planning_stage

December 3, 2024

[1]:

```
library(tidyverse)
library(repr)
library(infer)
library(cowplot)
library(broom)
library(GGally)
library(modelr)
library(car)
library (stats)
```

```
  Attaching core tidyverse packages               tidyverse
2.0.0
  dplyr     1.1.4      readr     2.1.5
  forcats   1.0.0      stringr   1.5.1
  ggplot2   3.5.1      tibble    3.2.1
  lubridate 1.9.3      tidyr     1.3.1
  purrr     1.0.2
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
  Use the conflicted package
(<http://conflicted.r-lib.org/>) to force all conflicts to
become errors

Attaching package: ‘cowplot’


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

    stamp


Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```

Attaching package: ‘modelr’


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

    bootstrap


Loading required package: carData


Attaching package: ‘car’


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

    recode


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

    some
```

1 Introduction to data and EDA

The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The data was obtained from Kaggle. The data was collected from patients that visit the center. All patients included in the study were females at least 21 years or older and of Pima Indian heritage. The dataset includes a total of 768 observations. The goal of the study was to predict whether a patient has diabetes or not based on the variables included in the study. The response variable in the study was the outcome, whether the patient will have diabetes or not. The outcome was represented as a binary variable, with 1 representing Yes and 0 representing No. For additional information, the diabetes pedigree function represents a score indicating the probabiltiy of having diabetes considering family history and age. There were 8 explanatory variables represented in the table below:

Variable	Representation
Age	Continuous Variable
Diabetes Pedigree Function	Diabetes % (ranging from 0.0-1.0)
BMI (body mass index)	Continuous variable
Insulin	Continuous variable
Skin Thickness	Continuous variable
Blood Pressure	Continuous variable
Glucose	Continuous variable
Pregnancies	Continuous variable

Question: When predicting the outcome for diabetes, how do variables like Age, Insulin, BMI and Glucose interact with each other to provide a more accurate outcome?

Response variables: Outcome

Explanatory variables: Age, Insulin, BMI, Glucose

The data will help answer the question of interest as all the explanatory variables will be used to build an additive model as well as an interactive model. This question is more focused on prediction as we are trying to predict diabetes from selecting a few known diabetes risk factors. Investigating this question of interest will aid in understanding what are the key factors that help in predicting the diagnosis of diabetes and whether certain variables need to be understood together in order to obtain a more accurate predictive model. Various analysis methods can be used to determine whether the additive model (looking at all the selected variables separately) or the interactive model (looking at variables in conjunction with other variables) provides a better model. In order to make an additive and interactive model, the BMI variable will be converted from a continuous variable into a categorical variable. A BMI of 18.5 or below will be considered underweight, between 18.5 - 24.9 will be considered healthy weight, between 25-29.9 will be considered overweight, and over 29.9 will be considered obese. Hence, the BMI variable will have 4 levels, with underweight as the reference level.

Potential visualization technique: A set of plots would be better for visualizing all the selected variables. A pairplot can be generated to determine whether any of the selected variables are correlated with each other. If some correlation is observed, that can aid in determining whether an additive or interactive model should be used. Since the outcome variable is a binary variable, a logistic regression will be used in order to plot data. The continuous variables will be plotted on the x-axis with the outcome on the y-axis.

[2]:

```r
diabetes <- read_csv("data/diabetes.csv", col_types = cols())
```

[3]:

```r
diabetes
```

A spec_tbl_df: 768 × 9

Pregnancies <dbl>	Glucose <dbl>	BloodPressure <dbl>	SkinThickness <dbl>	Insulin <dbl>	BMI <dbl>	DiabetesPe <dbl>
6	148	72	35	0	33.6	0.627
1	85	66	29	0	26.6	0.351
8	183	64	0	0	23.3	0.672
1	89	66	23	94	28.1	0.167
0	137	40	35	168	43.1	2.288
5	116	74	0	0	25.6	0.201
3	78	50	32	88	31.0	0.248
10	115	0	0	0	35.3	0.134
2	197	70	45	543	30.5	0.158
8	125	96	0	0	0.0	0.232
4	110	92	0	0	37.6	0.191
10	168	74	0	0	38.0	0.537
10	139	80	0	0	27.1	1.441
1	189	60	23	846	30.1	0.398
5	166	72	19	175	25.8	0.587
7	100	0	0	0	30.0	0.484
0	118	84	47	230	45.8	0.551
7	107	74	0	0	29.6	0.254
1	103	30	38	83	43.3	0.183
1	115	70	30	96	34.6	0.529
3	126	88	41	235	39.3	0.704
8	99	84	0	0	35.4	0.388
7	196	90	0	0	39.8	0.451
9	119	80	35	0	29.0	0.263
11	143	94	33	146	36.6	0.254
10	125	70	26	115	31.1	0.205
7	147	76	0	0	39.4	0.257
1	97	66	15	140	23.2	0.487
13	145	82	19	110	22.2	0.245
5	117	92	0	0	34.1	0.337
2	99	60	17	160	36.6	0.453
1	102	74	0	0	39.5	0.293
11	120	80	37	150	42.3	0.785
3	102	44	20	94	30.8	0.400
1	109	58	18	116	28.5	0.219
9	140	94	0	0	32.7	0.734
13	153	88	37	140	40.6	1.174
12	100	84	33	105	30.0	0.488
1	147	94	41	0	49.3	0.358
1	81	74	41	57	46.3	1.096
3	187	70	22	200	36.4	0.408
6	162	62	0	0	24.3	0.178
4	136	70	0	0	31.2	1.182
1	121	78	39	74	39.0	0.261
3	108	62	24	0	26.0	0.223
0	181	88	44	510	43.3	0.222
8	154	78	32	0	32.4	0.443
1	128	88	39	110	36.5	1.057
7	137	90	41	0	32.0	0.391
0	123	72	0	0	36.3	0.258

[4]:

```
diabetes_filtered <- diabetes %>%
select (Age, Outcome, Glucose, BMI, Insulin)
```

[5]:

```
diabetes_filtered
```

A tibble: 768 × 5

Age <dbl>	Outcome <dbl>	Glucose <dbl>	BMI <dbl>	Insulin <dbl>
50	1	148	33.6	0
31	0	85	26.6	0
32	1	183	23.3	0
21	0	89	28.1	94
33	1	137	43.1	168
30	0	116	25.6	0
26	1	78	31.0	88
29	0	115	35.3	0
53	1	197	30.5	543
54	1	125	0.0	0
30	0	110	37.6	0
34	1	168	38.0	0
57	0	139	27.1	0
59	1	189	30.1	846
51	1	166	25.8	175
32	1	100	30.0	0
31	1	118	45.8	230
31	1	107	29.6	0
33	0	103	43.3	83
32	1	115	34.6	96
27	0	126	39.3	235
50	0	99	35.4	0
41	1	196	39.8	0
29	1	119	29.0	0
51	1	143	36.6	146
41	1	125	31.1	115
43	1	147	39.4	0
22	0	97	23.2	140
57	0	145	22.2	110
38	0	117	34.1	0
21	0	99	36.6	160
42	1	102	39.5	0
48	1	120	42.3	150
26	0	102	30.8	94
22	0	109	28.5	116
45	1	140	32.7	0
39	0	153	40.6	140
46	0	100	30.0	105
27	1	147	49.3	0
32	0	81	46.3	57
36	1	187	36.4	200
50	1	162	24.3	0
22	1	136	31.2	0
28	0	121	39.0	74
25	0	108	26.0	0
26	1	181	43.3	510
45	1	154	32.4	0
37	1	128	36.5	110
39	0	137	32.0	0
52	1	123	36.3	0

