Project Proposal Resubmission

due October 11, 2021 by 11:59 PM

Shelby Brown, Katie Lam, Kaeden Hill

10/17/2021

Load Packages

```
library(tidyverse)
```

Load Data

```
heart <- readr::read_csv("heart.csv")
```

Changes Made for Resubmission

The research question was updated, along with the explanation of the statistical tests that will be used to analyze the data and answer the question (the last four paragraphs).

Introduction and Data, including Research Questions

For this data visualization, we will be analyzing chest pain types and their relation to other physiological factors. We will be looking for an association between factors, such as blood pressure, cholesterol, and exercise, and the type of chest pain a patient experiences. We also plan on using this dataset to look for a possible association between chest pain type and whether that patient experiences heart disease. The dataset we have chosen is the Heart Failure Prediction Dataset. We retrieved it from Kaggle, and it is a compiled dataset from five sets with common variables. These sources are the Hungarian Institute of Cardiology. Budapest, University Hospital, Zurich, Switzerland, University Hospital, Basel, Switzerland, the V.A. Medical Center, Long Beach, and the Cleveland Clinic Foundation. The variables of interest include the age of the patient, gender, their resting blood pressure (mm Hg), their serum cholesterol (mm/d), and whether they have heart disease. We will also be using the categorical variables compiled on the chest pain type (typical angina, atypical angina, non-anginal pain, or asymptomatic), and whether the angina was exercise induced (yes or no).

Research Question: Is either blood pressure, cholesterol level, or whether or not chest pain was exercise induced, a good predictor for the type of chest pain experienced? Is a certain chest pain type more associated with heart disease?

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved 10/09/2021 from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Glimpse

```
glimpse(heart)
```

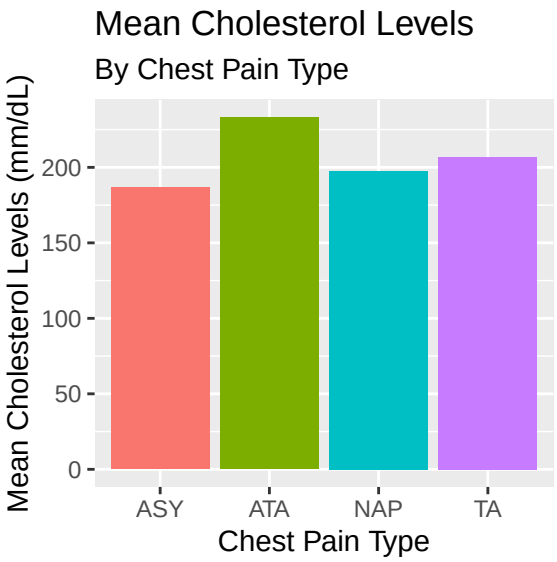
```
## Rows: 918
## Columns: 12
## $ Age            <dbl> 40, 49, 37, 48, 54, 39, 45, 54, 37, 48, 37, 58, 39, 49,~
## $ Sex            <chr> "M", "F", "M", "F", "M", "M", "F", "M", "M", "F", "F", ~
## $ ChestPainType  <chr> "ATA", "NAP", "ATA", "ASY", "NAP", "NAP", "ATA", "ATA",~
## $ RestingBP      <dbl> 140, 160, 130, 138, 150, 120, 130, 110, 140, 120, 130, ~
## $ Cholesterol    <dbl> 289, 180, 283, 214, 195, 339, 237, 208, 207, 284, 211, ~
## $ FastingBS      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ RestingECG     <chr> "Normal", "Normal", "ST", "Normal", "Normal", "Normal",~
## $ MaxHR          <dbl> 172, 156, 98, 108, 122, 170, 170, 142, 130, 120, 142, 9~
## $ ExerciseAngina <chr> "N", "N", "N", "Y", "N", "N", "N", "N", "Y", "N", "N", ~
## $ Oldpeak        <dbl> 0.0, 1.0, 0.0, 1.5, 0.0, 0.0, 0.0, 0.0, 1.5, 0.0, 0.0, ~
## $ ST_Slope       <chr> "Up", "Flat", "Up", "Flat", "Up", "Up", "Up", "Up", "Fl~
## $ HeartDisease   <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1~
```

Data Analysis Plan

```
mean_cholesterol <- heart %>%
  group_by(ChestPainType) %>%
  summarize(mean_cholesterol = mean(Cholesterol))%>%
  print()
```

```
## # A tibble: 4 x 2
##   ChestPainType mean_cholesterol
##   <chr>                    <dbl>
## 1 ASY                       187.
## 2 ATA                       233.
## 3 NAP                       197.
## 4 TA                        207.
```

```
mean_cholesterol %>%
  ggplot()+
  geom_col(mapping = aes(x = ChestPainType, y = mean_cholesterol,  fill = ChestPainType), position = "d
    theme(legend.position = "none")+
    labs(title = "Mean Cholesterol Levels",
         subtitle = "By Chest Pain Type",
         x = "Chest Pain Type",
         y = "Mean Cholesterol Levels (mm/dL)")
```




```
mean_rbp <- heart %>%
  group_by(ChestPainType) %>%
  summarize(mean_rbp = mean(RestingBP)) %>%
  print()
```

```
## # A tibble: 4 x 2
##   ChestPainType mean_rbp
##   <chr>            <dbl>
## 1 ASY               133.
## 2 ATA               131.
## 3 NAP               131.
## 4 TA                136.
```

```
count_excersize <- heart%>%
  group_by(ChestPainType) %>%
  filter(ExerciseAngina == "Y") %>%
  summarize(n()) %>%
  print()
```

```
## # A tibble: 4 x 2
##   ChestPainType `n()`
##   <chr>         <int>
## 1 ASY             297
## 2 ATA              17
## 3 NAP              51
## 4 TA                6
```

When determining whether certain factors are associated with the type of angina experienced, the explanatory variables will be cholesterol levels, resting blood pressure, or whether the angina was exercise induced. The association of each variable with the type of angina will be analyzed separately. The response variable will be the type of chest pain experienced by the patient.

Then, for the second part of the study, the explanatory variable will be the type of chest pain experienced, and the response variable will be whether the patient had heart disease. This will allow for an analysis of whether one type of chest pain is more associated with heart disease than others.

A Chi Squared test for independence can be used to determine whether there is a statistically significant association between one the predicting variable and the categorical outcome variable. This can be applied

to this study to determine if one of the predicting variables (cholesterol, blood pressure, and whether the angina was exercise induced) is related to the categorical outcome (type of angina). This will allow one to see if cholesterol levels or resting blood pressure might be associated more with a certain type of chest pain. The Chi Squared test will be done three times, once for each predicting variable, based on a table with the rows representing the type of angina experienced, and the columns representing different levels of the predicting variable. For example, the different levels for cholesterol levels will be "low cholesterol," "intermediate cholesterol," and "high cholesterol," each constituting of a range of values chosen through a literature analysis. For all of the Chi squared tests run, if the X^2 statistic is greater than the critical value (the value bordering the range at which the H0 is rejected) the predicting variable will be concluded to have a statistically significant effect on the outcome variable.

The means and bar graph shown above will be used in our study to show why the predicting variable (cholesterol levels, in the case of these preliminary graphics/calculations) was hypothesized to have an effect on the type of chest pain experienced, as atypical anginas had the highest mean cholesterol levels.