Applied Machine Learning and Predictive Modelling I: Modelling Stroke Data

Authors: Larissa Eisele, Fabian Lüthard, Yves Maillard

Module: Applied Machine Learning and Predictive Modelling I

Submitted on 10th of June, 2022

Supervisor: Matteo Tanadini and Daniel Meister

Lucerne University of Applied Sciences and Arts

Table of Contents

1 Introduction

Use case: We are a smart watch manufacturer working on a new feature for stroke prevention. We are going to analyze survey data that we plan to ask our users, complementing it with HR (Heart Rate) and CGM (Continuous Glucose Monitoring) data that our product already measures. We hope that our feature can prevent serious health issues and motivate our users to adopt healthier lifestyles.

2 Importing Data

```r
stroke_data <- read_csv('./data/healthcare-dataset-stroke-data.csv')
```

```
## Rows: 5110 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): gender, ever_married, work_type, Residence_type, bmi, smoking_status
## dbl (6): id, age, hypertension, heart_disease, avg_glucose_level, stroke
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
s <- spec(stroke_data)
s
```

```
## cols(
##   id = col_double(),
##   gender = col_character(),
##   age = col_double(),
##   hypertension = col_double(),
##   heart_disease = col_double(),
##   ever_married = col_character(),
##   work_type = col_character(),
##   Residence_type = col_character(),
##   avg_glucose_level = col_double(),
##   bmi = col_character(),
##   smoking_status = col_character(),
##   stroke = col_double()
## )
```

```r
cols_condense(s)
```

```
## cols(
##   .default = col_character(),
##   id = col_double(),
##   age = col_double(),
##   hypertension = col_double(),
##   heart_disease = col_double(),
##   avg_glucose_level = col_double(),
##   stroke = col_double()
## )
```

```
stroke_data$bmi <- as.numeric(stroke_data$bmi)
```

```
## Warning: NAs introduced by coercion
```

```
stroke_data$bmi[is.na(stroke_data$bmi)] <- mean(stroke_data$bmi, na.rm = TRUE)
stroke_data$stroke <- as.factor(stroke_data$stroke)
stroke_data$age <- as.integer(stroke_data$age)
stroke_data$smoking_status <- as.factor(stroke_data$smoking_status)
stroke_data$work_type <- as.factor(stroke_data$work_type)
stroke_data$gender <- as.factor(stroke_data$gender)
stroke_data$ever_married <- as.factor(stroke_data$ever_married)
stroke_data$Residence_type <- as.factor(stroke_data$Residence_type)


stroke_data
```

```
## # A tibble: 5,110 x 12
##       id gender   age hypertension heart_disease ever_married work_type
##    <dbl> <fct>  <int>        <dbl>         <dbl> <fct>        <fct>
##  1  9046 Male      67            0             1 Yes          Private
##  2 51676 Female    61            0             0 Yes          Self-employed
##  3 31112 Male      80            0             1 Yes          Private
##  4 60182 Female    49            0             0 Yes          Private
##  5  1665 Female    79            1             0 Yes          Self-employed
##  6 56669 Male      81            0             0 Yes          Private
##  7 53882 Male      74            1             1 Yes          Private
##  8 10434 Female    69            0             0 No           Private
##  9 27419 Female    59            0             0 Yes          Private
## 10 60491 Female    78            0             0 Yes          Private
## # ... with 5,100 more rows, and 5 more variables: Residence_type <fct>,
## #   avg_glucose_level <dbl>, bmi <dbl>, smoking_status <fct>, stroke <fct>
```

2.1 Summaries

```
#testing the effect of non smokers and smokers
count_by_smoke_status <- stroke_data %>%
  select(smoking_status, stroke) %>%
  group_by(smoking_status, stroke) %>%
  summarise( N = n())
```

```
## `summarise()` has grouped output by 'smoking_status'. You can override using the
## `.groups` argument.
```

```
#testing the effect of work type
 count_by_work_type <- stroke_data %>%
   select(work_type, stroke) %>%
   group_by(work_type, stroke) %>%
   summarise( N = n())
```

```
## `summarise()` has grouped output by 'work_type'. You can override using the
## `.groups` argument.
```

```
 # testing the effects of residence type
count_by_Residence_type <- stroke_data %>%
   select(Residence_type, stroke) %>%
   group_by(Residence_type, stroke) %>%
   summarise( N = n())
```

```
## `summarise()` has grouped output by 'Residence_type'. You can override using the
## `.groups` argument.
```

```
 # testing the effects of gender
count_by_gender <- stroke_data %>%
   select(gender, stroke) %>%
   group_by(gender, stroke) %>%
   summarise( N = n())
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.
```

```
 # testing the effects of hypertension
count_by_hypertension <- stroke_data %>%
   select(hypertension, stroke) %>%
   group_by(hypertension, stroke) %>%
   summarise( N = n())
```

```
## `summarise()` has grouped output by 'hypertension'. You can override using the
## `.groups` argument.
```

```
# testing the effects of heart disease
count_by_heart_disease <- stroke_data %>%
   select(heart_disease, stroke) %>%
   group_by(heart_disease, stroke) %>%
   summarise( N = n())
```

```
## `summarise()` has grouped output by 'heart_disease'. You can override using the
## `.groups` argument.
```

