Univariate Analysis

Shukry Zablah

05 December, 2018

Contents

Imports

```r
library(dplyr)
library(mosaic)
library(ggplot2)
```

Load Data

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

Global Characteristics

```r
names(PIMA)
```

```
## [1] "pregnancies"              "glucoseConcentration"
## [3] "bloodPressure"            "skinThickness"
## [5] "insulin"                  "bmi"
## [7] "diabetesPedigreeFunction" "age"
## [9] "hasDiabetes"
```

Refer to the codebook at references/codebook.md

```r
dim(PIMA)
```

```
## [1] 768   9
```

We have 768 observations with 8 features and our response variable.

Let's take a look at some of the observations to familiarize ourselves with the dataset.

```
head(PIMA)
```

```
##   pregnancies glucoseConcentration bloodPressure skinThickness insulin
## 1           6                  148            72            35      NA
## 2           1                   85            66            29      NA
## 3           8                  183            64            NA      NA
## 4           1                   89            66            23      94
## 5           0                  137            40            35     168
## 6           5                  116            74            NA      NA
##    bmi diabetesPedigreeFunction age hasDiabetes
## 1 33.6                    0.627  50           1
## 2 26.6                    0.351  31           0
## 3 23.3                    0.672  32           1
## 4 28.1                    0.167  21           0
## 5 43.1                    2.288  33           1
## 6 25.6                    0.201  30           0
```

Our data is not clean, it is important to choose early on the technique we will use to deal with missing values. We could consider creating new features for modeling.

Response Variable

The response variable is RESPONSE. It is a binary variable. 0 = No Diabetes and 1 = Diabetes.

```
tally(~ hasDiabetes, data = PIMA)
```

```
## hasDiabetes
##   0   1
## 500 268
```

We have 268 observations with Diabetes. This is around 35% of our dataset.

An interesting question: Why is the presence of diabetes in our population so high?

We have to be wary of generalizing our predictions if our dataset is not representative of the population that we are trying to predict. And if it is representative, then what is causing the PIMA Native American females to of 21+ years of age to test positive for diabetes?

Other Features

Descriptive statistics for the variables in our dataset are provided below:

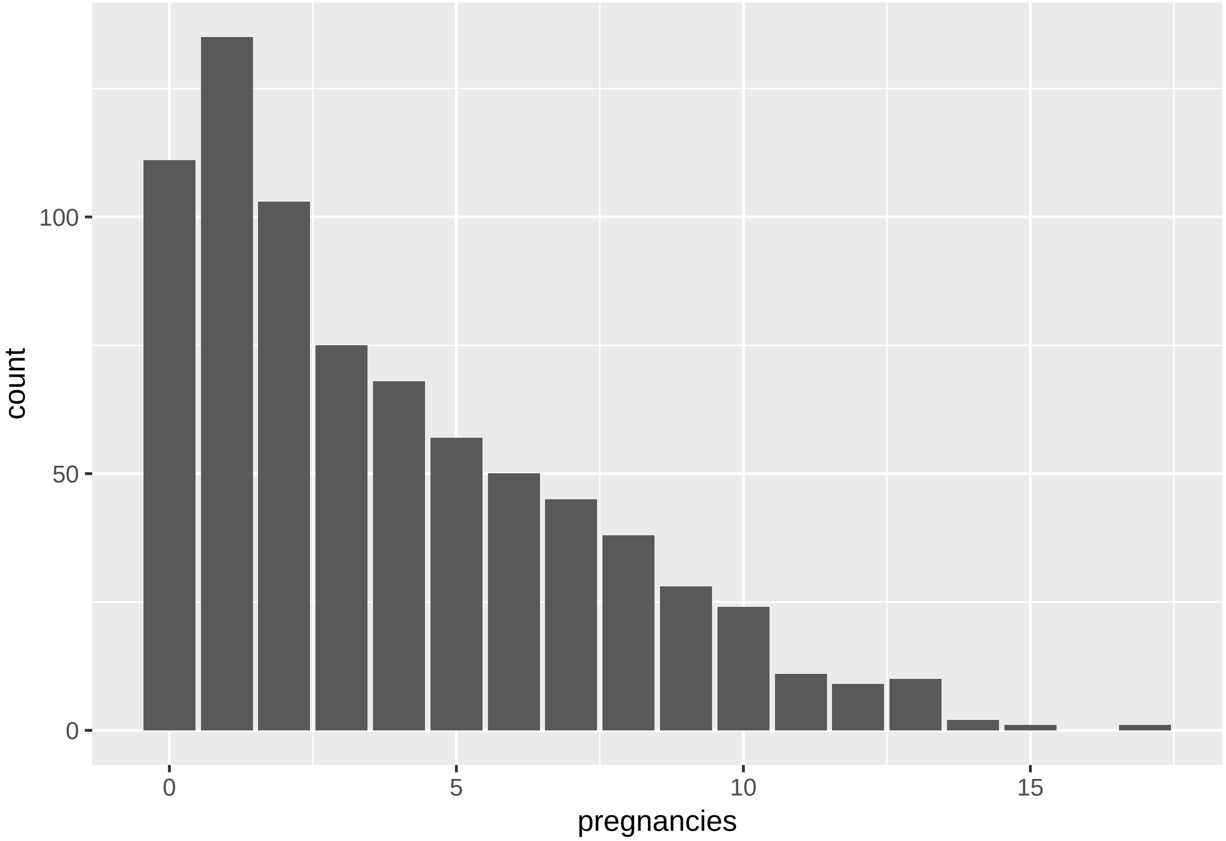
```
summary(PIMA)
```

```
##   pregnancies     glucoseConcentration bloodPressure    skinThickness
##  Min.   : 0.000   Min.   : 44.0        Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0        1st Qu.: 64.00   1st Qu.:22.00
##  Median : 3.000   Median :117.0        Median : 72.00   Median :29.00
##  Mean   : 3.845   Mean   :121.7        Mean   : 72.41   Mean   :29.15
##  3rd Qu.: 6.000   3rd Qu.:141.0        3rd Qu.: 80.00   3rd Qu.:36.00
##  Max.   :17.000   Max.   :199.0        Max.   :122.00   Max.   :99.00
##                   NA's   :5            NA's   :35       NA's   :227
##     insulin            bmi        diabetesPedigreeFunction      age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0780           Min.   :21.00
##  1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437           1st Qu.:24.00
##  Median :125.00   Median :32.30   Median :0.3725           Median :29.00
##  Mean   :155.55   Mean   :32.46   Mean   :0.4719            Mean   :33.24  
##  3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262            3rd Qu.:41.00  
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200            Max.   :81.00  
##  NA's   :374      NA's   :11                                               
##  hasDiabetes
##  0:500      
##  1:268      
##             
##             
##             
##             
##             
```

Missing data is not included in the visualizations below:

Pregnancies

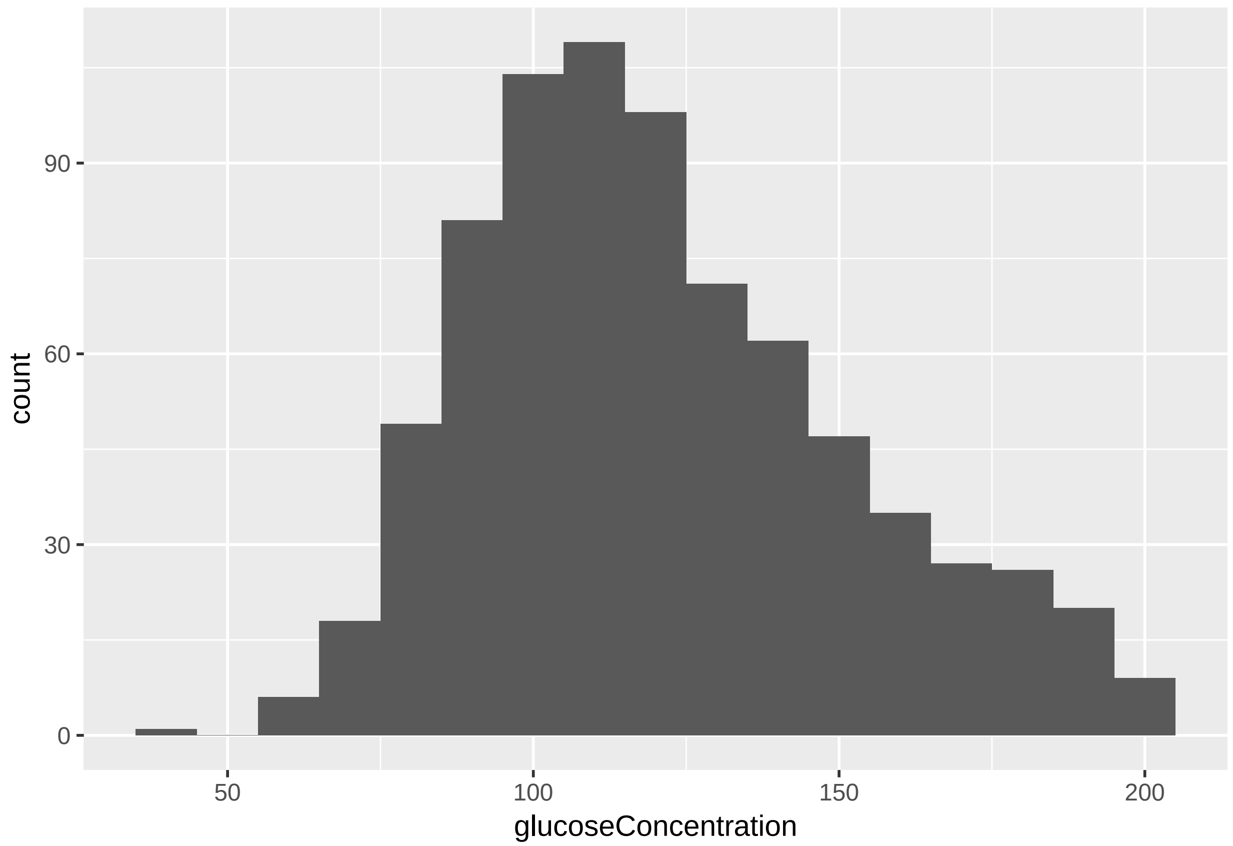
```
ggplot(PIMA, aes(x = pregnancies)) + geom_bar()
```



