Project 2

Folorunsho Atanda

2023-10-04

Load required library

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(RMySQL)
```

```
## Loading required package: DBI
```

```
library(dotenv)
```

Create database connection

```
dotenv::load_dot_env(file = "sql_pw.env")
my_sql_pw <- Sys.getenv("MYSQL_PW")

db <- dbConnect(
  MySQL(),
  user = "root",
  #password = my_sql_pw,
  password = "0830spscuny2023!!!",
  dbname = "project_2",
  host = "localhost",
  port = 3306
)
```

Load table from database as a data frame

```
diabetes <- db %>%
  dbGetQuery("select * from project_2.diabetes")

diabetes_df <- as_tibble(diabetes)

# Disconnect database
#db_status <- dbDisconnect(db)
```

Because **0** can't exist in the columns **Glucose, BloodPressure, SkinThickness, Insulin, BMI, DPF and Age**, going to change does to **NA**. Then also change **0** to **FALSE** and **1** to **TRUE** in **Outcome**

```
columns_to_modify <- c("Glucose", "BloodPressure", "SkinThickness", "Insulin",
                       "BMI", "DPF", "Age")

diabetes_df <- diabetes_df %>%
  mutate(across(all_of(columns_to_modify), ~if_else(. == 0, NA, .))) %>%
  mutate("Outcome" = ifelse(`Outcome` == 0, "Negative", "Positive"))
```

Summarise data

```
glimpse(diabetes_df)
```

```
## Rows: 768
## Columns: 10
## $ ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ Pregnancies   <int> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, 5, 7, 0, 7,~
## $ Glucose       <int> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125, 110, 168,~
## $ BloodPressure <int> 72, 66, 64, 66, 40, 74, 50, NA, 70, 96, 92, 74, 80, 60, ~
## $ SkinThickness <int> 35, 29, NA, 23, 35, NA, 32, NA, 45, NA, NA, NA, NA, 23, ~
## $ Insulin       <int> NA, NA, NA, 94, 168, NA, 88, NA, 543, NA, NA, NA, NA, 84~
## $ BMI           <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.3, 30.5, NA~
## $ DPF           <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.248, 0.134, ~
## $ Age           <int> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 34, 57, 59, ~
## $ Outcome       <chr> "Positive", "Negative", "Positive", "Negative", "Positiv~
```

```
summary(diabetes_df)
```

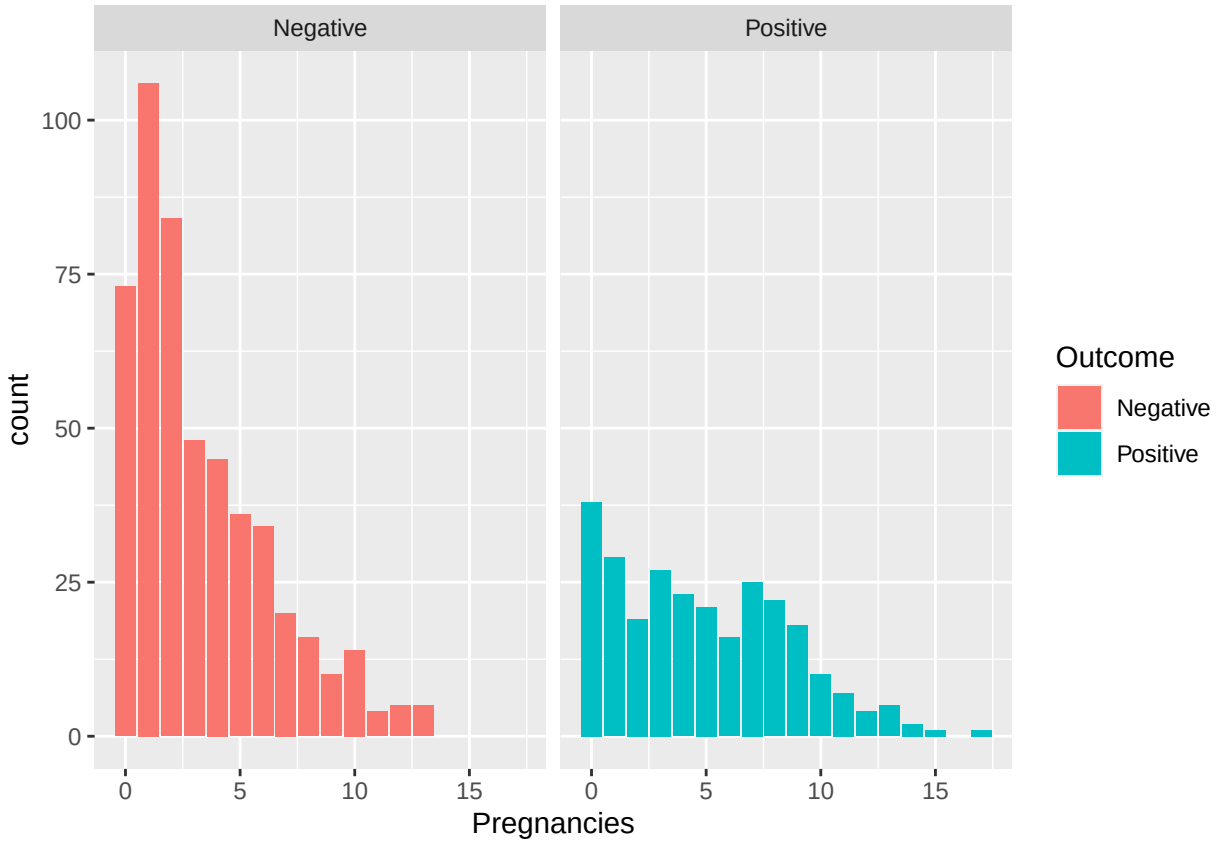
```
##        ID         Pregnancies        Glucose      BloodPressure
##  Min.   :  1.0   Min.   : 0.000   Min.   : 44.0   Min.   : 24.00
##  1st Qu.:192.8   1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 64.00
##  Median :384.5   Median : 3.000   Median :117.0   Median : 72.00
##  Mean   :384.5   Mean   : 3.845   Mean   :121.7   Mean   : 72.41
##  3rd Qu.:576.2   3rd Qu.: 6.000   3rd Qu.:141.0   3rd Qu.: 80.00
##  Max.   :768.0   Max.   :17.000   Max.   :199.0   Max.   :122.00
##                                   NA's   :5       NA's   :35
##  SkinThickness      Insulin           BMI             DPF
```

```
##  Min.   : 7.00   Min.   : 14.00   Min.   :18.20   Min.   :0.0780
##  1st Qu.:22.00   1st Qu.: 76.25   1st Qu.:27.50   1st Qu.:0.2437
##  Median :29.00   Median :125.00   Median :32.30   Median :0.3725
##  Mean   :29.15   Mean   :155.55   Mean   :32.46   Mean   :0.4719
##  3rd Qu.:36.00   3rd Qu.:190.00   3rd Qu.:36.60   3rd Qu.:0.6262
##  Max.   :99.00   Max.   :846.00   Max.   :67.10   Max.   :2.4200
##  NA's   :227     NA's   :374      NA's   :11
##       Age          Outcome
##  Min.   :21.00   Length:768
##  1st Qu.:24.00   Class :character
##  Median :29.00   Mode  :character
##  Mean   :33.24
##  3rd Qu.:41.00
##  Max.   :81.00
##
```

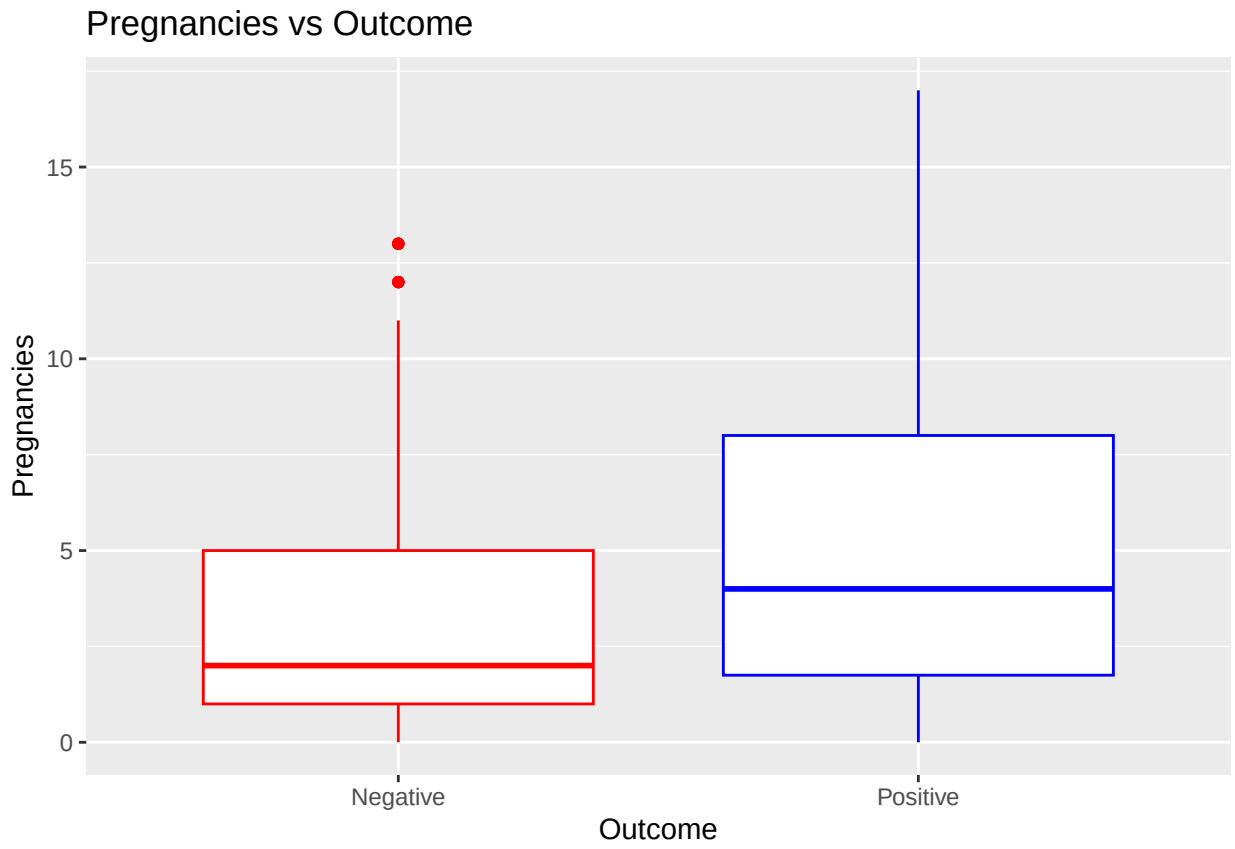
Compare Variables

Pregnancies vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `Pregnancies`, fill = `Outcome`)) +
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `Pregnancies`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    ) +
  labs(title = "Pregnancies vs Outcome")
```




```
diabetes_df %>%
  select(c(`Pregnancies`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    Pregnancies_mean = mean(`Pregnancies`),
    Pregnancies_median = median(`Pregnancies`),
    Pregnancies_Q1 = quantile(`Pregnancies`, probs = 0.25),
    Pregnancies_Q2 = quantile(`Pregnancies`, probs = 0.75)
    )
```

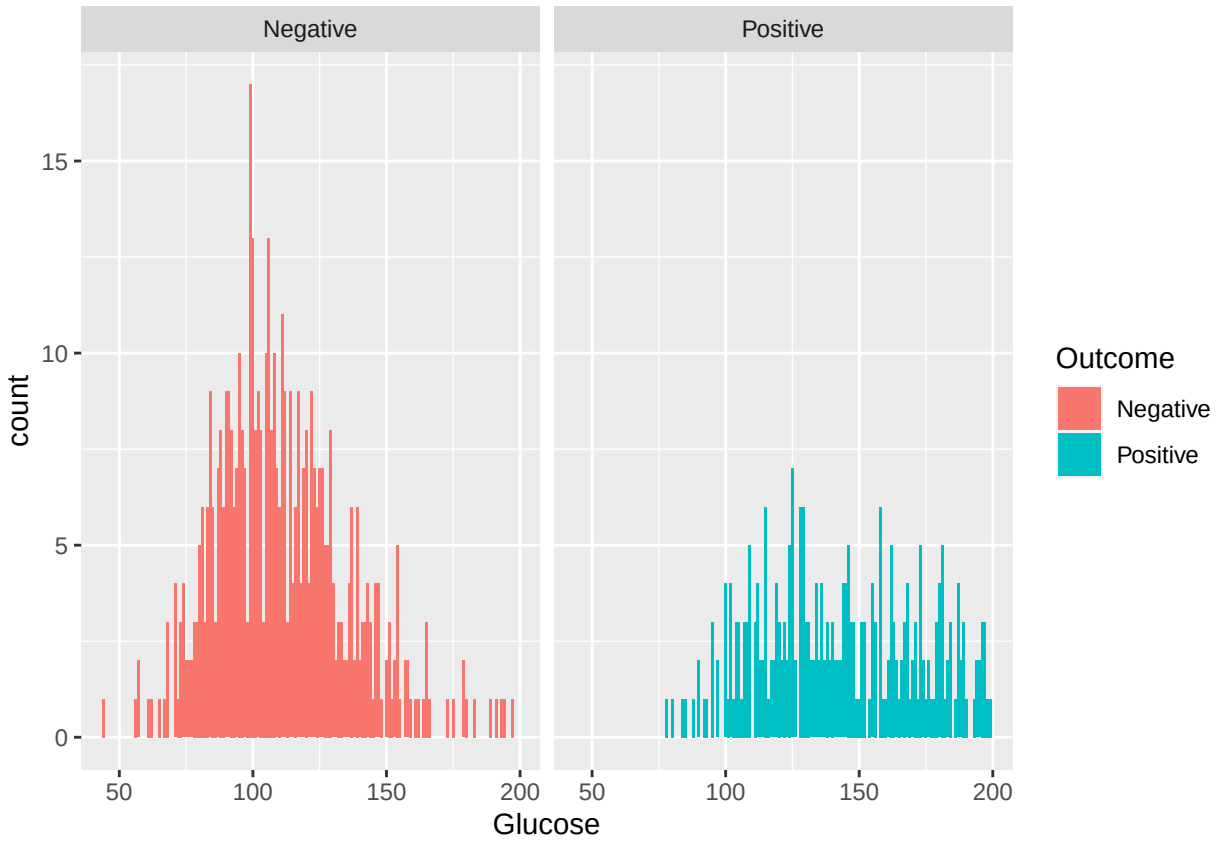
```
## # A tibble: 2 x 5
##   Outcome  Pregnancies_mean Pregnancies_median Pregnancies_Q1 Pregnancies_Q2
##   <chr>               <dbl>              <dbl>          <dbl>          <dbl>
## 1 Negative             3.30                  2           1                 5
## 2 Positive             4.87                  4           1.75              8
```

Analysis: The distribution of `Pregnancies` to `Outcome` is right skewed. This tells us that the median and mean are not the same. Which is confirmed in the box plot and the stat table.

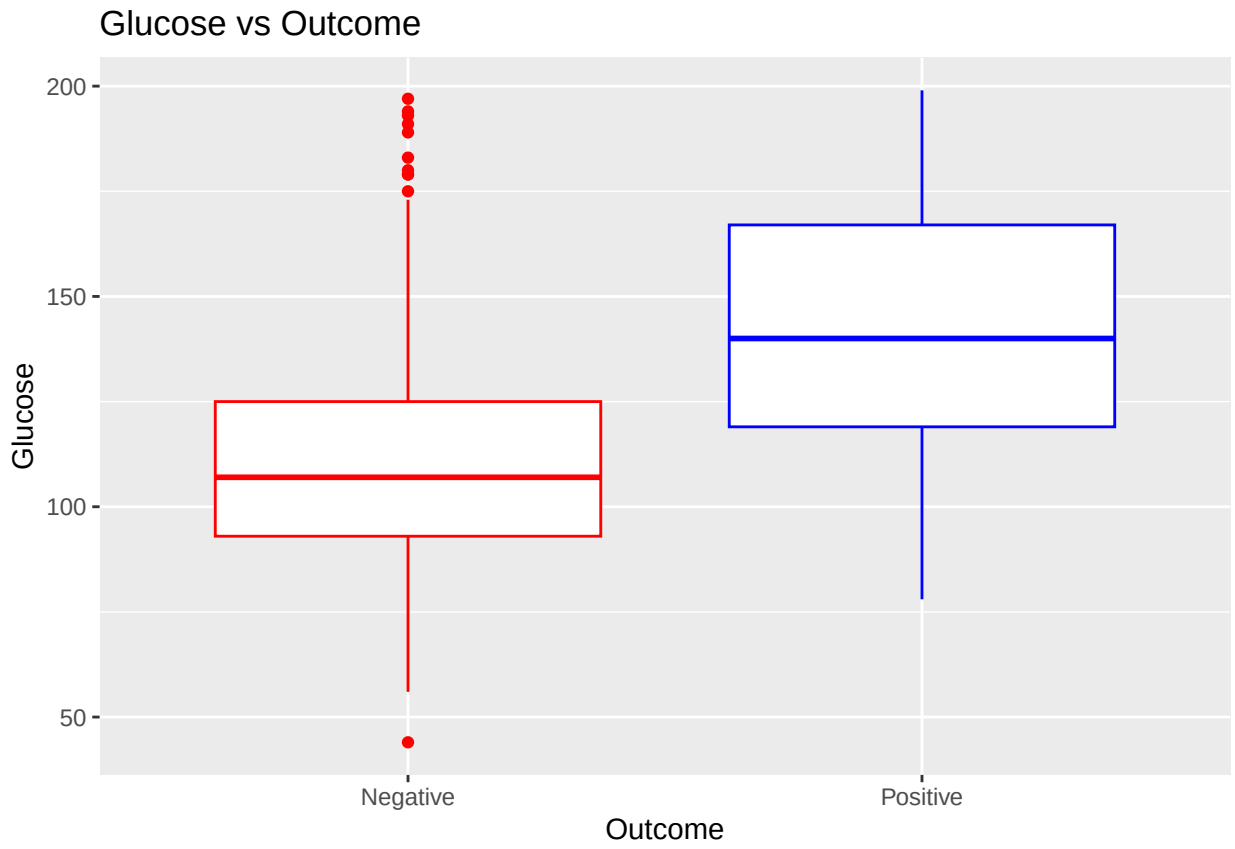
Glucose vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `Glucose`, fill = `Outcome`)) +
```