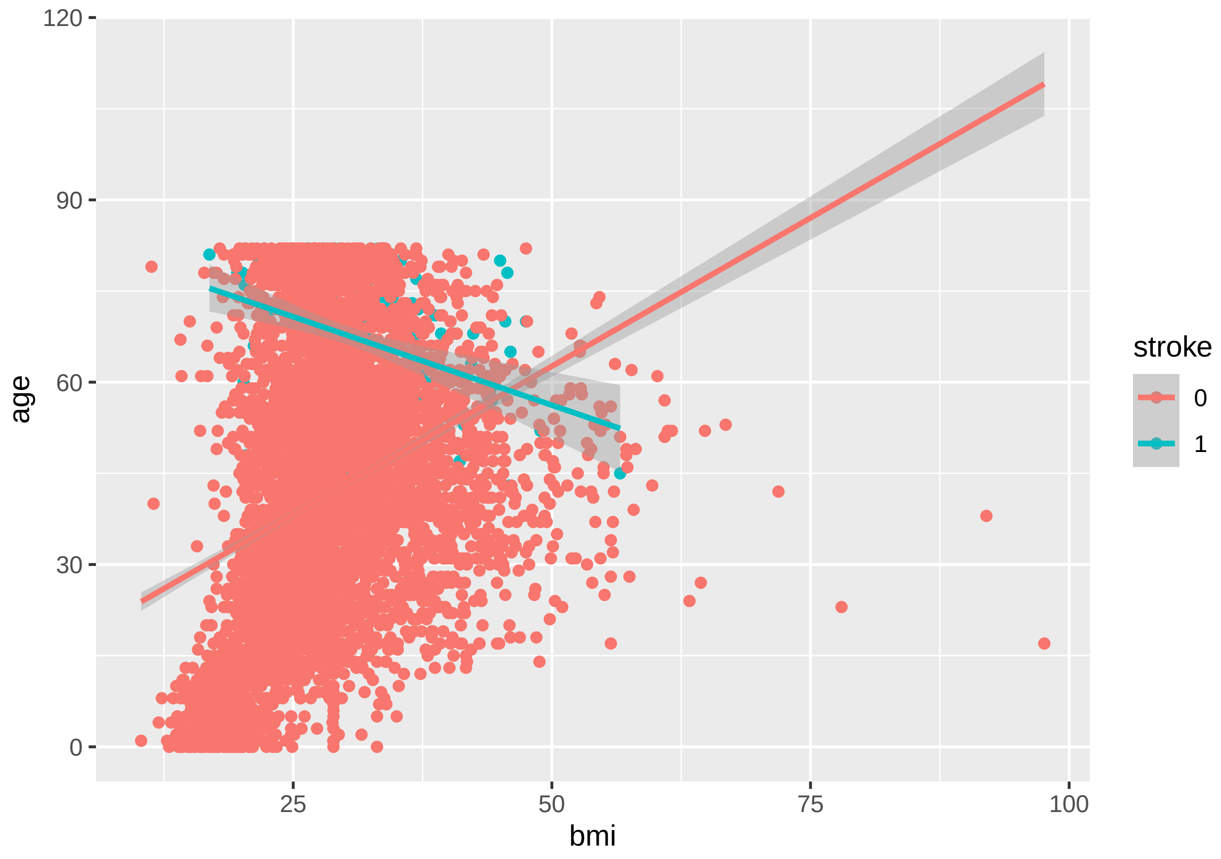
```
# testing the effects of marriage status
count_by_marriage <- stroke_data %>%
   select(ever_married, stroke) %>%
   group_by(ever_married, stroke) %>%
   summarise( N = n())
```

```
## `summarise()` has grouped output by 'ever_married'. You can override using the
## `.groups` argument.
```

2.2 Scatterplots

```
stroke_data %>%
  ggplot(mapping = aes(x = bmi, y = age, color = stroke)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

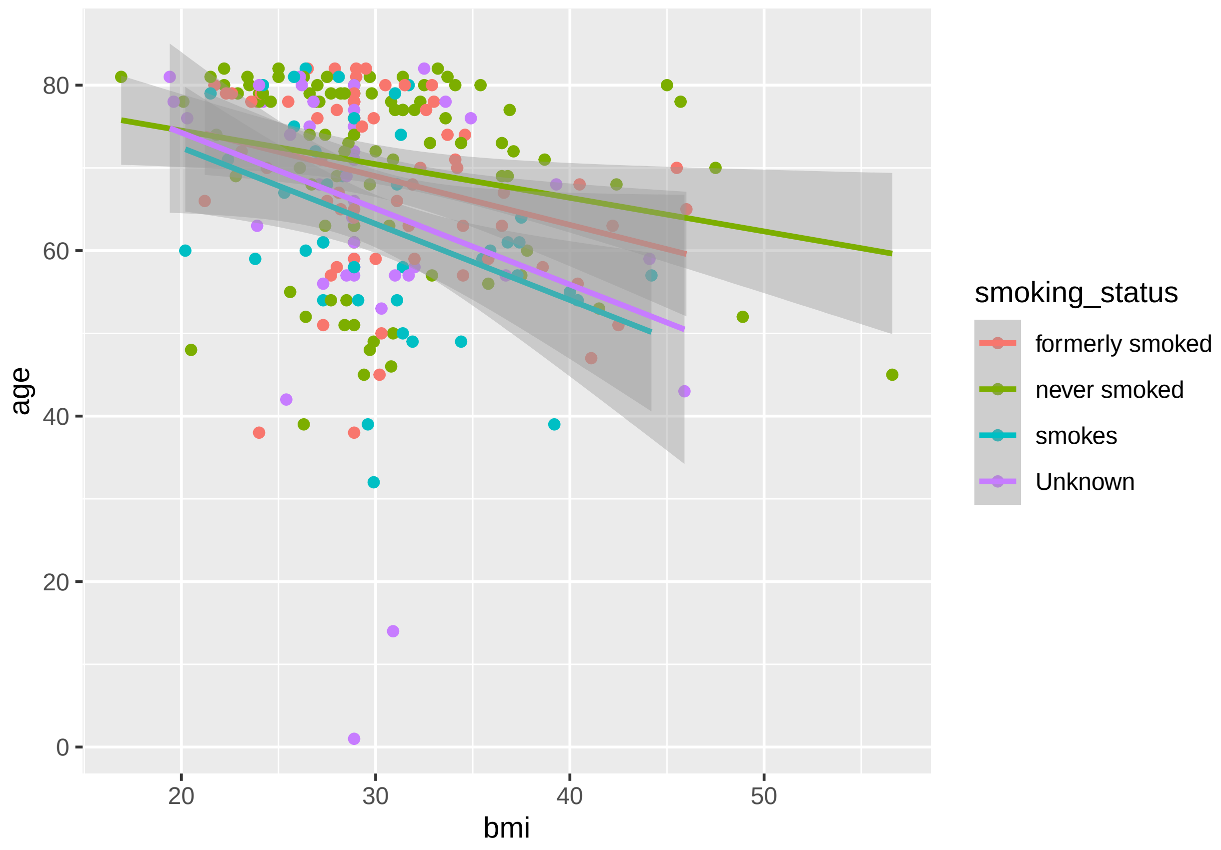
```
## `geom_smooth()` using formula 'y ~ x'
```