With the exception of 0 pregnancies, the count of women decreases as the number of pregnancies increases. The count for women that had 10 pregnancies is still around 25.

Plasma Glucose Concentration in Saliva

```
ggplot(PIMA, aes(x = glucoseConcentration)) + geom_histogram(binwidth = 10)
```



```
favstats(~ glucoseConcentration, data = PIMA)
```

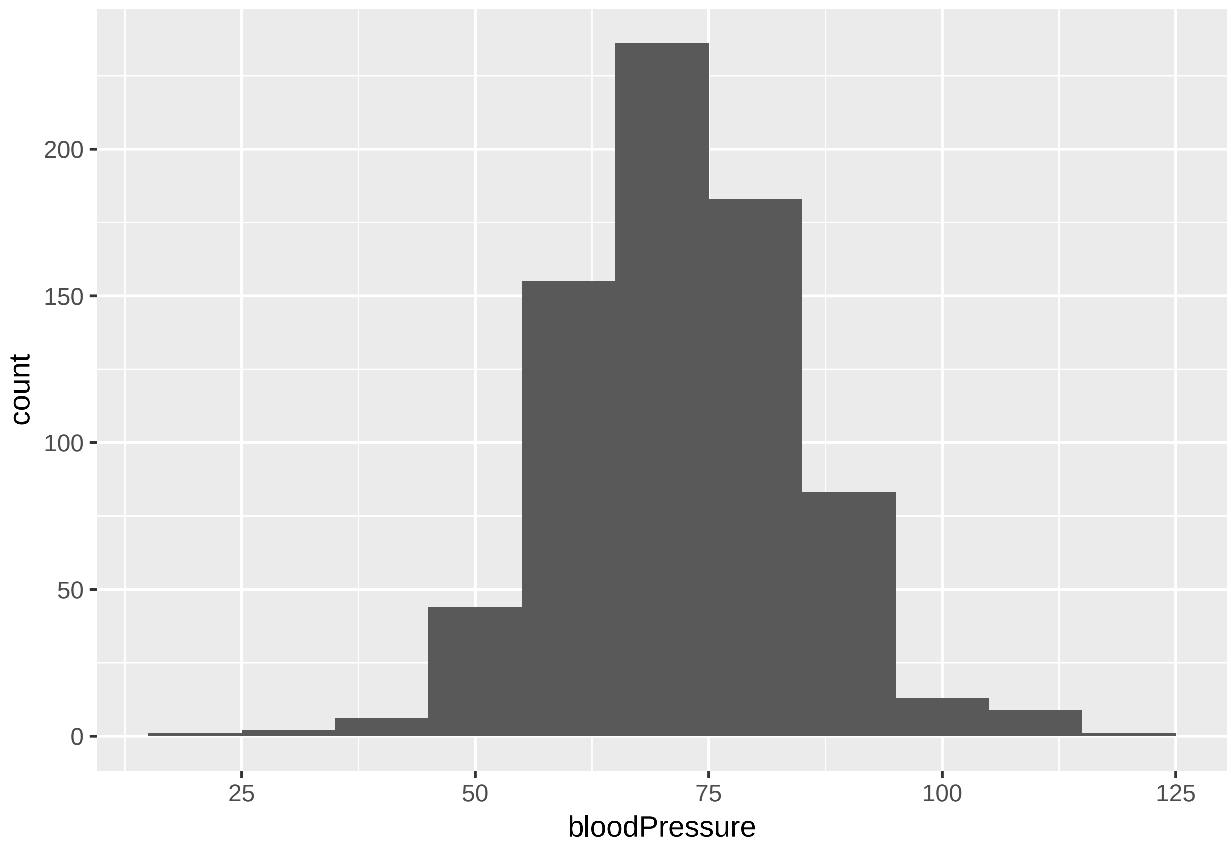
```
##  min Q1 median  Q3 max     mean       sd   n missing
##   44 99    117 141 199 121.6868 30.53564 763       5
```

