Diabetes Dataset Statistical Analysis

Applied Analytics Assessment 2

Astha Bathla (S3999096) and Snigdha Mathur(S4017572)

Last updated: 02 June, 2024

Loading libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(ggplot2)
library(purrr)
```

```
##
## Attaching package: 'purrr'

## The following object is masked from 'package:magrittr':
##
##     set_names
```

Introduction

Diabetes, a chronic disease, affects numerous individuals and is occurred due to high levels of glucose in the blood. Understanding the factors that contribute to diabetes is crucial for prevention, diagnosis, and management. This report focuses on a comprehensive analysis of a diabetes dataset, which includes various attributes related to individuals' health and medical history.

The primary goal of this analysis is to answer specific problem statements related to the relationships between glucose, bloodpressure, BMI, insulin, age, diabetic predigree function and skin thickness levels, diabetes outcome, and the number of pregnancies. Through various statistical techniques, we aim to uncover patterns and significant predictors of diabetes within this dataset.

Problem Statement

In this presentation we will ty to answer the following problem statements.

- Does the number of pregnancies affect the risk of developing diabetes? We will check this association using categorical association.
- Does Glucose, Bloodpressure, BMI, Insulin, Age, Diabetic Predigree function and SkinThickness levels play a significant role in detecting the risk of diabetes? We will check this using logical regression.

Data

This dataset, obtained from the open source site Kaggle, is from the National Institute of Diabetes and Digestive and Kidney Diseases. The first objective is to anticipate the type of diabetes in a patient using the numerous diagnostic parameters provided. In particular, the data set includes obervations about female patients of Pima Indian hertiage of age 21 years or older.

URL of the dataset. https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset.

The dataset has 9 attributes. Information about dataset attributes -

Pregnancies: An ordinal variable, used to express the Number of pregnancies.

Glucose: To express the Glucose level in blood. It has a ratio scale.

BloodPressure: A ratio variable , utilized to express the Blood pressure measurement(mm Hg).

SkinThickness: A ratio scaled variable, provides information about the thickness of the skin (TSFT) (mm).

Insulin: It demonstrates the Insulin level in blood (µU/ml). It's a ratio scale.

BMI: Denotes the Body mass index. It has an interval scale.

DiabetesPedigreeFunction: Diabetes pedigree function (a function which measures likelihood of diabetes by judging family history). It's a ratio level.

Age: To express the age of the female. Age is measured on a ratio scale.

Outcome: To express the final result. Outcome has a nomial scale.

The data has 7 numeric values and non of these columns has missing values. Outcome and Pregnancies are categorical variable which is converted in factors. Later, we detect outliers in the dataset and remove them before performing our analysis.

```
diabetes_dataframe <-
  read.csv(
    "C:/Users/AsthaBathla/Desktop/RMIT/SEM1/Applied Analytics/week2/Data-Applied Analytics/diabetes.csv


diabetes_dataframe$Outcome <- factor(x = diabetes_dataframe$Outcome, levels = c(0,1),
                                     labels = c( "No", "Yes"))

diabetes_dataframe$Pregnancies <- factor(x = diabetes_dataframe$Pregnancies,
                                     levels = sort(unique(diabetes_dataframe$Pregnancies)))

#Checking str of data
summary(diabetes_dataframe)
```

```
##   Pregnancies     Glucose      BloodPressure    SkinThickness      Insulin
##  1      :135   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0
##  0      :111   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0
##  2      :103   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5
##  3      : 75   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8
##  4      : 68   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2
##  5      : 57   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0
##  (Other):219
##       BMI        DiabetesPedigreeFunction      Age        Outcome
##  Min.   : 0.00   Min.   :0.0780           Min.   :21.00   No :500
##  1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00   Yes:268
##  Median :32.00   Median :0.3725           Median :29.00
##  Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :67.10   Max.   :2.4200           Max.   :81.00
##
```

```
#Checking the missing values
Missing_Values_diabetes <- sapply(diabetes_dataframe, function(x) sum(is.na(x)))
Missing_Values_diabetes
```

```
##              Pregnancies                  Glucose            BloodPressure
##                        0                        0                        0
##            SkinThickness                  Insulin                      BMI
##                        0                        0                        0
## DiabetesPedigreeFunction                      Age                  Outcome
##                        0                        0                        0
```

Descriptive Statistics and Visualisation

The key variables inspected in this study are glucose, blood pressure, BMI, insulin, age, skin thickness, outcome, and pregnancies. Among these, outcome and pregnancies are qualitative variables, for which we use summary functions and bar plots to demonstrate descriptive statistics. For the remaining variables, we use boxplots or histograms along with summary functions to produce their descriptive statistical analysis.

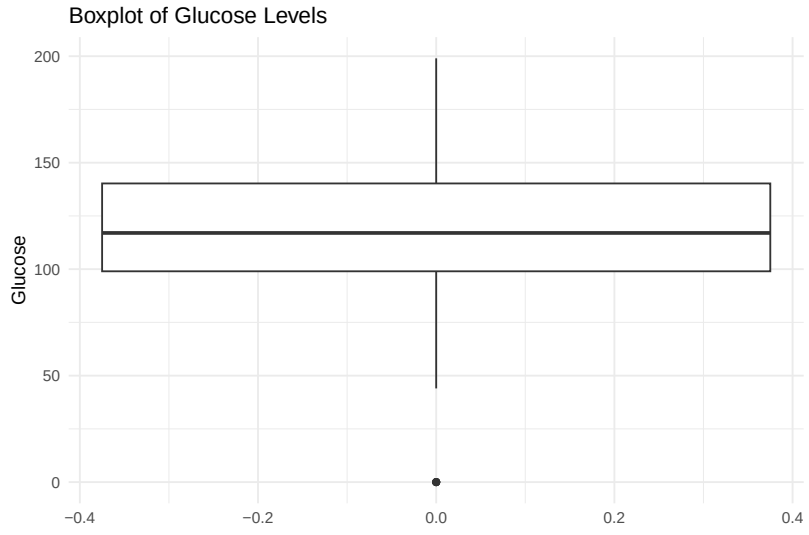
```r
# Analysis of glucose using summarise function.

