Analysis of Stroke Prdiction Dataset

Author: Jared Murphy

Introduction

The data set being used is titled "Stroke Prediction Dataset" and it was obtained from the URL https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. This data set was released by McKinsey and Company for a hack-a-thon in 2018. It has since been used for several studies, two of which can be found on both Science Direct (https://www.sciencedirect.com/science/article/pii/S2772442522000090?via%3Dihub#fn3) and the NIH websites (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8641997/). The veracity of the data has been confirmed beyond a reasonable doubt.

Pre-processing

Data Import

```
df_stroke <- read.csv("~/Desktop/healthcare-dataset-stroke-data.csv")
str(df_stroke)
```

```
'data.frame':   5110 obs. of  12 variables:
 $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
 $ gender           : chr  "Male" "Female" "Male" "Female" ...
 $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
 $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
 $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

Data Cleaning

```r
suppressPackageStartupMessages(library(dplyr))
df_stroke$id <- NULL
df_stroke <- df_stroke |>
  filter(gender != "Other")
df_stroke$gender <- as.factor(df_stroke$gender)
df_stroke$hypertension <- as.factor(df_stroke$hypertension)
levels(df_stroke$hypertension) <- c("No", "Yes")
df_stroke$heart_disease <- as.factor(df_stroke$heart_disease)
levels(df_stroke$heart_disease) <- c("No", "Yes")
df_stroke$ever_married <- as.factor(df_stroke$ever_married)
df_stroke$work_type <- as.factor(df_stroke$work_type)
levels(df_stroke$work_type) <- c("Children", "Govt", "Never", "Private", "Self")
df_stroke$Residence_type <- as.factor(df_stroke$Residence_type)
df_stroke <- df_stroke |>
  filter(bmi != "N/A")
df_stroke$bmi <- as.numeric(df_stroke$bmi)
df_stroke$smoking_status <- factor(df_stroke$smoking_status)
levels(df_stroke$smoking_status) = c("Unknown", "Former", "Never", "Smokes")
df_stroke$stroke <- as.factor(df_stroke$stroke)
levels(df_stroke$stroke) <- c("No", "Yes")
str(df_stroke)
```

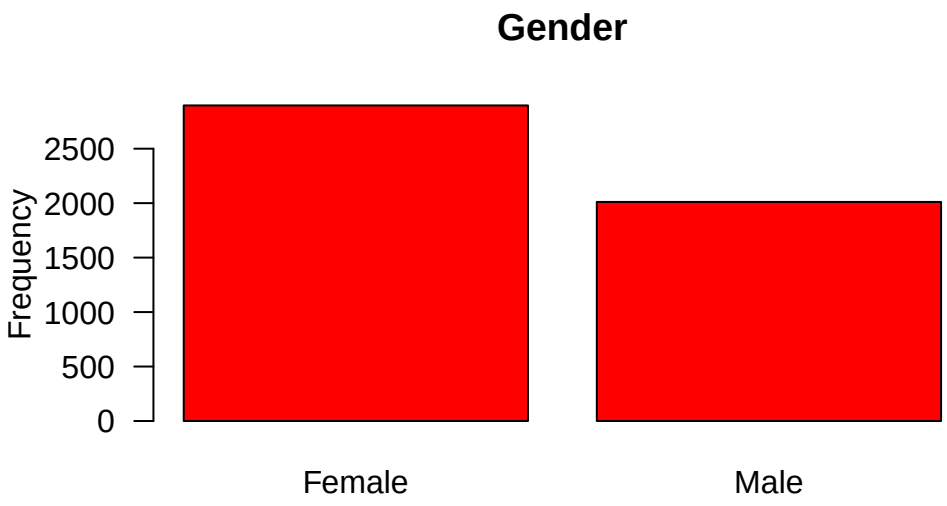
```
'data.frame':   4908 obs. of  11 variables:
 $ gender           : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 2 1 1 1 1 ...
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
 $ heart_disease    : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 2 1 1 1 2 ...
 $ ever_married     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 2 2 2 ...
 $ work_type        : Factor w/ 5 levels "Children","Govt",..: 4 4 4 5 4 4 4 4 4 2 ...
 $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 1 2 2 1 1 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
 $ smoking_status   : Factor w/ 4 levels "Unknown","Former",..: 1 2 3 2 1 2 2 4 2 3 ...
 $ stroke           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
```

Summary Statistics