We can see that the plasma glucose concentrations in saliva have a unimodal and symmetric distribution with a mean of about 121.7 and a standard deviation of 30.5.

There were 5 observations missing a value for glucose concentration.

Diastolic Blood Pressure

```
ggplot(PIMA, aes(x = bloodPressure)) + geom_histogram(binwidth = 10)
```



```
favstats(~ bloodPressure, data = PIMA)
```

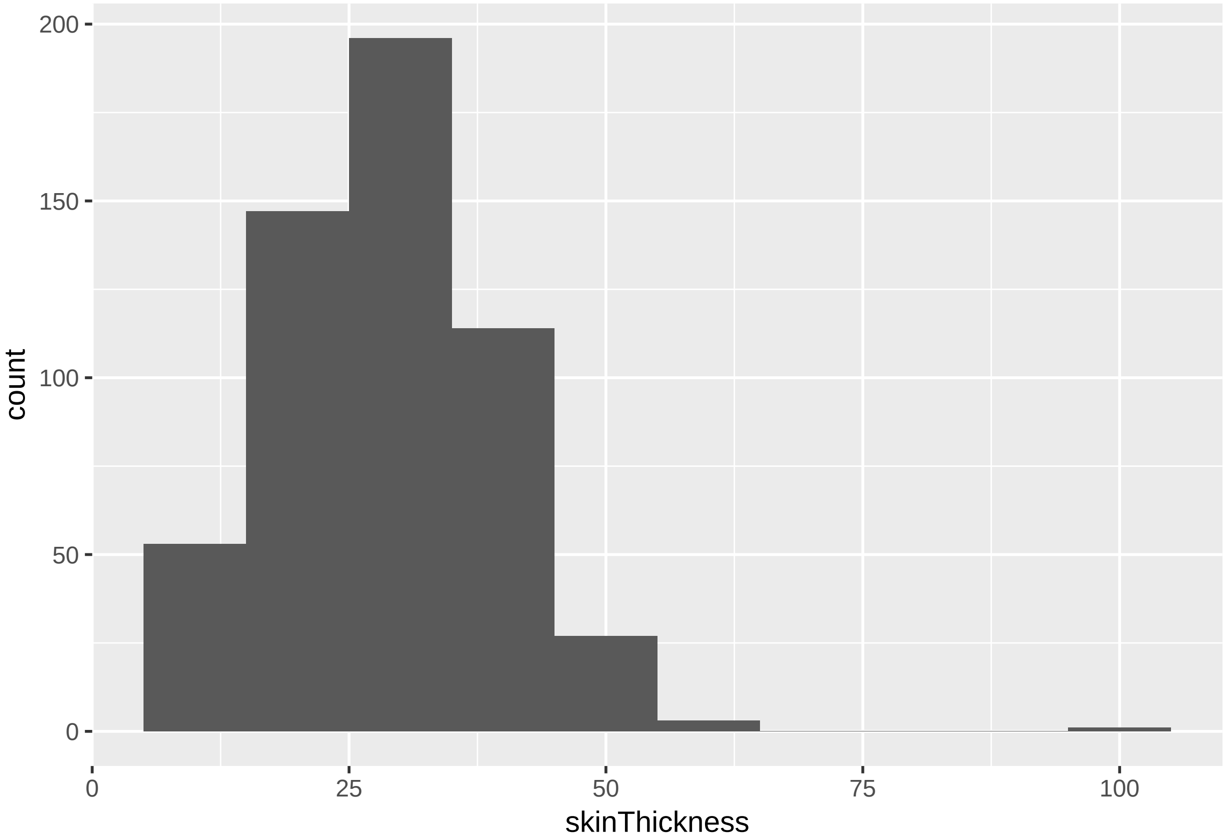
```
##  min Q1 median Q3 max     mean       sd   n missing
##   24 64     72 80 122 72.40518 12.38216 733      35
```