```
stroke_data %>%
  filter(stroke == 1) %>%
  ggplot(mapping = aes(x = bmi, y = age, color = smoking_status)) +
  geom_point(method = 'lm') +
  geom_smooth(method = 'lm')
```

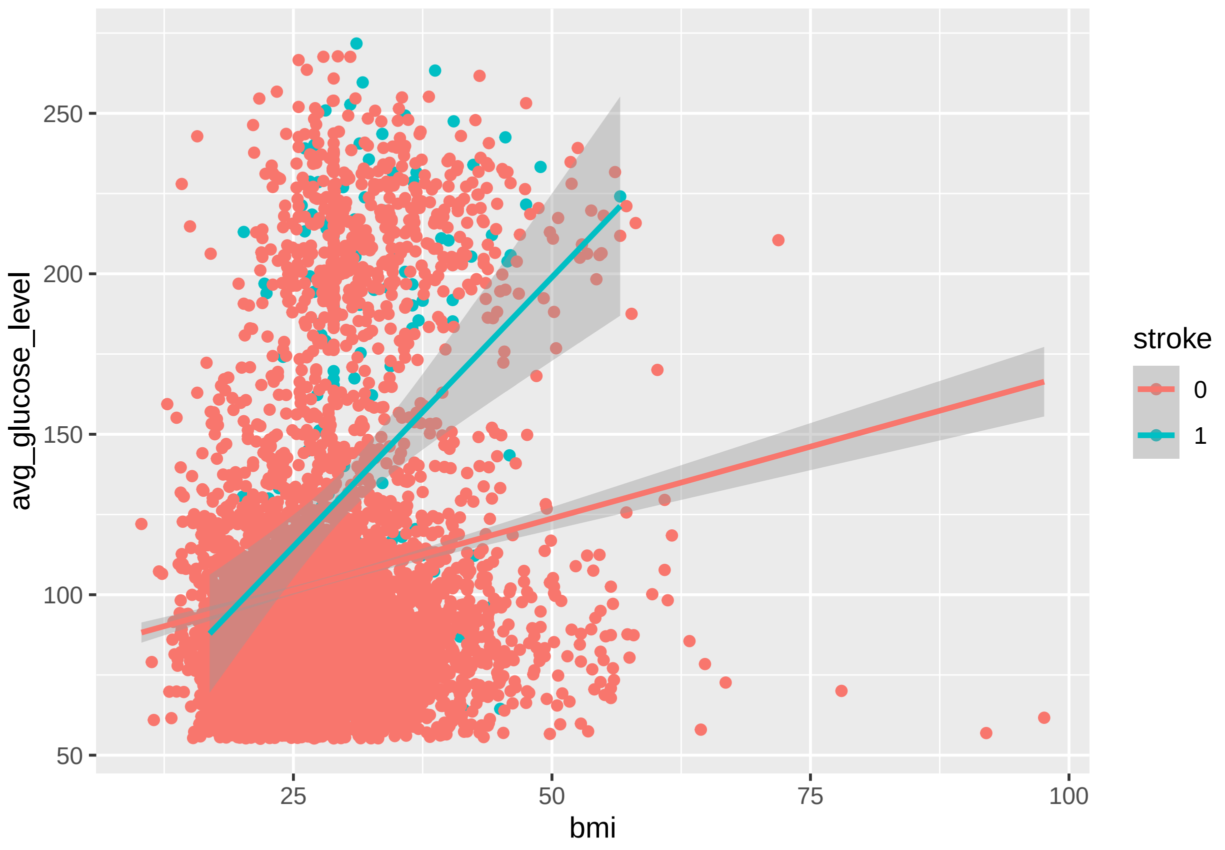
```
## Warning: Ignoring unknown parameters: method
```

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



```
stroke_data %>%
  ggplot(mapping = aes(x = bmi, y = avg_glucose_level, color = stroke)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

