##Project Requirement

# 1. Preliminary analysis:

#

# Perform preliminary data inspection and report the findings as to the structure of the data, missing values, duplicates, etc.

# Based on the findings from the previous question remove duplicates (if any) , treat missing values using an appropriate strategy.

# 2. Prepare an informative report about the data explaining the distribution of the disease and the related factors. You could use the below approach to achieve the objective

#

# Get a preliminary statistical summary of the data. Explore the measures of central tendencies and the spread of the data overall.

# Identify the data variables which might be categorical in nature. Describe and explore these variables using appropriate tools e.g. count plot

# Study the occurrence of CVD across Age.

# Study the composition of overall patients w.r.t. Gender.

# Can we detect a heart attack based on anomalies in the Resting Blood Pressure of the patient?

# Describe the relationship between Cholesterol levels and our target variable.

# What can be concluded about the relationship between peak exercising and the occurrence of a heart attack.

# Is thalassemia a major cause of CVD?

# How are the other factors determining the occurrence of CVD?

# Use a pair plot to understand the relationship between all the given variables.

# 3. Build a baseline model to predict using a Logistic Regression and explore the results.

library("readxl")

library("writexl")

library("dplyr")

library("ggplot2")

library("Hmisc")

library("treemapify")

library("scales")

library("ggthemes")

library("corrr")

library("GGally")

library("caret")

setwd("C:/Users/anbha/OneDrive/Desktop/Purdue University/Course6-Capstone Project/Project Data Set/1582800613\_project3datadictionary")

getwd()

data <- data.frame(read\_excel("data.xlsx"))

data

View(data)

head(data)

tail(data)

dim(data)

class(data)

str(data)##Notice all the data types are in Numeric and some needs to be converted to char/factors wherever applicable

##1.Perform preliminary data inspection and report the findings as the structure of the data, missing values, duplicates, etc.

##2.Based on the findings from the previous question remove duplicates (if any) and treat missing values using an appropriate strategy.

names(data)

names(data)<- gsub(' ','\_',names(data)) ##Removing WS in col names if any and replacing with '\_'

names(data)

colSums(is.na(data))###no missing values

sum(duplicated(data))##1 duplicated data

data <-unique(data)

dim(data)##duplicate removed

# age age in years

# sex (1 = male; 0 = female)

# cp chest pain type

# trestbps resting blood pressure (in mm Hg on admission to the hospital)

# chol serum cholestoral in mg/dl

# fbs (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

# restecg resting electrocardiographic results

# thalach maximum heart rate achieved

# 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

# ca number of major vessels (0-3) colored by flourosopy

# thal 3 = normal; 6 = fixed defect; 7 = reversable defect

# target 1 or 0

##changing column names into meaningful names

data <- rename(data,

chest\_pain\_type=cp,

resting\_blood\_pressure = trestbps,

cholestrol = chol,

fasting\_blood\_sugar = fbs,

resting\_ecg = restecg,

max\_heart\_rate = thalach,

exercise\_induced\_angina = exang,

st\_depression = oldpeak,

st\_slope = slope,

major\_vessels = ca,

thalessimia = thal

)

names(data)

# 3.Get a preliminary statistical summary of the data. Explore the measures of central tendencies and the spread of the data overall.

summary(data)

# Identify the data variables which might be categorical in nature. Describe and explore these variables using appropriate tools e.g., count plot.

##It is clear that columns such as sex,chest\_pain\_type,fasting\_blood\_sugar,

##resting\_ecg,exercise\_induced\_angina,st\_slope,major\_vessels,thalessimia,target

##shouldnt be of numeric type and can be converted to factors for better analysis

##converting numeric to factors

rapply(data,function(x)length(unique(x)))

data %>% summarise\_all(funs(n\_distinct(.)))##summarizing cat vars

catcols <- c("sex","chest\_pain\_type","fasting\_blood\_sugar","resting\_ecg",

"exercise\_induced\_angina","st\_slope","thalessimia")

data[catcols] <- lapply(data[catcols], factor)

##changing gender to male/female

##frequency table

data$sex

table(data$sex)

data$sex <- recode\_factor(data$sex, '0' = "female", '1' = "male")

table(data$sex)

##chart

# Describe and explore these variables using appropriate tools e.g., count plot.

# Male percentage is very high as comapred to female in this dataset

countsgender<-table(data$sex)

countsgender

pct<- round(countsgender/sum(countsgender)\*100)

labels <- paste(" ",c("Female","Male"),"-",pct,"%" )

pie(countsgender,labels = labels,main = "Gender Wise Distribution of Data",

col = c("red","green"))