The distribution of diastolic blood pressure is unimodal and symmetric. The distribution has a mean of about 72.4 (mm Hg) and a standard deviation of 12.4 (mm Hg). How does this relate to the average diastolic blood pressure for females 21+ in general?

There were 35 observations missing a value for blood pressure.

Triceps skin fold thickness

```
ggplot(PIMA, aes(x = skinThickness)) + geom_histogram(binwidth = 10)
```



```
favstats(~ skinThickness, data = PIMA)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##    7 22     29 36  99 29.15342 10.47698 541     227
```

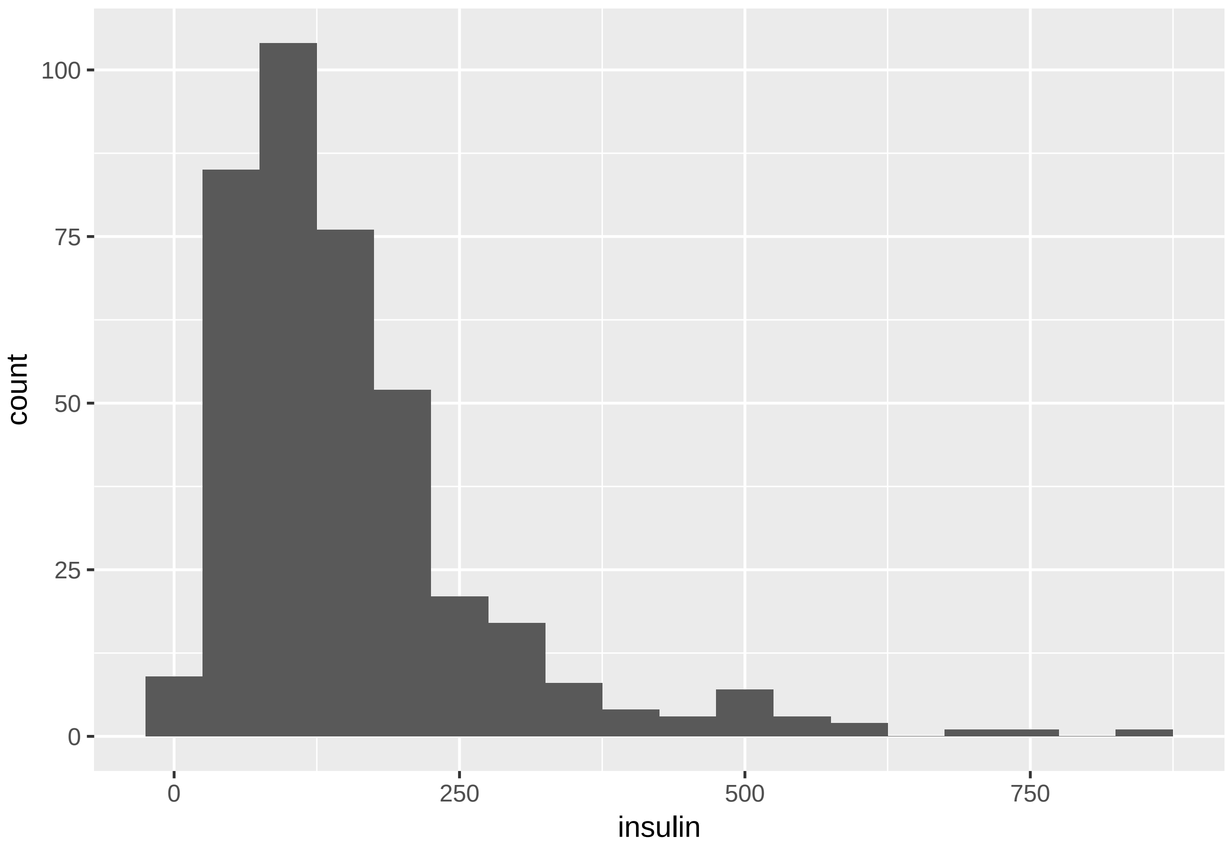
The distribution for triceps skin fold thickness is unimodal and symmetric with a mean of 29.2 (mm) and a standard deviation of 10.5 (mm).

Segen's Medical Dictionary has the following entry for triceps skin-fold thickness: A value used to estimate body fat, measured on the right arm halfway between the olecranon process of the elbow and the acromial process of the scapula. Normal thickness in males is 12 mm; in females, 23 mm.

There were 227 observations with a missing value for skin thickness.

Two Hours Serum Insulin

```
ggplot(PIMA, aes(x = insulin)) + geom_histogram(binwidth = 50)
```



```
favstats(~ insulin, data = PIMA)
```

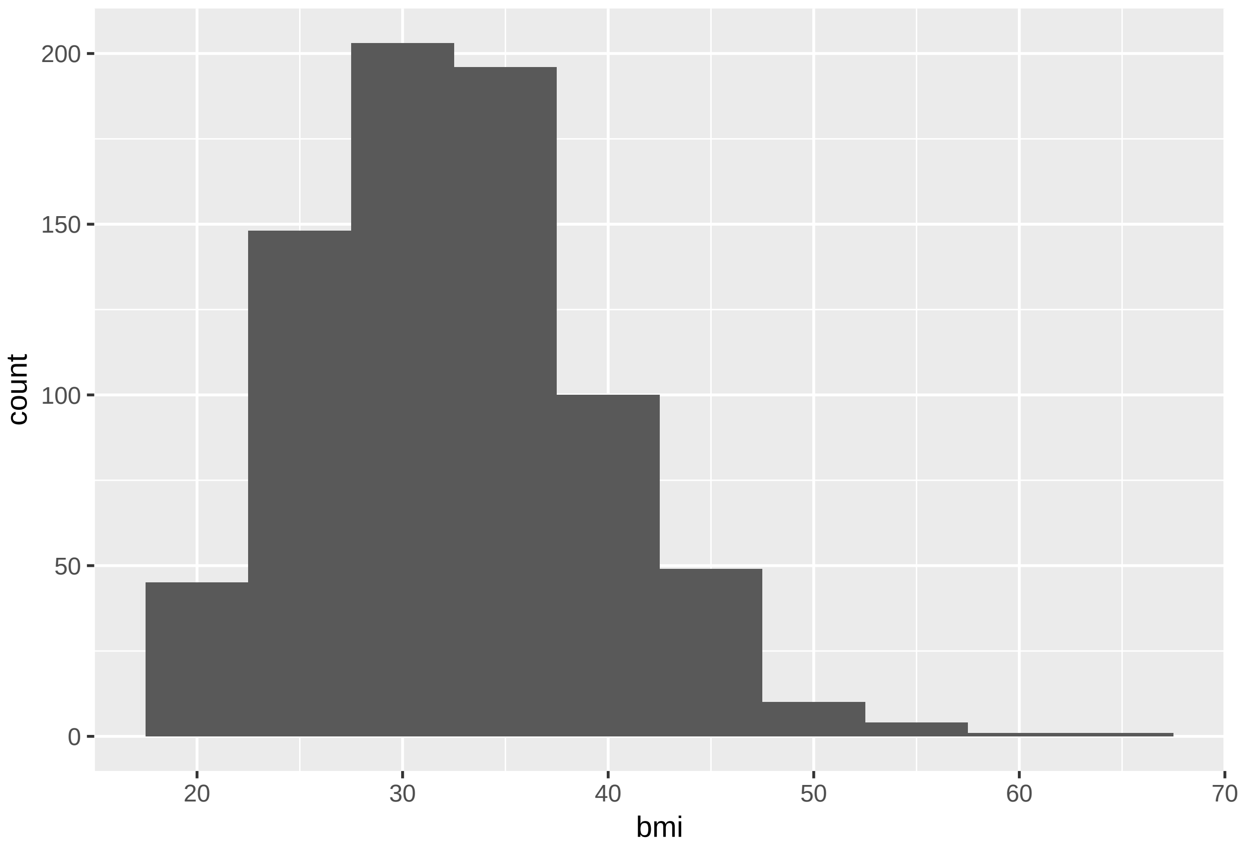
```
##  min    Q1 median  Q3 max     mean       sd   n missing
##   14 76.25    125 190 846 155.5482 118.7759 394     374
```

The distribution for the two hours test is clearly skewed to the right and unimodal. It has a median of 125 (mu U/ml) and a IQR of 113.8 (mu U/ml). This means that 50% of the data lies between 76.25 (mu U/ml) and 190 (mu U/ml). There are some outliers in the far right of the tail with a maximum of 846 (mu U/ml).