```
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `Glucose`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    ) +
  labs(title = "Glucose vs Outcome")
```



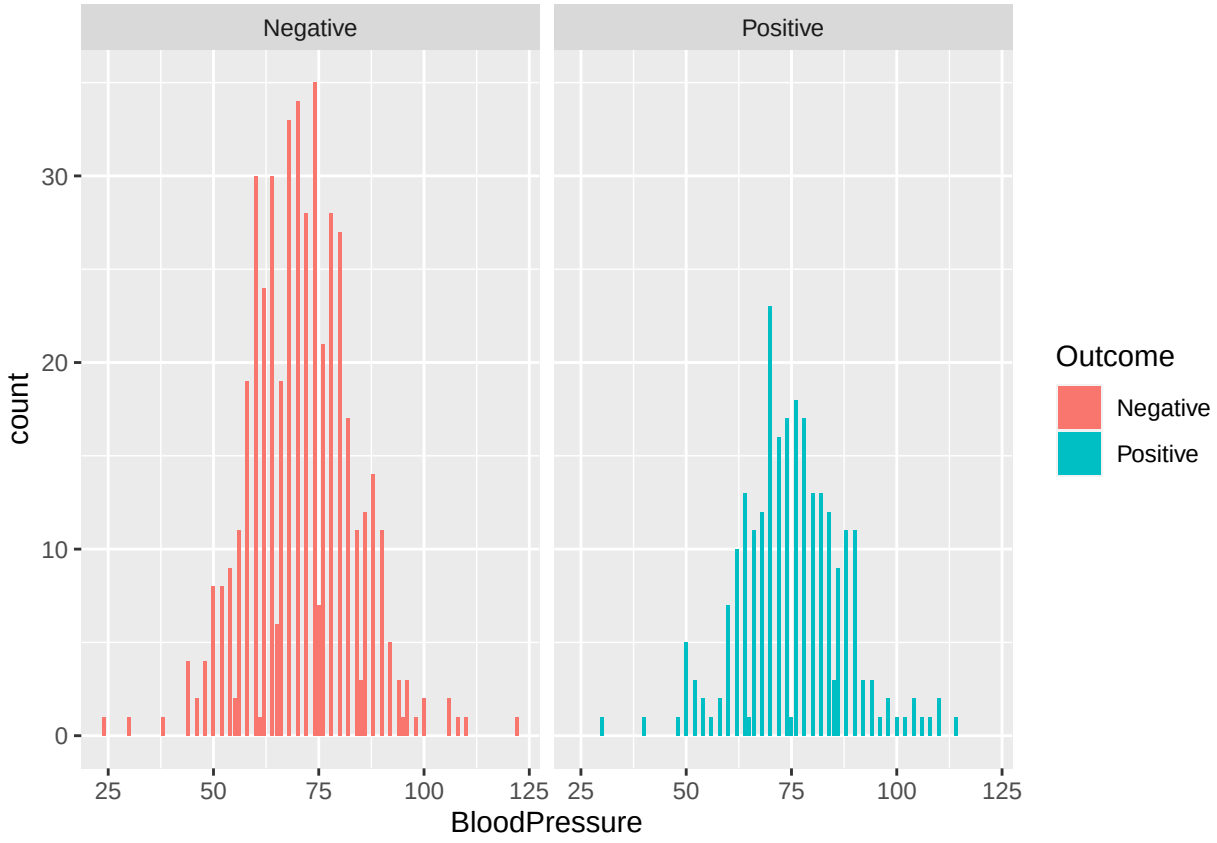

```
diabetes_df %>%
  select(c(`Glucose`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    Glucose_mean = mean(`Glucose`, na.rm = TRUE),
    Glucose_median = median(`Glucose`, na.rm = TRUE),
    Glucose_Q1 = quantile(`Glucose`, probs = 0.25, na.rm = TRUE),
    Glucose_Q2 = quantile(`Glucose`, probs = 0.75, na.rm = TRUE)
    )
```

```
## # A tibble: 2 x 5
##   Outcome  Glucose_mean Glucose_median Glucose_Q1 Glucose_Q2
##   <chr>           <dbl>          <dbl>      <dbl>      <dbl>
## 1 Negative         111.            107         93        125
## 2 Positive         142.            140        119        167
```

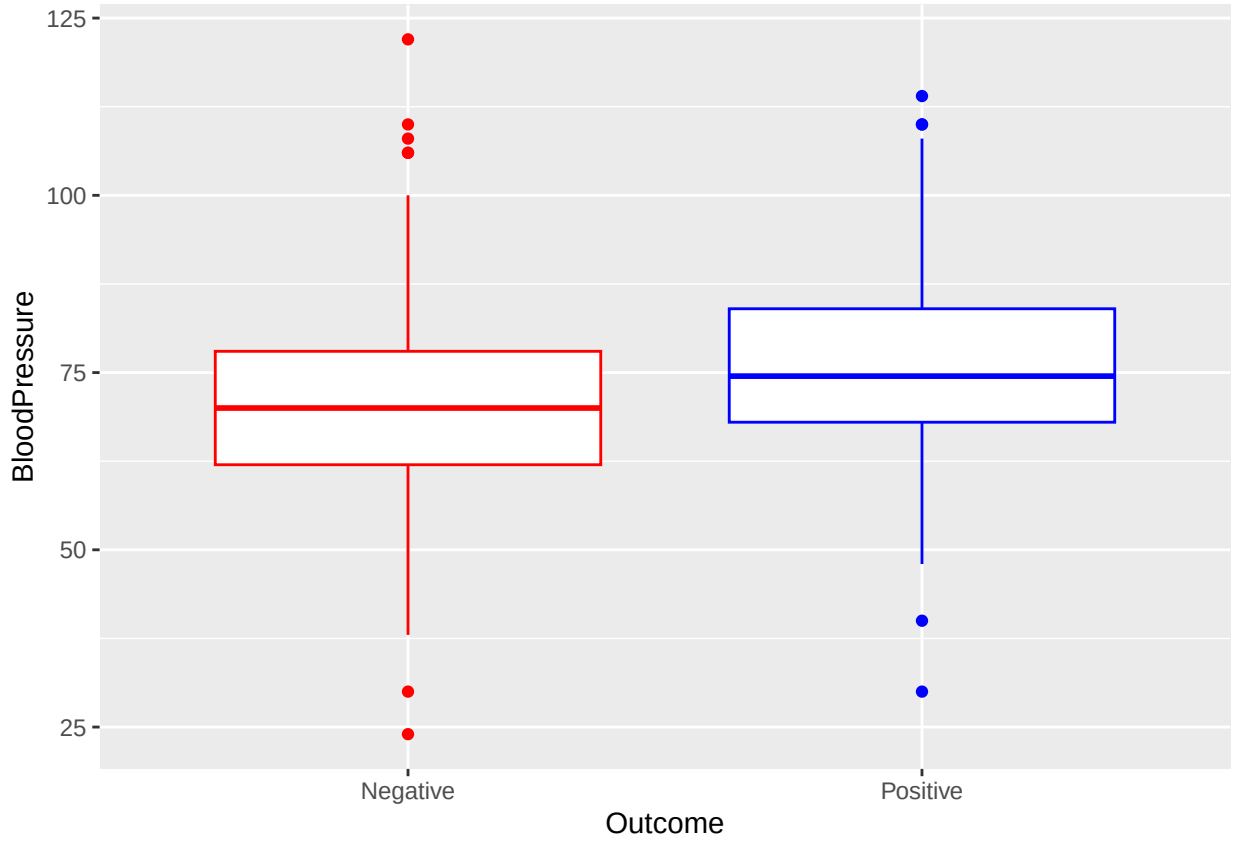
Analysis: The distribution of `Glucose` to `Outcome` is symmetrical. This tells us that the median and mean are very close to one another. Which is confirmed in the box plot and the stat table. From the stat table we can infer that most of the data is between 93 and 167.

Blood Pressure vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `BloodPressure`, fill = `Outcome`)) +
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `BloodPressure`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    )
```



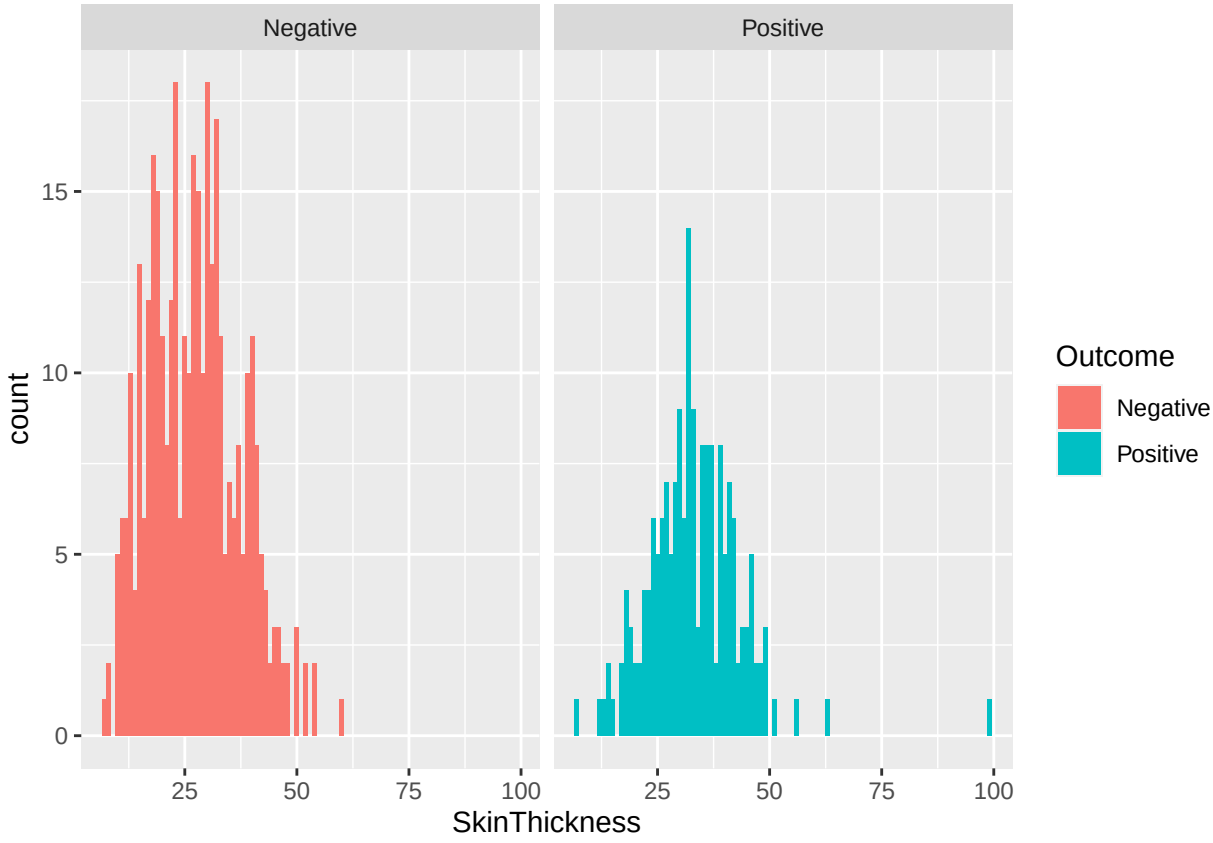

```
diabetes_df %>%
  select(c(`BloodPressure`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    BloodPressure_mean = mean(`BloodPressure`, na.rm = TRUE),
    BloodPressure_median = median(`BloodPressure`, na.rm = TRUE),
    BloodPressure_Q1 = quantile(`BloodPressure`, probs = 0.25, na.rm = TRUE),
    BloodPressure_Q2 = quantile(`BloodPressure`, probs = 0.75, na.rm = TRUE)
    )
```

```
## # A tibble: 2 x 5
##   Outcome  BloodPressure_mean BloodPressure_median BloodPressure_Q1
##   <chr>                 <dbl>                <dbl>            <dbl>
## 1 Negative               70.9                 70                62
## 2 Positive               75.3                 74.5              68
## # i 1 more variable: BloodPressure_Q2 <dbl>
```

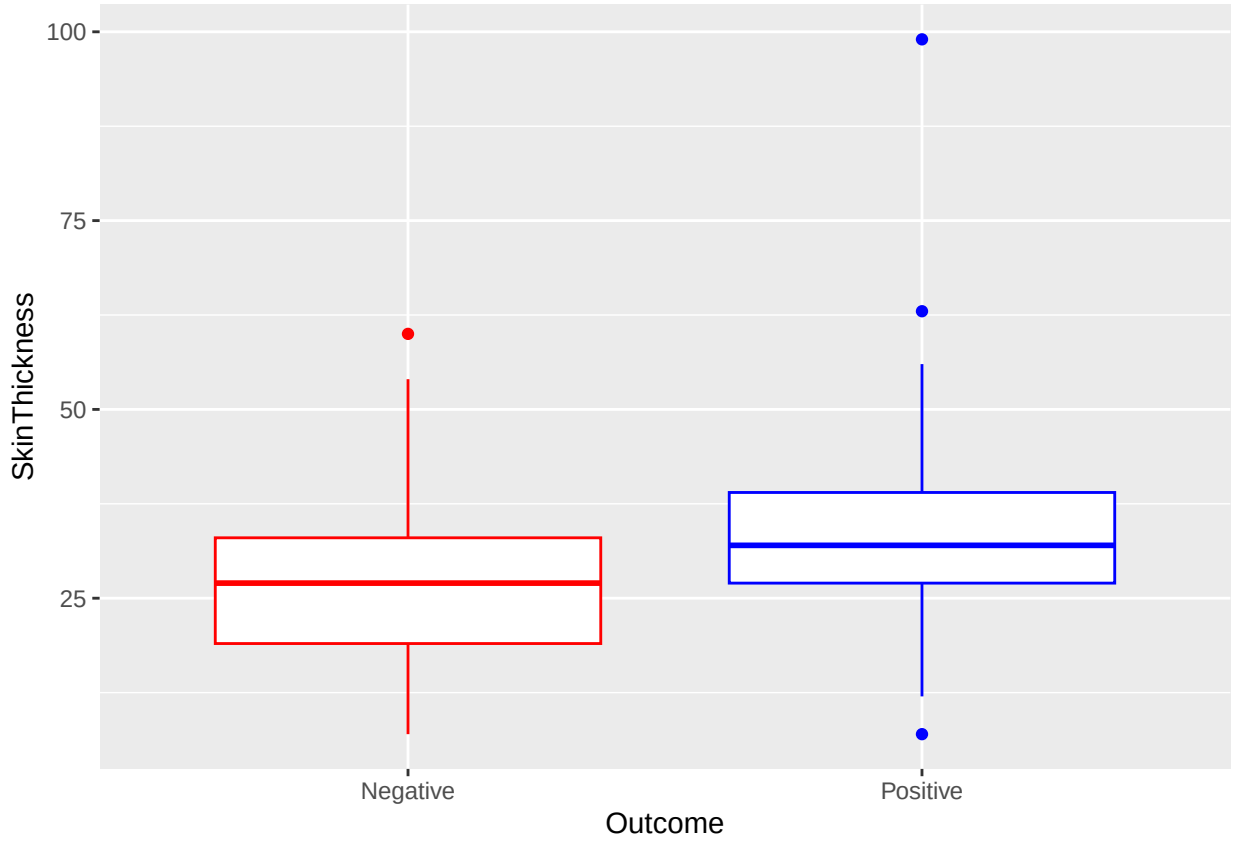
Analysis: The distribution of `BloodPressure` to `Outcome` is symmetrical. This tells us that the median and mean are very close to one another. Which is confirmed in the box plot and the stat table. From the stat table we can infer that most of the data is between 62 and 84.