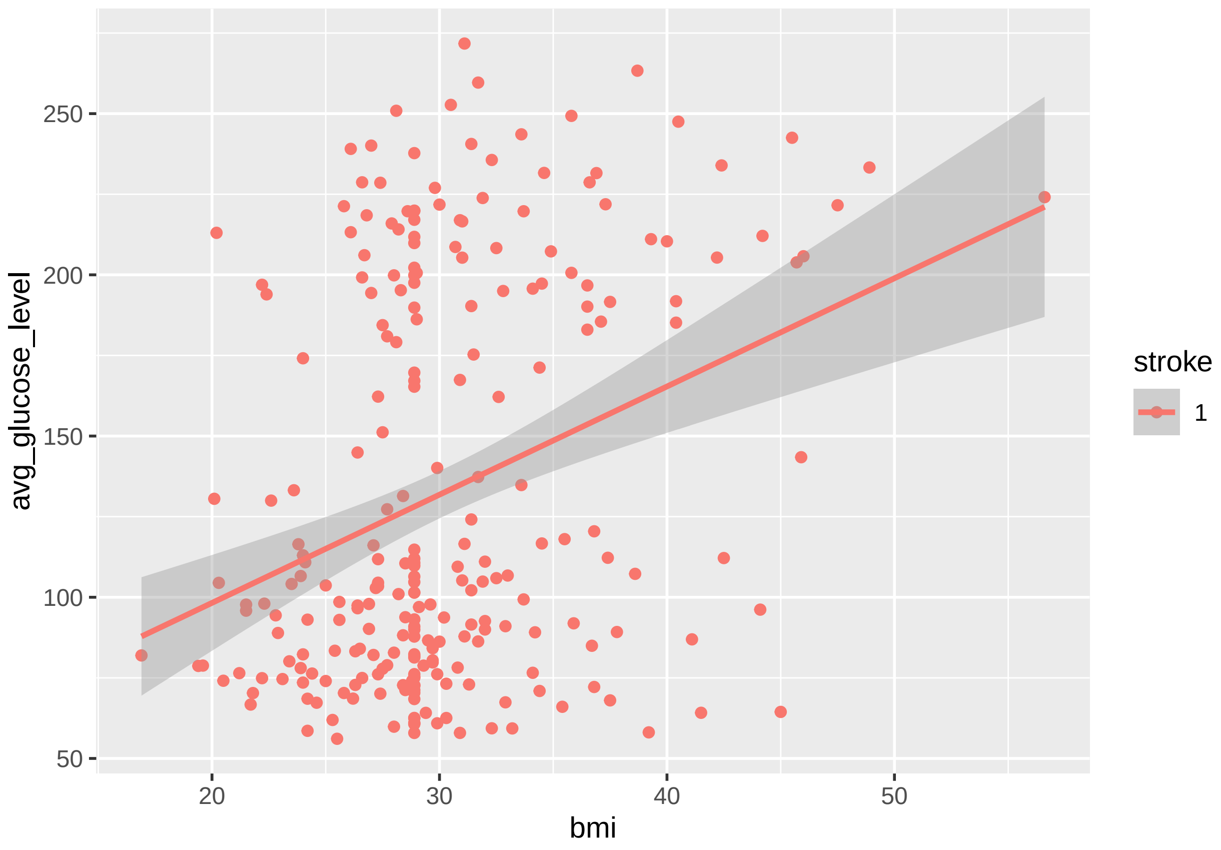
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
stroke_data %>%
  filter(stroke == 1) %>%
  ggplot(mapping = aes(x = bmi, y = avg_glucose_level, color = stroke)) +
  geom_point(method = 'lm') +
  geom_smooth(method = 'lm')
```

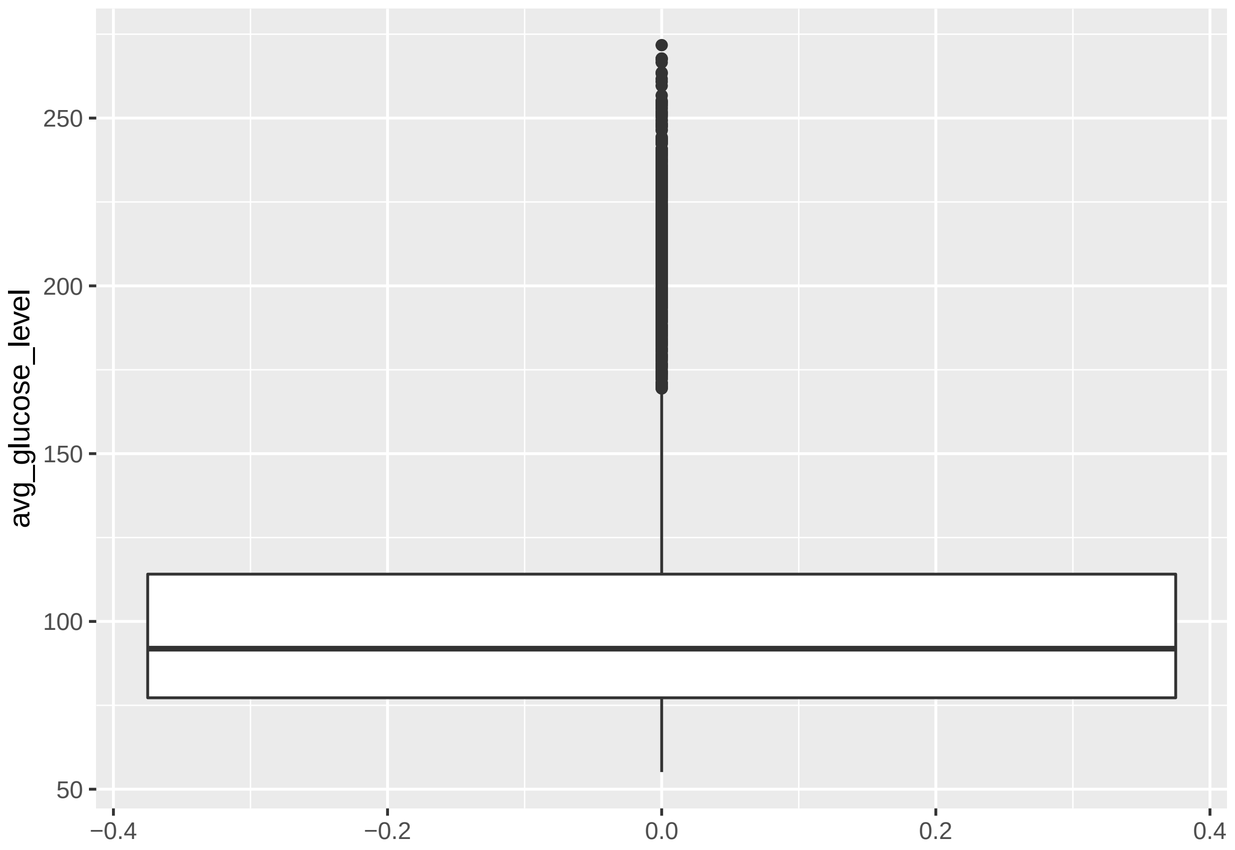
```
## Warning: Ignoring unknown parameters: method
```

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

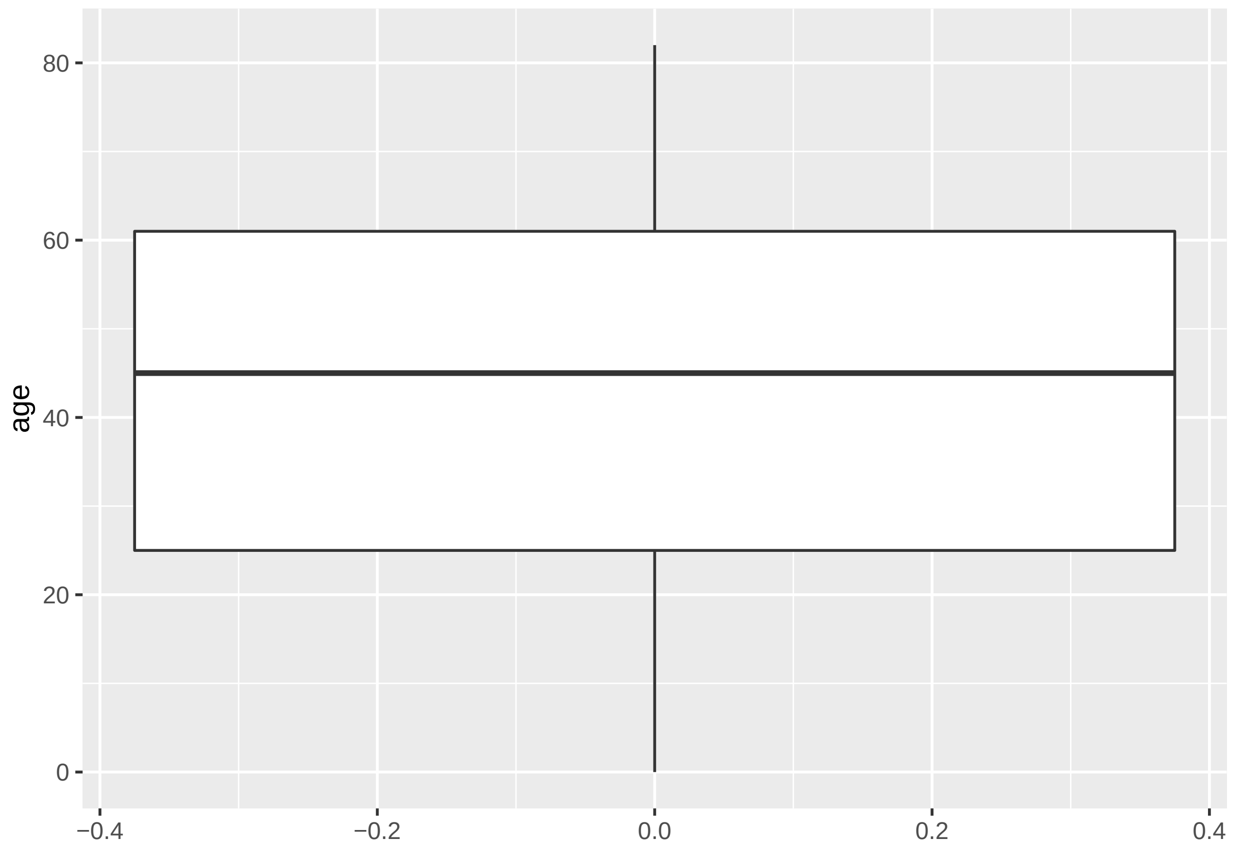


2.3 Boxplots

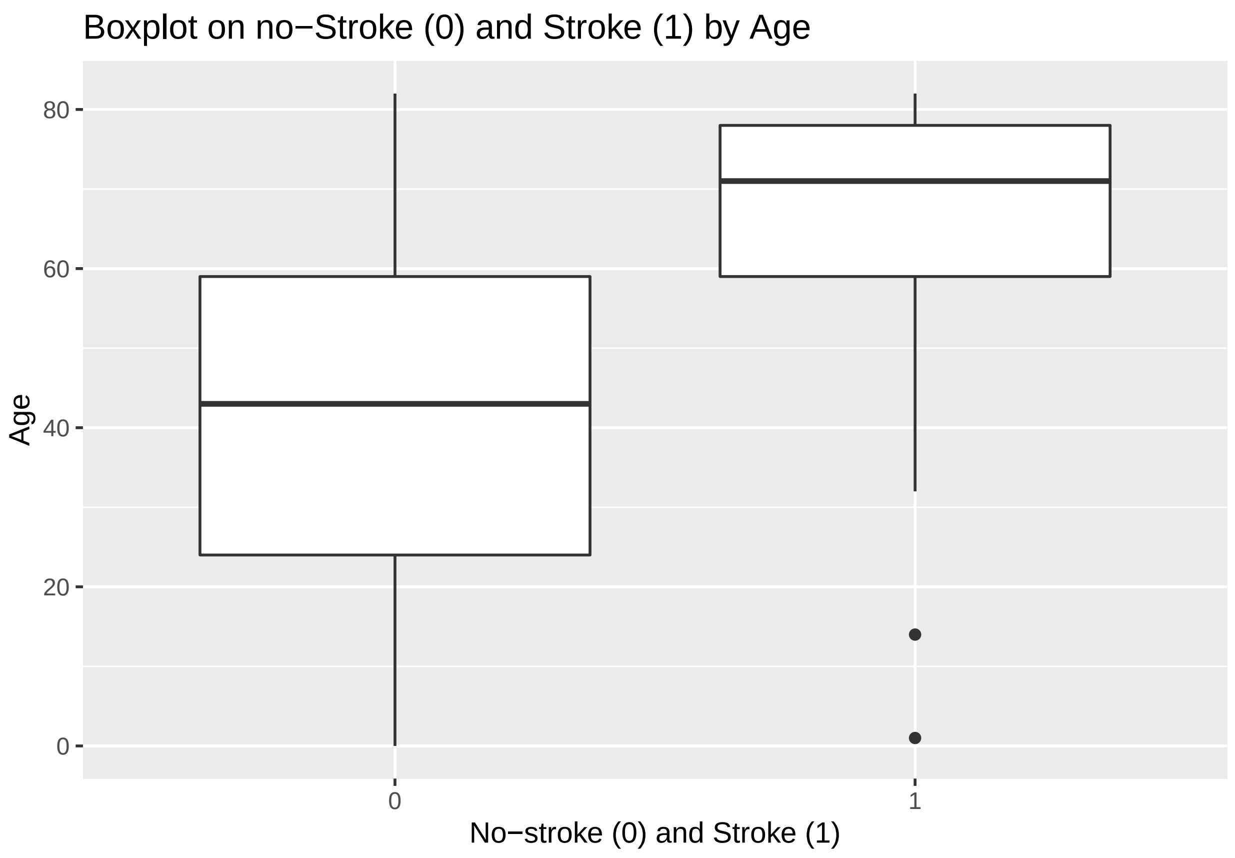
```r