[6]:

```
diabetes_filtered$bmi_levels <- cut(diabetes_filtered$BMI,
                           breaks=c(-Inf, 18.5, 25, 29.9, 67.1),
                           labels=c('underweight', 'healthy weight', 'overweight',␣
 ↪'obese'))
```

[7]:

```
diabetes_filtered
```

A tibble: 768 × 6

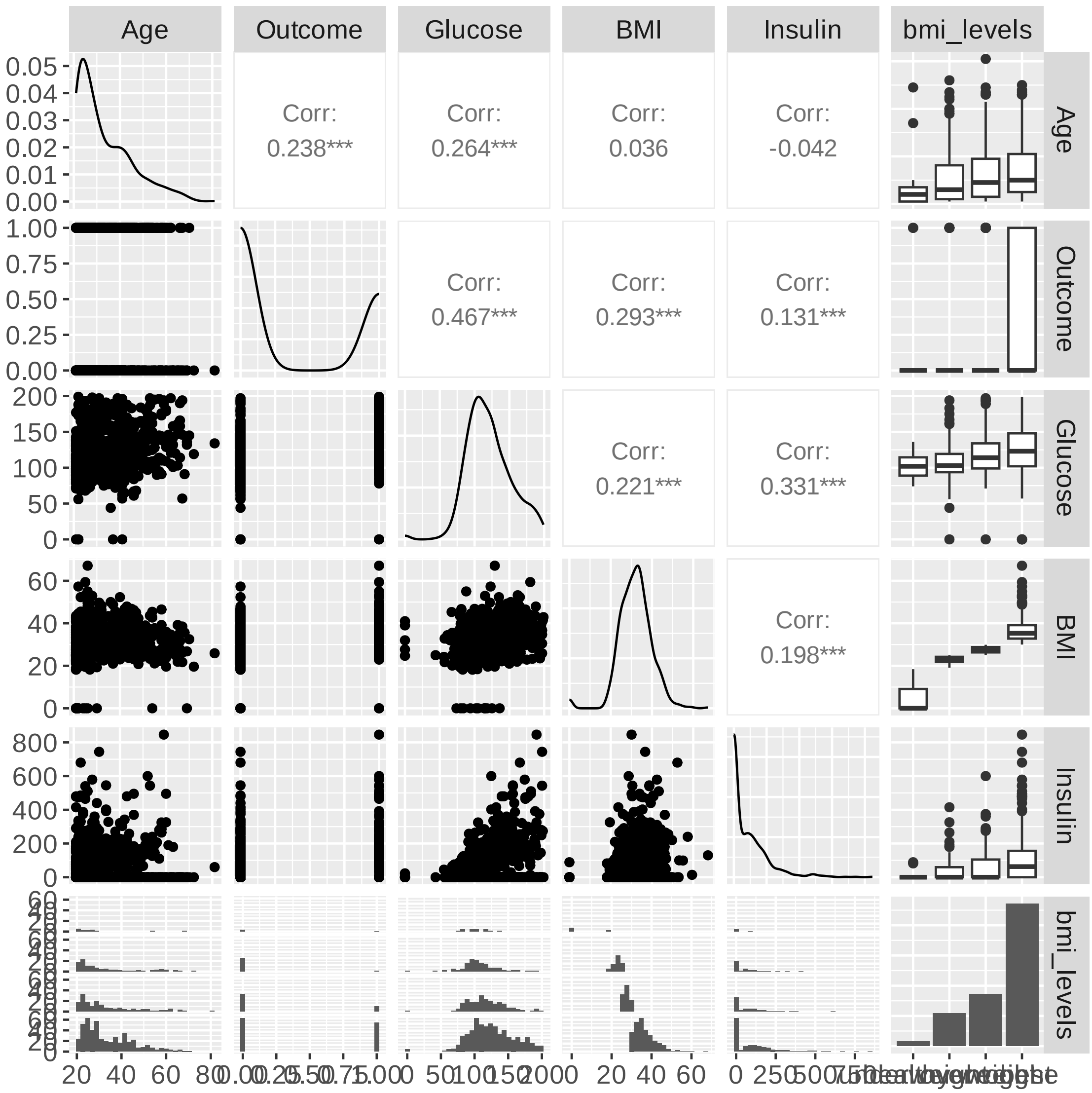
Age	Outcome	Glucose	BMI	Insulin	bmi_levels
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
50	1	148	33.6	0	obese
31	0	85	26.6	0	overweight
32	1	183	23.3	0	healthy weight
21	0	89	28.1	94	overweight
33	1	137	43.1	168	obese
30	0	116	25.6	0	overweight
26	1	78	31.0	88	obese
29	0	115	35.3	0	obese
53	1	197	30.5	543	obese
54	1	125	0.0	0	underweight
30	0	110	37.6	0	obese
34	1	168	38.0	0	obese
57	0	139	27.1	0	overweight
59	1	189	30.1	846	obese
51	1	166	25.8	175	overweight
32	1	100	30.0	0	obese
31	1	118	45.8	230	obese
31	1	107	29.6	0	overweight
33	0	103	43.3	83	obese
32	1	115	34.6	96	obese
27	0	126	39.3	235	obese
50	0	99	35.4	0	obese
41	1	196	39.8	0	obese
29	1	119	29.0	0	overweight
51	1	143	36.6	146	obese
41	1	125	31.1	115	obese
43	1	147	39.4	0	obese
22	0	97	23.2	140	healthy weight
57	0	145	22.2	110	healthy weight
38	0	117	34.1	0	obese
21	0	99	36.6	160	obese
42	1	102	39.5	0	obese
48	1	120	42.3	150	obese
26	0	102	30.8	94	obese
22	0	109	28.5	116	overweight
45	1	140	32.7	0	obese
39	0	153	40.6	140	obese
46	0	100	30.0	105	obese
27	1	147	49.3	0	obese
32	0	81	46.3	57	obese
36	1	187	36.4	200	obese
50	1	162	24.3	0	healthy weight
22	1	136	31.2	0	obese
28	0	121	39.0	74	obese
25	0	108	26.0	0	overweight
26	1	181	43.3	510	obese
45	1	154	32.4	0	obese
37	1	128	36.5	110	obese
39	0	137	32.0	0	obese
52	1	123	36.3	0	obese