Skin Thickness vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `SkinThickness`, fill = `Outcome`)) +
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `SkinThickness`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    )
```



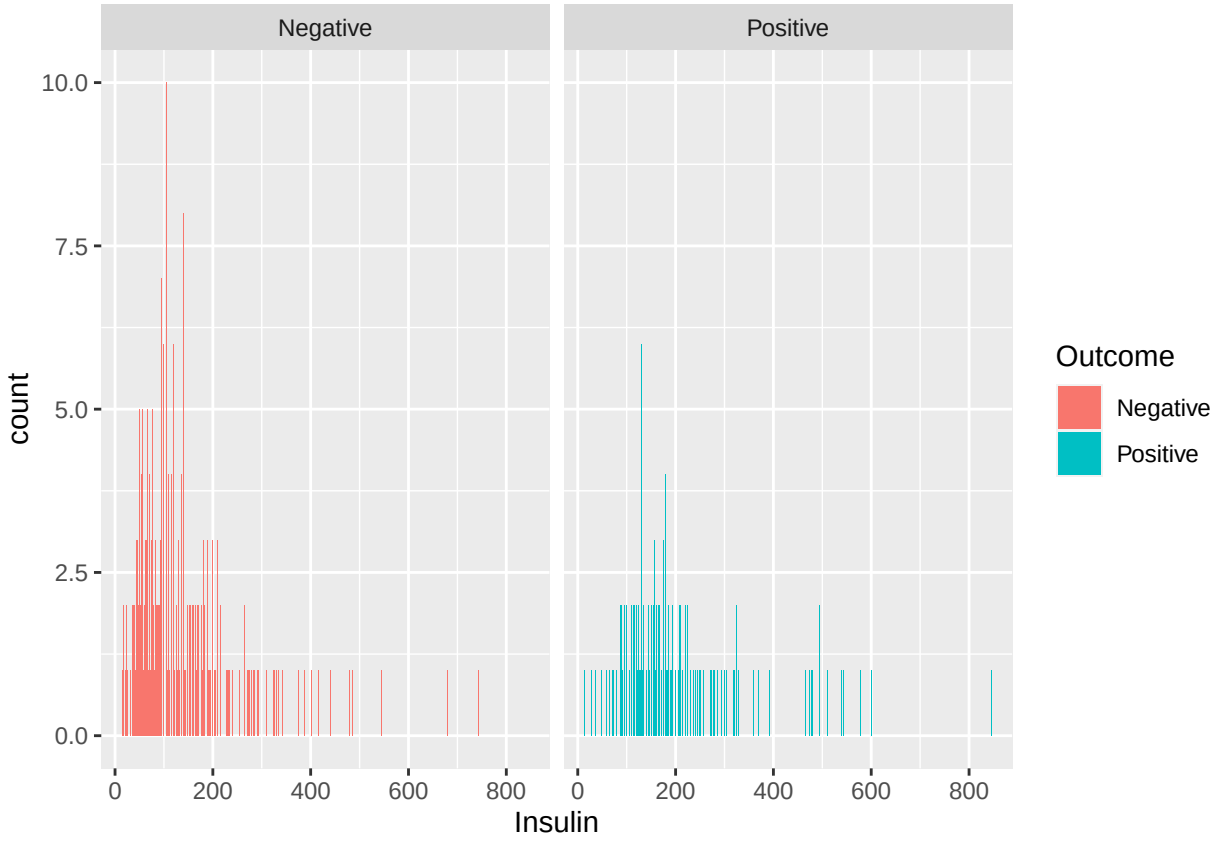

```
diabetes_df %>%
  select(c(`SkinThickness`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    SkinThickness_mean = mean(`SkinThickness`, na.rm = TRUE),
    SkinThickness_median = median(`SkinThickness`, na.rm = TRUE),
    SkinThickness_Q1 = quantile(`SkinThickness`, probs = 0.25, na.rm = TRUE),
    SkinThickness_Q2 = quantile(`SkinThickness`, probs = 0.75, na.rm = TRUE)
    )
```

```
## # A tibble: 2 x 5
##   Outcome  SkinThickness_mean SkinThickness_median SkinThickness_Q1
##   <chr>                 <dbl>                <dbl>            <dbl>
## 1 Negative               27.2                   27               19
## 2 Positive               33                     32               27
## # i 1 more variable: SkinThickness_Q2 <dbl>
```

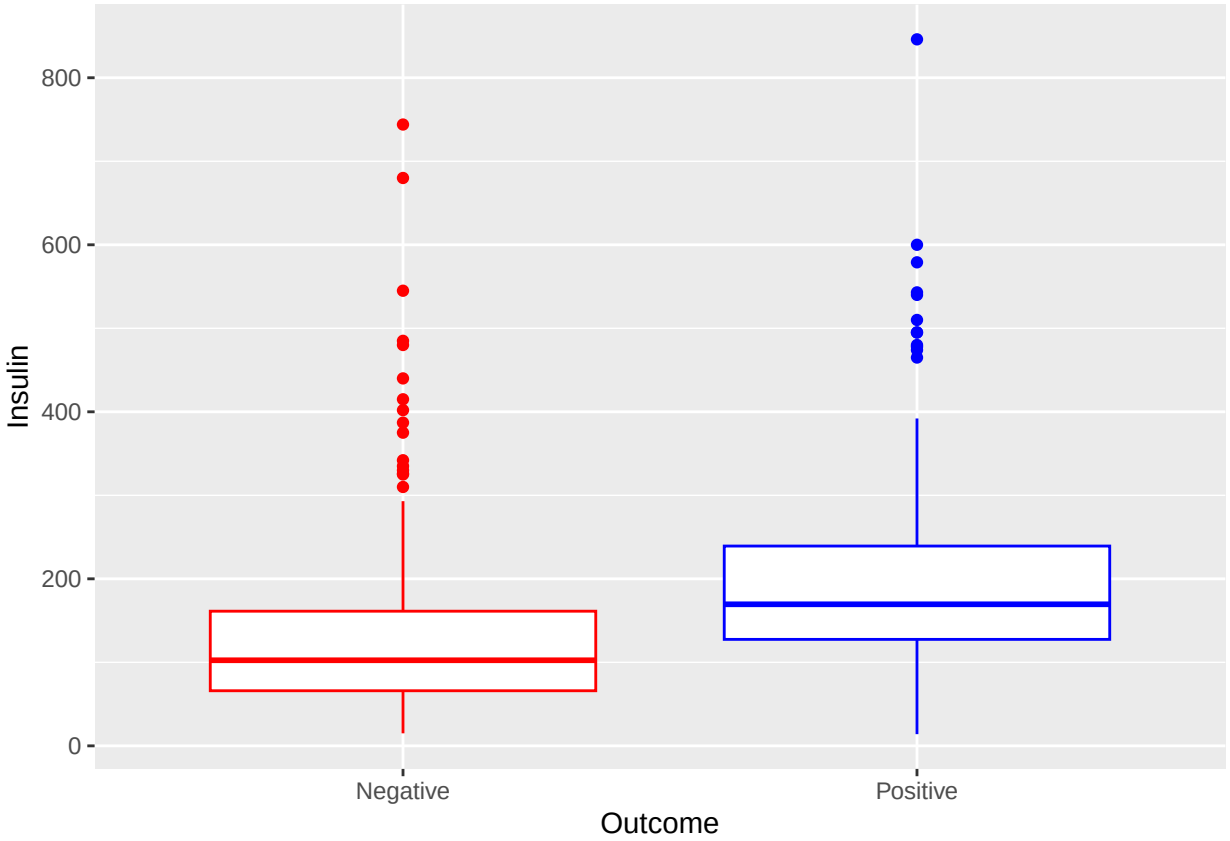
Analysis: The distribution of `SkinThickness` to `Outcome` is symmetrical. This tells us that the median and mean are very close to one another. Which is confirmed in the box plot and the stat table. From the stat table we can infer that most of the data is between 19 and 39.

Insulin vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `Insulin`, fill = `Outcome`)) +
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `Insulin`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    )
```




```
diabetes_df %>%
  select(c(`Insulin`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    Insulin_mean = mean(`Insulin`, na.rm = TRUE),
    Insulin_median = median(`Insulin`, na.rm = TRUE),
    Insulin_Q1 = quantile(`Insulin`, probs = 0.25, na.rm = TRUE),
    Insulin_Q2 = quantile(`Insulin`, probs = 0.75, na.rm = TRUE)
    )
```

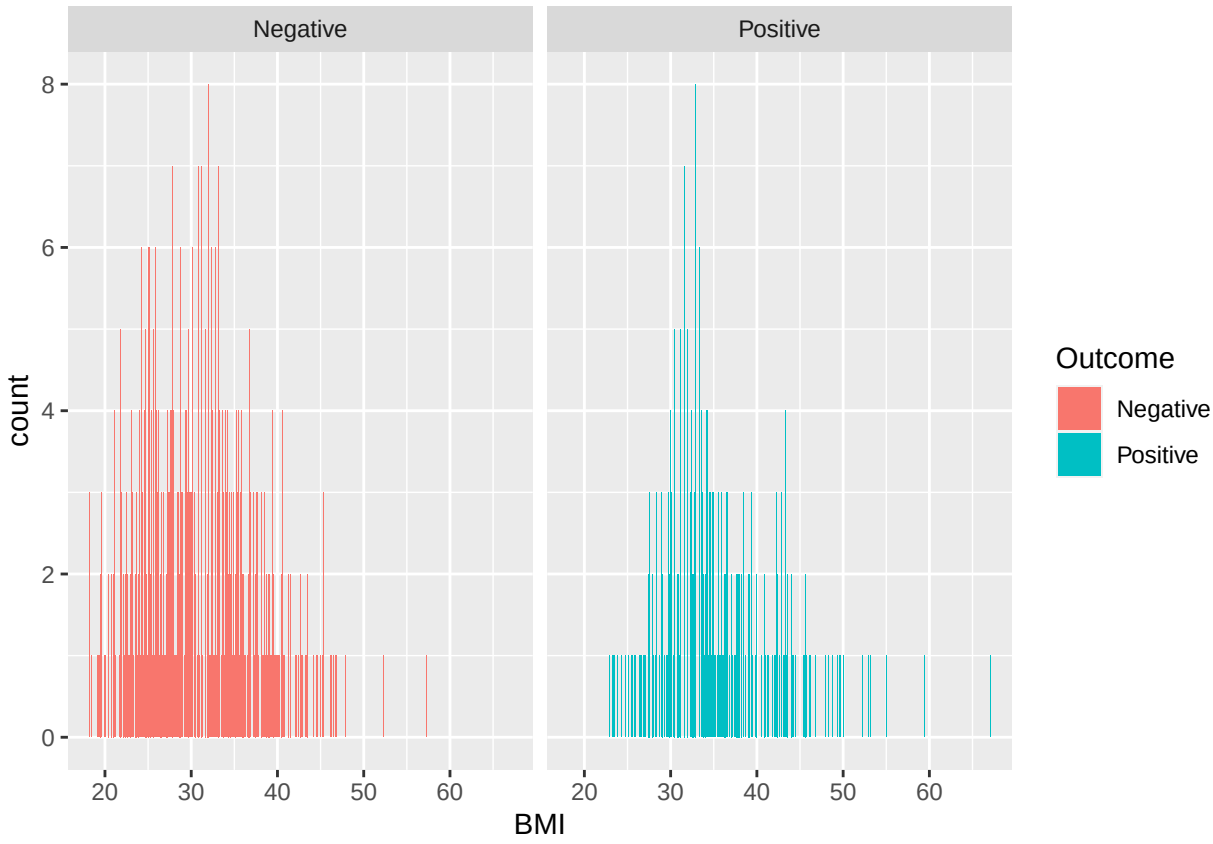
```
## # A tibble: 2 x 5
##   Outcome  Insulin_mean Insulin_median Insulin_Q1 Insulin_Q2
##   <chr>           <dbl>          <dbl>      <dbl>      <dbl>
## 1 Negative         130.           102.         66        161.
## 2 Positive         207.           170.        128.       239.
```

Analysis: The distribution of `Insulin` to `Outcome` is right skewed. This tells us that the median and mean are not the same. Which is confirmed in the box plot and the stat table.

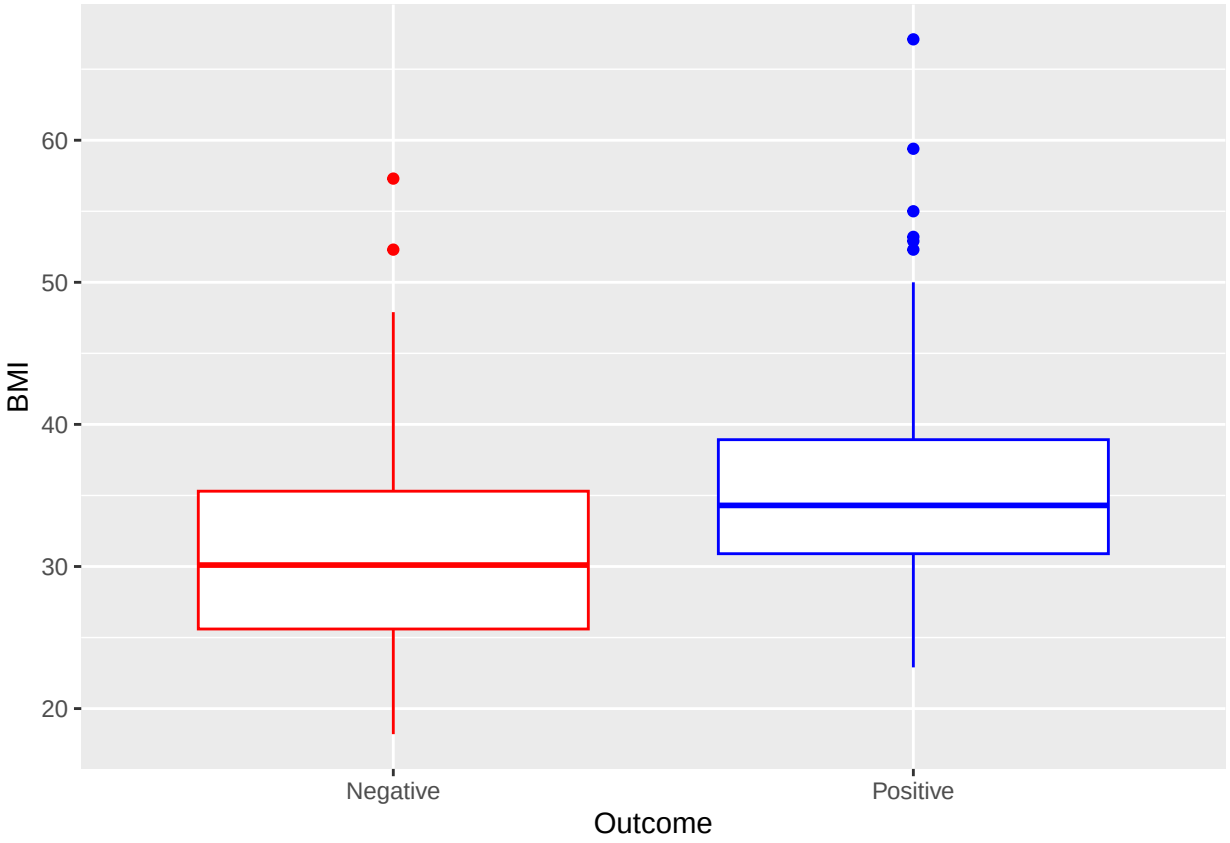
BMI vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `BMI`, fill = `Outcome`)) +
```