```
factor_summary <- function(Factors, title, color, las_value = 1, make_mean = TRUE){
  counts <- table(Factors)
  if (make_mean == TRUE){
    MEAN <- mean(as.numeric(Factors))
  } else {
    MEAN <- "N/A"
  }
  mean_mode <- data.frame(MEAN = MEAN, MODE = names(which.max(counts)))
  barplot(counts, col = color, main = title, ylab = "Frequency", las = las_value)
  str(Factors)
  return(list(counts,mean_mode))
}

numeric_summary <- function(column, title, color) {
  str(column)
  summary(column)
  boxplot(column,
        main = title,
        ylab = "Values",
        col = color)
  return(summary(column))
}
```

Gender

```
x <- factor_summary(df_stroke$gender, "Gender", "red")
```




```
 Factor w/ 2 levels "Female","Male": 2 2 1 1 2 2 1 1 1 1 ...
```

```
  x[[1]]
```

```
Factors
Female   Male
  2897   2011
```

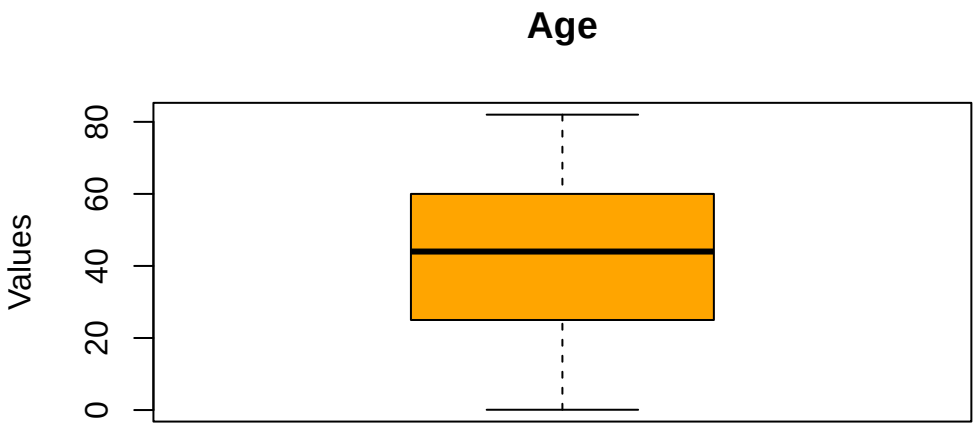
```
  x[[2]]
```

```
      MEAN   MODE
1 1.409739 Female
```

Age

```
numeric_summary(df_stroke$age, "Age", "orange")
```

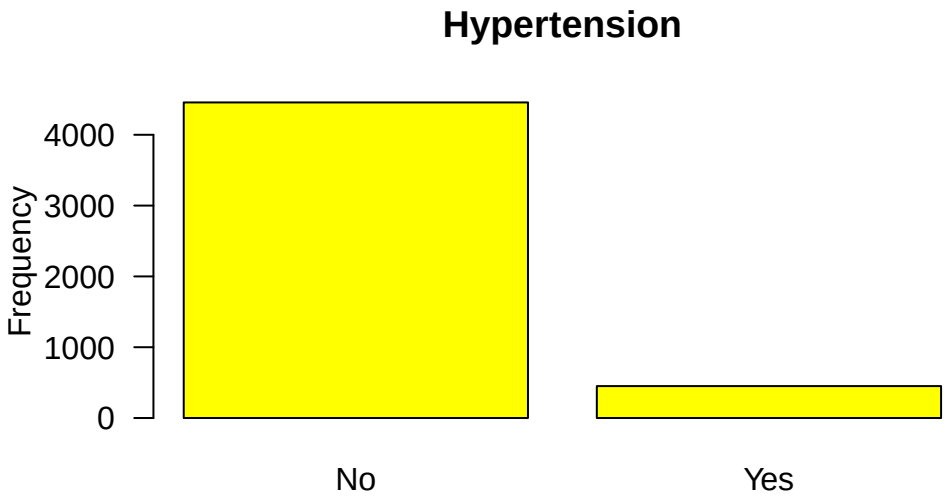
```
num [1:4908] 67 80 49 79 81 74 69 78 81 61 ...
```




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.08   25.00   44.00   42.87   60.00   82.00
```

