Project Part 1

Name: blank for part 1

Data Set

```
##   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
```

Description of Data Set

Background information

This data set contains information related to the presence of heart disease in patients. The data was compiled by four separate contributors, Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano, all of whom are MDs. The data was donated on 1988-07-01, and represents a sample 303 patients. The data was found on the UCI Machine Learning repository, which is linked in references below. However, the csv for the dataset was found on Kaggle, which was primarily used in this part of the project for ease of transformation.

While these creators have this dataset stored in multiple different databases, ranging from VA Beach to Hungary and Switzerland, Clevand is the most widely accepted database. To this day, only the Cleveland database has been used by machine learning researchers. While not directly stated, it is extremely likely that these patients were randomly sampled from clinics where the databases are based on, such as the Cleveland Clinic Foundation and the V.A. Medical Center. It is unsure how the data was collected, but the source of the dataset claims that names and social security numbers were removed from the dataset recently. Therefore, it is implied that the contributors to the dataset collected the data from medical records of the randomly selected patients.

The dataset consists of 76 attributes, but the datset being used only contains 14 of these, since all published experiments refer mainly to these 14 attributes.

Explanation of rows and variables

Each row contains a single observation, an observation being a patient. Each patient then has 14 attributes selected to describe them. These are age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num.

Age is the age in years, sex is the sex of the patient, 1 = male, 0 = female.

CP stands for chest pain type, with value 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic.

trestbps is resting blood pressure in mm Hg.

chol is serum cholesterol in mg/dl.

fbs is fasting blood sugar > 120 mg/dl, with 1 = true, 0 = false.

restecg is resting electrocardiographic results, with value 0 = normal, 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), and 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria.

thalach is maximum heart rate achieved.

exang is exercise induced angina, with 1 = yes and 0 = no.

oldpeak is ST depression induced by exercise relative to rest.

slope is the slope of the peak exercise ST segmet, with 1 = upsloping, 2 = flat, 3 = downsloping.

ca is number of major vessels (0-3) colored by flouroscopy.

thal stands for thalium, a radioactive tracer injected during a stress test. The tracer usually has a detector attached which can sense a coronary stenosis, also known as a "reversible defect". 1 = fixed defect, 2 = normal, 3 = reversible defect.

target is whether the patient has heart disease, with 0 = heart disease and 1 = no heart disease.

Potential issues

I came across a plethora of issues with the data when I first encountered it. I initially downloaded the data off of Kaggle, which I thought would be identical to the data in the UCI Machine Learning databse. However upon closer look the two sets had major differences. Firstly, several values were inputted incorrectly. The ca attribute has 4 possible values ranging from 0-3, but in the heart.csv dataset downloaded, data #93, 159, 164, 165 and 252 had ca = 4 which was out of range. Upon further inspection, it turns out that these values are listed as NaN in UCI's database, but was somehow transformed into 4 into kaggle's csv. A similar issue was encountered in the thal column where values listed as NaN in UCI's dataset was listed as 0 in the csv.

Another issues was that the descriptions of the variables and dataset were not clear. For example, in the Kaggle description all the variables are named but their values and what they represent are not made clear, such as for chest pain type or restecg. These had to be separately researched in forums to figure out what each value meant. Additionally, some of the values listed for the variables are incorrect. For example, thal is listed as 3 = normal, 6 = fixed defect, and 7 = reversible defect, when the values are actually listed as 1, 2, 3 in the csv. This was confirmed by cross referencing the two datasets and checking out posts in discussions where these issues were fixed. Perhaps the greatest error was in the description of target, the most important attribute, which was not described in the Kaggle description. The UCI database has it listed as 0 having no disease and 1 having a disease, while the kaggle dataset flips this around and lists 0 as having heart disease and 1 not having heart disease.

Lastly, there were multiple issues with spelling that I encountered. These were small things like medical terminology or processes such as "fluoroscopy". However, these typos were all in descriptions and not in the dataset itself so it is unlikely it will cause any issues in data manipulation.