```
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `BMI`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    )
```



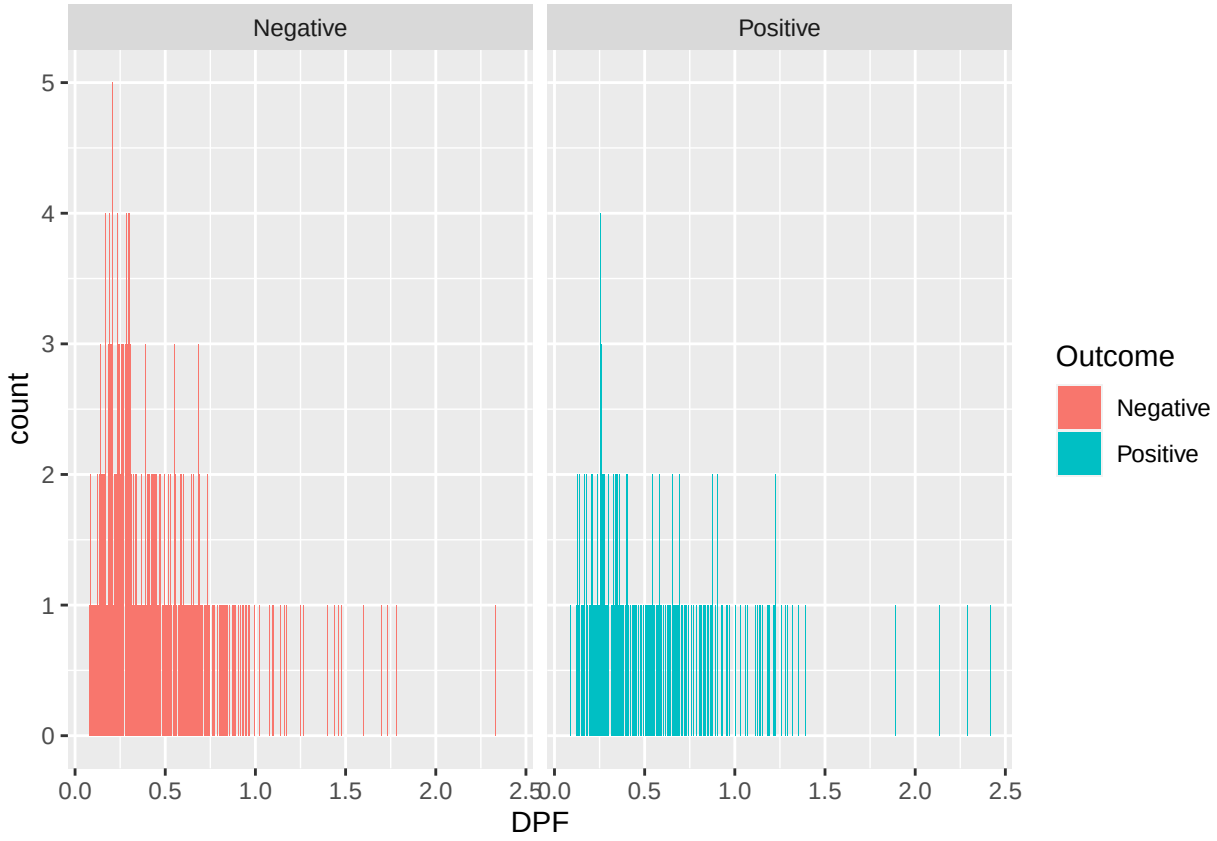

```
diabetes_df %>%
  select(c(`BMI`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    BMI_mean = mean(`BMI`, na.rm = TRUE),
    BMI_median = median(`BMI`, na.rm = TRUE),
    BMI_Q1 = quantile(`BMI`, probs = 0.25, na.rm = TRUE),
    BMI_Q2 = quantile(`BMI`, probs = 0.75, na.rm = TRUE)
    )
```

```
## # A tibble: 2 x 5
##   Outcome  BMI_mean BMI_median BMI_Q1 BMI_Q2
##   <chr>       <dbl>      <dbl>  <dbl>  <dbl>
## 1 Negative     30.9       30.1   25.6   35.3
## 2 Positive     35.4       34.3   30.9   38.9
```

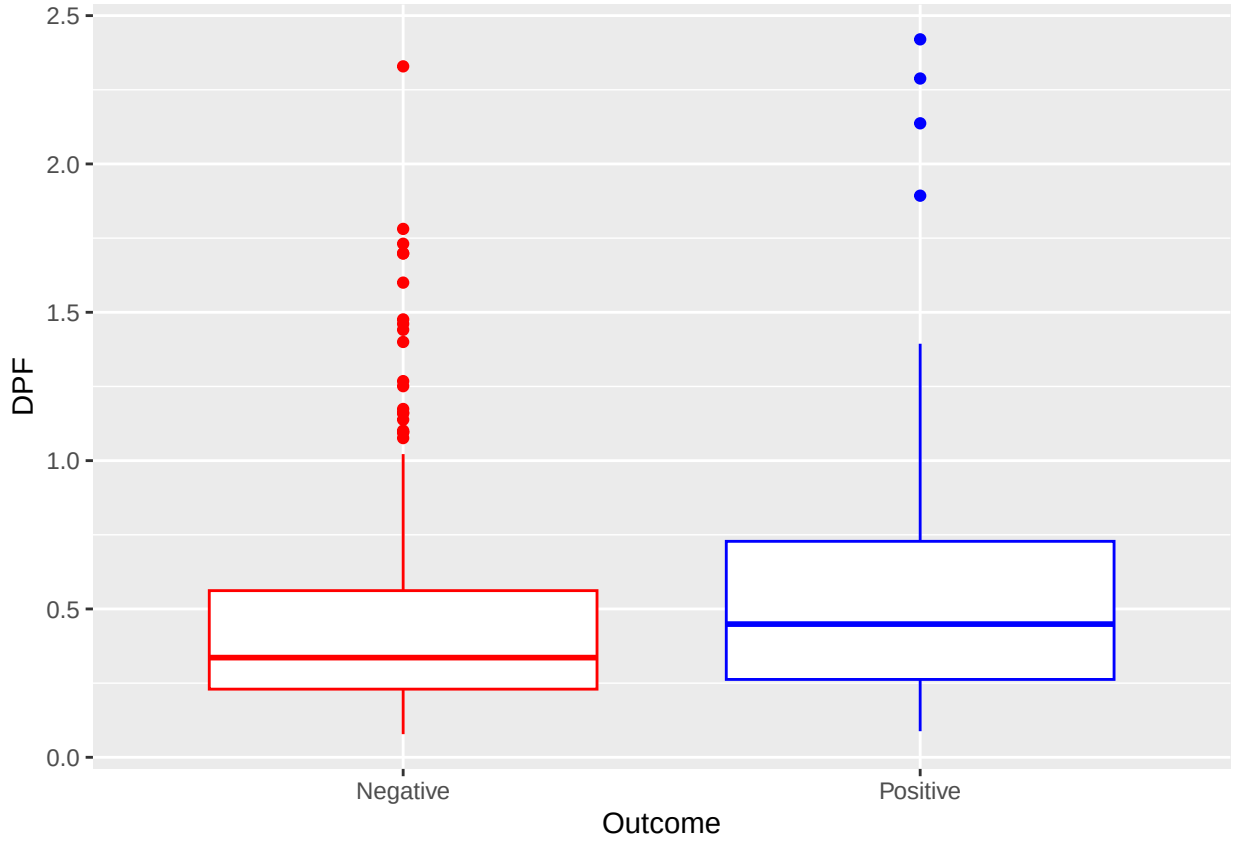
Analysis: The distribution of `BMI` to `Outcome` is symmetrical. This tells us that the median and mean are very close to one another. Which is confirmed in the box plot and the stat table. From the stat table we can infer that most of the data is between 19 and 39.

DPF vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `DPF`, fill = `Outcome`)) +
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `DPF`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    )
```




```
diabetes_df %>%
  select(c(`DPF`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    DPF_mean = mean(`DPF`, na.rm = TRUE),
    DPF_median = median(`DPF`, na.rm = TRUE),
    DPF_Q1 = quantile(`DPF`, probs = 0.25, na.rm = TRUE),
    DPF_Q2 = quantile(`DPF`, probs = 0.75, na.rm = TRUE)
    )
```

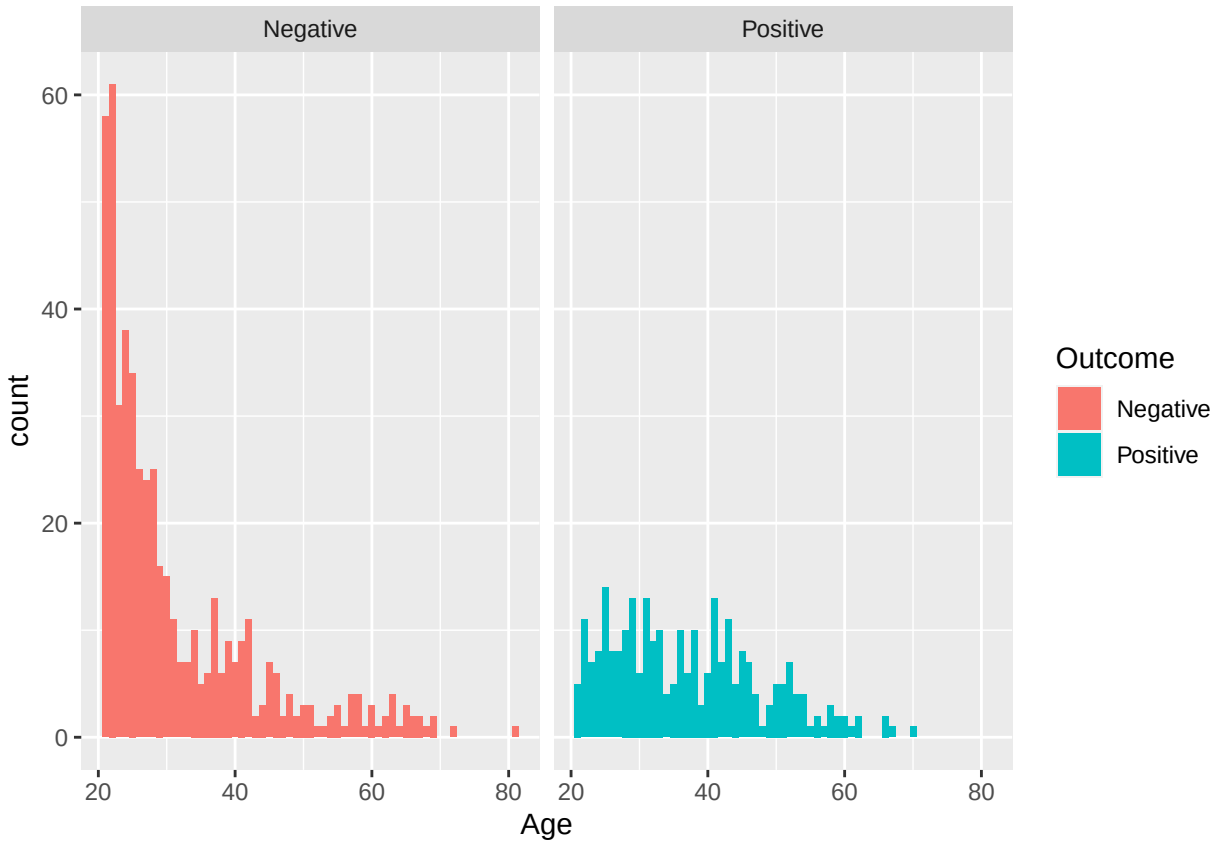
```
## # A tibble: 2 x 5
##   Outcome  DPF_mean DPF_median DPF_Q1 DPF_Q2
##   <chr>       <dbl>      <dbl>  <dbl>  <dbl>
## 1 Negative    0.430      0.336  0.230  0.562
## 2 Positive    0.550      0.449  0.262  0.728
```

Analysis: The distribution of `DPF` to `Outcome` is right skewed. This tells us that the median and mean are not the same. Which is confirmed in the box plot and the stat table.

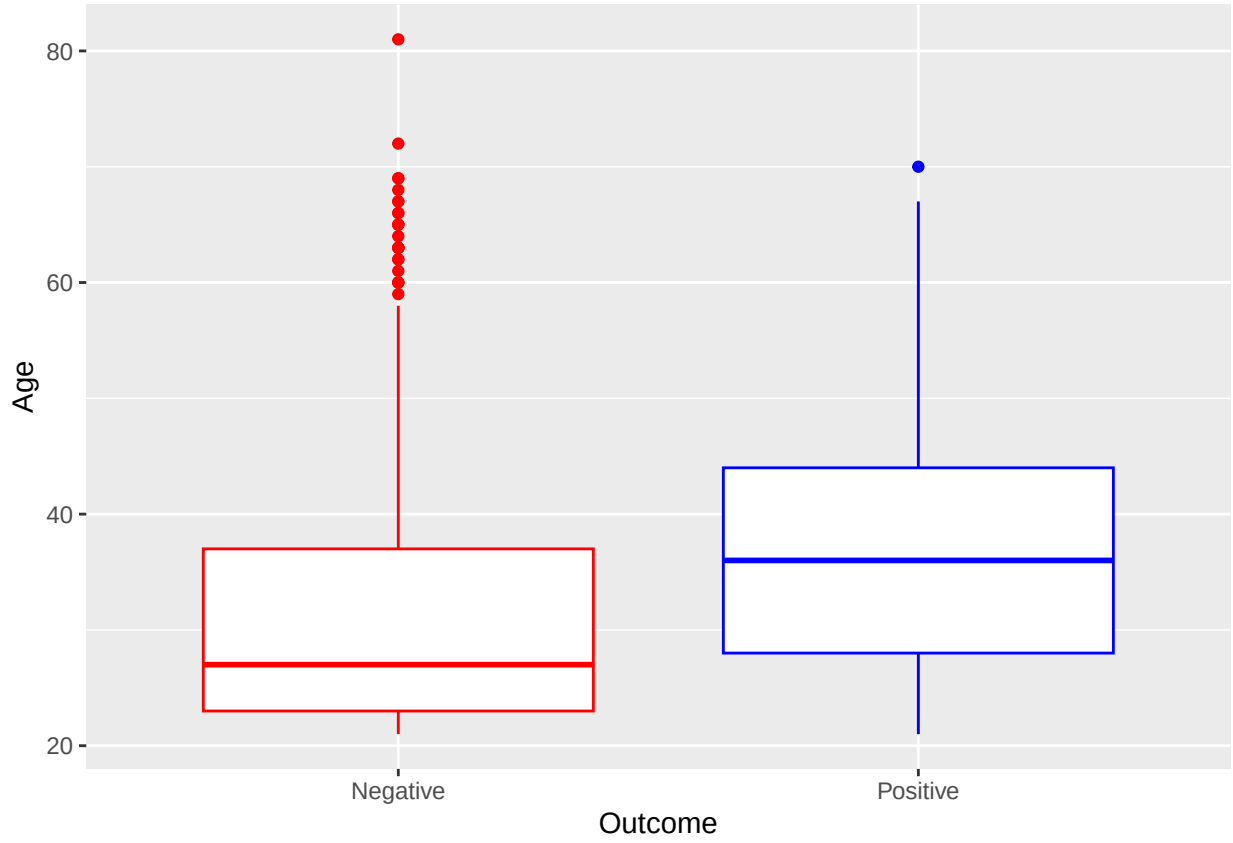
Age vs Outcome

```
diabetes_df %>%
  ggplot(aes(x = `Age`, fill = `Outcome`)) +
```

```
  geom_bar(
    stat = "count",
    position = "dodge",
    na.rm = TRUE
    ) +
  facet_wrap(~`Outcome`)
```




```
diabetes_df %>%
  ggplot(aes(x = `Outcome`, y = `Age`)) +
  geom_boxplot(
    stat = "boxplot",
    color = c("red","blue"),
    na.rm = TRUE
    )
```




```
diabetes_df %>%
  select(c(`Age`, `Outcome`)) %>%
  group_by(`Outcome`) %>%
  summarise(
    Age_mean = mean(`Age`),
    Age_median = median(`Age`),
    Age_Q1 = quantile(`Age`, probs = 0.25),
    Age_Q2 = quantile(`Age`, probs = 0.75)
    )
```

```
## # A tibble: 2 x 5
##   Outcome  Age_mean Age_median Age_Q1 Age_Q2
##   <chr>       <dbl>      <dbl>  <dbl>  <dbl>
## 1 Negative     31.2         27     23     37
## 2 Positive     37.1         36     28     44
```

Analysis: The distribution of `Age` to `Outcome` is right skewed. This tells us that the median and mean are not the same. Which is confirmed in the box plot and the stat table.

Correlation

Because `Glucose`, `BloodPressure`, and `BMI` have normal distribution, I want to see if there is a correlation between them, focusing on a **positive** diabetes `Outcome`.

```r
diabetes_cor <- diabetes_df %>%
  select(c(`Glucose`, `BloodPressure`, `BMI`, `Outcome`)) %>%
  filter(`Outcome` == "Positive")

Glucose_BloodPressure_cor <- cor(
  x = as.numeric(diabetes_cor$Glucose),
  y = as.numeric(diabetes_cor$BloodPressure),
  use = "na.or.complete",
  method = "pearson"
  )

Glucose_BMI_cor <- cor(
  x = as.numeric(diabetes_cor$Glucose),
  y = as.numeric(diabetes_cor$BMI),
  use = "na.or.complete",
  method = "pearson"
  )

BloodPressure_BMI_cor <- cor(
  x = as.numeric(diabetes_cor$BloodPressure),
  y = as.numeric(diabetes_cor$BMI),
  use = "na.or.complete",
  method = "pearson"
  )

correlation_df <- tibble(
  Relationship = c("Glucose vs Blood Pressure", "Glucose vs BMI",
                   "Blood Pressure vs BMI"),
  Correlation = c(Glucose_BloodPressure_cor, Glucose_BMI_cor,
                  BloodPressure_BMI_cor)
)

correlation_df
```

```
## # A tibble: 3 x 2
##   Relationship              Correlation
##   <chr>                           <dbl>
## 1 Glucose vs Blood Pressure      0.101
## 2 Glucose vs BMI                 0.0566
## 3 Blood Pressure vs BMI          0.249
```

Analysis: From the table we see that there is no correlation between these three variables.