
```

```r
data_3 <- data_3 %>% rename(max_heart_rate = thalach)
colnames(data_3)[8]
```

```
## [1] "max_heart_rate"
```

```r
data_1 <- data_1 %>% rename(rest_ecg = RestingECG)
colnames(data_1)[7]
```

```
## [1] "rest_ecg"
```

```r
data_2 <- data_2 %>% rename(rest_ecg = restecg)
colnames(data_2)[7]
```

```
## [1] "rest_ecg"
```

```r
data_3 <- data_3 %>% rename(rest_ecg = restecg)
colnames(data_3)[7]
```

```
## [1] "rest_ecg"
```

```r
data_1 <- data_1 %>% rename(heart_disease = HeartDisease)
colnames(data_1)[12]
```

```
## [1] "heart_disease"
```

```r
data_2 <- data_2 %>% rename(heart_disease = target)
colnames(data_2)[14]
```

```
## [1] "heart_disease"
```

```r
data_3 <- data_3 %>% rename(heart_disease = target)
colnames(data_3)[14]
```

```
## [1] "heart_disease"
```

```r
head(data_1)
```

```
##   age sex chest_pain rest_bp chol fasting_bs rest_ecg max_heart_rate
## 1  40   M        ATA     140  289          0   Normal            172
## 2  49   F        NAP     160  180          0   Normal            156
## 3  37   M        ATA     130  283          0       ST             98
## 4  48   F        ASY     138  214          0   Normal            108
## 5  54   M        NAP     150  195          0   Normal            122
## 6  39   M        NAP     120  339          0   Normal            170
##   ExerciseAngina Oldpeak ST_Slope heart_disease
## 1              N     0.0       Up             0
## 2              N     1.0     Flat             1
## 3              N     0.0       Up             0
## 4              Y     1.5     Flat             1
## 5              N     0.0       Up             0
## 6              N     0.0       Up             0
```

```r
head(data_2)
```

```
##   age sex chest_pain rest_bp chol fasting_bs rest_ecg max_heart_rate exang
## 1  52   1          0     125  212          0        1            168     0
## 2  53   1          0     140  203          1        0            155     1
## 3  70   1          0     145  174          0        1            125     1
## 4  61   1          0     148  203          0        1            161     0
```

```
## 5  62   0          0     138  294          1        1            106     0
## 6  58   0          0     100  248          0        0            122     0
##   oldpeak slope ca thal heart_disease
## 1     1.0     2  2    3             0
## 2     3.1     0  0    3             0
## 3     2.6     0  0    3             0
## 4     0.0     2  1    3             0
## 5     1.9     1  3    2             0
## 6     1.0     1  0    2             1
```

```
head(data_3)
```

```
##   age sex chest_pain rest_bp chol fasting_bs rest_ecg max_heart_rate exang
## 1  63   1          3     145  233          1        0            150     0
## 2  37   1          2     130  250          0        1            187     0
## 3  41   0          1     130  204          0        0            172     0
## 4  56   1          1     120  236          0        1            178     0
## 5  57   0          0     120  354          0        1            163     1
## 6  57   1          0     140  192          0        1            148     0
##   oldpeak slope ca thal heart_disease
## 1     2.3     0  0    1             1
## 2     3.5     0  0    2             1
## 3     1.4     2  0    2             1
## 4     0.8     2  0    2             1
## 5     0.6     2  0    2             1
## 6     0.4     1  0    1             1
```