Representations

```
# Numerical summary 1

heartGender <- heart

heartGender$sex[heartGender$sex == "1"] <- "Male"
heartGender$sex[heartGender$sex == "0"] <- "Female"

group_by(heartGender, sex) %>%
  summarize(Disease_Likeliness=mean(target))
```

```
## # A tibble: 2 x 2
##   sex    Disease_Likeliness
##   <chr>               <dbl>
## 1 Female              0.75
## 2 Male                0.449
```

This is the likelihood of each gender having heart disease. It's relatively simple and does not take into account other factors, only considering the proportion of that gender which has heart disease to the entire gender population.

```
# Numerical summary 2

group_by(heart, age) %>%
  summarize(Average_RestingBPS =mean(trestbps))
```

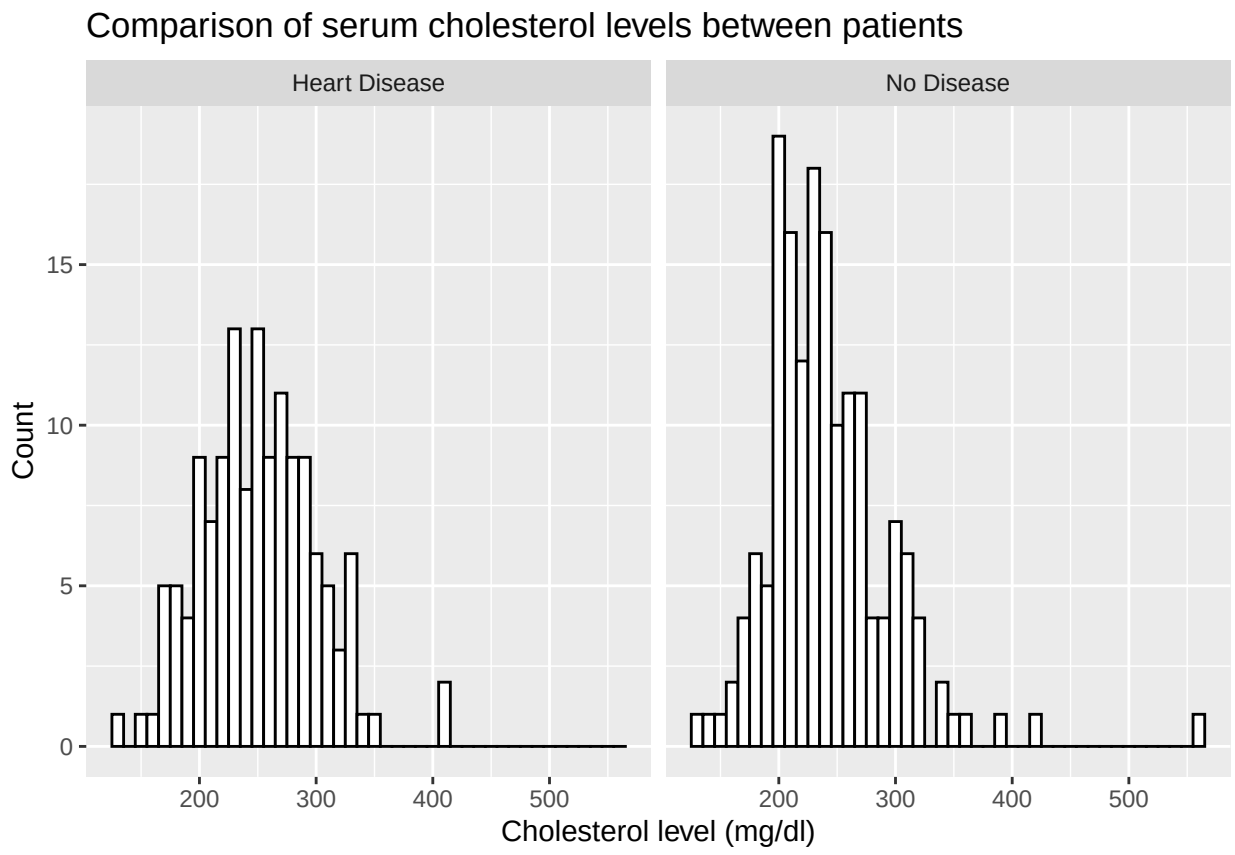
```
## # A tibble: 41 x 2
##      age Average_RestingBPS
##    <int>              <dbl>
##  1    29               130
##  2    34               118
##  3    35               126.
##  4    37               125
##  5    38               132
##  6    39               122.
##  7    40               134
```

```
##  8    41               119
##  9    42               127
## 10    43               126.
## # ... with 31 more rows
```

This is a summary of resting heart rate over age, and was graphed to see if there was any correlation between resting BPS and age. A graphical summary better represents the data later on.

