Assignment 3: Methods and Plan

Marcela Flaherty 97930697

Data:

Source: https://www.kaggle.com/datasets/whenamancodes/predict-diabities?resource=download

The Predict Diabetes dataset, found on Kaggle, originates from a larger database by the National Institute of Diabetes and Digestive and Kidney Diseases with the overarching aim of predicting diabetes. The dataset contains 9 variables, 768 instances, and no missing values. There were multiple constraints placed on the original database to select these instances. The dataset contains data for Pima Indian females that are atleast 21 years old.

Independent variables include:

Pregnancies - Integer

- Number of pregnancies
- [0,17]

Glucose - Integer

- Glucose level in blood
- [0,199]

BloodPressure - Integer

- Blood pressure measurement
- [0,122]

SkinThickness - Integer

- Thickness of the skin
- [0,99]

Insulin - Integer

- Insulin level in blood
- [0,846]

BMI - Float (decimal)

- Body mass index
- [0,67.1]

DiabetesPedigreeFunction - Float (decimal)

- Diabetes percentage
- [0,2.42]

Age - Integer

- Age
- [0,81]

The dependent variable in this dataset is Outcome, a binary variable determining if the patient has diabetes where 1 indicates "yes" and 0 indicates "no."

Question:

Is the occurrence of diabetes in Pima Indian females above the age of 21 predictively associated with their body mass index (BMI) and glucose level?

How the data will help you address the question of interest:

The data will help me directly address my question of interest through descriptive statistics, inferential insights, and predictive modeling in order to predict diabetes outcome based on two measurable numerical variables. Further, the data will allow for a level of generalization.

Explain whether your question is focused on prediction, inference, or both:

My question has both prediction and inference, but with an emphasis on prediction. From an inference standpoint, I will be gathering insight and understanding relationship strength between BMI, Glucose Levels, and diabetes outcome. Whereas from a prediction standpoint, I will be making forecasts about future data points based on patterns observed in this sample, predicting diabetes occurrence based on BMI and glucose levels.

Demonstrate that the dataset can be read from the web into R.

In [1]:
```
library(tidyverse)
library(tidymodels)
library(ggplot2)
library(gridExtra)
library(grid)
library(leaps)
```

```
print("Loaded")
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.
0 ──
✔ dplyr     1.1.3     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.3     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts
() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all c
onflicts to become errors
── Attaching packages ─────────────────────────────────── tidymodels 1.1.
1 ──

✔ broom        1.0.5     ✔ rsample      1.2.0
✔ dials        1.2.0     ✔ tune         1.1.2
✔ infer        1.0.5     ✔ workflows    1.1.3
✔ modeldata    1.2.0     ✔ workflowsets 1.0.1
✔ parsnip      1.1.1     ✔ yardstick    1.2.0
✔ recipes      1.0.8

── Conflicts ────────────────────────────────────── tidymodels_conflicts
() ──
✖ scales::discard() masks purrr::discard()
✖ dplyr::filter()   masks stats::filter()
✖ recipes::fixed()  masks stringr::fixed()
✖ dplyr::lag()      masks stats::lag()
✖ yardstick::spec() masks readr::spec()
✖ recipes::step()   masks stats::step()
• Use tidymodels_prefer() to resolve common conflicts.


Attaching package: ‘gridExtra’


The following object is masked from ‘package:dplyr’:

    combine


[1] "Loaded"
```

In [2]:
```
#Github URL and read CSV file from web into a data frame.

URL <- 'https://raw.githubusercontent.com/marcesf/project-pima/main/diabetes

df_original <- read.csv(URL)

head(df_original)
```

A data.frame: 6 × 9

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigre
	<int>	<int>	<int>	<int>	<int>	<dbl>	
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	

Clean and wrangle your data into a tidy format.

In [3]:
```
#Change Outcome variable from binary (for visualization purposes),
#remove STANDOUT irrelevant variables and preview data frame, there are no m

df <- df_original %>%
    mutate(Outcome = as_factor(case_when(Outcome == 1 ~ "Diabetic", TRUE ~ "
    select(-Pregnancies, -SkinThickness)

head(df)
```