##changing chest\_pain\_type to typical angina , atypical angina , non-anginal pain, asymptomatic

#frequency table

data$chest\_pain\_type

table(data$chest\_pain\_type)

data$chest\_pain\_type <- recode\_factor(data$chest\_pain\_type, '0' = "typical angina", '1' = "angina",

'2' = 'non-anginal pain',

'3' = 'asymptomatic' )

table(data$chest\_pain\_type)

#chart

##typical angina occurs most frequently

##and asymptomatic is the least occuruing of the chest pain types

countChestPainType <- table(data$chest\_pain\_type)

countChestPainType

pct<- round(countChestPainType/sum(countChestPainType)\*100)

labels <- paste(pct," %" )

p <- barplot(countChestPainType,

main = "Chest Pain Type Distribution",

xlab = "Types of Chest Pain",

ylab = "Number of Patients",

ylim = c(0, max(countChestPainType) + 100),

legend = rownames(countChestPainType),

args.legend = list(x ="topright",inset = c(-0.1,-0.25),cex=0.5),

col = c("red","blue","green","yellow"))

text(x = p,y = countChestPainType + 25,labels = labels)

##changing fasting blood sugar

#frequency table

data$fasting\_blood\_sugar

table(data$fasting\_blood\_sugar)

data$fasting\_blood\_sugar <- recode\_factor(data$fasting\_blood\_sugar, '0' = "non-diabetic",

'1' = "diabetic")

table(data$fasting\_blood\_sugar)

#chart

##to see spread of diabetic people

##high percentage of people(85%) in this data set are non diabetic

countsdiabetic<-table(data$fasting\_blood\_sugar)

countsdiabetic

pct<- round(countsdiabetic/sum(countsdiabetic)\*100)

pct

labels <- paste(pct,"%" )

xx<-barplot(countsdiabetic,

width = 1,

main = "Diabetic and Non diabetic patients",

xlab = "Diabetic/Non-Diabetic",

ylab = "Number of Patients",

ylim = c(0, max(countsdiabetic) + 100),

legend = rownames(countsdiabetic),

args.legend = list(x ="topright",inset = c(-0.1,-0.25),cex=0.5),

col = rainbow(2))

text(x = xx,labels = labels,y=countsdiabetic+20,col = "black")

##changing resting\_ecg

##frequency table

data$resting\_ecg

table(data$resting\_ecg)

data$resting\_ecg <- recode\_factor(data$resting\_ecg, '0' = "normal",

'1' = "abnormal",

'2' = 'hyper' )

table(data$resting\_ecg)

##chart

##TO SEE resting\_ecg spread of its categories

## we can see hyper is very negligible quantity

##and both normal and abnormal are in equal quantity

plotdata <- data %>%

count(resting\_ecg)

ggplot(plotdata,

aes(fill = resting\_ecg,

area = n,

label = resting\_ecg)) +

geom\_treemap() +

geom\_treemap\_text(colour = "white",

place = "centre") +

labs(title = "Resting ECG Spread") +

theme(legend.position = "none")

##changing exercise\_induced\_angina

##frequency table

data$exercise\_induced\_angina

table(data$exercise\_induced\_angina)

data$exercise\_induced\_angina <- recode\_factor(data$exercise\_induced\_angina, '0' = "no",

'1' = "yes")

table(data$exercise\_induced\_angina)

##chart

##TO SEE exercise\_induced\_angina spread of its categories

## Almost 67% of data doesnt have exercise\_induced\_angina

plotdata <- data %>%

count(exercise\_induced\_angina) %>%

arrange(desc(exercise\_induced\_angina)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$exercise\_induced\_angina, "\n",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = exercise\_induced\_angina)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y",

start = 0,

direction = -1) +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Excercise Induced Angina Spread")

##changing st\_slope

##frequency table

data$st\_slope

table(data$st\_slope)

data$st\_slope <- recode\_factor(data$st\_slope, '0' = "unsloping",

'1' = "flat",

'2' = "downsloping")

##chart

##st\_slope

##downsloping and flat are of equal proportion but unsloping is very less

plotdata <- data %>%

count(st\_slope) %>%

arrange(desc(st\_slope)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata

plotdata$label <- paste0(plotdata$st\_slope, "-",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = st\_slope)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "ST Slope Spread")