#Boxplot on avg. glucose_level
ggplot(stroke_data, aes(y = avg_glucose_level)) +
  geom_boxplot()
```



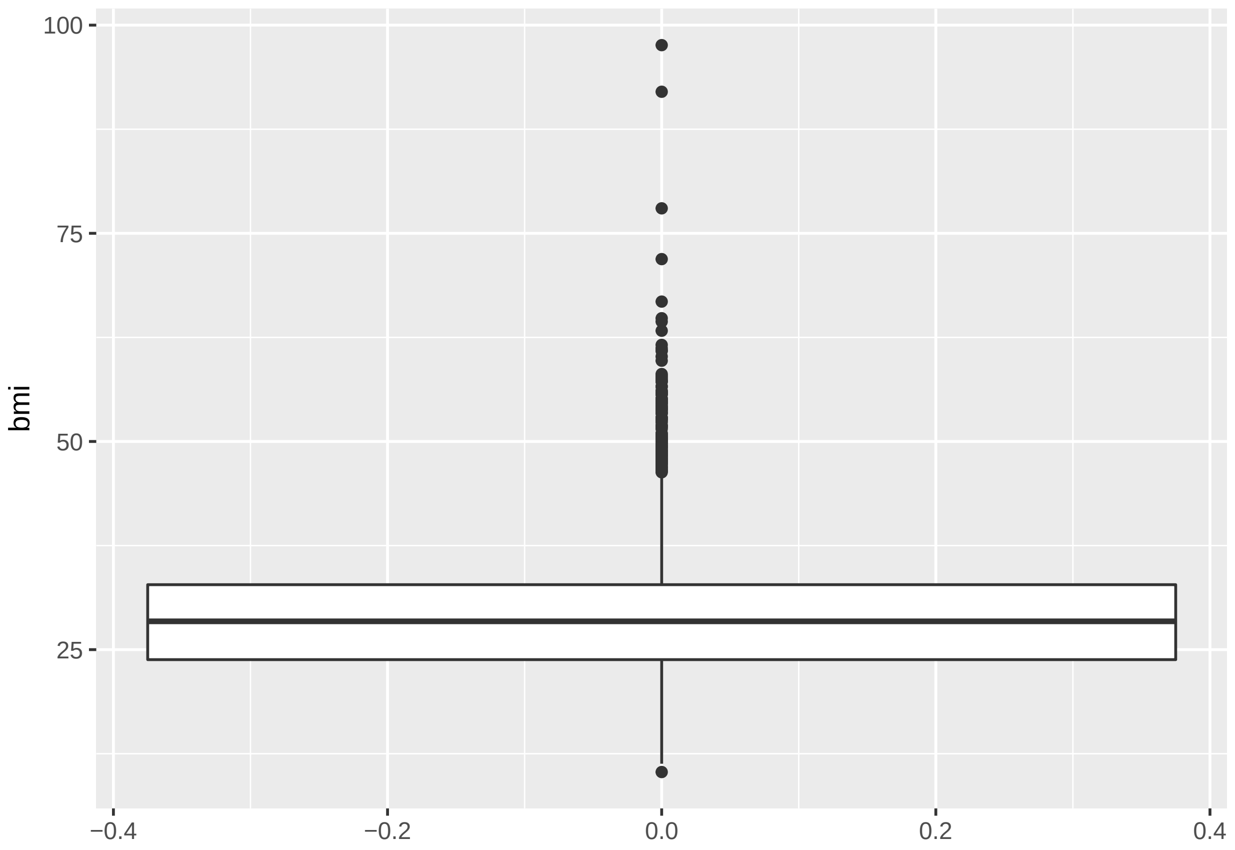
```r
#Boxplot on Age
ggplot(stroke_data, aes(y = age)) +
  geom_boxplot()
```



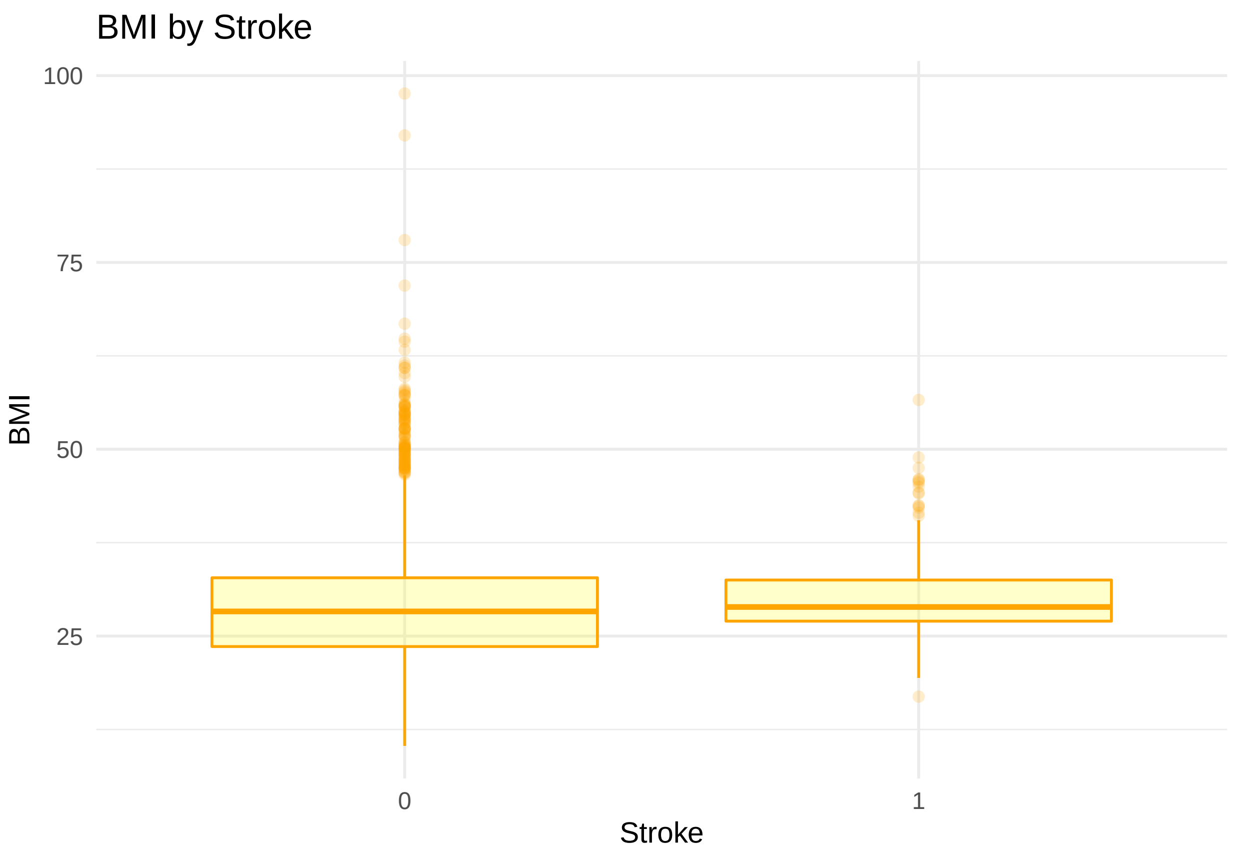
```
stroke_by_age <- stroke_data %>%
# dplyr::filter(stroke == 1) %>%
 ggplot(aes(x = stroke,
            y = age)) +
  geom_boxplot() +
  labs(title = "Boxplot on no-Stroke (0) and Stroke (1) by Age",
       x = "No-stroke (0) and Stroke (1)",
       y = "Age")
       color = "green"

stroke_by_age
```



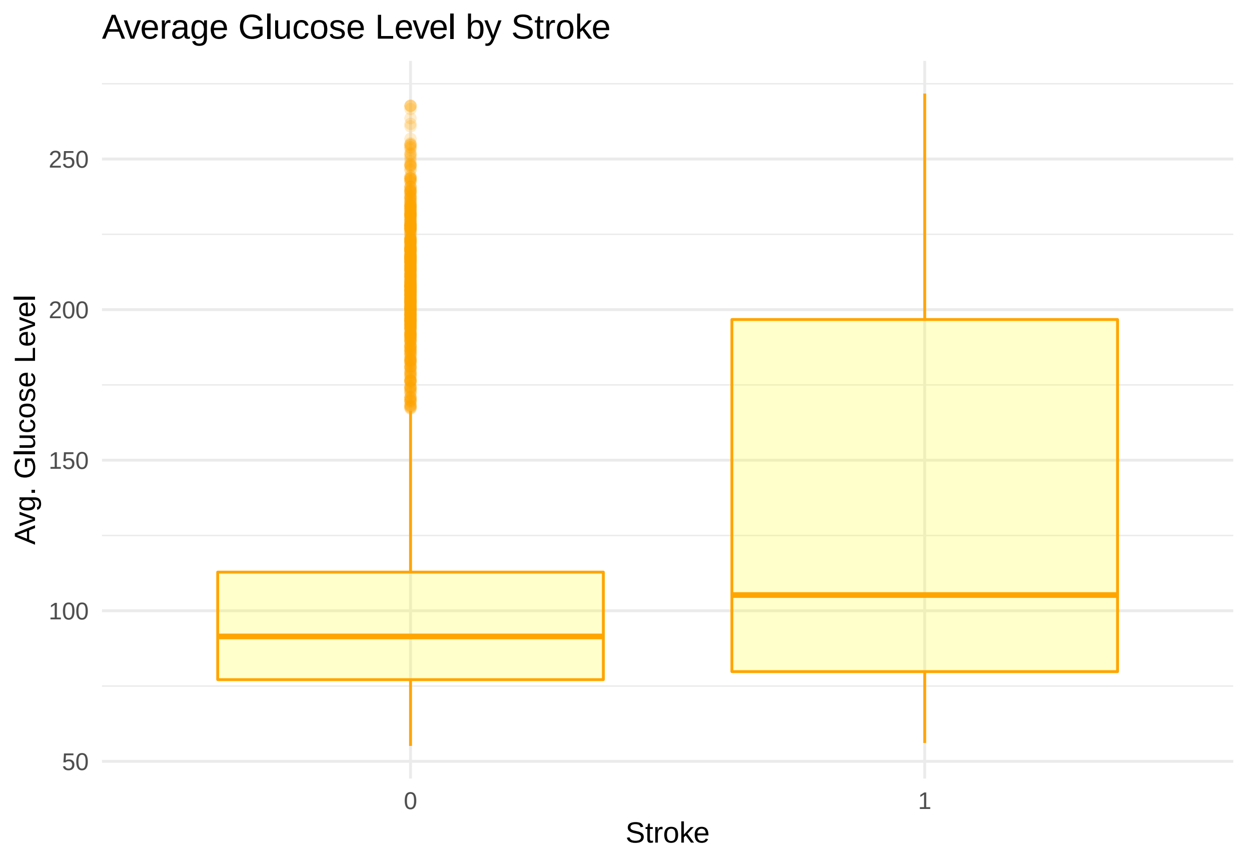
```
#Boxplot on BMI
ggplot(stroke_data, aes(y = bmi)) +
geom_boxplot()
```



```
stroke_by_bmi <- stroke_data %>%
# dplyr::filter(stroke == 1) %>%
 ggplot(aes(x = stroke,
            y = bmi)) +
  geom_boxplot(color="orange", fill="yellow", alpha=0.2) +
  ggtitle("BMI by Stroke") +
  xlab("Stroke") + ylab("BMI") +
  theme_minimal() + theme(axis.text.x = element_text(angle = 0))
stroke_by_bmi
```



```r
#boxplot by avg_glucose_level and stroke
stroke_by_avg_glucose_level <- stroke_data %>%
# dplyr::filter(stroke == 1) %>%
 ggplot(aes(x = stroke,
            y = avg_glucose_level)) +
  geom_boxplot(color="orange", fill="yellow", alpha=0.2) +
  ggtitle("Average Glucose Level by Stroke") +
  xlab("Stroke") + ylab("Avg. Glucose Level") +
  theme_minimal() + theme(axis.text.x = element_text(angle = 0))

stroke_by_avg_glucose_level
```



3 Methodology

```
set.seed(7406)
n=dim(stroke_data[1])  # number of observations in dataset
n_train=0.70*n  # training set is 70%
flag = sort(sample(1:n, size=n_train, replace=FALSE))
```

```
## Warning in 1:n: numerical expression has 2 elements: only the first used
```

```
# Use df (all data points without ID column) df_train, and df_test
# Gender, hypertension, heart disease, ever married, work type, residence type, smoking status, and str
# This should allow for the best modeling options possible for our methods.
df_train = stroke_data[flag,]
df_test = stroke_data[-flag,]
```

```
head(df_test)
```

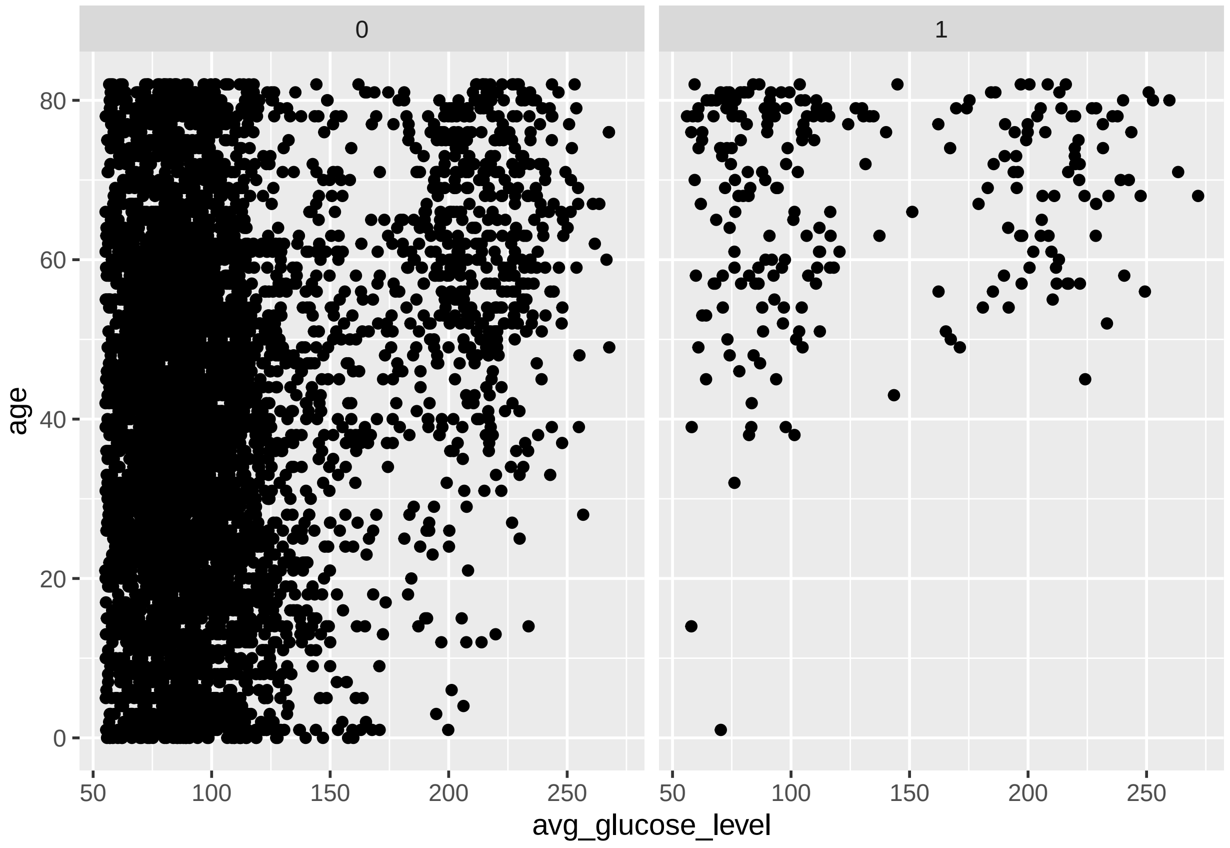
```
## # A tibble: 6 x 12
##      id gender   age hypertension heart_disease ever_married work_type
##   <dbl> <fct>  <int>        <dbl>         <dbl> <fct>        <fct>
## 1 51676 Female    61            0             0 Yes          Self-employed
## 2 10434 Female    69            0             0 No           Private
## 3 60491 Female    78            0             0 Yes          Private
## 4 12095 Female    61            0             1 Yes          Govt_job
## 5 58202 Female    50            1             0 Yes          Self-employed
## 6 56112 Male      64            0             1 Yes          Private
## # ... with 5 more variables: Residence_type <fct>, avg_glucose_level <dbl>,
## #   bmi <dbl>, smoking_status <fct>, stroke <fct>
```

```
head(df_train)
```

```
## # A tibble: 6 x 12
##      id gender   age hypertension heart_disease ever_married work_type
##   <dbl> <fct>  <int>        <dbl>         <dbl> <fct>        <fct>
## 1  9046 Male      67            0             1 Yes          Private
## 2 31112 Male      80            0             1 Yes          Private
## 3 60182 Female    49            0             0 Yes          Private
## 4  1665 Female    79            1             0 Yes          Self-employed
## 5 56669 Male      81            0             0 Yes          Private
## 6 53882 Male      74            1             1 Yes          Private
## # ... with 5 more variables: Residence_type <fct>, avg_glucose_level <dbl>,
## #   bmi <dbl>, smoking_status <fct>, stroke <fct>
```

4 Linear Model

```
# qplot -> age/glucose level per stroke
library(ggplot2)
 qplot(y = age, x = avg_glucose_level,
        data = stroke_data,
        facets = ~ stroke)
```



```
# fitting models for simple regression model
 lm.stroke <- lm(age ~ avg_glucose_level, data = stroke_data)
 summary(lm.stroke)
```

```
##
## Call:
## lm(formula = age ~ avg_glucose_level, data = stroke_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.362 -16.762   1.127  16.638  44.662
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       30.584911   0.783855   39.02   <2e-16 ***
## avg_glucose_level  0.118988   0.006792   17.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.99 on 5108 degrees of freedom
## Multiple R-squared:  0.05667,    Adjusted R-squared:  0.05649
## F-statistic: 306.9 on 1 and 5108 DF,  p-value: < 2.2e-16
```

```
# Generalised Linear Model with family set to Poisson
```

```
‘‘‘r
glm.stroke <- glm(age ~ smoking_status,
family = "poisson",
data = stroke_data)
summary(glm.stroke)
```

```
##
## Call:
## glm(formula = age ~ smoking_status, family = "poisson", data = stroke_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.7708  -3.0398  -0.1092   2.1971   7.7616
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   4.006059   0.004535  883.27   <2e-16 ***
## smoking_statusnever smoked -0.161357   0.005646  -28.58   <2e-16 ***
## smoking_statussmokes       -0.153864   0.006891  -22.33   <2e-16 ***
## smoking_statusUnknown      -0.598470   0.006482  -92.32   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 72137  on 5109  degrees of freedom
## Residual deviance: 61972  on 5106  degrees of freedom
## AIC: 89244
##
## Number of Fisher Scoring iterations: 5
```

```r
#glm.stroke <- glm(bmi ~ stroke,
#family = "poisson",
#data = stroke_data)
#summary(glm.stroke)
```

5 Generalised Linear Model with family set to Binomial Fabian

6 Generalised Additive Model Fabian

7 Neural Network Yves

8 Support Vector Machine (Larissa)

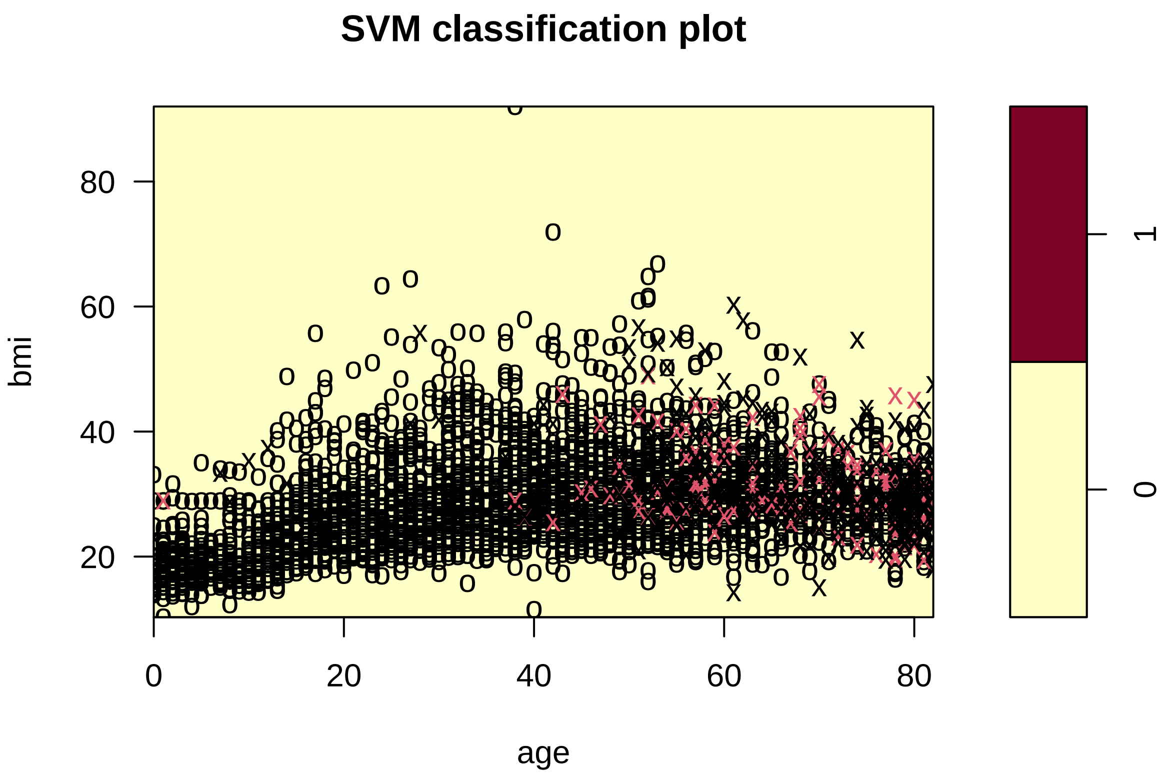
Stroke Data Classification using a Support Vector Machine.

```r
ytrain = df_train$stroke
ytest = df_test$stroke
```

```r
svm_model <- svm(stroke ~. , data = df_train, type = "C-classification", kernel = "radial", cost = "5")
summary(svm_model)
```

```
##
## Call:
## svm(formula = stroke ~ ., data = df_train, type = "C-classification",
##     kernel = "radial", cost = "5")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  5
##
## Number of Support Vectors:  600
##
##  ( 182 418 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

```r
plot(svm_model, data = df_train, bmi ~ age, slice = list(avg_glucose_level = 3))
```



```
#svm_training_prediction <- predict(svm_model, newdata = df_train)
#svm_training_error <- mean(svm_training_prediction != ytrain)
#draw_confusion_matrix(confusionMatrix(svm_training_prediction,df_train$stroke), "Stroke", "No Stroke")
#svm_training_error
#confusionMatrix(svm_training_prediction,df_train$stroke)
```

```
#svm_prediction <- predict(svm_model, newdata = df_test)
#svm_test_error <- mean(svm_prediction != ytest)
#draw_confusion_matrix(confusionMatrix(svm_prediction,df_test$stroke), "Stroke", "No Stroke")
#svm_test_error
```

9 OPTIONAL solve an optimisation problem