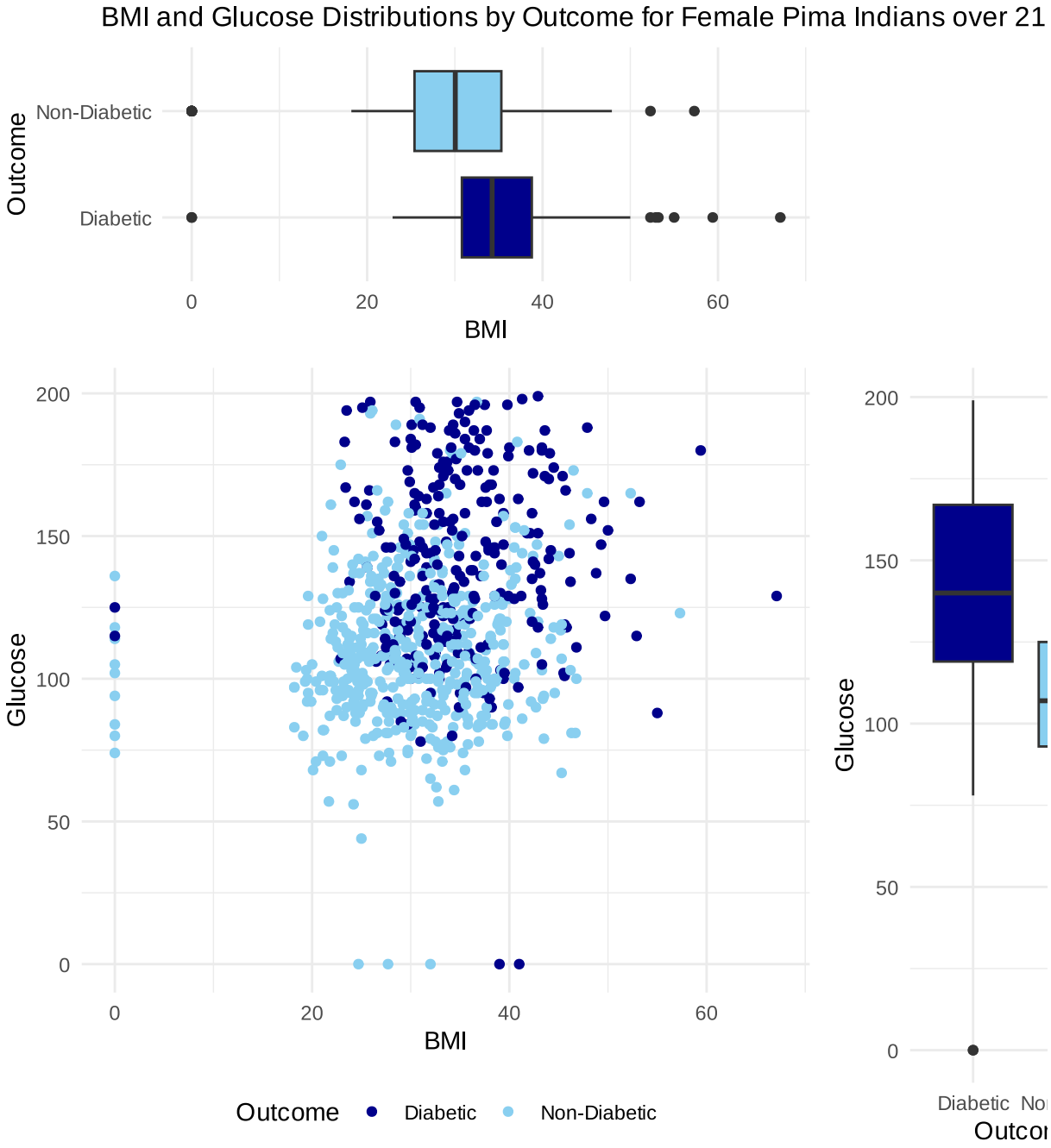
A data.frame: 6 × 7

	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<fct>
1	148	72	0	33.6	0.627	50	Diabetic
2	85	66	0	26.6	0.351	31	Non-Diabetic
3	183	64	0	23.3	0.672	32	Diabetic
4	89	66	94	28.1	0.167	21	Non-Diabetic
5	137	40	168	43.1	2.288	33	Diabetic
6	116	74	0	25.6	0.201	30	Non-Diabetic

I chose to exclude Pregnancies due to its irrelevance of my exploration and SkinThickness due to questionable reliability. These variables do not appear to be relevant to my question, and narrowing allows for a clear exploration of my predictive question. I did not isolate my dataset to only Glucose, BMI, Age, and Outcome, as I want to retain the possibility of including other variables in the future. This ensures a clean and flexible data frame, with the option to use additional variables as controls or confounders in the future if needed.

Propose a high quality plot or a combo plot sharing a common story.