[8]:

```
diabetes_pairplots <-
    diabetes_filtered %>%
    ggpairs(progress = FALSE) +
    theme(
        text = element_text(size = 15),
        plot.title = element_text(face = "bold"),
        axis.title = element_text(face = "bold")
    )

diabetes_pairplots
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



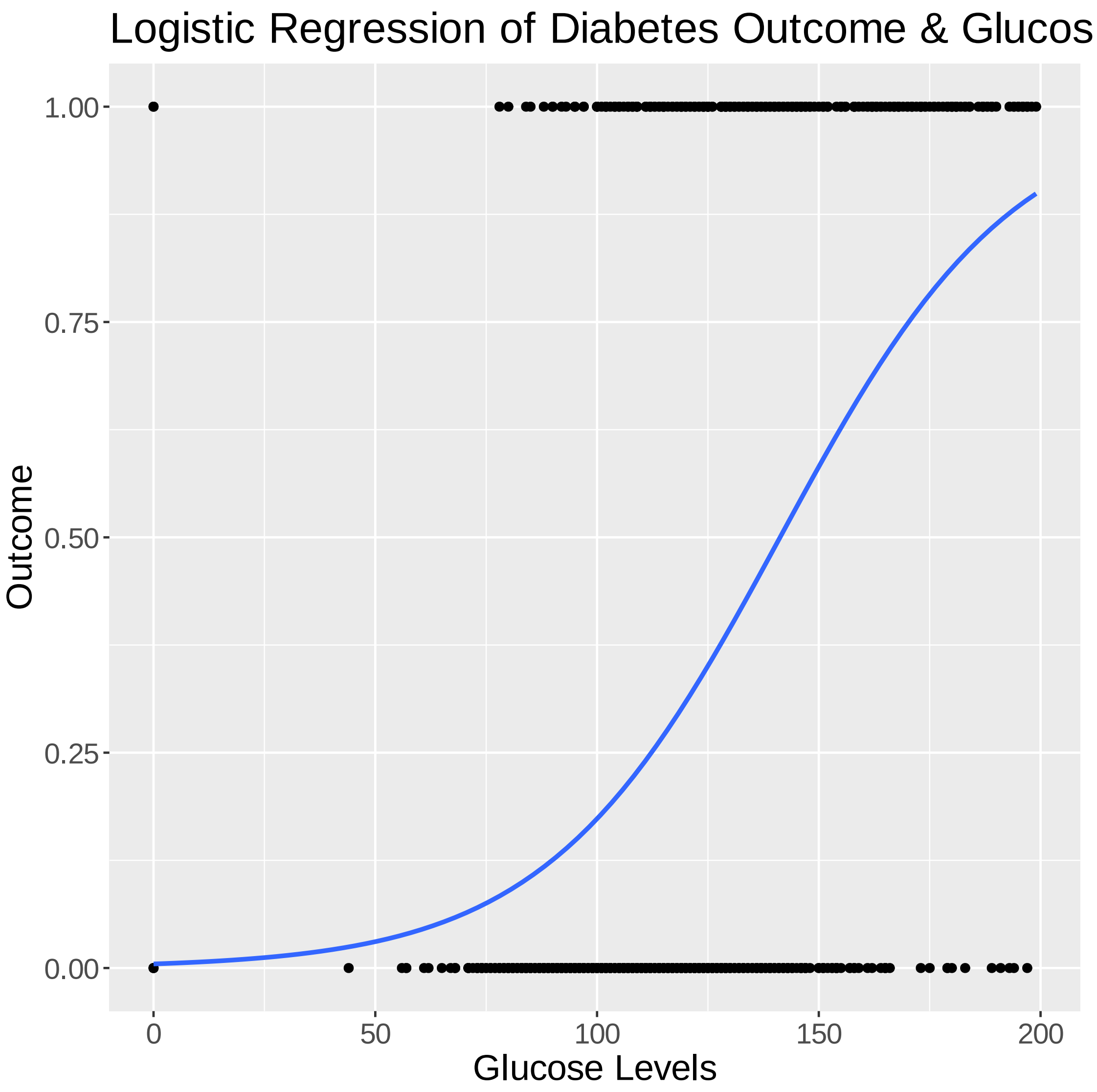
This pairplot displays all the selected variables in the filtered dataset (Outcome, Age, Insulin, Glucose, and BMI). Mostly, all variables seem to be positively correlated with one another, with insulin and age being negatively correlated. Overall, the correlation coefficients are not too large, indicating a weak correlation between the variables. One set of variables, glucose and outcome seem to have a slightly stronger correlation compared to the rest of the variable pairs. Overall, since none of the variable pairs seem to be strongly correlated with each other, it reduces chances of multicollinearity, perhaps making the models more accurate.

[9]:

```
outcome_glucose <- diabetes_filtered %>%
mutate(diabetes_filtered = if_else(Outcome == "Yes", 1, 0)) %>%
    ggplot () +
  geom_point (aes(x = Glucose, y = Outcome)) +
  geom_smooth (aes(x = Glucose, y = Outcome), method = glm, method.args =␣
 ↪c(family = binomial), se = FALSE) +
  labs(y = "Outcome", x = "Glucose Levels") +
  ggtitle("Logistic Regression of Diabetes Outcome & Glucose") +
  theme(text = element_text(size = 16.5))

outcome_glucose
```

```
`geom_smooth()` using formula = 'y ~ x'
```



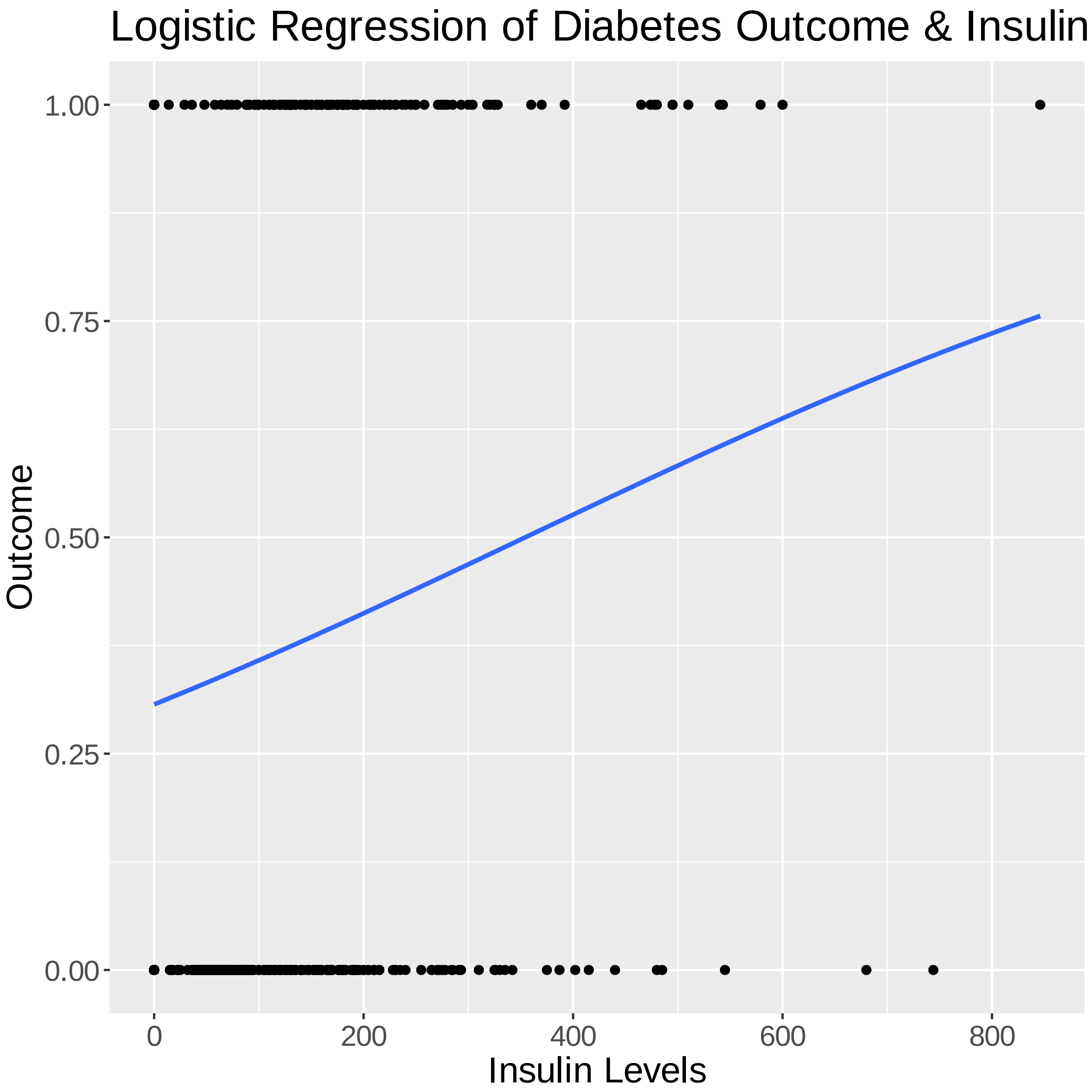
This graph shows the logistic regression for Outcome and Glucose. There is slight relationship between the two variables. There is an S shaped curve depicted by the blue line, indicating some relationship between the variables. Generally, higher levels of glucose result in diabetes (with an outcome of 1) and lower levels of glucose resulting in no diabetes (with an outcome of 0). However, this relationship is still weak with no strong trend.

[10]:

```
outcome_insulin <- diabetes_filtered %>%
mutate(diabetes_filtered = if_else(Outcome == "Yes", 1, 0)) %>%
    ggplot () +
  geom_point (aes(Insulin, Outcome)) +
  geom_smooth (aes(x = Insulin, y = Outcome), method = glm, method.args = c(family = binomial), se = FALSE) +
  labs(y = "Outcome", x = "Insulin Levels") +
  ggtitle("Logistic Regression of Diabetes Outcome & Insulin") +
  theme(text = element_text(size = 16.5))

outcome_insulin
```

```
`geom_smooth()` using formula = 'y ~ x'
```

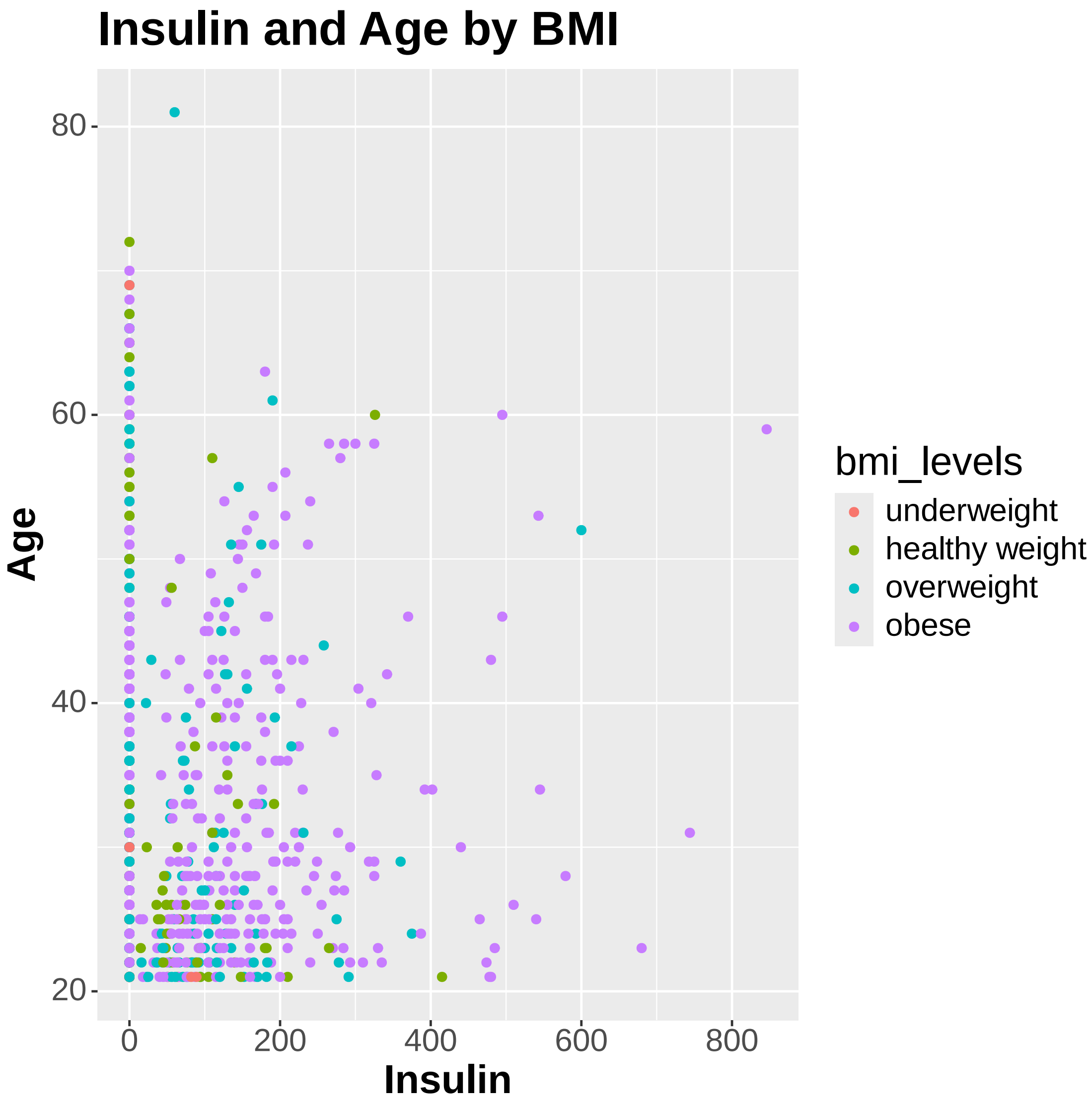


This graph shows the logistic regression for Outcome and Insulin. There is a very weak relationship between the two variables as seen by the almost linear relationship depicted by the blue line. Despite maintaining insulin levels being a key factor in controlling diabetes, insulin levels do not seem to predict the outcome of diabetes accurately. Low insulin levels are more common compared to high insulin levels. Additionally, both insulin levels result in an outcome of 1 and 0 somewhat equally with no apparent trend.

[11]:

```r
insulin_age_bmi <-
  diabetes_filtered %>%
  ggplot() +
  geom_point (aes(Insulin, Age, color = bmi_levels)) +
  ggtitle("Insulin and Age by BMI") +
  xlab("Insulin") +
  ylab("Age") +
  theme(
    text = element_text(size = 18),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold")
  )

insulin_age_bmi
```

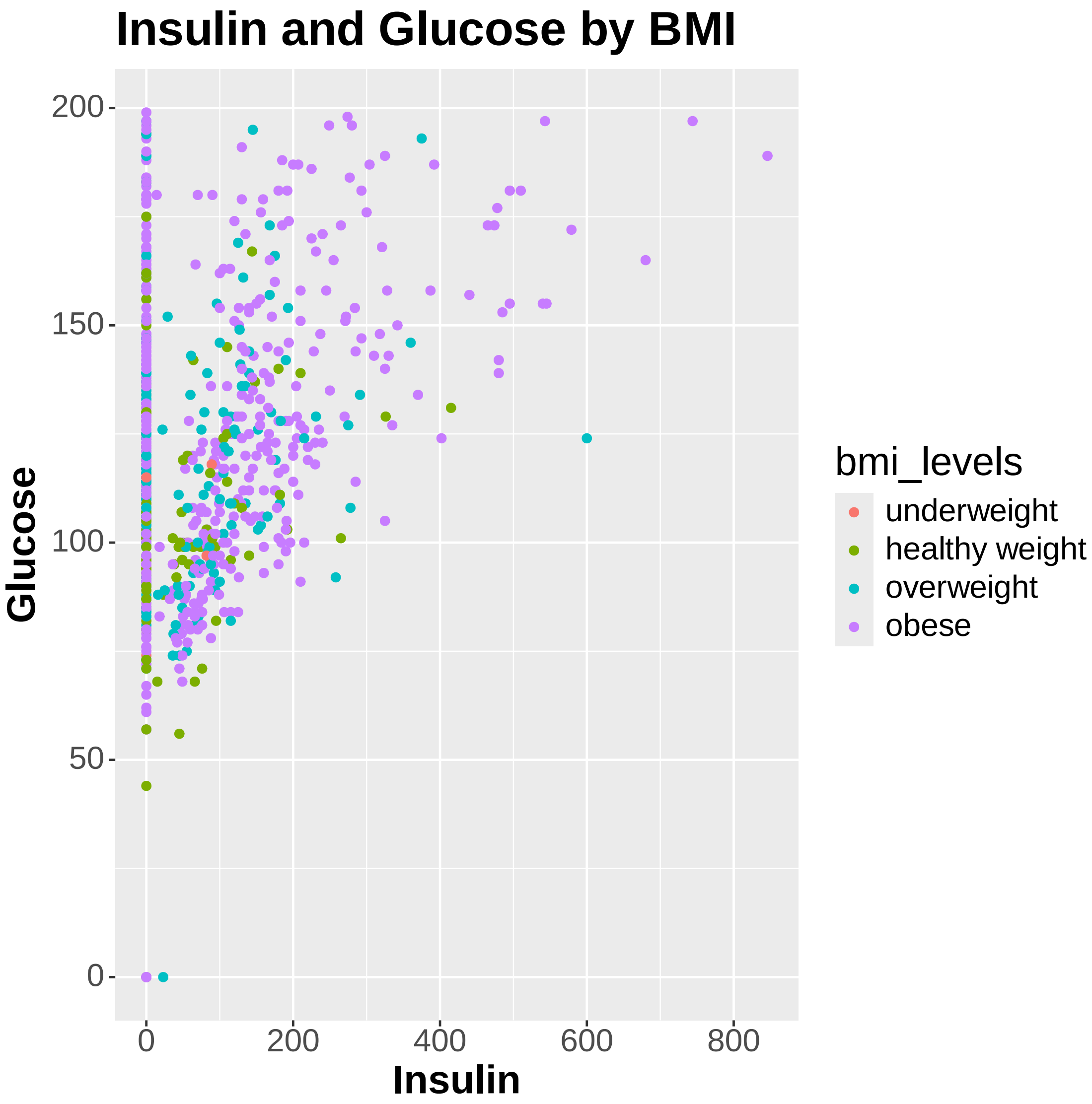


This graph shows the relationship between Insulin and Age by each BMI group. There does not seem to be ay obvious relationship between all the variables. Generally, people with a obese BMI group (purple) had higher insulin levels, with no apparent correlation to age. There are multiple observations from all BMI groups seen with a 0 insulin level with varying ages.

[12]:

```
insulin_glucose_bmi <-
  diabetes_filtered %>%
  ggplot() +
  geom_point (aes(Insulin, Glucose, color = bmi_levels)) +
  ggtitle("Insulin and Glucose by BMI") +
  xlab("Insulin") +
  ylab("Glucose") +
  theme(
    text = element_text(size = 18),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold")
  )

insulin_glucose_bmi
```

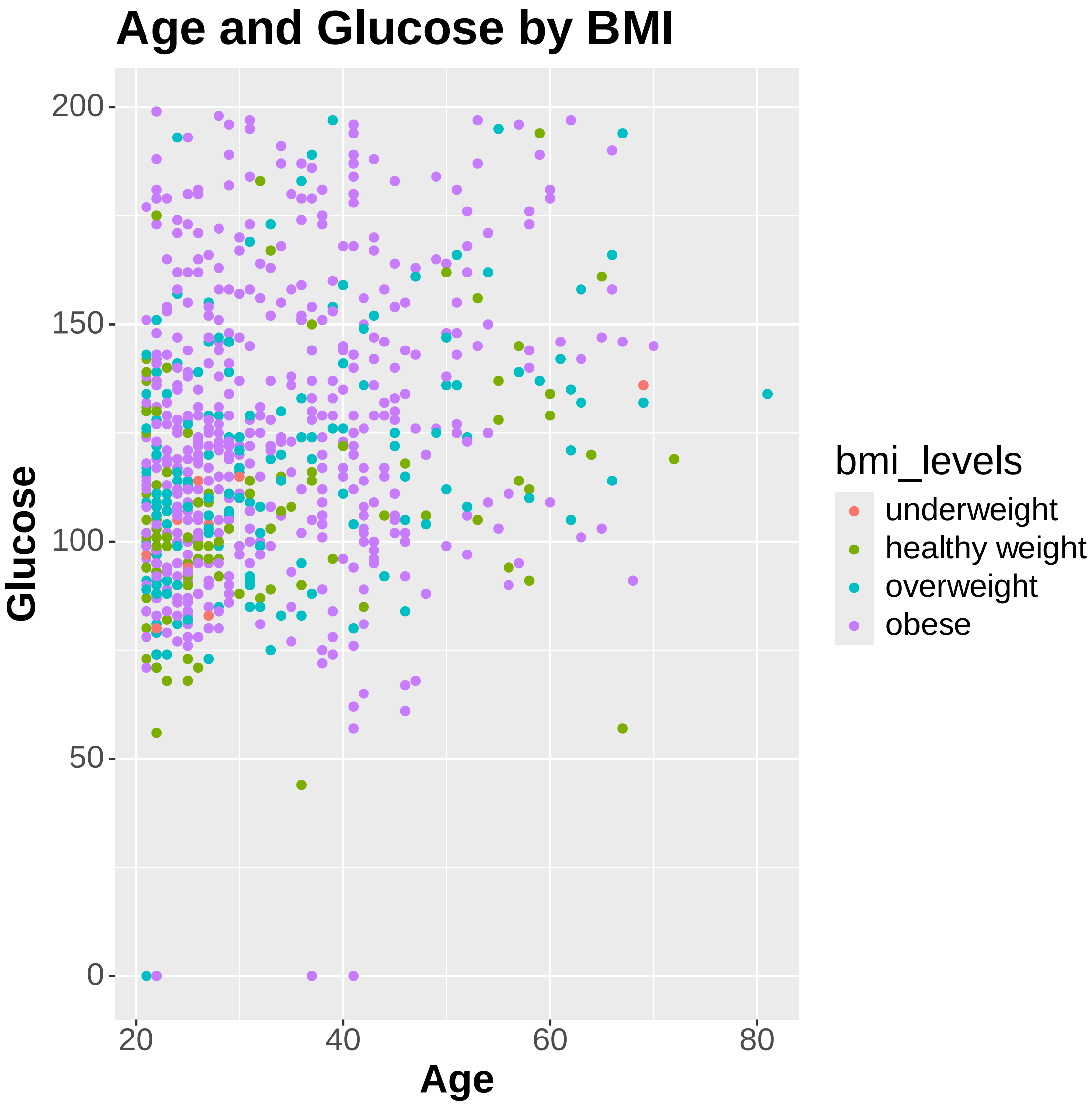


This graph shows the relationship between Insulin and Glucose by each BMI group. There is a slight positive relationship between insulin and glucose with no apparent trend within the BMI groups. As Insulin levels increase, generally there is also an increase in glucose levels as well. Above insulin levels of about 200, most of the observations are from the obese or overweight BMI groups. Similar to the age and insulin graph, there are multiple observations from all BMI groups seen with a 0 insulin level with varying levels of glucose.

[13]:

```
age_glucose_bmi <-
  diabetes_filtered %>%
  ggplot() +
  geom_point (aes(Age, Glucose, color = bmi_levels)) +
  ggtitle("Age and Glucose by BMI") +
  xlab("Age") +
  ylab("Glucose") +
  theme(
    text = element_text(size = 18),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold")
  )

age_glucose_bmi
```



This graph shows the relationship between Age and Glucose by each BMI group. There does not seem to be ay obvious relationship between all the variables. Generally, there is a concentration of observations under age 40. For glucose levels above 150, most of the observations are from the obese or overweight BMI group.

2 Methods and Plan

Question: When predicting the outcome for diabetes, how does Age interact with other variables like Insulin, BMI and Glucose to provide a more accurate outcome?

The proposed model for answering the above question includes a logistic regression approach with variables Glucose and Age to predict the Outcome. To answer the original question of whether Age interacts with other variables to predict Outcome, several models were generated. Since the main goal of this question was to predict Outcome, the variable that was highly correlated with Outcome was chosen to be included (Glucose). Additionally, ensuring the chosen input variables were not too highly correlated with each other was important to prevent any inflation of error estimates. A logistic regression was appropriate since the response variable is a binary variable.

To perform a logistic regression, a few assumptions were made. * The errors were assumed to be independent * The chosen input variables were assumed to not be correlated or have low correlation * The relationship between the input variables and the log odds of the response variable was linear

In order to ensure the input variables were not highly correlated with each other, a correlation matrix was generated in Assignment 1. Additionally, it was assumed that the sample size was large enough, which in this dataset was sufficient (768 observations) to generate models. In order to test for the interaction of variables with Age, 5 variables were individually interacted with Age, along with Glucose. These variables were tested individually in order to keep the generated models simple for interpretation. 5 logistic models were generated, each testing one variable individually interacting with Age. Each model's VIF values were tested in order to ensure the model's stability and measure the amount of multicollinearity in the logistic regression analysis. The calculation of VIF values was prioritized due to the presence of interaction terms in all models. Interaction terms could introduce correlation, making it highly inaccurate and unstable to predict Outcome. Once all models were generated, VIF values were used to select one, best model. To find the better fitting model, the chosen model was then compared with an additive model, using AIC values.

One limitation of choosing a logistic regression is that it cannot be used to model complex relationships with multiple variables, or to model nonlinear relationships. Due to these limitations, only 3 variables were used at once in the generated models, preventing highly complex models. Another limitation of using only VIF values to measure multicollinearity is that VIF can only measure pairwise correlation, hindering the ability to detect any higher order multicollinearity. Lastly, one limitation of using AIC as a measure to detect a model's performance is that AIC only compares models on a relative scale. Hence, AIC values should not be used for absolute predictive performance.

Word Count: 461

3 Implementation of a proposed model

In order to create a model, the first step was to filter the data and take out any values of 0 that were inappropriate in some columns (ex: BMI, Glucose etc).

[25]:

```r
BMI_count <- sum(diabetes$BMI == 0)
Glucose_count <- sum(diabetes$Glucose == 0)
BloodPressure_count <- sum(diabetes$BloodPressure == 0)
SkinThickness_count <- sum(diabetes$SkinThickness == 0)



BMI_count
Glucose_count
BloodPressure_count
SkinThickness_count
```

11

5

35

227

All of the above columns should not have any values of 0, which is why the number of 0s were counted in these columns. Since there are 768 observations, having 227 counts of 0 values in SkinThickness severely affects analysis. Hence, SkinThickness will be removed from the analysis and will not be included in any of the proposed models. For the remaining 3 variables, since there are only a few observations for 0s, these values will be removed from the dataset, instead of completely removing the variable.

[26]:

```r
diabetes_filter <- diabetes %>%
select (-SkinThickness) %>%
filter (Glucose != 0) %>%
filter (BMI != 0) %>%
filter (BloodPressure !=0)

diabetes_filter
```

A tibble: 724 × 8

Pregnancies	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Age
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
6	148	72	0	33.6	0.627	50
1	85	66	0	26.6	0.351	31
8	183	64	0	23.3	0.672	32
1	89	66	94	28.1	0.167	21
0	137	40	168	43.1	2.288	33
5	116	74	0	25.6	0.201	30
3	78	50	88	31.0	0.248	26
2	197	70	543	30.5	0.158	53
4	110	92	0	37.6	0.191	30
10	168	74	0	38.0	0.537	34
10	139	80	0	27.1	1.441	57
1	189	60	846	30.1	0.398	59
5	166	72	175	25.8	0.587	51
0	118	84	230	45.8	0.551	31
7	107	74	0	29.6	0.254	31
1	103	30	83	43.3	0.183	33
1	115	70	96	34.6	0.529	32
3	126	88	235	39.3	0.704	27
8	99	84	0	35.4	0.388	50
7	196	90	0	39.8	0.451	41
9	119	80	0	29.0	0.263	29
11	143	94	146	36.6	0.254	51
10	125	70	115	31.1	0.205	41
7	147	76	0	39.4	0.257	43
1	97	66	140	23.2	0.487	22
13	145	82	110	22.2	0.245	57
5	117	92	0	34.1	0.337	38
5	109	75	0	36.0	0.546	60
3	158	76	245	31.6	0.851	28
3	88	58	54	24.8	0.267	22
2	99	60	160	36.6	0.453	21
1	102	74	0	39.5	0.293	42
11	120	80	150	42.3	0.785	48
3	102	44	94	30.8	0.400	26
1	109	58	116	28.5	0.219	22
9	140	94	0	32.7	0.734	45
13	153	88	140	40.6	1.174	39
12	100	84	105	30.0	0.488	46
1	147	94	0	49.3	0.358	27
1	81	74	57	46.3	1.096	32
3	187	70	200	36.4	0.408	36
6	162	62	0	24.3	0.178	50
4	136	70	0	31.2	1.182	22
1	121	78	74	39.0	0.261	28
3	108	62	0	26.0	0.223	25
0	181	88	510	43.3	0.222	26
8	154	78	0	32.4	0.443	45
1	128	88	110	36.5	1.057	37
7	137	90	0	32.0	0.391	39
0	123	72	0	36.3	0.258	52

The following filtered dataset will be used to conduct analysis and generate a model. 5 models will be generated with Glucose and a variable interacting with Age, since the main research question is to investigate the interaction of different variables with Age.

Model 1: Outcome ~ Glucose + Age*Pregnancies

Model 2: Outcome ~ Glucose + Age*BloodPressure

Model 3: Outcome ~ Glucose + Age*Insulin

Model 4: Outcome ~ Glucose + Age*BMI

Model 5: Outcome ~ Glucose + Age*DiabetesPedigreeFunction

The VIF values for each model and the input variables will be checked, in order to select the best model with the least amount of multicollinearity.

[30]:

```r
model_1 <- glm (formula = Outcome ~ Glucose + Age*Pregnancies,
              data = diabetes_filter,
              family = "binomial")
vif_1 <- vif(model_1)
aic_model_1 <- AIC (model_1)


tidy (model_1)
tidy (vif_1)
tidy (aic_model_1)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

A tibble: 5 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-7.341041247	0.638731510	-11.493157	1.427969e-30
Glucose	0.037996998	0.003495785	10.869374	1.613033e-27
Age	0.044137588	0.014290374	3.088624	2.010859e-03
Pregnancies	0.435528535	0.110655705	3.935889	8.288940e-05
Age:Pregnancies	-0.008515166	0.002726836	-3.122727	1.791836e-03