diabetes_dataframe %>% summarise(Min = min(Glucose),
                First_quantile = quantile(Glucose, 0.25),
                Median = median(Glucose),
                Mean = mean(Glucose),
                Third_quantile = quantile(Glucose,0.75),
                Max = max(Glucose),
                Min = min(Glucose),
                Range = Max - Min,
                Standard_Deviation = sd(Glucose),
                n = n(),
                Mode = mode(Glucose),
                Missing_Values = sum(is.na(Glucose)))
```

```
##   Min First_quantile Median     Mean Third_quantile Max Range
## 1   0             99    117 120.8945         140.25 199   199
##   Standard_Deviation   n    Mode Missing_Values
## 1           31.97262 768 numeric              0
```

```r
# Boxplot of GLucose levels
glucose_boxplot <- ggplot(diabetes_dataframe, aes(y = Glucose)) +
  geom_boxplot() +
  labs(title = "Boxplot of Glucose Levels",
       y = "Glucose",
       x = "") +
  theme_minimal()

glucose_boxplot
```



The glucose level boxplot summarizes the distribution of glucose measurements. The median glucose level is 117. The interquartile range (IQR), from 99 to 140, contains the middle 50% of values. The same can be seen through the results of summarize function of dplyr package. Whiskers expand within the smallest and largest values inside 1.5 times the IQR, capturing most data points. An outlier below 50 indicates an

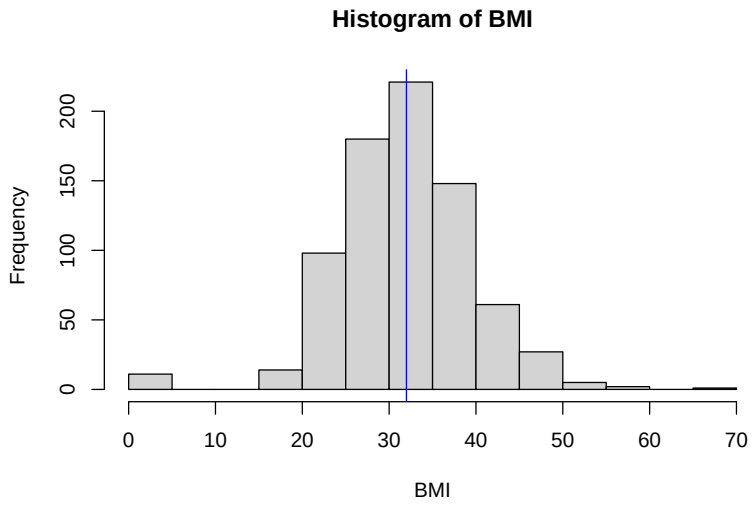
unusually low value. The range of this variable is 0 to 199. Overall, the distribution is relatively symmetric with a central tendency around 110 and a moderate spread.

```
# Summarise function for descriptive analysis

diabetes_dataframe %>% summarise(Min = min(BMI),
                First_quantile = quantile(BMI, 0.25),
                Median = median(BMI),
                Mean = mean(BMI),
                Third_quantile = quantile(BMI,0.75),
                Max = max(BMI),
                Min = min(BMI),
                Range = Max - Min,
                Standard_Deviation = sd(BMI),
                n = n(),
                Mode = mode(BMI),
                Missing_Values = sum(is.na(BMI)))
```

```
##   Min First_quantile Median     Mean Third_quantile  Max Range
## 1   0           27.3     32 31.99258           36.6 67.1  67.1
##   Standard_Deviation   n    Mode Missing_Values
## 1            7.88416 768 numeric              0
```

```
# BMI Histogram
BMI_hist <- hist(x= diabetes_dataframe$BMI, col="lightgrey",
                         xlab="BMI",
                         main="Histogram of BMI")
abline(v = mean(diabetes_dataframe$BMI, na.rm = TRUE), col="blue")
```



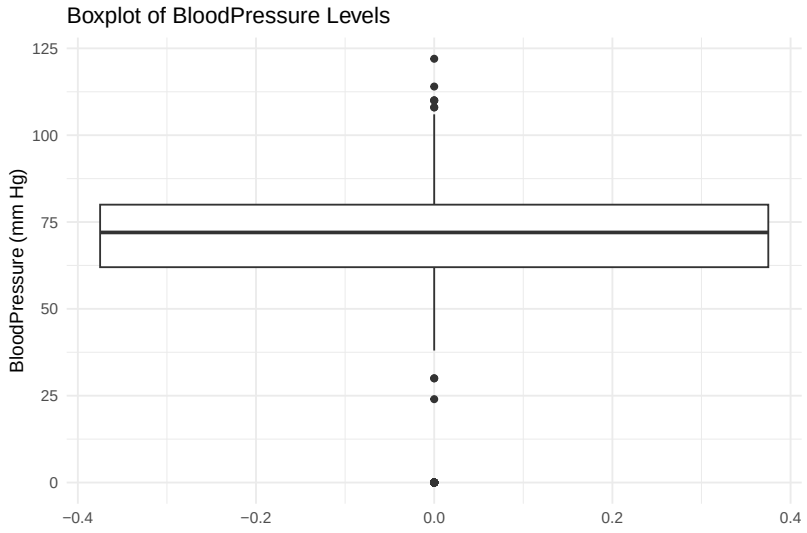
The histogram of BMI (Body Mass Index) demonstrates the distribution of BMI values within the data. The distribution is almost normal, peaking near the center and tapering off at both the ends. This depicts that most individuals have BMI values near to the central range, with fewer females at the extremes. The summary function reveals that the mean BMI is 31.99 and the median is 32, pointing that they are almost identical. The BMI values range from a minimum of 0 to a maximum of 67.1, with a standard deviation of 7.89.