In [4]:
```
#Scatter plot and box plot combo and sources used.

#http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-autom
#https://cran.r-project.org/web/packages/egg/vignettes/Ecosystem.html
#https://ggplot2.tidyverse.org/reference/ggtheme.html

box_bmi <- ggplot(df, aes(x = Outcome, y = BMI, fill = Outcome)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Non-Diabetic" = "#89CFF0" , "Diabetic" = "#0
  coord_flip() +
  theme_minimal() +
  theme(legend.position = "none")

box_glucose <- ggplot(df, aes(x = Outcome, y = Glucose, fill = Outcome)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Non-Diabetic" = "#89CFF0" , "Diabetic" = "#0
  theme_minimal() +
  theme(legend.position = "none")

scatter <- ggplot(df, aes(x = BMI, y = Glucose, color = Outcome)) +
  geom_point() +
  scale_color_manual(values = c("Non-Diabetic" = "#89CFF0" , "Diabetic" = "#
  theme_minimal() +
  theme(legend.position = "bottom")

grid.arrange(box_bmi, NULL, scatter, box_glucose,
             ncol = 2, nrow = 2,
             widths = c(2.5, 1), heights = c(1, 2.5),
             top=textGrob("BMI and Glucose Distributions by Outcome for Fema
```



Clearly explain why you consider this plot relevant to address your question or to explore the data.

I chose a combination of a scatter plot of BMI and Glucose and individual box plots of BMI and Glucose all by Outcome of diabetes for Pima Indian females over the age of 21. This combination plot visualizes the association between BMI and Glucose levels in relation to outcome while allowing individual analysis of the two by outcome through the boxplots. Through this, the relationship between these three variables can be easily visualized. Outcome has values "Diabetic" and "Non-Diabetic" and they are made distinct through the use of different colours (light blue and dark blue) consistent in all three plots.

The plot is relevant to my question as it directly explores associations and relationships between the variables my question directly addresses. The scatterplot is used to visualize the relationship and any patterns, whereas the boxplots are used to visualize distributions of BMI and Glucose values with relation to outcome to be able to visualize the strength of these individual relationships to further explore the scatter plot suggestions, where there is a more in-depth analysis of insights such as outliers or skewness.

With this combination, there is a strong preliminary examination of the strength of a predictive exploration. Interpreting the visualizations, we see that there are strong positive relationships between BMI & Outcome and Glucose & Outcome. The relationship between Glucose & Outcome is visually stronger. This is evident in the scatterplot where the Diabetic points tend strong with Glucose, and with the individual boxplots where Glucose & Outcome where the boxplot for Diabetic overlaps less with Non-Diabetic than in BMI & Outcome and also tends higher. Outliers are evident in all boxplots, with less extremely high outliers in Glucose & Outcome for Diabetic. Overall the distributions visualizing BMI and Glucose by Outcome for Pima Indian females over 21 allow for an initial exploration of the relationship and strengths between the three variables.

Propose one method to address your question of interest using the selected dataset and explain why it was chosen.

We will extend and relax the analysis of the association between body mass index, glucose levels, and diabetes outcome in Pima Indian females over the age of 21, opening additional variables for inferential analysis. We will construct an additive linear regression model, targeting both a predictive and inferential focus. Our chosen method, forward selection, will be used to determine our model. Forward selection is a step-wise approach that adds variables to the regression model, beginning with a null model. We propose splitting our dataset evenly: 50% training and 50% testing for model selection and inference. The data split, along with a potential application of cross-validation, will address overfitting and ensure that our model is generalizable and accurate.

Which assumptions are required, if any, to apply the method selected?

For our linear regression model to produce valid and reliable results, it must follow the assumptions of linear regression discussed in class. We will assume that there's a linear relation between variables, errors are independent, the conditional distribution of error terms is normal, error terms have equal variance, and there is minimal multicollinearity. These assumptions uphold the use of our linear regression model.

What are potential limitations or weaknesses of the method selected?

The method of forward selection has potential limitations or weaknesses. Forward selection, a greedy algorithm, doesn't explore all possible predictor combinations, potentially overlooking more accurate models. It considers associations rather than causal relationships and is highly dependent on the order of variable selection. To offset these limitations, the data will be divided into training and testing sets proposed by a 50/50 split with a potential consideration of cross-validation alongside strong visualizations. These approaches will strengthen the model's reliability and generalizability with a thorough analysis.

Why is this method appropriate?

This method, an additive linear regression model implemented by forward selection, thoroughly explores the factors and associations contributing to diabetes in Pima Indian females over the age of 21. The splitting of the data and the potential use of cross-validation will contribute to the model's reliability, while the linear regression assumptions ensure the validity of our analysis. Our method is therefore appropriate.

Implementation of a proposed model.