##changing thalessimia

##only three categories of thalessimia given

##1 = normal; 2 = fixed defect; 3 = reversable defect

##so converting 4th category i.e 0 to 2 (since 2 has the max no of values)

##frequency table

data$thalessimia

table(data$thalessimia)

str(data$thalessimia)

##chart

plotdata <- data %>%

count(thalessimia)

ggplot(plotdata,

aes(fill = thalessimia,

area = n,

label = thalessimia)) +

geom\_treemap() +

geom\_treemap\_text(colour = "white",

place = "centre") +

labs(title = "Thalessimia Spread") +

theme(legend.position = "none")

##since we dont have any factor for 0 in thalessimia,

##we will convert the 2 rows where thalessimia = 0 as thalessimia = 2

data$thalessimia <- recode\_factor(data$thalessimia, '0' = "2")

data$thalessimia <- recode\_factor(data$thalessimia, '1' = "normal",

'2'= "fixed defect",

'3'= "reversable defect")

##plotting again after converting thalessimia = 2

##frequency table

table(data$thalessimia)

##chart

plotdata <- data %>%

count(thalessimia)

ggplot(plotdata,

aes(fill = thalessimia,

area = n,

label = thalessimia)) +

geom\_treemap() +

geom\_treemap\_text(colour = "white",

place = "centre") +

labs(title = "Thalessimia ECG Spread") +

theme(legend.position = "none")

##5.Study the occurrence of CVD across different ages

## To analyze the CVD, let's explore the target variable first

##frequency table

str(data$target)

table(data$target)

## 0 - Disease- & 1 - disease+

data$target2 <- recode\_factor(data$target, '0' = "Disease-",

'1' = "Disease+")

str(data$target2)

table(data$target2)

##chart for disease- and disease+

plotdata <- data %>%

count(target2)

ggplot(plotdata,

aes(x = target2,

y = n)) +

geom\_bar(stat = "identity") +

geom\_text(aes(label = n),

vjust=-0.5) +

labs(x = "Target",

y = "Number Of Patients",

title = "Target Variable Distribution ")

##This shows there are more people with CVD in this Data Set

##Now let us compare Age vs CVD

##Dividing the data set based on Target Variable

## as dataHealthy and dataDiseased for better understanding

dataHealthy <- data %>% filter(target2 == 'Disease-')

View(dataHealthy)

dataDiseased <- data %>% filter(target2 == 'Disease+')

View(dataDiseased)

##Study the occurrence of CVD across different ages

##bar plot (group) for health and Diseased

ggplot(data,

aes(x = age ,

fill = target2 )) +

geom\_bar(position = position\_dodge(preserve = "single" ))+

labs(title = "Age Distribution of Diseased and Healthy", y="Number Of Patients",x="Age")

##kernel density plot for diseased

##the graph

ggplot(dataDiseased, aes(x=age))+

geom\_density(color="darkblue", fill="lightblue")+

geom\_vline(aes(xintercept=mean(age)),

color="blue", linetype="dashed", size=1)+

labs(title = "Age Density Plot For Diseased",y="Density",x="Age")

ggplot(dataDiseased,

aes(x = as.factor(age))) +

geom\_bar(fill = "indianred3",

color = "black")+

labs(title = "Frequency by age for diseased",x="Age",y="Number Of Patients")

##from the graph it is evident that CVD increases from 47 and peaks at 54

##so 47-54 is the riskiest age band for getting CVD

##6.Can we detect heart attack based on anomalies in resting blood pressure of the patient?

##box plot for resting\_blood\_pressure for diseased and healthy

ggplot(data, aes(x=target2, y=resting\_blood\_pressure, fill=target2)) +

geom\_boxplot(alpha=0.3) +

labs(title = "Distribution of Resting Blood Pressure with Target Variable",y="Resting Blood Pressure",x="Target Variable")+

theme(legend.position="none")

ggplot(data, aes(x=sex, y=resting\_blood\_pressure, fill=target2)) +

geom\_boxplot(alpha=0.3) +

labs(title = "Gender Based Distribution of Resting Blood Pressure with Target Variable",y="Resting Blood Pressure",x="Gender")+

theme(legend.title = element\_blank())

##male patients have the same range but female patients range vary w.r.t Blood pressure

#Male range : 120-140(both diseased and Healthy)

#Female range : 120-140(diseased) , 130-155(healthy)

#Density plot to identify relationship between resting\_blood\_pressure and CVD

ggplot(data, aes(x=resting\_blood\_pressure,fill=target2))+

geom\_density(color="red")+

geom\_vline(aes(xintercept=mean(resting\_blood\_pressure)),

color="blue", linetype="dashed", size=1)+

labs(title = "Density Distribution of Blood Pressure with Target Variable",x="Resting Blood Pressure",y="Density")

##We cannot say Resting Blood Pressure has connection to CDV in this case,

##as both the box plot and density plot suggest they have the same range

##7. Study the composition of overall patients w.r.t . gender.

##stacked bar for overall percentage of gender in data

plotdata <- data %>%

count(sex) %>%

arrange(desc(sex)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata

plotdata$label <- paste0(plotdata$sex, "-",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = sex)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Gender Spread")

##donut chart for percentage of male and female in diseased

plotdata <- dataDiseased %>%

count(sex) %>%

arrange(desc(sex)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$sex, "\n",

round(plotdata$prop), "%")

plotdata

ggplot(plotdata,

aes(x = 1,

y = prop,

fill = sex)) +

geom\_col() +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y") +

xlim(c(0.2, 1 + 0.5))+

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Gender spread of Patients in Diseased Data")

##pie chart for percentage of male and female in healthy

plotdata <- dataHealthy %>%

count(sex) %>%

arrange(desc(sex)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$sex, "\n",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = sex)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y",

start = 0,

direction = -1) +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Gender spread of Patients in Healthy Data")

##donut chart to calcualte diseased in female population

plotdata <- data %>% filter(sex == 'female') %>%

count(target2) %>%

arrange(desc(target2)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$target2, "\n",

round(plotdata$prop), "%")

plotdata

ggplot(plotdata,

aes(x = 1,

y = prop,

fill = target2)) +

geom\_col() +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y") +

xlim(c(0.2, 1 + 0.5))+

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Diseased and Healthy in Female Patients")

##pie chart to calcualte diseased in male population

plotdata <- data %>% filter(sex == 'male') %>%

count(target2) %>%

arrange(desc(target2)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$target2, "\n",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = target2)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y",

start = 0,

direction = -1) +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Dieseased and Healthy in Male patients")

##from these graphs we are able to infer that

# 1.In overall data, male population is more than women

# but large percentage of women seem to be diseased (75%)

# as compared to men (45%). So 3/4th of female seeems to have CVD. So women in this data seems to be at high risk

# 8.Describe the relationship between cholesterol levels and our target variable.

my\_data <- data %>%

group\_by(target2) %>%

summarise(mean = mean(cholestrol),

std = sd(cholestrol),

min = min(cholestrol),

max = max(cholestrol),

med = median(cholestrol))

my\_data

e <- ggplot(data, aes(x = target2, y = cholestrol))

e + geom\_violin(aes(fill = target2), trim = FALSE) +

geom\_boxplot(width = 0.2)+

scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#FC4E07"))+

theme(legend.position = "none")+labs(title="Cholestrol Distibution with Target Variable",x="Target Variable",y="Cholestrol")

##the graph and stat functions shows disease+ data has quite a lot of outliers

# and also the violin plot suggest data for diseased and healthy

# are distributed in the same range to an extent and hence it is inconclusive with this data set

# 9. What can be concluded about the relationship between peak exercising and occurrence of heart attack?

plotdata <- data %>%

group\_by(target2) %>%

summarise(n = n(),

mean = mean(max\_heart\_rate),

sd = sd(max\_heart\_rate),

se = sd / sqrt(n))

plotdata

ggplot(plotdata,

aes(x = target2,

y = mean,

group = 1)) +

geom\_point(size = 3) +

geom\_line() +

geom\_errorbar(aes(ymin = mean - se,

ymax = mean + se),

width = .1)+

labs(title="Mean Value Comparison of Max Heart Rate with Target Variable",x="Target Variable",y="Mean")

ggplot(data, aes(x=max\_heart\_rate,fill = target2))+

geom\_density(color="darkblue")+

geom\_vline(aes(xintercept=mean(max\_heart\_rate)),

color="blue", linetype="dashed", size=1)+

labs(title = "Density plot for Max Heart Rate with Target Variable",x="Max Heart Rate",y="Density")

##lets explore more

ggplot(data,

aes(x = max\_heart\_rate ,

fill = target2 )) +

geom\_bar(position = position\_dodge(preserve = "single" ))+

labs(title = "Max Heart Rate comparison of with Target",x="Max Heart Rate",y="Number Of Patients")

##dieseased seems to have max heart rate greater than mean value i.e 150

##max density concentration at 162 for diseased.

##max density concentration at 148 for healthy

#10.Is thalassemia a major cause of CVD? How are the other factors determining the occurrence of CVD?

ggplot(data,

aes(x = target2 ,

fill = thalessimia )) +

geom\_bar(position = position\_dodge(preserve = "single" ))+

labs(title = "Thalessimia comparison with Target Variable",x="Target",y="Number Of Patients")+

geom\_text(aes(label =..count..),stat="count",position = position\_dodge(width =1 ),vjust=0.01)

#thalessimia appears to be a major cause,accounting to max percentage of diseased patients

# out of 164 diseased, more than 125 seems to have irreversable thalessimia

##reversable defect and normal doesnt seem to have great impact on CVD

## lets see how diabetes relate to CVD

plotdata <- data %>% filter(fasting\_blood\_sugar == 'diabetic') %>%

count(target2) %>%

arrange(desc(target2)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$target2, "\n",

round(plotdata$prop), "%")

plotdata

ggplot(plotdata,

aes(x = 1,

y = prop,

fill = target2)) +

geom\_col() +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y") +

xlim(c(0.2, 1 + 0.5))+

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Diabetic patients in Healthy and Dieseased")

## lets see how diabetes relate to CVD

plotdata <- dataDiseased %>%

count(fasting\_blood\_sugar) %>%

arrange(desc(fasting\_blood\_sugar)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$fasting\_blood\_sugar, "\n",

round(plotdata$prop), "%")

plotdata

ggplot(plotdata,

aes(x = 1,

y = prop,

fill = fasting\_blood\_sugar)) +

geom\_col() +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y") +

xlim(c(0.2, 1 + 0.5))+

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Spread of Diabetic/Non Diabetic patients in diseased")

plotdata <- dataHealthy %>%

count(fasting\_blood\_sugar) %>%

arrange(desc(fasting\_blood\_sugar)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$fasting\_blood\_sugar, "\n",

round(plotdata$prop), "%")

plotdata

ggplot(plotdata,

aes(x = 1,

y = prop,

fill = fasting\_blood\_sugar)) +

geom\_col() +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y") +

xlim(c(0.2, 1 + 0.5))+

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Spread of Diabetic/Non Diabetic patients in Healthy")

##diabetic doesnt play an important role in this Dataset

##relationship between chest pain type and target var

ggplot(data, aes(fill=target2, x=chest\_pain\_type)) +

geom\_bar(position="stack", stat="count") +

ggtitle("Chest Pain type distribution between Diseased and Healthy patients") +

ylab("Number Of Patients")+coord\_flip()

# '1' = "angina",

# '2' = 'non-anginal pain',

# '3' = 'asymptomatic' all seem to contribute more to diseased than typical angina

# i.e a patient would more likely to be healthy if cp type is typical angina

#than compared to other CP types

##in other words angina contributes to cvd the most among cp types

##resting\_ecg relation with target

plotdata <- data %>% filter(resting\_ecg == 'normal') %>%

count(target2) %>%

arrange(desc(target2)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$target2, "\n",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = target2)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y",

start = 0,

direction = -1) +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Normal ECG Spread across Target Variable")

plotdata <- data %>% filter(resting\_ecg == 'abnormal') %>%

count(target2) %>%

arrange(desc(target2)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$target2, "\n",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = target2)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y",

start = 0,

direction = -1) +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Abnormal ECG Spread across Target Variable")

##Abnormal ECG seems to be associated more with Disease+ patients

plotdata <- dataDiseased %>%

count(resting\_ecg) %>%

arrange(desc(resting\_ecg)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata$label <- paste0(plotdata$resting\_ecg, "\n",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = resting\_ecg)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

coord\_polar("y",

start = 0,

direction = -1) +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "ECG Spread Across Diseased")

##abnormal accounts to almost 60% of positive cases

#11.Use a pair plot to understand the relationship between all the given variables.

str(data)

d <- data[,sapply(data,class)=="numeric"]

d$target2 <- data$target2

ggpairs(d,

columns = 1:(ncol(d)-1) ,

aes(color = target2, alpha = 0.5))

##ST\_Depression and Major\_Vessels seems to have a tight co relation

##with the Target Var

res <- cor.test(data$st\_depression, data$target,

method = "pearson")

res

##box plot for St depression

ggplot(data, aes(x=target2, y=st\_depression, fill=target2)) +

geom\_boxplot(alpha=0.3) +

labs(title = "Distribution of ST Depression for Target Variable",y="ST Depression",x="Target")+

theme(legend.position="none")