```r
diabetes_dataframe %>% summarise(Min = min(BloodPressure),
                First_quantile = quantile(BloodPressure, 0.25),
                Median = median(BloodPressure),
                Mean = mean(BloodPressure),
                Third_quantile = quantile(BloodPressure,0.75),
                Max = max(BloodPressure),
                Min = min(BloodPressure),
                Range = Max - Min,
                Standard_Deviation = sd(BloodPressure),
                n = n(),
                Mode = mode(BloodPressure),
                Missing_Values = sum(is.na(BloodPressure)))
```

```
##   Min First_quantile Median     Mean Third_quantile Max Range
## 1   0             62     72 69.10547             80 122   122
##   Standard_Deviation   n    Mode Missing_Values
## 1           19.35581 768 numeric              0
```

```r
#  BloodPressure Boxplot
BloodPressure_boxplot <- ggplot(diabetes_dataframe, aes(y = BloodPressure)) +
  geom_boxplot() +
  labs(title = "Boxplot of BloodPressure Levels",
       y = "BloodPressure (mm Hg)",
       x = "") +
  theme_minimal()

BloodPressure_boxplot
```

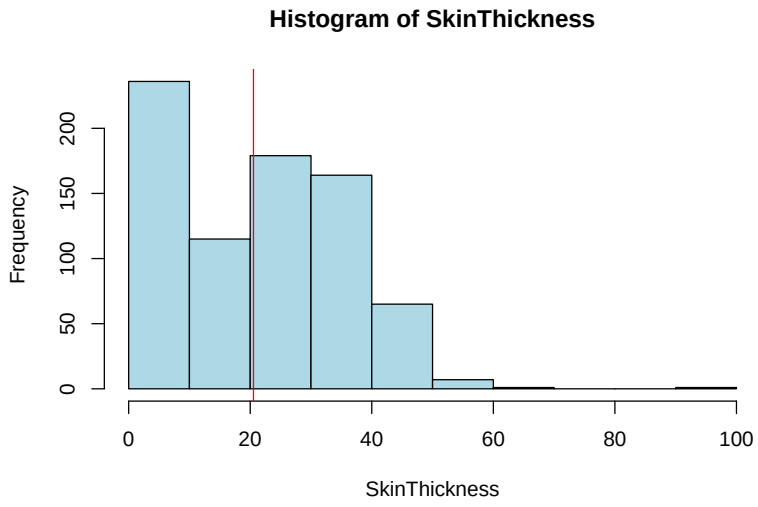


Blood pressure values range from 0 to 122, with a mean of 69.1 (mm Hg) and a standard deviation of 19.36. The boxplot is relatively symmetric, denoting a balanced distribution around the median blood pressure level of roughly 72 (mm Hg). The interquartile range (IQR) spans from about 62.5 (mm Hg) to 82.5 (mm Hg), showing moderate variability in the middle 50% of the data. Several outliers prevail, with values significantly greater or lesser than the majority, possibly evidenting measurement errors or females with a typical blood pressure levels. In comprehension, the distribution is nearly symmetric, suggesting that blood pressure levels are evenly distributed around the median.

```
diabetes_dataframe %>% summarise(Min = min(SkinThickness),
                First_quantile = quantile(SkinThickness, 0.25),
                Median = median(SkinThickness),
                Mean = mean(SkinThickness),
                Third_quantile = quantile(SkinThickness,0.75),
                Max = max(SkinThickness),
                Min = min(SkinThickness),
                Range = Max - Min,
                Standard_Deviation = sd(SkinThickness),
                n = n(),
                Mode = mode(SkinThickness),
                Missing_Values = sum(is.na(SkinThickness)))
```

```
##   Min First_quantile Median     Mean Third_quantile Max Range
## 1   0              0     23 20.53646             32  99    99
##   Standard_Deviation   n    Mode Missing_Values
## 1           15.95222 768 numeric              0
```

```
# Skin Thickness Histogram
diabetes_dataframe$SkinThickness %>%  hist(col="lightblue",
                          xlab="SkinThickness",
                          main="Histogram of SkinThickness")
abline(v = mean(diabetes_dataframe$SkinThickness, na.rm = TRUE), col="red")
```



The SkinThickness variable ranges from 0 to 99, with a mean of 20.54 and a median of 23. The median being larger than the mean indicates that most of thevalues lie on the lower end. The histogram visualizes the distribution, showing that SkinThickness values are right-skewed, meaning they are concentrated on the lower end with fewer higher values. This suggests that most individuals have lower SkinThickness values, but a few have much higher values Summarizefunction depicts that the standard deviation of 15.95 indicates that SkinThickness values typically vary by about 15.95 units from the mean.

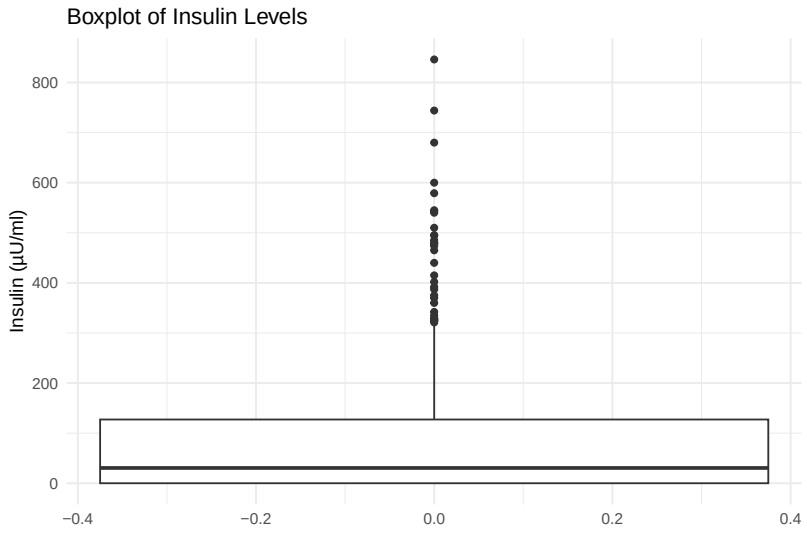
```
diabetes_dataframe %>% summarise(Min = min(Insulin),
                First_quantile = quantile(Insulin, 0.25),
                Median = median(Insulin),
```