In [5]:
```
#create new data frame maintaining binary values for outcome, split data

df_model <- df_original %>%
    select(-Pregnancies, -SkinThickness)

split <- initial_split(df_model, prop = 0.5, strata = Outcome)
train <- training(split)
test <- testing(split)

head(train)
head(test)
```

A data.frame: 6 × 7

	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<int>
1	85	66	0	26.6	0.351	31	0
2	89	66	94	28.1	0.167	21	0
3	116	74	0	25.6	0.201	30	0
4	115	0	0	35.3	0.134	29	0
5	103	30	83	43.3	0.183	33	0
6	109	75	0	36.0	0.546	60	0

A data.frame: 6 × 7

	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<int>
1	148	72	0	33.6	0.627	50	1
2	183	64	0	23.3	0.672	32	1
3	137	40	168	43.1	2.288	33	1
4	78	50	88	31.0	0.248	26	1
5	125	96	0	0.0	0.232	54	1
6	110	92	0	37.6	0.191	30	0

In [6]:
```
#forward selection algorithm with leaps

forward <- regsubsets(Outcome ~ ., data=train, nvmax=10, method="forward")
stats <- summary(forward)

stats
```

```
Subset selection object
Call: regsubsets.formula(Outcome ~ ., data = train, nvmax = 10, method = "fo
rward")
6 Variables  (and intercept)
                         Forced in Forced out
Glucose                      FALSE      FALSE
BloodPressure                FALSE      FALSE
Insulin                      FALSE      FALSE
BMI                          FALSE      FALSE
DiabetesPedigreeFunction     FALSE      FALSE
Age                          FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: forward
         Glucose BloodPressure Insulin BMI DiabetesPedigreeFunction Age
1  ( 1 ) "*"     " "           " "     " " " "                      " "
2  ( 1 ) "*"     " "           " "     "*" " "                      " "
3  ( 1 ) "*"     " "           " "     "*" " "                      "*"
4  ( 1 ) "*"     "*"           " "     "*" " "                      "*"
5  ( 1 ) "*"     "*"           " "     "*" "*"                      "*"
6  ( 1 ) "*"     "*"           "*"     "*" "*"                      "*"
```

In [7]:
```
#adjusted R squared for all 6 models

stats$adjr2
```

0.159475645645376 · 0.212639941031193 · 0.236914048629585 · 0.246707422210569 · 0.254204246878895 · 0.256217633308687

In [8]:
```
#train chosen model with the highest adjusted R squared

model5 <- lm(Outcome ~ Glucose + BloodPressure + BMI + DiabetesPedigreeFunct
```

```
tidy(model5)
summary(model5)
```

A tibble: 6 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-0.807313124	0.1261696532	-6.398632	4.641786e-10
Glucose	0.004284467	0.0007034773	6.090413	2.772155e-09
BloodPressure	-0.002893144	0.0011669725	-2.479188	1.360345e-02
BMI	0.016051038	0.0030854733	5.202131	3.236846e-07
DiabetesPedigreeFunction	0.144112326	0.0657112617	2.193115	2.890777e-02
Age	0.007527297	0.0018452016	4.079390	5.509905e-05

```
Call:
lm(formula = Outcome ~ Glucose + BloodPressure + BMI + DiabetesPedigreeFunct
ion +
    Age, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0582 -0.2994 -0.1154  0.3448  1.0523

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.8073131  0.1261697  -6.399 4.64e-10 ***
Glucose                   0.0042845  0.0007035   6.090 2.77e-09 ***
BloodPressure            -0.0028931  0.0011670  -2.479   0.0136 *
BMI                       0.0160510  0.0030855   5.202 3.24e-07 ***
DiabetesPedigreeFunction  0.1441123  0.0657113   2.193   0.0289 *
Age                       0.0075273  0.0018452   4.079 5.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4122 on 378 degrees of freedom
Multiple R-squared:  0.2639,    Adjusted R-squared:  0.2542
F-statistic: 27.11 on 5 and 378 DF,  p-value: < 2.2e-16
```