There were 374 observations with a missing value for the 2 hour insulin test.

Body Mass Index

```
ggplot(PIMA, aes(x = bmi)) + geom_histogram(binwidth = 5)
```



```
favstats(~ bmi, data = PIMA)
```

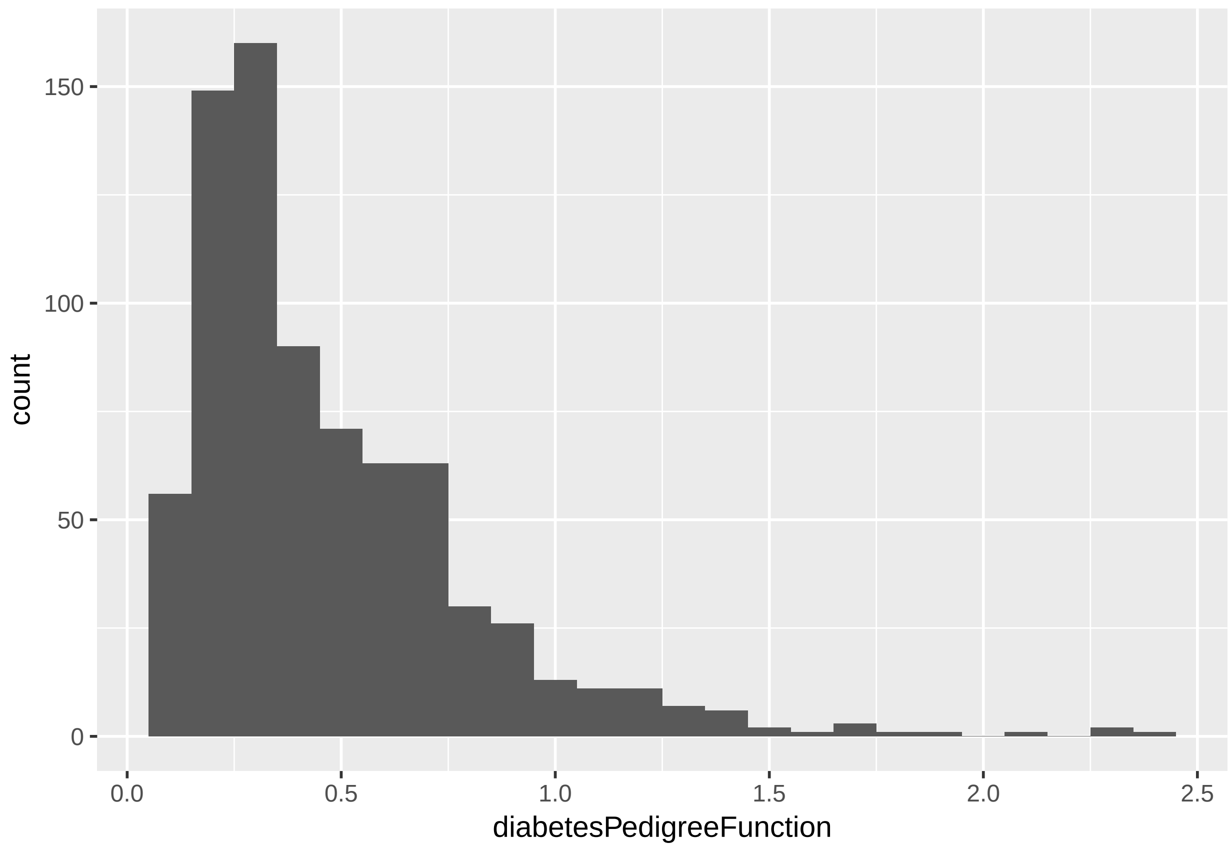
```
##   min   Q1 median   Q3  max     mean       sd   n missing
##  18.2 27.5   32.3 36.6 67.1 32.45746 6.924988 757      11
```

The distribution for bmi is unimodal and slightly skewed to the right. The distribution has mean of 32.5 kg/mˆ2 and a standard deviation of 6.9 kg/mˆ2. There are some outliers in the far right of the distribution representing obese people (the maximum is 67.1 kg/mˆ2).

There were 11 observations missing a value for bmi.