```
#more data cleaning: making sure all types are integers for easy use


typeof(data_1$age)
```

```
## [1] "integer"
```

```
typeof(data_1$sex)
```

```
## [1] "character"
```

```
typeof(data_1$chest_pain)
```

```
## [1] "character"
```

```
typeof(data_1$rest_bp)
```

```
## [1] "integer"
```

```
typeof(data_1$chol)
```

```
## [1] "integer"
```

```
typeof(data_1$fasting_bs)
```

```
## [1] "integer"
```

```
typeof(data_1$rest_ecg)
```

```
## [1] "character"
```

```
typeof(data_1$max_heart_rate)
```

```
## [1] "integer"
```

```
typeof(data_1$heart_disease)
```

```
## [1] "integer"
```

```
typeof(data_2$age)
```

```
## [1] "integer"
```

```
typeof(data_2$sex)
```

```
## [1] "integer"
```

```
typeof(data_2$chest_pain)
```

```
## [1] "integer"
```

```
typeof(data_2$rest_bp)
```

```
## [1] "integer"
```

```
typeof(data_2$chol)
```

```
## [1] "integer"
```

```
typeof(data_2$fasting_bs)
```

```
## [1] "integer"
```

```
typeof(data_2$rest_ecg)
```

```
## [1] "integer"
```

```
typeof(data_2$max_heart_rate)
```

```
## [1] "integer"
```

```
typeof(data_2$heart_disease)
```

```
## [1] "integer"
```

```
typeof(data_3$age)
```

```
## [1] "integer"
```

```
typeof(data_3$sex)
```

```
## [1] "integer"
```

```
typeof(data_3$chest_pain)
```

```
## [1] "integer"
```

```
typeof(data_3$rest_bp)
```

```
## [1] "integer"
```

```
typeof(data_3$chol)
```

```
## [1] "integer"
```

```
typeof(data_3$fasting_bs)
```

```
## [1] "integer"
```

```
typeof(data_3$rest_ecg)
```

```
## [1] "integer"
```

```
typeof(data_3$max_heart_rate)
```

```
## [1] "integer"
```

```
typeof(data_3$heart_disease)
```

```
## [1] "integer"
```

```
data_1$chest_pain <- as.integer(data_1$chest_pain)
```

```
## Warning: NAs introduced by coercion
```

```
data_1$sex <- as.integer(data_1$sex)
```

```
## Warning: NAs introduced by coercion
```

```
data_1$rest_ecg <- as.integer(data_1$rest_ecg)
```

```
## Warning: NAs introduced by coercion
```

```
#more data cleaning: making sure no "impossible" data. IE, male = 1, female = 0, and the value is a 3.

