9.3_Final_Project_Step_2

Moshe Burnstein

`r Sys.Date()`

{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

R Markdown

Diabetes Data

I explore four datasets to address the question of whether high blood pressure causes Type 2 diabetes. T

```
library(heplots)
data(Diabetes, package="heplots")
str(Diabetes)
library(pastecs)
stat.desc(Diabetes, norm = TRUE)
```

The sspg variable(steady-state plasma glucose level) indicates the average of the last four blood glucos

```
weight_glucose_vs_diagnosis <- Diabetes%>%
                               select(group,relwt, glufast, sspg)%>%
                               filter(group!= "Normal")
weight_glucose_vs_diagnosis
mean(weight_glucose_vs_diagnosis$relwt)
mean(weight_glucose_vs_diagnosis$glufast)
mean(weight_glucose_vs_diagnosis$sspg)
```

I have created a dataframe of all diabetics compared to relative weight, fasting glucose, and sspg. The

```
error_bars_relwt <- bar + stat_summary(fun.data = mean_cl_normal, geom = "errorbar", position = positio
error_bars_relwt
weight_glucose_vs_diagnosis%>%
  group_by(group)%>%
  summarise_at(vars(relwt, glufast, sspg), mean)
normal_nums <- Diabetes%>%
     select(group,relwt,glufast,sspg)%>%
     filter(group== "Normal")%>%
     summarise_at(vars(relwt,glufast,sspg),mean)
normal_nums
```

The plot shows that relative weight exceeds normal for the chemically diabetic, but reverts to below average for the overtly diabetic. Maybe chemically diabetic patients are counseled by their physicians to lose weight, and it still does not help to prevent the onset of full diabetes? Or maybe the disease itself causes weight loss? I then compare the diabetics' stats to the normal ones, and normals were in the normal range for all three variables.

The Jordan dataset comes from The Humanitarian Data Exchange(HDX). It is in an excel file and needs read.xl package to access the data.

```
library(readxl)
jordan_df <- read_excel('jordandataset.xlsx')
```

Use dplyr to select variables to explore.

```
library(dplyr)
jord_df <- jordan_df%>%
           select(Sex, Age,  bp_systolic, bp_diastolic, hba1c, MartialStatus, BMI,
           starts_with("Comorbid"))
```

Check for NAs.

```
is.na(jord_df)
```

Remove NAs. Check that all values = FALSE.

```
cleaned_jord <- na.omit(jord_df)
is.na(cleaned_jord)
head(cleaned_jord)
```

Look at correlation matrix of variables impacting diabetes.

```
cor(cleaned_jord[, c("Age", "bp_systolic", "bp_diastolic", "hba1c", "BMI")])
```

The systolic and diastolic blood pressure readings are understandably correlated against each other. Otherwise, only systolic blood pressure is correlated with age. # Visualize correlations with ggpairs. One must import GGally library.

```
library(GGally)
ggpairs(jordan_df%>%
        select(Age, bp_systolic, bp_diastolic, hba1c))
```

Look at averages for bp, a1c, BMI, and Age.

```
mean(cleaned_jord$bp_systolic)
mean(cleaned_jord$bp_diastolic)
mean(cleaned_jord$Age)
mean(cleaned_jord$hba1c)
mean(cleaned_jord$BMI)
```

The mean for diastolic pressure is 90 mm Hg, higher than 80 mm Hg, which is the normal upper limit. The

#The PimaIndiansDiabetes2 dataset is accessible through the mlbench library in R. Look at average Pima Indians pressure, using the dplyr package.

```
pima_diastolic <- PimaIndiansDiabetes2%>%
                  group_by(diabetes)%>%
                  filter(pressure != "NA")%>%
                  summarise_at(vars(pressure), mean)
pima_diastolic
```

The diastolic pressures for non-diabetic Pimas was 71 mm Hg, and 75 mm Hg for diabetics, both below the 80 mm Hg threshold. In the Pima Indian population, hypertension was not associated with Type 2 diabetes. There is an increased pressure associated with diabetes.

```
diastolic_boxplot <- ggplot(PimaIndiansDiabetes2, aes(diabetes, pressure, color = diabetes)) + geom_boxp
diastolic_boxplot
histogram_pressure <- ggplot(PimaIndiansDiabetes2, aes(pressure)) + geom_histogram(binwidth = 1.75)
histogram_pressure
```

The histogram shows a normal distribution of pressures in the normal range, with a few outliers. The boxplot illustrates higher, albeit normal pressures for the diabetic group.

Remove NAs to produce a proper correlation matrix.

```
clean_Pima <- na.omit(PimaIndiansDiabetes2)
cor(clean_Pima[, c("pregnant", "glucose", "pressure", "triceps", "insulin", "mass", "pedigree", "age")]
```

```
The 2-hour serum insulin is understandably positively correlated with the glucose tolerance test. The ag

'''{r}
pcor((clean_Pima[, c("pressure", "pedigree", "pregnant", "triceps", "insulin", "mass", "glucose", "age"]
```

The partial correlation between pressure and pedigree is slightly negative. This is after controlling for all other variables. The r^2 is 0.00913777. However, one might even suspect this, because it has a p-value more than 0.05(6.061597e-02),

The NCSU stat diabetics dataset is available online.

```
library(dplyr)
data_diabetics <- read.delim('https://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt')
diabetic_nums <- data_diabetics%>%
                 group_by(SEX)%>%
                 summarise_at(vars(BMI,BP), mean)
head(diabetic_nums)
```

This tibble presents a slightly higher BMI(26.8 to 26.0) for women relative to men, and a higher bp(98.2 to 91.5) for women relative to men.

```
library(ggplot2)
scatter <- ggplot(data_diabetics, aes(BMI,BP)) + geom_point()
scatter
```

The scatterplot points to a positive correlation between BMI and BP. The correlation matrix confirms this(0.3954109).

```
cor(data_diabetics[, c("AGE", "SEX", "BMI", "BP")])
```

The data shows that as people age, women tend to have more children. That need not be proven. The older
The Pima Indians manifest Type 2 diabetes not associated with hypertension. The incidence of diabetes ar
I do not yet possess skills to employ efficient machine learning models to learn from all of these data