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 4 × 2

names <chr>	x <dbl>
Glucose	1.035703
Age	3.383697
Pregnancies	16.704919
Age:Pregnancies	24.012926

[31]:

```
model_2 <- glm (formula = Outcome ~ Glucose + Age*BloodPressure,
              data = diabetes_filter,
              family = "binomial")
vif_2 <- vif(model_2)
aic_model_2 <- AIC (model_2)

tidy (model_2)
tidy (vif_2)
tidy (aic_model_2)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

A tibble: 5 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-7.6843164963	1.7865986528	-4.3010871	1.699622e-05
Glucose	0.0366774056	0.0034545382	10.6171660	2.479754e-26
Age	0.0576841793	0.0518527760	1.1124608	2.659401e-01
BloodPressure	0.0214400428	0.0236711720	0.9057449	3.650709e-01
Age:BloodPressure	-0.0004360262	0.0006732439	-0.6476497	5.172115e-01

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 4 × 2

names <chr>	x <dbl>
Glucose	1.033206
Age	46.199152
BloodPressure	9.939639
Age:BloodPressure	66.154107

[36]:

```
model_3 <- glm (formula = Outcome ~ Glucose + Age*Insulin,
              data = diabetes_filter,
              family = "binomial")
vif_3 <- vif(model_3)
aic_model_3 <- AIC (model_3)

tidy (model_3)
tidy (vif_3)
tidy (aic_model_3)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

A tibble: 5 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-5.7819541767	5.061606e-01	-11.423161	3.203663e-30
Glucose	0.0382541798	3.698495e-03	10.343176	4.493931e-25
Age	0.0105338006	9.326194e-03	1.129485	2.586931e-01
Insulin	-0.0075665293	2.617106e-03	-2.891181	3.837966e-03
Age:Insulin	0.0002268237	7.728755e-05	2.934802	3.337606e-03

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 4 × 2

names <chr>	x <dbl>
Glucose	1.153016
Age	1.441015
Insulin	11.004859
Age:Insulin	10.394570

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 1 × 1

x <dbl>
735.7293

[33]:

```
model_4 <- glm (formula = Outcome ~ Glucose + Age*BMI,
                data = diabetes_filter,
                family = "binomial")
vif_4 <- vif(model_4)
aic_model_4 <- AIC (model_4)

tidy (model_4)
tidy (vif_4)
tidy (aic_model_4)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

A tibble: 5 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-7.399223791	1.473321770	-5.0221370	5.109971e-07
Glucose	0.034960559	0.003518317	9.9367286	2.881343e-23
Age	-0.018539955	0.041551954	-0.4461873	6.554619e-01
BMI	0.036280154	0.044442975	0.8163304	4.143112e-01
Age:BMI	0.001585447	0.001288614	1.2303507	2.185658e-01

```
Warning message:
```

```
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 4 × 2

names	x
<chr>	<dbl>
Glucose	1.030558
Age	28.050391
BMI	9.433674
Age:BMI	32.558025

[34]:

```
model_5 <- glm (formula = Outcome ~ Glucose + Age*DiabetesPedigreeFunction,
                data = diabetes_filter,
                family = "binomial")
vif_5 <- vif(model_5)
aic_model_5 <- AIC (model_5)


tidy (model_5)
tidy (vif_5)
tidy (aic_model_5)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

A tibble: 5 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-6.781740366	0.663826838	-10.2161286	1.679135e-24
Glucose	0.036610181	0.003488775	10.4937065	9.233363e-26
Age	0.030263933	0.014595594	2.0734979	3.812596e-02
DiabetesPedigreeFunction	1.291206480	0.962562161	1.3414266	1.797820e-01
Age:DiabetesPedigreeFunction	-0.007250551	0.026569807	-0.2728869	7.849402e-01

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 4 × 2

names	x
<chr>	<dbl>
Glucose	1.024097
Age	3.610703
DiabetesPedigreeFunction	11.138264
Age:DiabetesPedigreeFunction	13.445681

From these VIF values, model 3 and model 5 have the lowest VIF values, with model 3 having the lowest values. Hence, model 3 will be selected for the logistic regression model. For further comparison, 1 additional model will also be generated:

Model 6: Outcome ~ Glucose + Age

[37]:

```
model_6 <- lm (formula = Outcome ~ Glucose + Age,
              data = diabetes_filter)
vif_6 <- vif(model_6)
aic_model_6 <- AIC (model_6)

tidy (model_6)
tidy (vif_6)
tidy (aic_model_6)
```

A tibble: 3 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-0.683245434	0.0686746095	-9.949025	6.118822e-22
Glucose	0.007037239	0.0005157272	13.645273	7.066981e-38
Age	0.005080867	0.0013479047	3.769455	1.769887e-04

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 2 × 2

names <chr>	x <dbl>
Glucose	1.074649
Age	1.074649

```
Warning message:
"'tidy.numeric' is deprecated.
See help("Deprecated")"
```

A tibble: 1 × 1

x <dbl>
773.3088

The selected model (model 3) has an AIC value of 735.7, whereas the additive model has an AIC value of 773.3, indicating that the selected model with an interaction term is a better performing model. Hence, the selection methods performed earlier with VIF values proved to create a better model than just an additive model (a lower AIC will indicate a better fitting model). In the steps below, predicted probabilies from Outcome will be calculated and will be used to plot against Age and Glucose, separately.

[38]:

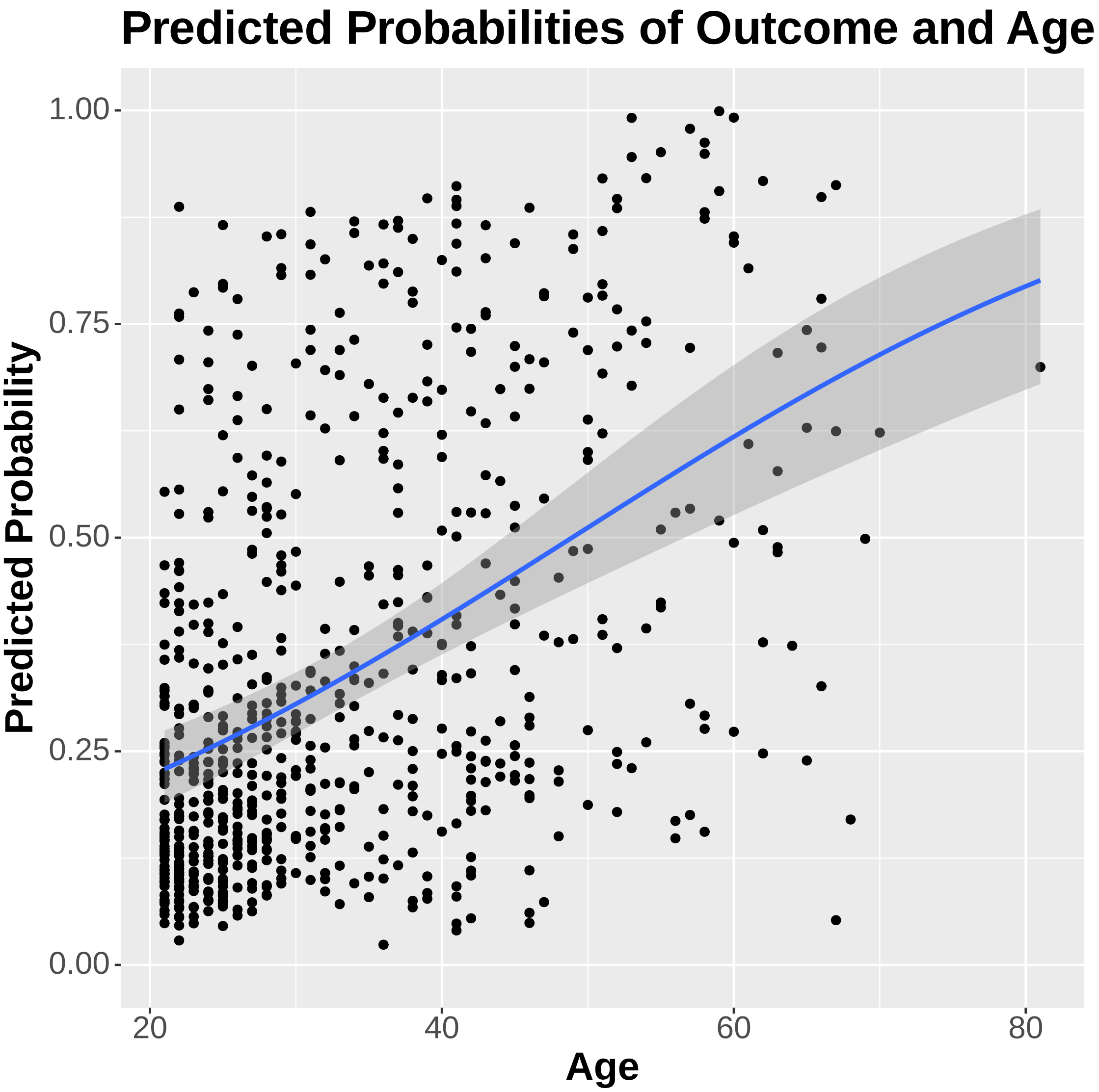
```
predicted_probabilities <- predict (model_3, type = "response")
```

[39]:

```
predicted_plot_age <- diabetes_filter %>%
ggplot (aes (x = Age, y = predicted_probabilities)) +
geom_point () +
geom_smooth(method = "glm", method.args = list(family = binomial)) +
ylim (0,1) +
ggtitle("Predicted Probabilities of Outcome and Age") +
  xlab("Age") +
  ylab("Predicted Probability") +
  theme(
    text = element_text(size = 18),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold")
  )

predicted_plot_age
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message in eval(family$initialize):
"non-integer #successes in a binomial glm!"
```



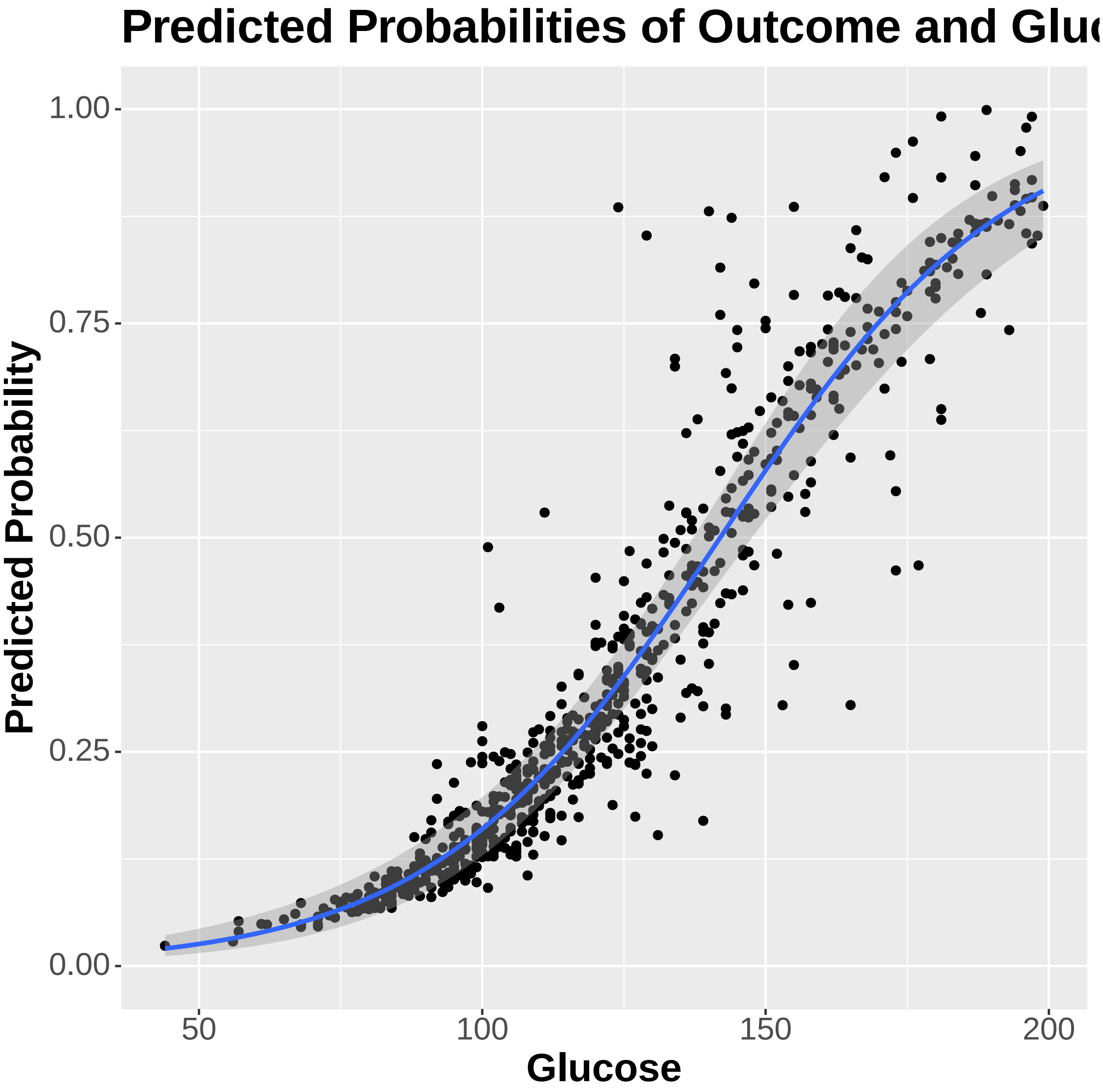
This graph indicates a scatterplot between Age and the Predicted Probabilities for the Outcome of diabetes. A logistic regression fit is added to the plot to indicate how closely age predicts the probabilities according to logistic regression. In this graph, the data does not fit to the chosen logistic regression and displays a more spread out distribution, indicating that a logistic regression relationship between Age and Predicted Probabilities is not appropriate.

[40]:

```r
predicted_plot_glucose <- diabetes_filter %>%
ggplot (aes (x = Glucose, y = predicted_probabilities)) +
geom_point () +
geom_smooth(method = "glm", method.args = list(family = binomial)) +
ylim (0,1) +
ggtitle("Predicted Probabilities of Outcome and Glucose") +
  xlab("Glucose") +
  ylab("Predicted Probability") +
  theme(
    text = element_text(size = 18),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold")
  )

predicted_plot_glucose
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message in eval(family$initialize):
"non-integer #successes in a binomial glm!"
```



This graph indicates a scatterplot between Glucose and the Predicted Probabilities for the Outcome of diabetes. A logistic regression fit is added to the plot to indicate how closely the glucose values predict the probabilities according to logistic regression. In this graph, the data fits closely to the chosen logistic regression, indicating that a logistic regression relationship between Glucose values and Predicted Probabilities is appropriate.

```
[ ]:
```