unique(data_1$sex)
```

```
## [1] NA
```

```
unique(data_2$sex)
```

```
## [1] 1 0
```

```
unique(data_3$sex)
```

```
## [1] 1 0
```

```
unique(data_1$heart_disease)
```

```
## [1] 0 1
```

```
unique(data_2$heart_disease)
```

```
## [1] 0 1
```

```
unique(data_3$heart_disease)
```

```
## [1] 1 0
```

What do you not know how to do right now that you need to learn to import and cleanup your dataset?

My only issues is in one column of one dataset it's characters but in the other two, it's integers but it's the same data. Might need to learn how to change all values of a character to a value. i.e. Chest_pain_type in the first data set as ATA for atypical angina but the other two have it as a 1. Also, it looks like I have a lot of cholesterol values that are 0? might need to remove those.

#https://towardsdatascience.com/data-cleaning-in-r-made-simple-1b77303b0b17 (personal note for future cleaning tools!)

age: age in years sex: sex [1 = male, 0 = female] chest_pain: chest pain type [Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic] rest_bp: resting blood pressure (in mm Hg) chol: serum cholestoral in mg/dl fasting_bs: (fasting blood sugar > 120 mg/dl) [1 = true; 0 = false] max_heart_rate: maximum heart rate achieved

rest_ecg: resting electrocardiographic results [Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria]

exang: exercise induced angina [1 = yes, 0 = no] oldpeak = ST depression induced by exercise relative to rest slope: the slope of the peak exercise ST segment [Value 0: upsloping, Value 1: flat, Value 2: downsloping] ca: number of major vessels (0-3) colored by flourosopy thal: [0 = error (in the original dataset 0 maps to NaN's), 1 = fixed defect, 2 = normal, 3 = reversable defect] heart_disease: [0 = no disease, 1 = disease]

Draft 5-10 Research questions that focus on the problem statement/topic.