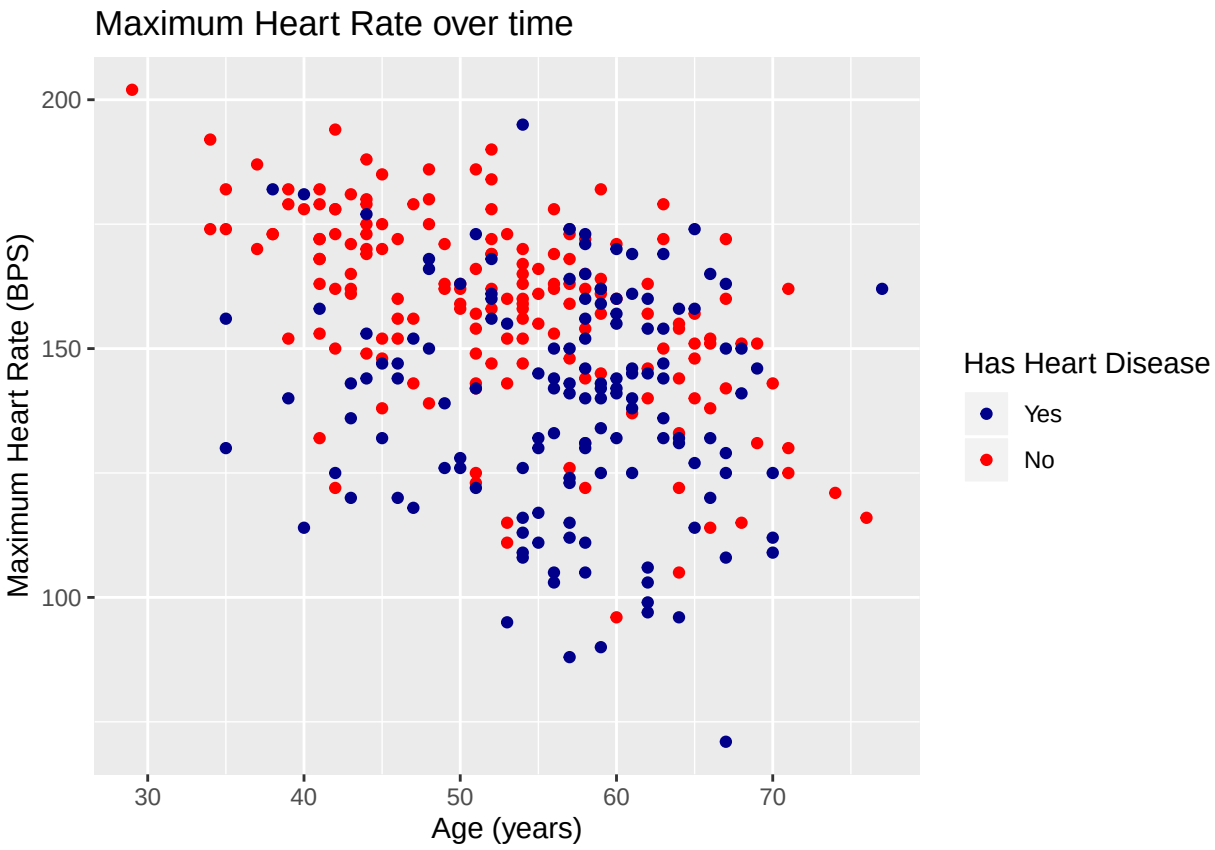
```
# Graphical representation 1

cholCompare <- ggplot(heart, aes(x=chol))
cholCompare + geom_histogram(aes(y=..count..),
                             binwidth=10, fill="white", color="black") +
  facet_grid(.~target, labeller = labeller(target =
                                           c('0'="Heart Disease", '1'="No Disease"))) +
  labs(title = "Comparison of serum cholesterol levels between patients",
               x="Cholesterol level (mg/dl)", y = "Count")
```



