Diabetes Analysis

Marina Huang

2023-04-21

The Data

In this project, I practice data analysis on health care data using R. I will dive into Machine Learning Algorithms and Generalized Linear Models.

The dataset can be downloaded from this link (click). It is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of this dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage.

This dataset consist of several medical predictor (independent) variables and one target (dependent) variable, `outcome`. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The summary of the dataset is displayed here:

```
summary(diabetes)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin          BMI       DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780          Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437          1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725          Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719          Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262          3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200          Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

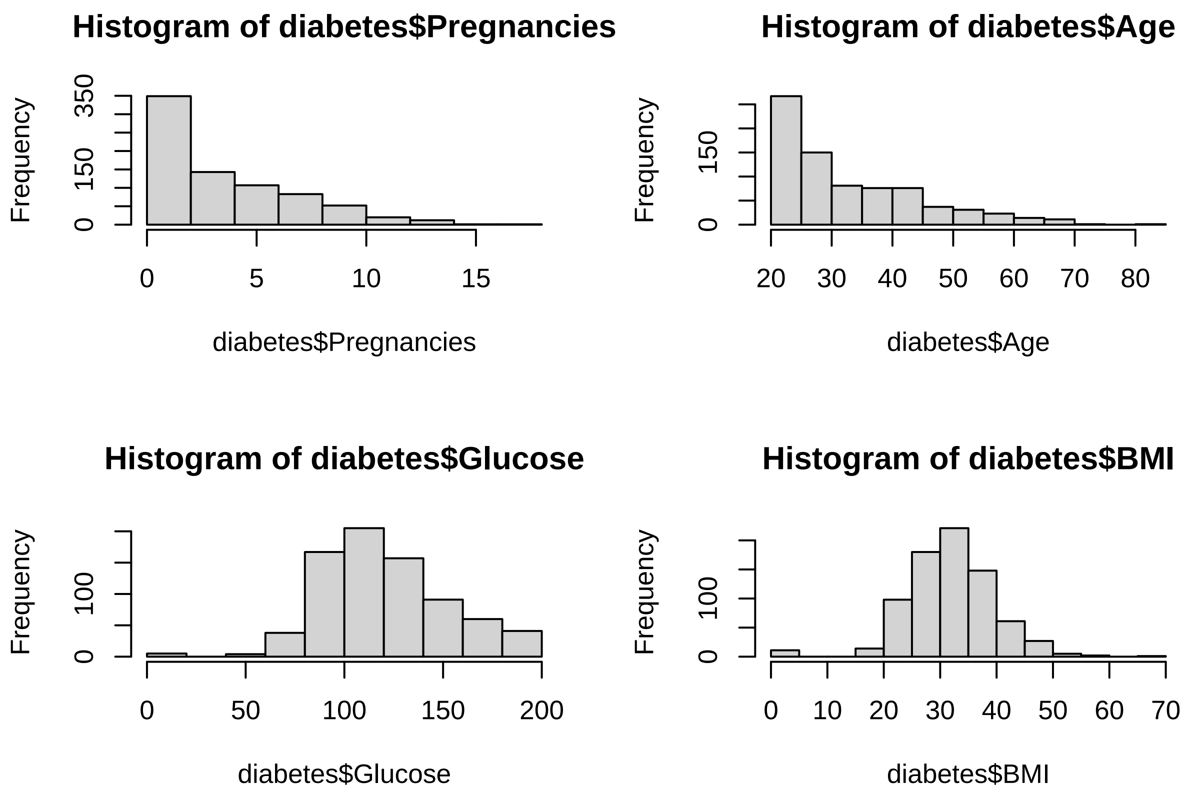
Explorative Data Analysis

Univariate Data Analysis

Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values.

Let's take a look at the effect of `Pregnancies`, `Age`, `Glucose`, and `BMI` on diabetes:

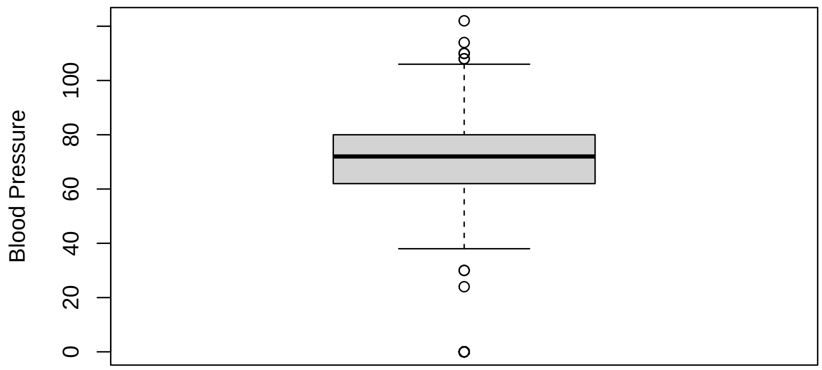
```
par(mfrow = c(2,2))
hist(diabetes$Pregnancies)
hist(diabetes$Age)
hist(diabetes$Glucose)
hist(diabetes$BMI)
```



From these distribution graphs, `Age` and `Pregnancies` area not in normal distribution as expected, since the underlying population should not be normally distributed either. Glucose level and BMI are following a normal distribution.

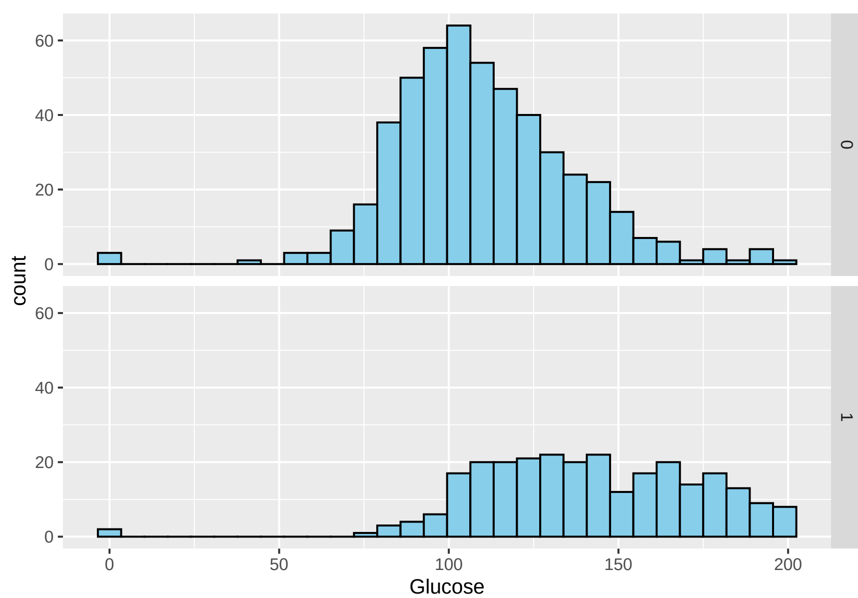
Now looking into `BloodPressure`:

```
boxplot(diabetes$BloodPressure,
        ylab = "Blood Pressure")
```



The impact of `glucose` on diabetes:

```
diabetes %>%
  ggplot(aes(x = Glucose)) +
  geom_histogram(fill = "sky blue", color = "black") +
  facet_grid(Outcome~.)
```



Goal: Assess the mean difference of glucose levels between the positive and negative groups.

Null Hypothesis: There is no significant difference between glucose levels in positive and negative groups.

Conditions

- Individuals are independent of each other.
- Distribution is skewed (not normal), but there is >30 samples.
- Both the groups are independent of each other and the sample size is lesser than 10% of the population.

```
t.test(Glucose ~ Outcome, diabetes)
```

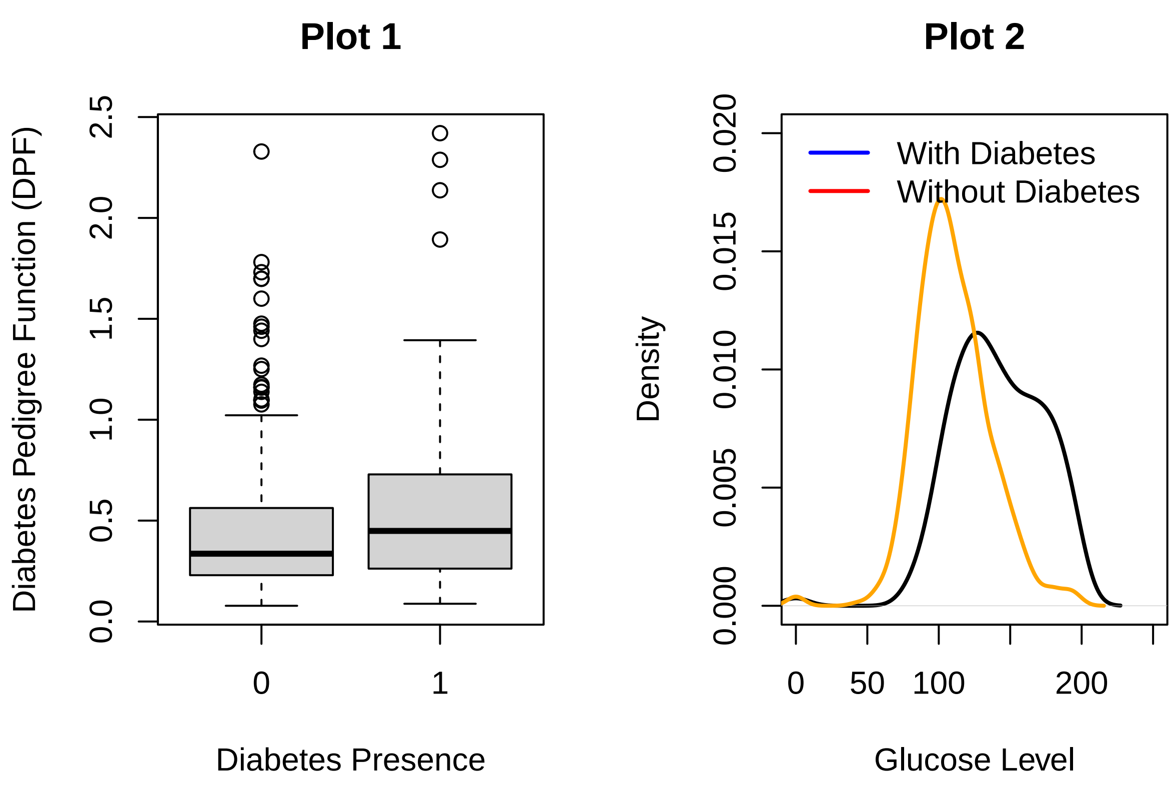
```
##
##  Welch Two Sample t-test
##
## data:  Glucose by Outcome
## t = -13.752, df = 461.33, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -35.74707 -26.80786
## sample estimates:
## mean in group 0 mean in group 1
##        109.9800        141.2575
```

The p-value is < 0.05 (the critical value), so we reject the null hypothesis for the alternate hypothesis. We can say that we are, 95% confident, that the average glucose levels for individuals with diabetes is > the people without diabetes.

Box Plot of the impact of Age on Diabetes Pedigree Function

```
par(mfrow = c(1,2))

# Boxplot

with_d <- diabetes[diabetes$Outcome == 1,]
without <- diabetes[diabetes$Outcome == 0,]

boxplot(diabetes$DiabetesPedigreeFunction ~ diabetes$Outcome,
                         ylab = "Diabetes Pedigree Function (DPF)",
                         xlab = "Diabetes Presence",
                         main = "Plot 1",
                         outline = TRUE)

# Density Plot

plot(density(with_d$Glucose),
     xlim = c(0, 250),
     ylim = c(0.00, 0.02),
     xlab = "Glucose Level",
     main = "Plot 2",
     lwd = 2)

lines(density(without$Glucose),
      col = "orange",
      lwd = 2)

legend("topleft",
```

```
       col = c("blue", "red"),
       legend = c("With Diabetes", "Without Diabetes"),
       lwd = 2,
       bty = "n")
```



Fromo Plot 2, the distribution is shifted towards the left for those without diabetes. This indicates those **without diabetes generally have a lower blood glucose level.**

Welch Two Sample t-Test

```
t.test(with_d$DiabetesPedigreeFunction, without$DiabetesPedigreeFunction)
```

Data Correlation Analysis

Scatter Matrix of All Columns:

```
ggcorr(diabetes[,-9], name = "corr", label = TRUE) +
  theme(legend.position = "none") +
  labs(title = "Correlation Plot of Variance") +
  theme(plot.title = element_text(face = "bold", color = "black", hjust = 0.5, size = 12))
```

Correlation Plot of Variance

Pregnancy, Age, Insulin, SkinThickness are having higher correlation.

Basic GLM

Logistic Regression

Decision Tree

Naive Bayes