1. Are people with diabetes more likely to develop heart disease at an early age?
2. Is there a certain demographic group that is at higher risk of heart disease than others?
3. Does frequent exercise lower the risk of developing heart disease?
4. Are smokers more likely to develop heart disease?
5. How many deaths does heart disease cause?

Provide a concise explanation of how you plan to address this problem statement.

We can look through datasets that have other medical data so that we can see what the most likely predictors of heart disease are. My project will simply be finding out the predictors for heart disease. Once this is known, the data can be used as a preventive measure for people who have this predictors.

Discuss how your proposed approach will address (fully or partially) this problem.

If we find out that everyone with heart disease has a certain feature in common, we can start educating people and helping to reduce the deaths caused by heart disease. This project is simply finding out what people with heart disease have in common so that we can detect people who are at a high risk for heart disease early on and maybe start taking preventive measures to prevent heart disease all together.

Do some digging and find at least 3 datasets that you can use to address the issue. (There is not a required number of fields or rows for these datasets) Original source where the data was obtained is cited and, if possible, hyperlinked.

(fedesoriano 2021; Lapp 1988; Belsonraja 2022)

Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. 12 variables are used

https://www.kaggle.com/datasets/belsonraja/heart-disease-prediction Its the heart disease dataset from this dataset we can derive various insights that help us know the weightage of each feature and how they are interrelated to each other but this time our sole aim is to detect the probability of person that will be affected by a savior heart problem or not. 13 variables used.