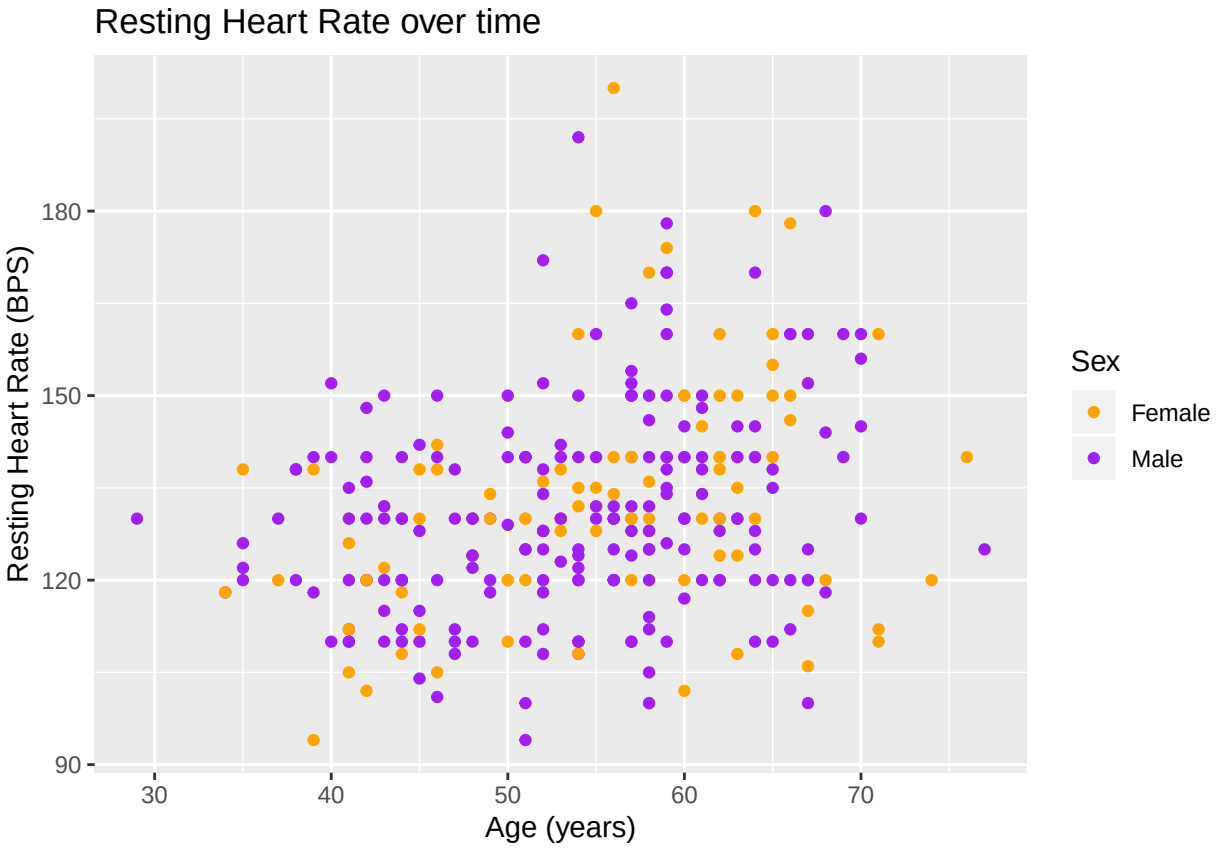
Interestingly enough, cholesterol levels are higher on average for those who have heart disease. Since this is a count histogram, more people without heart disease have levels closer to 200, showing the large spike in that general area, while the median for those with disease is closer to 300.

```
# Graphical Representation 2
rateCompare <- ggplot(heart, aes(x=age, y =thalach, color = as.factor(target)))
rateCompare + geom_point() +
  scale_color_manual(values = c("0" = "darkblue", "1" = "red"),
                                                    name="Has Heart Disease",
                                                breaks=c("0","1"), labels=c("Yes","No")) +
  labs(title="Maximum Heart Rate over time",
x="Age (years)", y="Maximum Heart Rate (BPS)")
```



This scatter plot demonstrates a negative linear trend. It appears that those with heart disease have lower maximum heart rates compared to those who do not have heart disease; however, the sample may be a little skewed. This is because a large portion of samples that are older than 50 have heart disease, while younger samples rarely have heart disease.

```
# Graphical Representation 3
sexCompare <- ggplot(heart, aes(x=age, y =trestbps, color = as.factor(sex)))
sexCompare + geom_point() +
  scale_color_manual(values = c("0" = "orange", "1" = "purple"),
                                                name="Sex", breaks=c("0","1"),
                                                labels=c("Female","Male")) +
  labs(title="Resting Heart Rate over time",
x="Age (years)", y="Resting Heart Rate (BPS)")
```



This was graphed to see if there was any correlation between resting heart rate and gender over time. While one could argue that the graph shows a little but of a positive trend near the 50-65 age range, the points are relatively scattered and show a constant trend.

References

1. https://archive.ics.uci.edu/ml/datasets/heart+Disease Origin of the dataset
2. https://www.kaggle.com/ronitf/heart-disease-uci Where the csv was taken from