```
                Mean = mean(Insulin),
                Third_quantile = quantile(Insulin,0.75),
                Max = max(Insulin),
                Min = min(Insulin),
                Range = Max - Min,
                Standard_Deviation = sd(Insulin),
                n = n(),
                Mode = mode(Insulin),
                Missing_Values = sum(is.na(Insulin)))
```

```
##   Min First_quantile Median     Mean Third_quantile Max Range
## 1   0              0   30.5 79.79948         127.25 846   846
##   Standard_Deviation   n    Mode Missing_Values
## 1            115.244 768 numeric              0
```

```
# Insulin Boxplot
Insulin_boxplot <- ggplot(diabetes_dataframe, aes(y = Insulin)) +
  geom_boxplot() +
  labs(title = "Boxplot of Insulin Levels",
       y = "Insulin (µU/ml)",
       x = "") +
  theme_minimal()

Insulin_boxplot
```

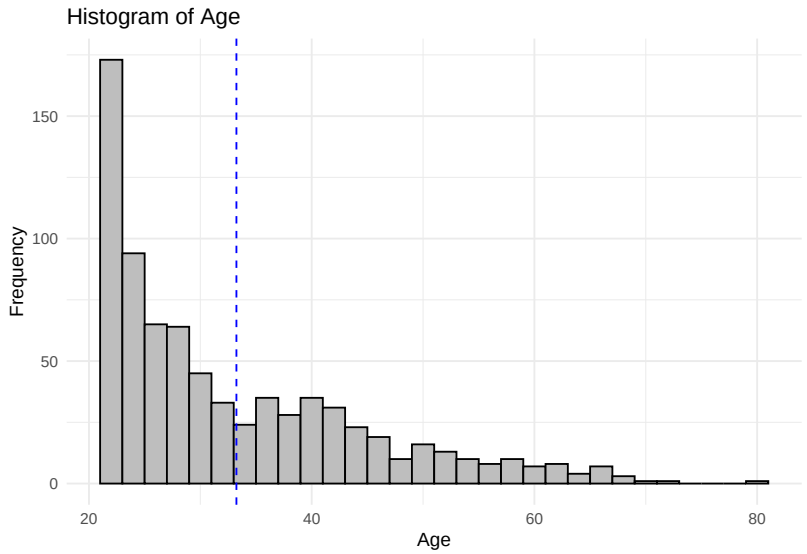


The boxplot of insulin levels denotes a highly right-skewed distribution with a number of outliers. The median insulin level, demonstrated by the line inside the box, is relatively low, indicating that half of the females have insulin levels less than this value. The interquartile range (IQR), representing the middle 50% of the data, is also low, from a range of 0 to 127.25, showing that the most insulin values are clustered toward the lower end. The whiskers extend to the highest values within 1.5 times the IQR from the lower and upper quartiles, capturing the majority of the data. However, the numerous points above the whiskers indicate a significant number of high outliers, suggesting that some individuals have substantially higher insulin levels compared to the rest of the dataset.

```r
diabetes_dataframe %>% summarise(Min = min(Age),
                 First_quantile = quantile(Age, 0.25),
                 Median = median(Age),
                 Mean = mean(Age),
                 Third_quantile = quantile(Age,0.75),
                 Max = max(Age),
                 Min = min(Age),
                 Range = Max - Min,
                 Standard_Deviation = sd(Age),
                 n = n(),
                 Mode = mode(Age),
                 Missing_Values = sum(is.na(Age)))
```

```
##   Min First_quantile Median     Mean Third_quantile Max Range
## 1  21             24     29 33.24089             41  81    60
##   Standard_Deviation   n    Mode Missing_Values
## 1           11.76023 768 numeric              0
```

```r
# Age histogram
ggplot(diabetes_dataframe, aes(x = Age)) +
geom_histogram(binwidth = 2, fill = "grey", color = "black") +
geom_vline(aes(xintercept = mean(Age)), color = "blue", linetype = "dashed") +
labs(title = "Histogram of Age", x = "Age", y = "Frequency") +
theme_minimal()
```



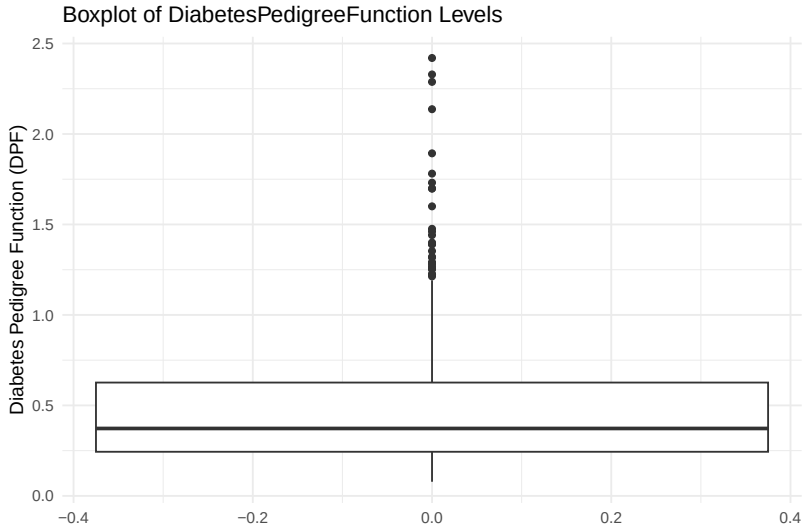
The histogram of age illustrates a right-skewed distribution of ages in the dataset. Most of individuals are in the younger age range, with a peak around the 20 to 25 age group, where the highest frequency is observed. As age increases, the frequency of individuals decreases, indicating fewer older individuals in the dataset. The vertical dashed blue line represents the mean age, which is around 30. This visualization highlights that while the data is primarily made of younger individuals, there is a smaller but significant observations of individuals of up to 70 years of age. This skewed distribution depicts that the population being analysed has a higher concentration of younger individuals.

```r
diabetes_dataframe %>% summarise(Min = min(DiabetesPedigreeFunction),
                 First_quantile = quantile(DiabetesPedigreeFunction, 0.25),
                 Median = median(DiabetesPedigreeFunction),
                 Mean = mean(DiabetesPedigreeFunction),
                 Third_quantile = quantile(DiabetesPedigreeFunction,0.75),
                 Max = max(DiabetesPedigreeFunction),
                 Min = min(DiabetesPedigreeFunction),
                 Range = Max - Min,
                 Standard_Deviation = sd(DiabetesPedigreeFunction),
                 n = n(),
                 Mode = mode(DiabetesPedigreeFunction),
                 Missing_Values = sum(is.na(DiabetesPedigreeFunction)))
```

```
##     Min First_quantile Median      Mean Third_quantile  Max Range
## 1 0.078        0.24375 0.3725 0.4718763        0.62625 2.42 2.342
##   Standard_Deviation   n    Mode Missing_Values
## 1          0.3313286 768 numeric              0
```

```r
# Checking outliers for DiabetesPedigreeFunction using boxplot
DiabetesPedigreeFunction_boxplot <-
  ggplot(diabetes_dataframe, aes(y = DiabetesPedigreeFunction)) +
  geom_boxplot() +
  labs(title = "Boxplot of DiabetesPedigreeFunction Levels",
       y = "Diabetes Pedigree Function (DPF)",
       x = "") +
  theme_minimal()

DiabetesPedigreeFunction_boxplot
```



The boxplot of diabetes pedigree function levels shows a right-skewed distribution with numerous outliers. The median value is around 0.3, and the interquartile range (IQR) ranges from approximately 0.2 to 0.6, depicting that the middle 50% of scores lie within this range. The whiskers capture the majority of the data, but numerous outliers above the upper whisker indicate higher values. This suggests significant variability and a highlighted presence of females with a strong genetic predisposition to diabetes.

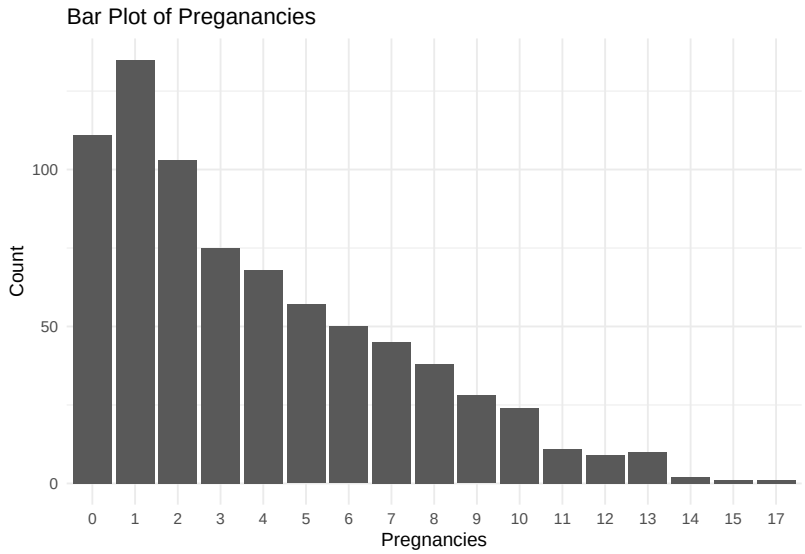
```r
summary(diabetes_dataframe$Pregnancies)
```

```
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  17
## 111 135 103  75  68  57  50  45  38  28  24  11   9  10   2   1   1
```

```r
#barplot for Preganancies

barplot_outcome <- ggplot(diabetes_dataframe, aes(x = factor(Pregnancies))) + geom_bar()+
  labs(title = "Bar Plot of Preganancies",
       x = "Pregnancies",
       y = "Count") +
  theme_minimal()

barplot_outcome
```



The bar plot illustrates the distribution of the number of pregnancies among the females in the dataset. The most frequent number of pregnancies is 1, with over 120 occurrences. The number of individuals decreases as the number of pregnancies increases. This decreasing trend is steep initially and becomes more gradual as the number of pregnancies increases. The plot includes a few outliers with very high numbers of pregnancies (e.g., 15 and 17), indicating rare cases in the dataset.The distribution is right-skewed, meaning that there are more individuals with a lower number of pregnancies, and fewer individuals with a high number of pregnancies.A total of 111 females haven't undergone a reporduction process.

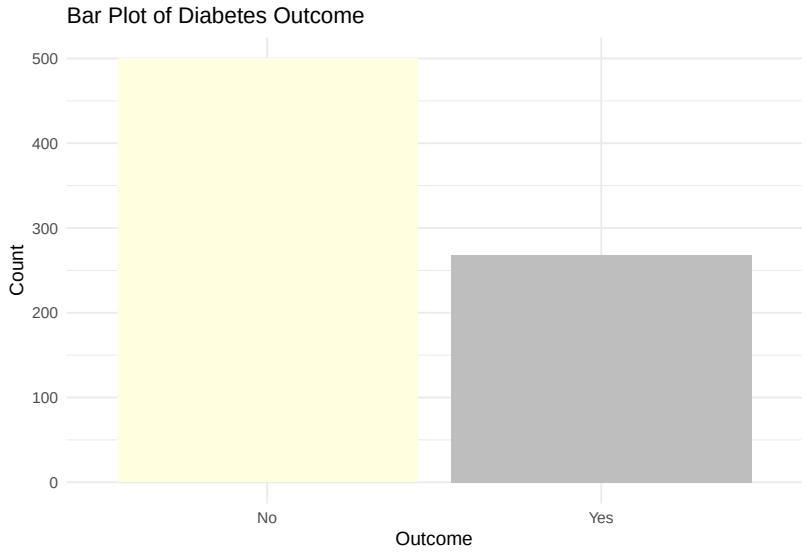
```r
summary(diabetes_dataframe$Outcome)
```

```
##  No Yes
## 500 268
```

```r
# Bar plot for outcome
barplot_outcome <- ggplot(diabetes_dataframe, aes(x = factor(Outcome))) +
  geom_bar(fill = c("lightyellow", "grey")) +
  labs(title = "Bar Plot of Diabetes Outcome",
       x = "Outcome",
```

```
       y = "Count") +
  scale_x_discrete(labels = c("0" = "Non-Diabetic", "1" = "Diabetic")) +
  theme_minimal()


print(barplot_outcome)
```



The bar plot represents the distribution of diabetes outcomes (Yes or No) among the individuals in the dataset. The bar labelled "No" is significantly taller, with a count of approximately 500 individuals, i.e, the dataset has more cases of non-diabetic females. The other bar is shorter, with a count of around 268 diabetic individuals.

Outliers Detection and Removal

In this section, we indentify the outiers in the numeric columns (new data frame created while excluding pregnancies and outcome column) by calculating values outside the lower bound and upper bound of the each numeric column. Lower bound is calculated by subtracting 1.5 times of IQR from first quantile, on the other hand upper bound is calculated using by adding 1.5 timings of IQR in third quantile. We remove the outliers by capping method. It is performed by replacing all values below lower_bound by l_bound and all values above u_bound are replaced by u_bound. After performing capping we produced a new dataset, i.e, diabetes_df_capped.

```
# Defining the function to calculate outliers
outliers_calculation_function <- function(m, column) {
  Quantitle3 <- quantile(m[, column], 0.75, na.rm = TRUE)
  Quantitle1 <- quantile(m[, column], 0.25, na.rm = TRUE)
  IQR <- Quantitle3 - Quantitle1
  u_bound <- Quantitle3 + 1.5 * IQR
  l_bound <- Quantitle1 - 1.5 * IQR
  outliers_in_data <- m[, column] < l_bound | m[, column] > u_bound
  list(l_bound = l_bound, u_bound = u_bound,
       outliers = sum(outliers_in_data))
}
```

```r
diabetes_df_numeric <- diabetes_dataframe[, !names(diabetes_dataframe) %in%
                                    c("Outcome", "Pregnancies")]


# Assuming diabetes_df_numeric is your dataframe with numeric columns

# Applying the function to each column
outliers__summary <-
  lapply(names(diabetes_df_numeric),
         function(col) outliers_calculation_function(diabetes_df_numeric, col))

# Assigning column names to the results
names(outliers__summary) <- names(diabetes_df_numeric)

outliers_tb <- do.call(rbind, lapply(outliers__summary, as.data.frame))
outliers_tb <- data.frame(Column = names(outliers__summary), outliers_tb)

# Displaying the resulting table
print(outliers_tb)
```

```
##                                          Column  l_bound u_bound outliers
## Glucose                                 Glucose   37.125 202.125        5
## BloodPressure                     BloodPressure   35.000 107.000       45
## SkinThickness                     SkinThickness  -48.000  80.000        1
## Insulin                                 Insulin -190.875 318.125       34
## BMI                                         BMI   13.350  50.550       19
## DiabetesPedigreeFunction DiabetesPedigreeFunction   -0.330   1.200       29
## Age                                         Age   -1.500  66.500        9
```

```r
# Removing Outliers
cap_outliers <- function(z, outlier_bounds) {
  for (col in names(outlier_bounds)) {
    bounds <- outlier_bounds[[col]]
    z[, col] <- ifelse(z[, col] < bounds$l_bound,
                       bounds$l_bound, z[, col])
    z[, col] <- ifelse(z[, col] > bounds$u_bound,
                       bounds$u_bound, z[, col])
  }
  return(z)
}

# Applying the function to cap outliers
diabetes_df_capped <- cap_outliers(diabetes_dataframe, outliers__summary)

# Displaying the capped dataset
head(diabetes_df_capped)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
```

```
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50     Yes
## 2                    0.351  31      No
## 3                    0.672  32     Yes
## 4                    0.167  21      No
## 5                    1.200  33     Yes
## 6                    0.201  30      No
```

Categorical Association

To analyze the categorical association between the number of pregnancies and the diabetes outcome, we implement a Chi-squared test of independence. This test evaluates whether there is a significant relationship between two categorical variables.

For this analysis, we use the chisq.test(contingency_table) function, which executes the Chi-squared test of independence on a contingency table. In this circumstance, the contingency table is generated using table(diabetes_df_capped$Pregnancies, diabetes_df_capped$Outcome).

The test determines whether the distribution of Outcome is independent of the number of pregnancies.

The Chi-Squared statistic is calculated using the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where: - O_i is the observed frequency for the i-th cell of the contingency table. - E_i is the expected frequency for the i-th cell, calculated as $E_i = \frac{(row\ total \times column\ total)}{grand\ total}$.

The statistical hypotheses for this Chi-square test of association can be written as follows:

H_0 :: *There is no association in the population between number of pregnancies and outcome*(*independence*)

H_A :: *There is association in the population between number of pregnancies and outcome*(*dependence*)

```
Pregnancy_ca <- chisq.test(table( diabetes_df_capped$Pregnancies, diabetes_df_capped$Outcome ))
```

```
## Warning in chisq.test(table(diabetes_df_capped$Pregnancies,
## diabetes_df_capped$Outcome)): Chi-squared approximation may be incorrect
```

```
Pregnancy_ca
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(diabetes_df_capped$Pregnancies, diabetes_df_capped$Outcome)
## X-squared = 64.595, df = 16, p-value = 8.648e-08
```

```
Pregnancy_ca$observed
```

```
##
##       No Yes
##   0   73  38
##   1  106  29
##   2   84  19
##   3   48  27
##   4   45  23
##   5   36  21
##   6   34  16
##   7   20  25
##   8   16  22
##   9   10  18
##   10  14  10
##   11   4   7
##   12   5   4
##   13   5   5
##   14   0   2
##   15   0   1
##   17   0   1
```

```
Pregnancy_ca$expected
```

```
##
##              No        Yes
##   0  72.2656250 38.7343750
##   1  87.8906250 47.1093750
##   2  67.0572917 35.9427083
##   3  48.8281250 26.1718750
##   4  44.2708333 23.7291667
##   5  37.1093750 19.8906250
##   6  32.5520833 17.4479167
##   7  29.2968750 15.7031250
##   8  24.7395833 13.2604167
##   9  18.2291667  9.7708333
##   10 15.6250000  8.3750000
##   11  7.1614583  3.8385417
##   12  5.8593750  3.1406250
##   13  6.5104167  3.4895833
##   14  1.3020833  0.6979167
##   15  0.6510417  0.3489583
##   17  0.6510417  0.3489583
```