##violin plot for Major Vessels

res <- cor.test(data$major\_vessels, data$target,

method = "pearson")

res

e <- ggplot(data, aes(x = target2, y = major\_vessels))

e + geom\_violin(aes(fill = target2), trim = FALSE) +

scale\_fill\_manual(values = c("#00AFBB", "#E7B800", "#FC4E07"))+

labs(title = "Distribution of Major Vessels Across Target Variable",x="Target",y="Major Vessels")+

theme(legend.position = "none")

##It is evident that

##ST\_Deperesssion and Major Vessels have a strong impact on the predictor var target

##st\_slope relation

plotdata <- dataDiseased %>%

count(st\_slope) %>%

arrange(desc(st\_slope)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata

plotdata$label <- paste0(plotdata$st\_slope, "-",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = st\_slope)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "ST Slope Spread across diseased")

plotdata <- dataHealthy %>%

count(st\_slope) %>%

arrange(desc(st\_slope)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata

plotdata$label <- paste0(plotdata$st\_slope, "-",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = st\_slope)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "ST Slope Spread Across Healthy")

##downsloping seems to contribute to CVD

##Relationship with exercise\_induced\_Angina

##No % is really high (67%), lets see how this plays with target var

plotdata <- dataDiseased %>%

count(exercise\_induced\_angina) %>%

arrange(desc(exercise\_induced\_angina)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata

plotdata$label <- paste0(plotdata$exercise\_induced\_angina, "-",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = exercise\_induced\_angina)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Exercise Induced Angina Spread Across Diseased")

plotdata <- dataHealthy %>%

count(exercise\_induced\_angina) %>%

arrange(desc(exercise\_induced\_angina)) %>%

mutate(prop = round(n\*100/sum(n), 1),

lab.ypos = cumsum(prop) - 0.5\*prop)

plotdata

plotdata$label <- paste0(plotdata$exercise\_induced\_angina, "-",

round(plotdata$prop), "%")

ggplot(plotdata,

aes(x = "",

y = prop,

fill = exercise\_induced\_angina)) +

geom\_bar(width = 1,

stat = "identity",

color = "black") +

geom\_text(aes(y = lab.ypos, label = label),

color = "black") +

theme\_void() +

theme(legend.position = "FALSE") +

labs(title = "Exercise Induced Angina Spread Across Healthy")

##

# The percentage of 'no' in Disease+ data is very high whereas healthy patients

# dataset is domainated by 'yes'

# and hence it plays a role in deciding the CVD outcome.

##Let us write this excel to analyze with Tableau

write\_xlsx(data,"cvd\_data\_latest.xlsx")

# 13.Visualize the variables using Tableau to create an understanding for attributes of a Diseased vs a Healthy person.

#

# 14. Demonstrate the variables associated with each other and factors to build a dashboard

##Model building and Testing

# 12. Perform logistic regression, predict the outcome for test data, and validate the results by using the confusion matrix.

data1 <- data %>% select(-target2)

data1

data1$target <- as.factor(data1$target)

train\_indices <- sample(1:nrow(data1),0.7\*nrow(data1))

train\_indices

train <- data1[train\_indices,]

test <- data1[-train\_indices,]

##building model by including all columns from the data (basemodel)

basemodel <- glm(target~.,data = train,family = 'binomial')

summary(basemodel)

pred\_prob <- predict(basemodel,test)

pred <- as.factor(ifelse(pred\_prob >= 0.5,1,0))

caret::confusionMatrix(pred,test$target) ##accuracy is 82%

##drilling down with columns having high co-relation suggested by R

model1 <- glm(target~ sex+chest\_pain\_type+resting\_blood\_pressure+resting\_ecg+max\_heart\_rate+thalessimia+major\_vessels,data = train,family = 'binomial')

summary(model1)

pred\_prob <- predict(model1,test)

pred <- as.factor(ifelse(pred\_prob >= 0.5,1,0))

caret::confusionMatrix(pred,test$target) ##accuracy reduces to 78%

##building model by including columns selected from my analysis (myModel)

myModel <- glm(target~ sex+age+chest\_pain\_type+resting\_ecg+max\_heart\_rate+exercise\_induced\_angina+st\_depression+st\_slope+thalessimia+major\_vessels,data = train,family = 'binomial')

summary(myModel)

pred\_prob <- predict(myModel,test)

pred <- as.factor(ifelse(pred\_prob >= 0.5,1,0))

caret::confusionMatrix(pred,test$target)