Discuss how you plan to uncover new information in the data that is not self-evident.

By running different equations and going over values like the p-value to draw conclusions from the data. ALso, running comparisons and removing outliers to have cleaner data.

What are different ways you could look at this data to answer the questions you want to answer? # I want to start by creating many different tables and graphs so that I can better visualize the data. This will also help me decide what I want to test next and what equations might be best to run. Going through three different dataesets all with more than 12 variables is definitely a lot of data and it isn't feasible to run every equation on every variable.

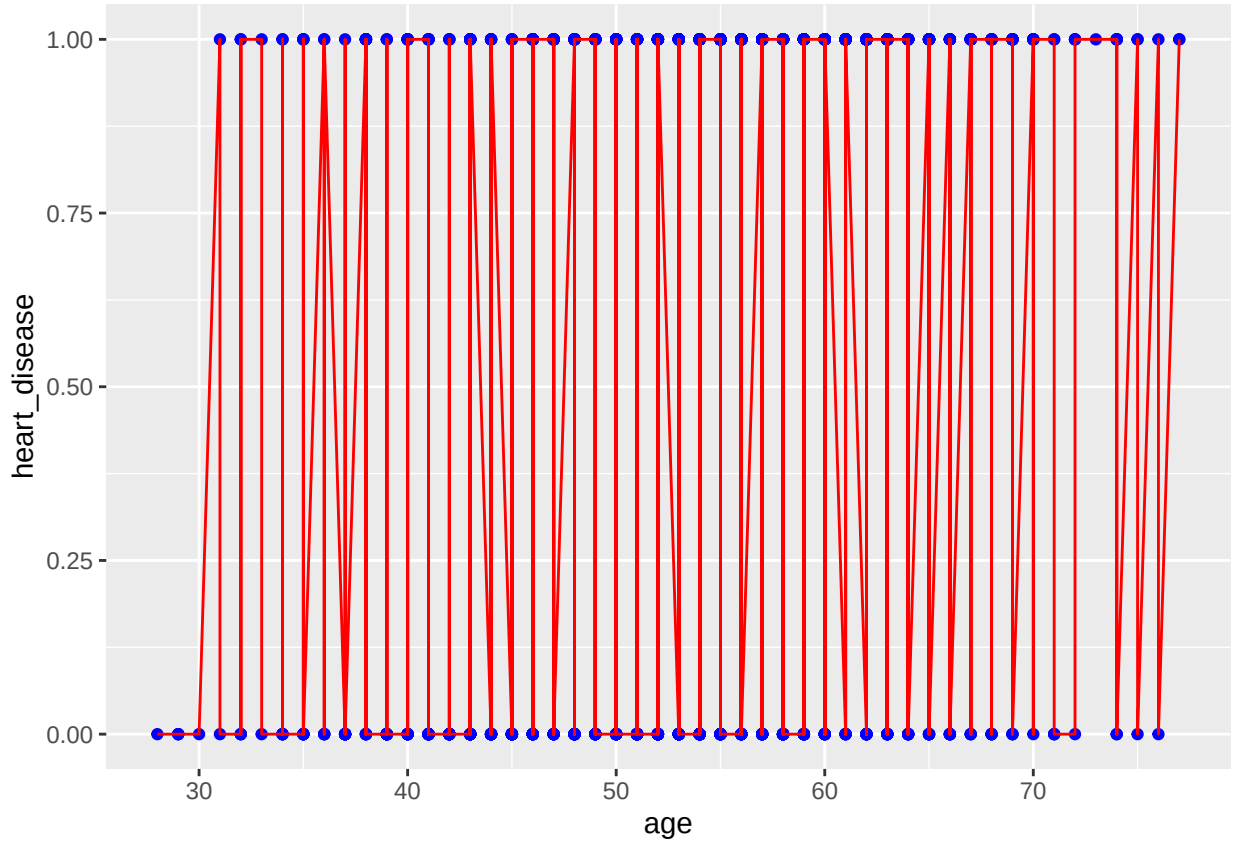
Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

It might make things easier to combine all three datasets into one, but it's also interseting to look at graphs of all three separted and how they differ. I might create a new variable for probability or prediction.

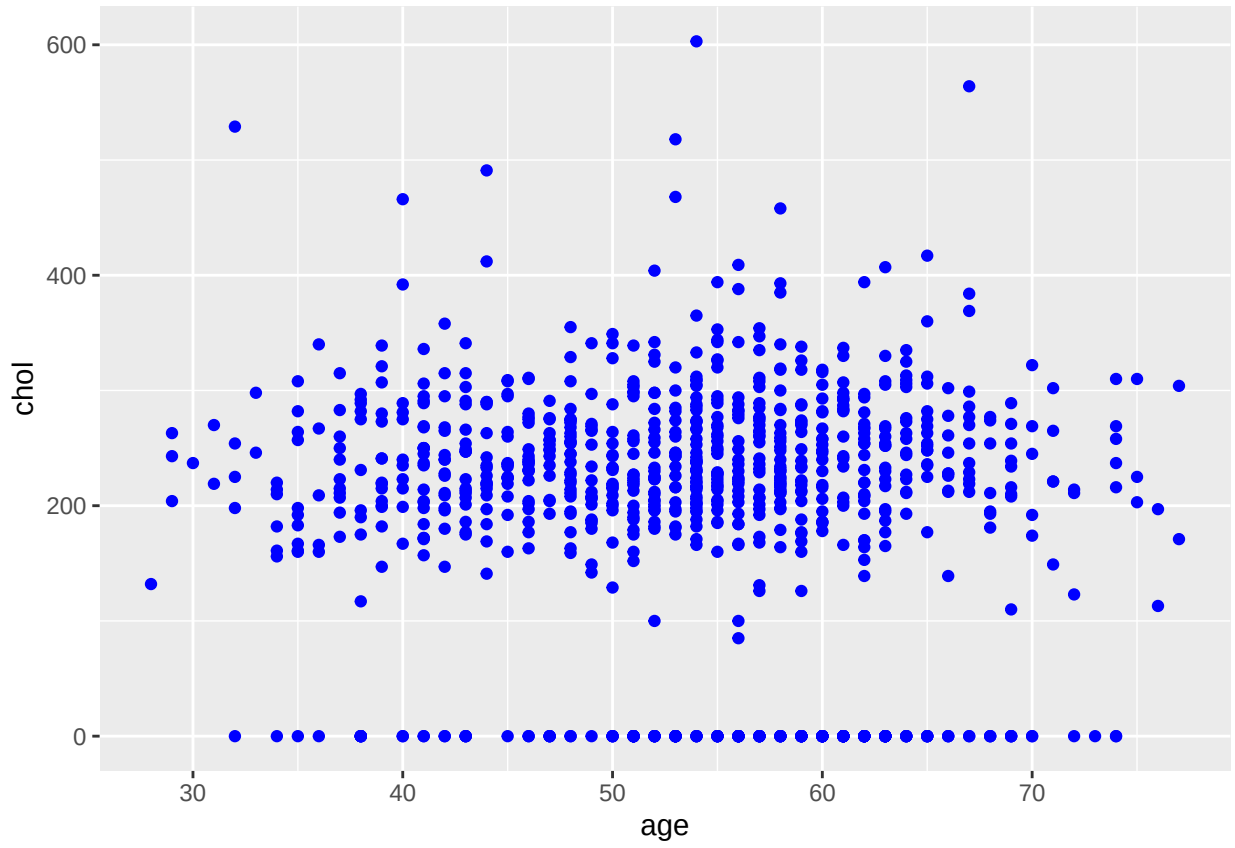
How could you summarize your data to answer key questions? # I think there are certain columns that I might not use in my calculations. Biggest things we need to know is if there is a correlation between heart disease and certain factors and if it's statistically significant.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

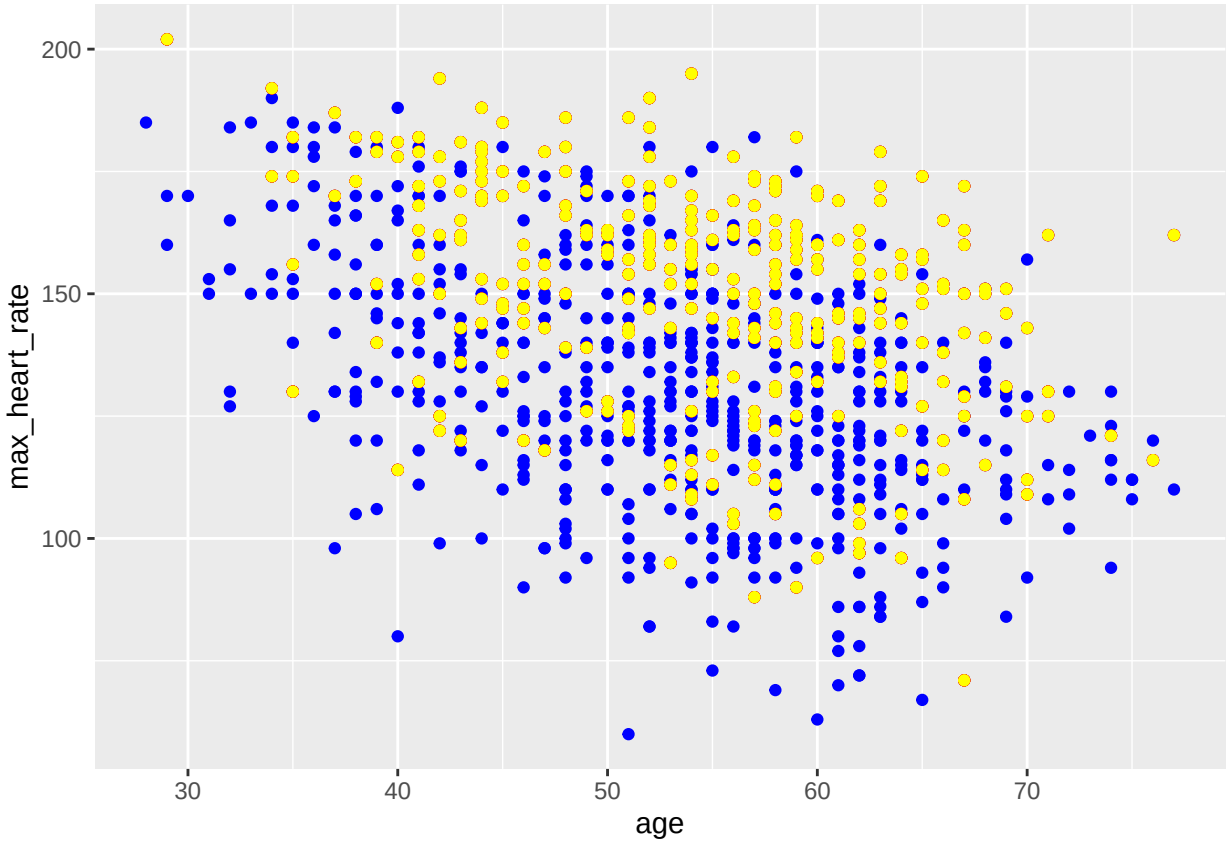
```
ggplot(data = data_1, aes(y = heart_disease, x = age)) +
  geom_point(color='blue') +
  geom_line(color='red',data = data_1, aes(y = heart_disease, x = age))
```



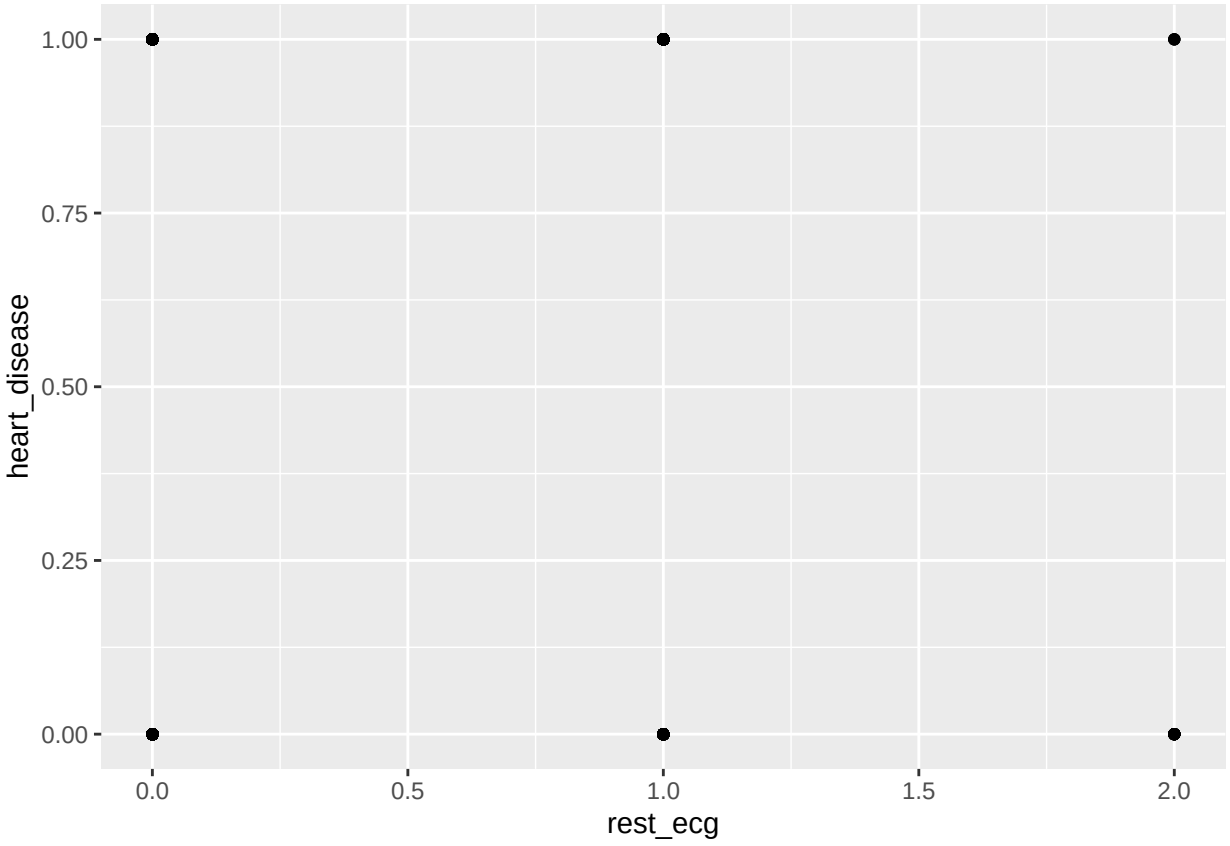

```
ggplot(data = data_1, aes(y = chol, x = age)) +
  geom_point(color='blue')
```




```
#not sure why the 2nd data set isn't showing on this one???
ggplot(data = data_1, aes(y = max_heart_rate, x = age)) +
  geom_point(color='blue') +
  geom_point(color='red',data = data_2, aes(y = max_heart_rate, x = age)) +
  geom_point(color='yellow',data = data_3, aes(y = max_heart_rate, x = age))
```




```
#below plot isn't working
ggplot(data_2, aes(rest_ecg, heart_disease)) + geom_point()
```



above is a few examples of some plots that I'm working on and then might incorporate more of these ones below: table() scatterplot() density plot geom_boxplot geom_bar pandoc table

What do you not know how to do right now that you need to learn to answer your questions?

I need to learn how to better read the summary function to draw conclusions on datasets. But I'll also need to run comparisons on the data to see what is causation vs correlation and how much weight each risk factor might have on the other. This should be easy to do with what we have learned in class.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I don't think I'll need to use any machine learning techniques for this project. I don't need to make any predictions using unsupervised or supervised learning that I'm aware of.

Some additional questions you may want to consider asking yourself as you work through this section of the project: What features could you filter on? How could arranging your data in different ways help? Can you reduce your data by selecting only certain variables? Could creating new variables add new insights? Could

summary statistics at different categorical levels tell you more? How can you incorporate the pipe (%>%) operator to make your code more efficient?

Bibliography

Belsonraja. 2022. *Heart Disease Prediction*. https://www.kaggle.com/datasets/belsonraja/heart-disease-prediction.

fedesoriano. 2021. *Heart Failure Prediction Dataset*. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

Lapp, David. 1988. *Heart Disease Dataset*. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.