```
Pregnancy_ca$p.value
```

```
## [1] 8.648349e-08
```

Interpretation of Analysis

Contingency Table calculates the frequency of diabetic (1) and non-diabetic (0) females for each number of pregnancies. The Chi-squared test statistic(Xsquared) is equivalent to 64.595. It quantifies how much the

observed frequencies deviate from the expected frequencies. Degrees of freedom is equivalent to 16. It is studied using the formula (number of rows - 1) * (number of columns - 1). In this case it is, df = (17 - 1) * (2 - 1) = 16. The p-value for the test is 8.648349e-08. Since, 8.648349e-08 < 0.05 the test provides strong evidence against the null hypothesis, suggesting that the number of pregnancies and the diabetes outcome depend on each other and are indeed associated.

Logical Regression

Logistic regression, a statistical analysis method, is utilized for binary categorization. It models the correlation between a binary attribute which is dependent (outcome) and more than or equal to 1 independent attributes (predictors). The output of logistic regression is the probability that a given instance belongs to a particular class.

For binary outcomes, the logistic function is used to map predicted values to probabilities. It is calculated by using the following formula:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 \cdot \text{Glucose}$$

where: - p is the probability of having diabetes. - β_0 is the intercept. - β_1 is the coefficient for glucose. - X_1 is the glucose level of an individual.

Analysis in R a logistic regression model was performed by fitting using Glucose, BloodPressure, SkinThickness, Insulin, BMI, Age as the predictor for the binary outcome variable.

model: glm(formula = BloodPressure + Insulin + SkinThickness + Glucose + BMI + DiabetesPedigreeFunction + Age , family = binomial, data = diabetes_df_capped, family = binomial, data = diabetes_df_capped)

For this logistical regression analysis, null hypothesis can be stated as:

$$H_0 :: \textit{There are none predicators that have an effect on the outcome variable } (\textit{independence})$$

and alternative hypothesis can be stated as:

$$H_A :: \textit{There are atleast one of the predicators that has an effect on the outcome variable } (\textit{dependence})$$

```
l_model_diabetes <- glm(Outcome ~  BloodPressure + Insulin + SkinThickness
                        + Glucose + BMI + DiabetesPedigreeFunction + Age ,
                        family = binomial, data = diabetes_df_capped)

# Display model summary
summary(l_model_diabetes)
```

```
##
## Call:
## glm(formula = Outcome ~ BloodPressure + Insulin + SkinThickness +
##     Glucose + BMI + DiabetesPedigreeFunction + Age, family = binomial,
##     data = diabetes_df_capped)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)              -8.789623   0.760617 -11.556  < 2e-16 ***
## BloodPressure            -0.014263   0.007149  -1.995 0.046049 *
## Insulin                  -0.001653   0.001180  -1.401 0.161202
## SkinThickness            -0.001085   0.007183  -0.151 0.879913
## Glucose                   0.035546   0.003706   9.592  < 2e-16 ***
## BMI                       0.097715   0.015829   6.173 6.70e-10 ***
## DiabetesPedigreeFunction  1.159297   0.326434   3.551 0.000383 ***
## Age                       0.033368   0.008429   3.959 7.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 728.54  on 760  degrees of freedom
## AIC: 744.54
##
## Number of Fisher Scoring iterations: 5
```

```r
# Calculate odds ratio
exp(coef(l_model_diabetes))
```

```
##              (Intercept)            BloodPressure                  Insulin
##             0.0001523054             0.9858384165             0.9983482674
##            SkinThickness                  Glucose                      BMI
##             0.9989153695             1.0361854614             1.1026482016
## DiabetesPedigreeFunction                      Age
##             3.1876906228             1.0339313152
```