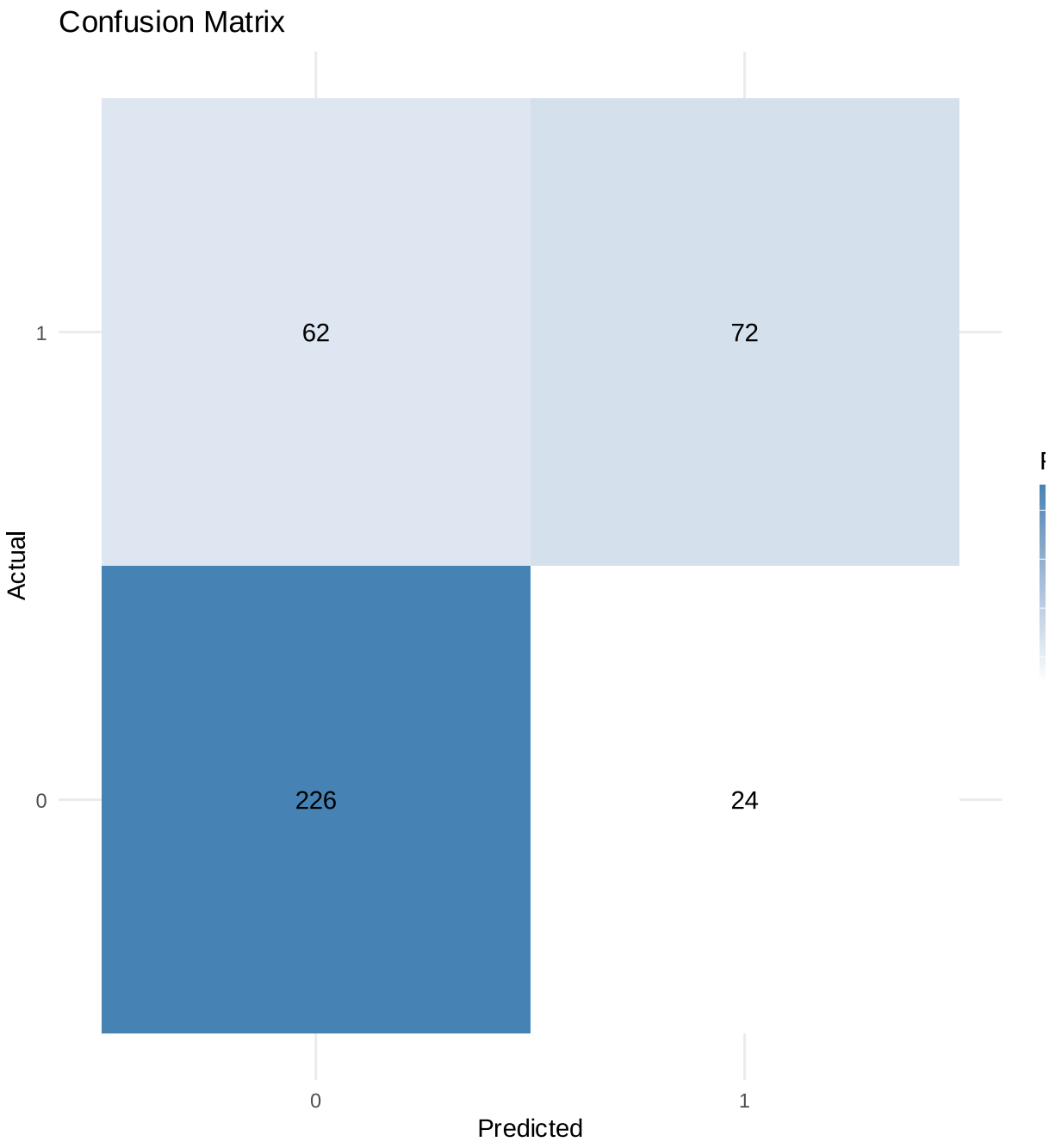
In [9]:
```
#fit predict with test

predictions <- predict(model5, test, type="response")

summary(predictions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.3639  0.1724  0.3308  0.3360  0.5003  1.2286
```

Report the results.

In [10]:
```
#visualization as confusion matrix

#https://ggplot2.tidyverse.org/reference/geom_text.html
```

```r
#https://www.statology.org/confusion-matrix-in-r/
#https://r-charts.com/correlation/heat-map-ggplot2/

test$predicted_class <- ifelse(predictions > 0.5, 1, 0)
confusion_matrix <- as.data.frame(table(Predicted = test$predicted_class, Ac

ggplot(confusion_matrix, aes(x = Predicted, y = Actual, fill = Freq)) +
    geom_tile() +
    geom_text(aes(label = Freq)) +
    scale_fill_gradient(low = "white", high = "steelblue") +
    theme_minimal() +
    labs(title = "Confusion Matrix", x = "Predicted", y = "Actual")
```



Interpretation of the results.

With forward selection and a split by training and testing data, a model with five variables was chosen: Glucose, BloodPressure, BMI, DiabetesPedigreeFunction, and Age. We discover that these are significant predictors of diabetes, aligning with the original focus on BMI and Glucose. Predicting with our chosen model, we then visualize with a confusion matrix to be able to demonstrate predictions vs actual, illustrating good performance. Overall, the chosen model, fitting, and visualization demonstrate the important factors in predicting diabetes outcome, backing the initial predictive goal and the expanded interest.

In [ ]: