Multivariate Analysis

Shukry Zablah

05 December, 2018

Contents

Imports 1
Load Data 1
Correlation of Variables 1
Multivariate Visualizations 2
Insulin and Glucose Concentration 2
Age and Pregnancies 2
BMI and Skin Thickness 2

Imports

```
library(dplyr)
library(mosaic)
library(ggplot2)
library(ggcorrplot)
```

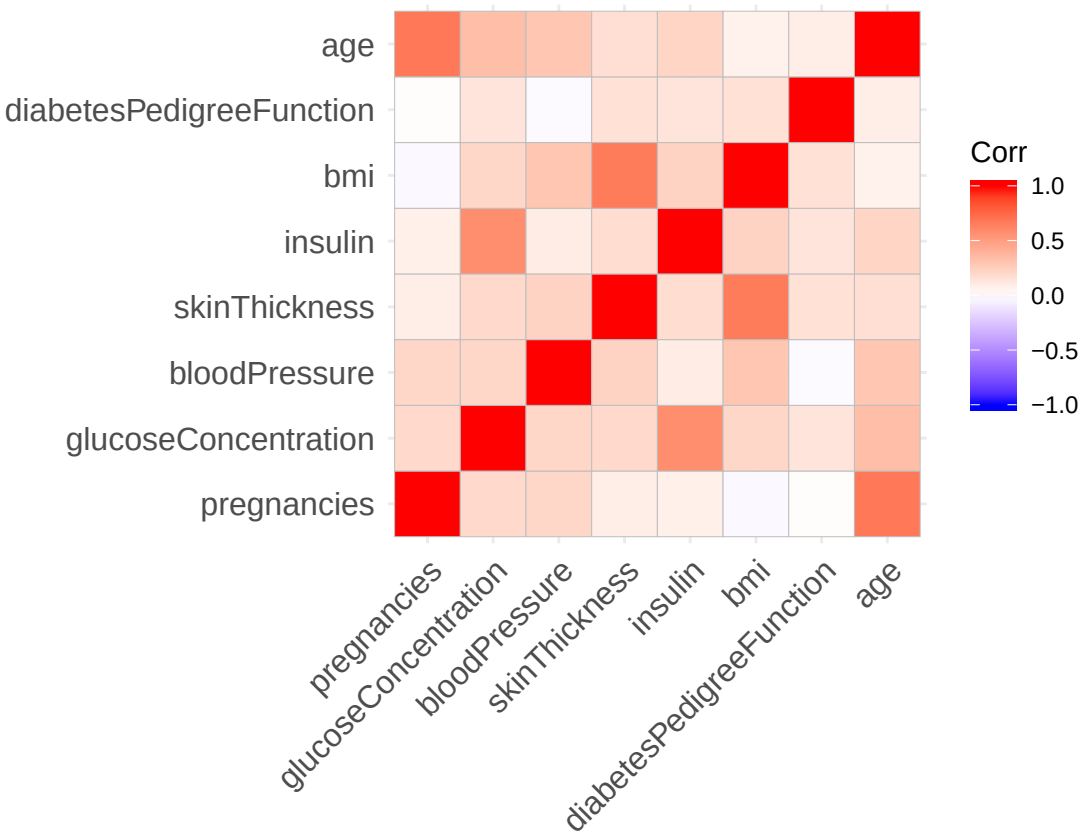
Load Data

```
PIMA <- readRDS(file = "../data/PIMA_noNAs.Rds")
```

Note that the following analysis is done for the dataset without any missing values.

Correlation of Variables

```
PIMA %>%
  select(-hasDiabetes) %>%
  cor() %>%
  ggcorrplot()
```



In the correlation plot above we can see that there are some featured that are correlated. This is a hint that we might now need both features of a correlated pair in our model as they are likely to not add valuable information. We can see that the stronger correlated features are:

- insulin and glucose concentration: 0.581223
- age and pregnancies: 0.6796085
- bmi and skin thickness: 0.6643549

Multivariate Visualizations

Insulin and Glucose Concentration

```
# viz for insulin vs glucoseConcentration (maybe include hasDiabetes?)
```

Age and Pregnancies

```
# viz for age vs pregnancies (maybe include hasDiabetes?)
```

BMI and Skin Thickness

```
# viz for bmi and skin thickness (maybe include hasDiabetes?)
```