```r
# Create prediction data
diabetes_df_capped$predicted_prob <- predict(l_model_diabetes, type = "response")
```

Interpretation of Analysis

1. *GLUCOSE:* The logistic regression model indicates a significant positive association between glucose levels and the likelihood of having diabetes. The estimated coefficient for glucose is 0.035546 (SE = 0.003706, z = 9.592, p < 2e-16). The odds ratio for glucose is 1.0362, indicating that for each one-unit increase in glucose level, the chance of having diabetes increase by approximately 3.62%. Results suggest that females with higher glucose levels are more vulnerable to diabetes (odds percentage formula is mentioned at the end of this section) .

2. *BLOOD PRESSURE:* The logistic regression model specifies a remarkable negative association between blood pressure and the risk of acquiring diabetes. The estimated coefficient for blood pressure is -0.014263 (SE = 0.007149, z = -1.995, p = 0.046049). The odds ratio for blood pressure is 0.9858, indicating that for each one-unit increase in blood pressure, the likelihood of having diabetes decrease by approximately 1.42%. This might indicate that in case of diabetes the blood pressure of the paitent drops.

3. *SKIN THICKNESS:* The logistic regression model reveals no significant association between skin thickness and the chances of contracting diabetes. The estimated coefficient for skin thickness is -0.001085 (SE = 0.007183, z = -0.151, p = 0.879913). P-value is greater than 0.05. The odds ratio for skin thickness is 0.9989, indicating a negligible effect on the potential of having diabetes.

4. *INSULIN:* The logistic regression model indicates no significant association between insulin levels and the potential for having diabetes. The estimated coefficient for insulin is -0.001653 (SE = 0.001180, z = -1.401, p = 0.161202). P-value is equivalent to 0.161 > 0.05. The odds ratio for insulin is 0.9983, indicating that for each one-unit increase in insulin, the odds of having diabetes decrease by approximately 0.17%.

5. *BMI:* The logistic regression model stipulates a notable positive association between BMI and the likelihood of having diabetes. The estimated coefficient for BMI is 0.097715 (SE = 0.015829, z = 6.173, p = 6.70e-10). The odds ratio for BMI is 1.1026, indicating that for each one-unit increase in BMI, the chances of having diabetes rises by by approximately 10.26%.

6. *AGE:* The logistic regression model evaluates a significant positive association between age and the possibility of getting diabetes. The estimated coefficient for age is 0.033368 (SE = 0.008429, z = 3.959, p = 7.53e-05). The odds ratio for age is 1.0339, indicating that for each one-year increase in age, the possibilty of having diabetes increase by approximately 3.39%. Hence, this implies that older people are more prone to diabetes than younger people.

7. *DIABETES PEDIGREE FUNCTION:* The logistic regression model indicates an impressive positive association between the diabetes pedigree function and the risk of experiencing diabetes. The estimated coefficient for diabetes pedigree function is 1.159297 (SE = 0.326434, z = 3.551, p = 0.000383). The odds ratio for diabetes pedigree function is 3.1877, indicating that with every one-unit rise in the diabetes pedigree function, the chances of having diabetes increases by about 218.77%. This suggests that people with family history of diabetes, are more suspectible to suffer from diabetes.

The model's AIC (Akaike Information Criterion) is 741.45, suggesting a good fit. AIC is a measure used in statistical modeling to compare different models and assess the quality of each model relative to the others. A lower AIC indicates a better-fitting model. Odds calculation is as follows:

Odds Ratio:

$$e^b$$

,

Odds Percentage Change:

$$(e^b - 1) * 100$$

where: - b is the estimated coefficent.

Discussion

Major Findings:The logistic regression analysis reveals significant associations between several predictors and the likelihood of having diabetes. Specifically, glucose levels, blood pressure, BMI, age, and diabetes pedigree function are statistically significant predictors of diabetes risk. For one unit increase in glucose level, BMI, age and DPF, risk of diabetes increases by 3.62%, 10.26%, 3.39% and 218.77% respectively, whereas for one unit increase in blood pressure, the potential for of having diabetes decreases by 1.42% approximately. Overall, females with a family history or with higher age, glucose levels or BMI levels, are more suspectible to diabetes and diabetic piatents have low blood pressure, insulin levels and skin thickness.

The categorical analysis through a Chi-squared test of association between the number of pregnancies and diabetes outcome also shows Chi-squared statistic is 64.595 with 16 degrees of freedom and a p-value of indicating a strong association between the number of pregnancies and diabetes outcome. This suggests that the number of pregnancies is significantly associated with the likelihood of having diabetes. In total, the risk of diabetes is higher for women with repeative reproductive history.

Strengths: The analysis includes both categorical and numeric predictors, providing a holistic view of the factors influencing diabetes risk.Multiple predictors show strong statistical significance, reinforcing the robustness of the findings.

Limitations: The findings are based on the specific dataset used and may not generalize to other populations or datasets. The dataset is limited to female and does not include data for male.

Directions for future investigations: Future studies can include additional predictors such as physical activity and dietary habits to provide a more comprehensive assessment of diabetes risk as well as can include a genderwise distribution of data. We can include time factor, as in change in different predicators and how it increased the risk of diabetes.Exploring interaction effects between different predictors.

The combined analysis underlines the importance of monitoring glucose levels, BMI, age, and diabetes pedigree function in predicting and managing diabetes risk. Additionally, the significant association between the number of pregnancies and diabetes outcome highlights the need to consider reproductive history in diabetes risk assessments.

Higher glucose levels, BMI, number of pregnancies, and diabetes pedigree function significantly increase the risk of diabetes.

References

Akshay Dattatray Khare(2022) *Diabetes Dataset*, Kaggle , accessed on 22nd May 2024, https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data

James Baglin (2016), *Module 1 - Statistics - Dealing Confidently with Uncertainty* , Applied Analytics, accessed on 22nd May 2024, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_02.html

James Baglin (2016), *Module 2 - Descriptive Statistics through Visualisation* , Applied Analytics, accessed on 22nd May 2024, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_02.html

James Baglin (2016), *Module 8 - Categorical Associations* , Applied Analytics, accessed on 25th May 2024, https://astral-theory-157510.appspot.com/secured/MATH1324_Module_08.html

Michy Alice (2015), *How to perform a Logistic Regression in R*, R - Bloggers, accessed on 25th May 2024,https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/

Data Preprocessing (Data Wrangling) *Module 5 - Scan: Missing Values* accessed on 23rd May 2024, http://rare-phoenix-161610.appspot.com/secured/Module_05.html#Overview

Data Preprocessing (Data Wrangling) *Module 6 - Scan: Outliers* accessed on 23rd May 2024, http://rare-phoenix-161610.appspot.com/secured/Module_06.html#Overview

R Bloggers (nd) *How to Remove Outliers in R* R Bloggers, accessed on 22nd May 2024, https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/


```
c("magrittr", "MASS", "ggplot2", "dplyr", "lubridate", "purrr") %>%
  map(citation) %>%
  print(style = "text")
```

```
## [[1]]
## Bache S, Wickham H (2022). _magrittr: A Forward-Pipe Operator for R_. R
## package version 2.0.3, <https://CRAN.R-project.org/package=magrittr>.
##
## [[2]]
## Venables WN, Ripley BD (2002). _Modern Applied Statistics with S_,
## Fourth edition. Springer, New York. ISBN 0-387-95457-0,
## <https://www.stats.ox.ac.uk/pub/MASS4/>.
##
## [[3]]
```

Wickham H (2016). _ggplot2: Elegant Graphics for Data Analysis_.
Springer-Verlag New York. ISBN 978-3-319-24277-4,
<https://ggplot2.tidyverse.org>.
##
[[4]]
Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A
Grammar of Data Manipulation_. R package version 1.1.4,
<https://CRAN.R-project.org/package=dplyr>.
##
[[5]]
Grolemund G, Wickham H (2011). "Dates and Times Made Easy with
lubridate." _Journal of Statistical Software_, *40*(3), 1-25.
<https://www.jstatsoft.org/v40/i03/>.
##
[[6]]
Wickham H, Henry L (2023). _purrr: Functional Programming Tools_. R
package version 1.0.2, <https://CRAN.R-project.org/package=purrr>.