Diabetes Pedigree Function

```
ggplot(PIMA, aes(x = diabetesPedigreeFunction)) + geom_histogram(binwidth = 0.1)
```



```
favstats(~ diabetesPedigreeFunction, data = PIMA)
```

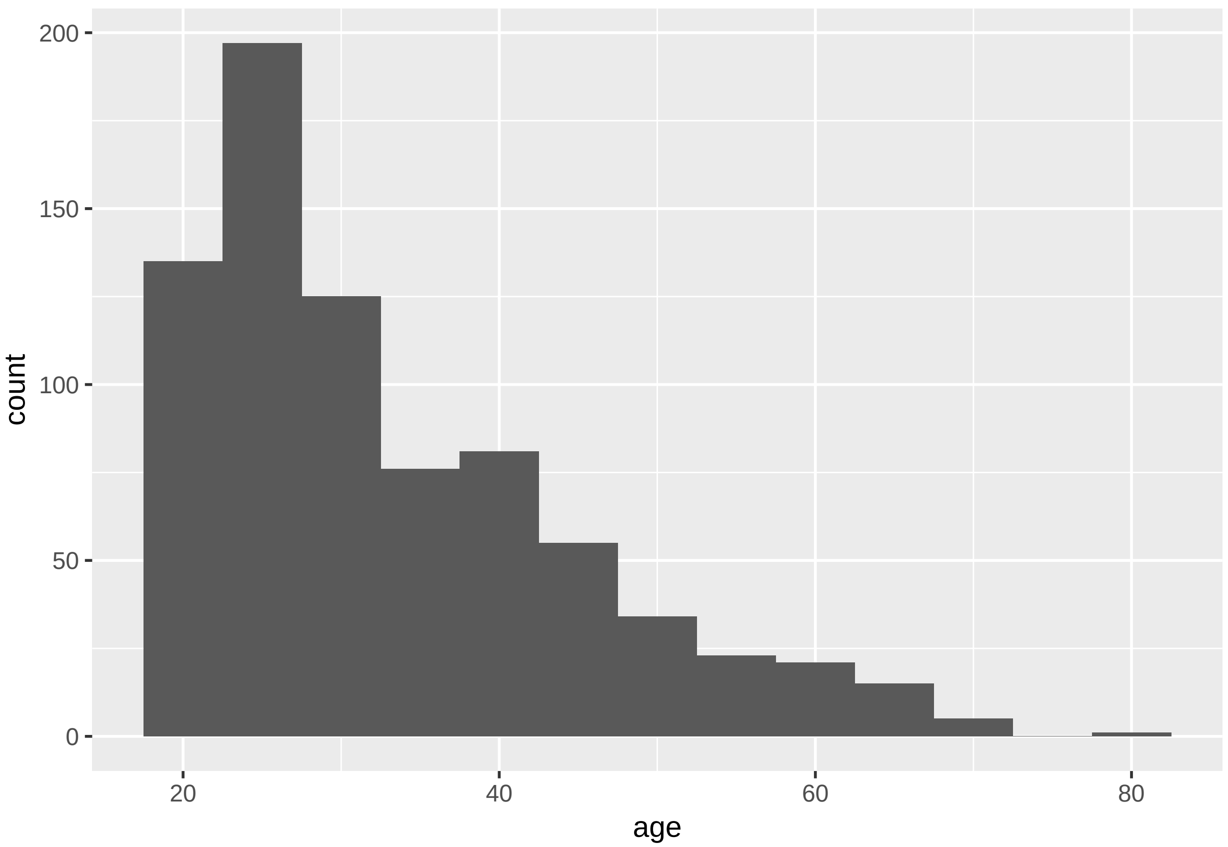
```
##    min      Q1 median      Q3  max      mean        sd   n missing
##  0.078 0.24375 0.3725 0.62625 2.42 0.4718763 0.3313286 768       0
```

The distribution for the values of the diabetes pedigree function is unimodal and skewed to the right. The distribution has a median score of 0.37 and the middle 50% of the observations had a score between 0.24 and 0.62. There are some outliers at the far right of the distribution.

There were 0 observations with missing values for this feature.

Age

```
ggplot(PIMA, aes(x = age)) + geom_histogram(binwidth = 5)
```



```r
favstats(~ age, data = PIMA)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##   21 24     29 41  81 33.24089 11.76023 768       0
```

The distribution for age is unimodal and skewed to the right. The median age is 29 years and the middle 50% of the observations are between 24 years and 41 years of age. The maximum value for age is 81 years.

There are 0 observations missing a value for age.