Hypertension

```
x <- factor_summary(df_stroke$hypertension, "Hypertension", "yellow")
```




```
 Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
```

```
x[[1]]
```

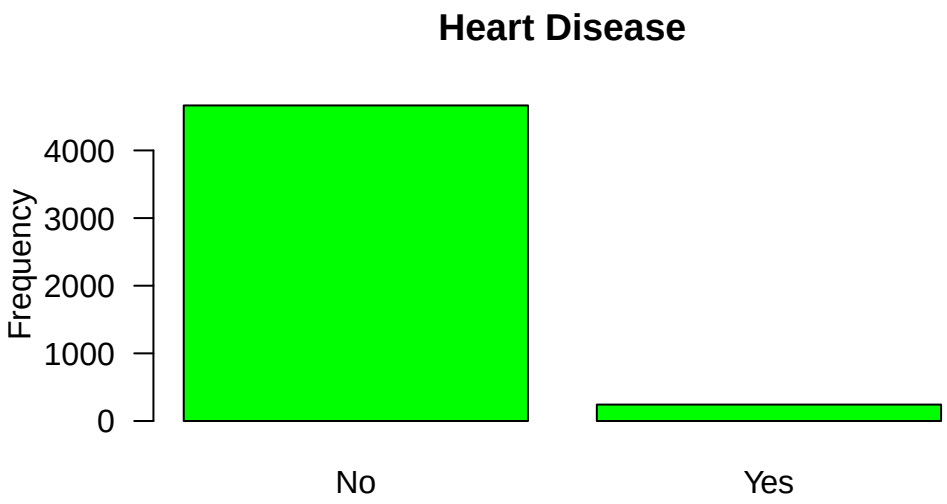
```
Factors
  No  Yes
4457  451
```

```
x[[2]]
```

```
      MEAN MODE
1 1.091891   No
```

Heart Disease

```
x <- factor_summary(df_stroke$heart_disease, "Heart Disease", "green")
```




```
 Factor w/ 2 levels "No","Yes": 2 2 1 1 1 2 1 1 1 2 ...
```

```
  x[[1]]
```

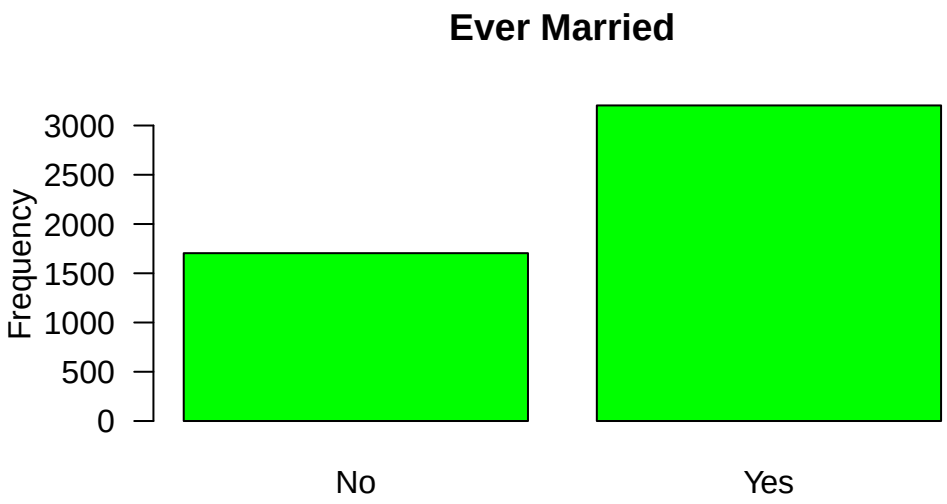
```
Factors
  No  Yes
4665  243
```

```
  x[[2]]
```

```
      MEAN MODE
1 1.049511   No
```

Ever Married

```
  x <- factor_summary(df_stroke$ever_married, "Ever Married", "green")
```




```
 Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 2 2 2 ...
```

```
x[[1]]
```

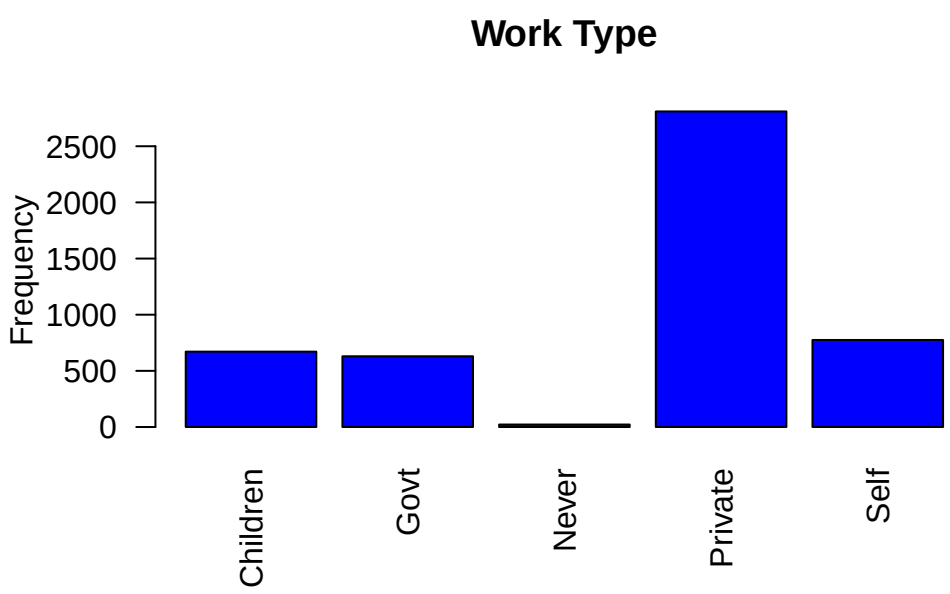
```
Factors
  No  Yes
1704 3204
```

```
x[[2]]
```

```
     MEAN MODE
1 1.652812  Yes
```

Work Type

```
x <- factor_summary(df_stroke$work_type, "Work Type", "blue",2, FALSE)
```




```
 Factor w/ 5 levels "Children","Govt",..: 4 4 4 5 4 4 4 4 4 2 ...
```

```
x[[1]]
```

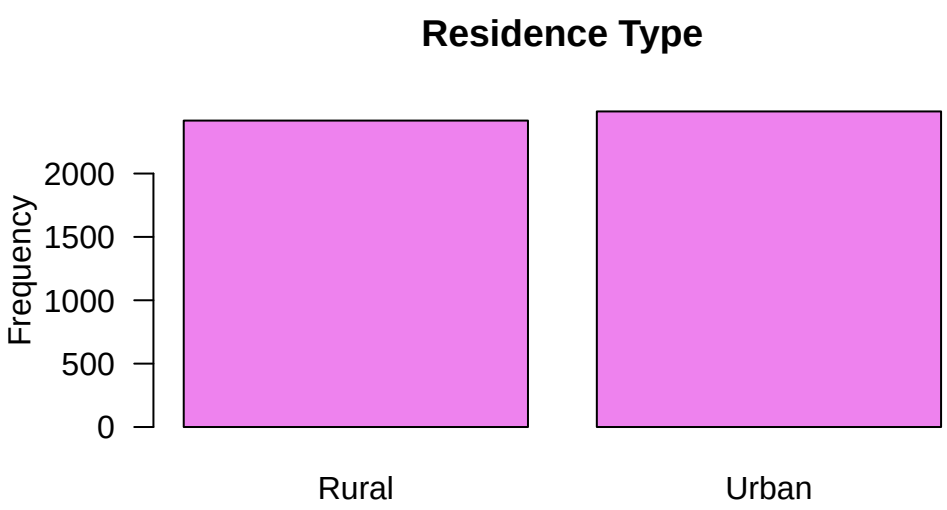
```
Factors
Children     Govt    Never  Private     Self
     671      630       22     2810      775
```

```
x[[2]]
```

```
  MEAN    MODE
1  N/A Private
```

Residence Type

```
x <- factor_summary(df_stroke$Residence_type, "Residence Type", "violet")
```




```
 Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 1 2 2 1 1 ...
```

```
  x[[1]]
```

```
Factors
Rural Urban
 2418  2490
```

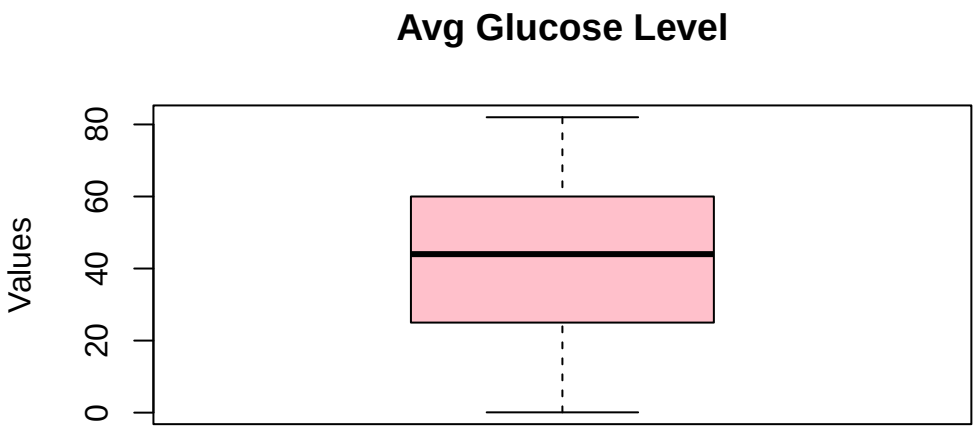
```
  x[[2]]
```

```
      MEAN  MODE
1 1.507335 Urban
```

Average Glucose Level

```
  numeric_summary(df_stroke$age, "Avg Glucose Level", "pink")
```

```
 num [1:4908] 67 80 49 79 81 74 69 78 81 61 ...
```

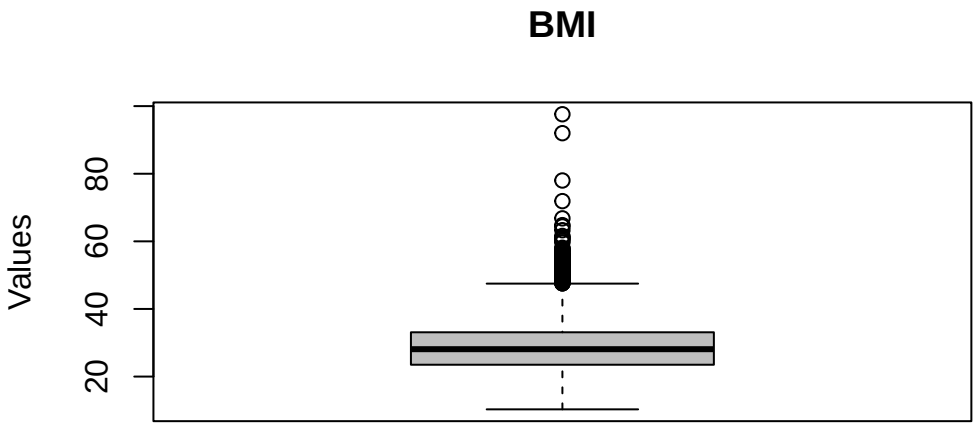



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.08   25.00   44.00   42.87   60.00   82.00
```

BMI

```
numeric_summary(df_stroke$bmi, "BMI", "grey")
```

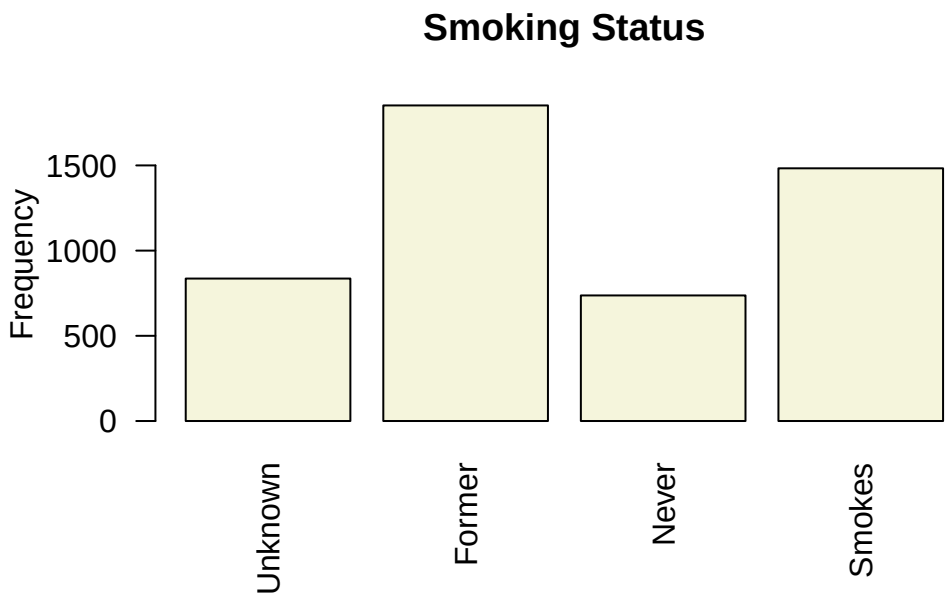
```
num [1:4908] 36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
```




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.30   23.50   28.10   28.89   33.10   97.60
```

Smoking Status

```
x <- factor_summary(df_stroke$smoking_status, "Smoking Status", "beige", 2, FALSE )
```




```
 Factor w/ 4 levels "Unknown","Former",..: 1 2 3 2 1 2 2 4 2 3 ...
```

```
x[[1]]
```

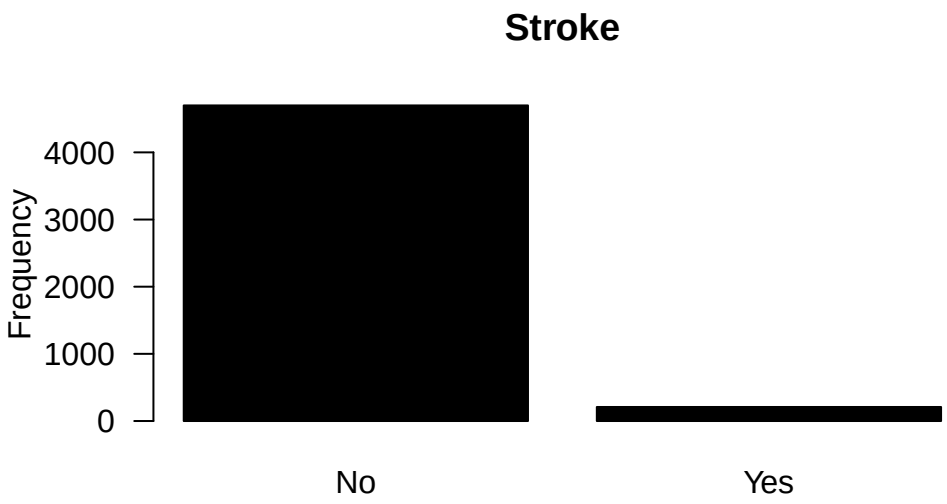
```
Factors
Unknown  Former   Never  Smokes
    836    1852     737    1483
```

```
x[[2]]
```

```
  MEAN   MODE
1  N/A Former
```

Stroke

```
x <- factor_summary(df_stroke$stroke, "Stroke", "black")
```




```
 Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
```

```
  x[[1]]
```

```
Factors
  No  Yes
4699  209
```

```
  x[[2]]
```

```
      MEAN MODE
1 1.042584   No
```

One Sample Hypothesis Test: BMI

The CDC stated in a National Health and Nutrition Examination Survey, found at URL https://www.cdc.gov/nchs/data/nhanes/databriefs/adultweight.pdf, that "The percent of persons who are overweight or obese, with a BMI of 25.0 or higher, increased from 56 percent in 1988 94 to 64 percent in 1999 2000." The stroke prediction data set that we are analyzing

was released in 2018, and we can assume it was collected sometime near the date of its release. Let us test whether the proportion of adults ages 20 and older with a BMI of 25.0 or higher is greater than the 1999-2000 proportion of 64 percent.

1. Claim: circa 2018, the proportion of adults ages 20 and older with a bmi above 25.0 is greater than .64
2. Parameter of interest: p = the proportion of adults with a bmi above 25
3. Ho: p <= .64, Ha: p > .64
4. alpha = .01 (medical)
5. X-squared = 345.87 (see below)
6. p-value < .0001 (see below)
7. Reject Ho
8. At the .01 level of support, there is sufficient evidence to claim that circa 2018 the true proportion of adults with bmi greater than 25 was greater than 64%. This suggests that the proportion of overweight or obese adults has risen since the 1999 2000 measurements.

```
po <- .64
n <- df_stroke |>
  filter(age >= 20) |>
  summarise(count = n())
n <- n[1,1]
x <- df_stroke |>
  filter(age >= 20 & bmi > 25.0) |>
  summarise(count = n())
x <- x[1,1]
result <- prop.test(x, n, p = po, alternative = "greater", conf.level = .99, correct = FAL
result
```

```
    1-sample proportions test without continuity correction

data:  x out of n, null probability po
X-squared = 345.87, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is greater than 0.64
99 percent confidence interval:
 0.7661461 1.0000000
sample estimates:
        p
0.7817861
```

Two Sample Hypothesis Test: BMI - Urban vs. Rural

Our data set is split between urban and rural residents. Given that sample sizes are large, CLT applies and we can assume normality. Given that this data set was collected by a reputable source, we can assume independence within and between groups. Below the group sample standard deviations can be seen to be approximately equal.

1. Claim: The average bmi among rural residents is not equal to that of urban residents
2. Parameter of interest: Mu_rural - Mu_urban
3. Ho: Mu_rural - Mu_urban = 0, Ha: Mu_rural - Mu_urban != 0
4. alpha = .01 (medical)
5. t = .020551 (see below)
6. p-value = .9836 (see below)
7. Fail to Reject Ho
8. At the .01 level of support, there is insufficient evidence to support the claim that bmi is different between urban and rural residents

```
urban <- df_stroke |>
  filter(Residence_type == "Urban")
rural <- df_stroke |>
  filter(Residence_type == "Rural")
paste("urban bmi std dev:", sd(urban$bmi))
```

```
[1] "urban bmi std dev: 7.79298476010291"
```

```
paste("rural bmi std dev:", sd(rural$bmi))
```

```
[1] "rural bmi std dev: 7.91859667154322"
```

```
t.test(rural$bmi, urban$bmi, mu=0, paired=FALSE, var.equal=TRUE, conf.level=0.99)
```

```
    Two Sample t-test

data:  rural$bmi and urban$bmi
t = 0.020551, df = 4906, p-value = 0.9836
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -0.5733051  0.5825233
```

```
sample estimates:
mean of x mean of y
 28.89690  28.89229
```

Multivariate Regression Analysis: bmi

Assumptions

Multicollinearity

After each model is created and fitted, we will measure the Variance Inflation Factor to see if multicollinearity is a problem. If a predictor VIF is above 4 we will remove it.

Homoscedasticity

After each model is created and fitted, we will visually inspect each the fitted values vs. residual plots. If any funnel shapes, curvatures or other patterns are present we will reevaluate the model.

Residuals Normally Distributed

After each model is created and fitted, we will visually analyze a residual histogram to make sure that residuals approximate a bell shaped curve.

Full Model

```
full_model_formula <- as.formula("bmi ~ gender + age + hypertension + heart_disease + ever
print(full_model_formula)
```

```
bmi ~ gender + age + hypertension + heart_disease + ever_married +
    work_type + Residence_type + avg_glucose_level + smoking_status +
    stroke
```

```
full_model <- lm(full_model_formula, data = df_stroke)
summary(full_model)
```

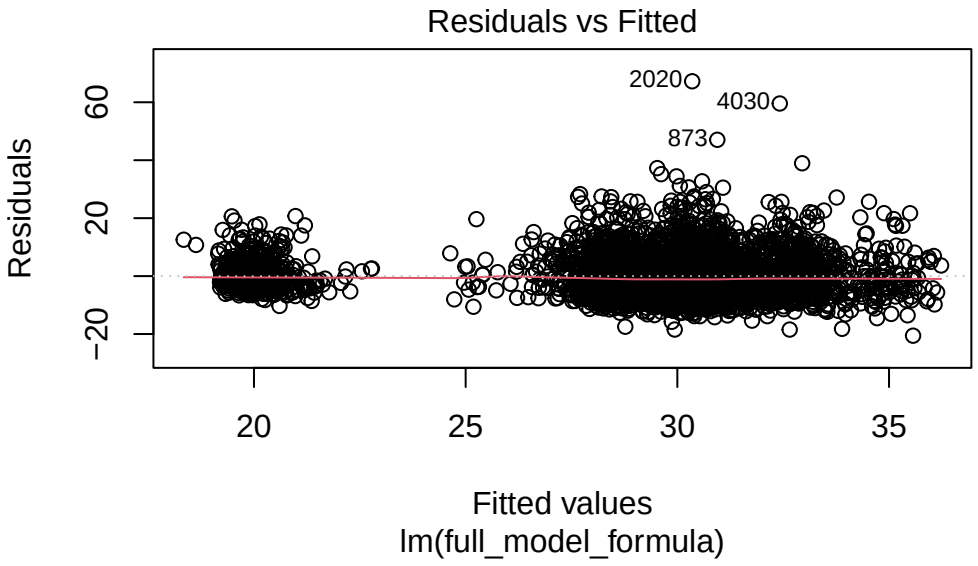
```
Call:
lm(formula = full_model_formula, data = df_stroke)

Residuals:
    Min      1Q  Median      3Q     Max
-20.568  -4.499  -1.137   3.285  67.248

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          18.923235   0.493042  38.381  < 2e-16 ***
genderMale            0.091274   0.201683   0.453 0.650886
age                  -0.014982   0.007471  -2.005 0.044988 *
hypertensionYes       2.534437   0.357859   7.082 1.62e-12 ***
heart_diseaseYes     -0.847645   0.476034  -1.781 0.075033 .
ever_marriedYes       2.116334   0.289364   7.314 3.02e-13 ***
work_typeGovt         8.565594   0.500714  17.107  < 2e-16 ***
work_typeNever        5.419310   1.495213   3.624 0.000293 ***
work_typePrivate      8.586440   0.414796  20.700  < 2e-16 ***
work_typeSelf         8.092861   0.512783  15.782  < 2e-16 ***
Residence_typeUrban  -0.002428   0.196055  -0.012 0.990119
avg_glucose_level     0.020337   0.002309   8.809  < 2e-16 ***
smoking_statusFormer -0.419661   0.290359  -1.445 0.148432
smoking_statusNever  -0.176761   0.350429  -0.504 0.613994
smoking_statusSmokes -0.783291   0.330159  -2.372 0.017708 *
strokeYes            -0.767752   0.505670  -1.518 0.129006
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.858 on 4892 degrees of freedom
Multiple R-squared:  0.2399,    Adjusted R-squared:  0.2375
F-statistic: 102.9 on 15 and 4892 DF,  p-value: < 2.2e-16
```

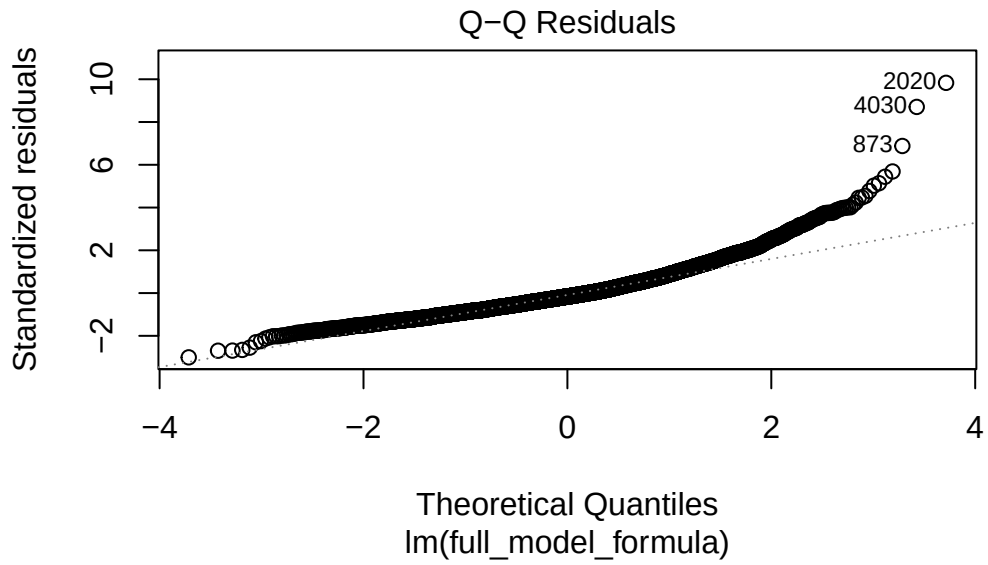
```r
suppressPackageStartupMessages(library(car))
vif(full_model)
```

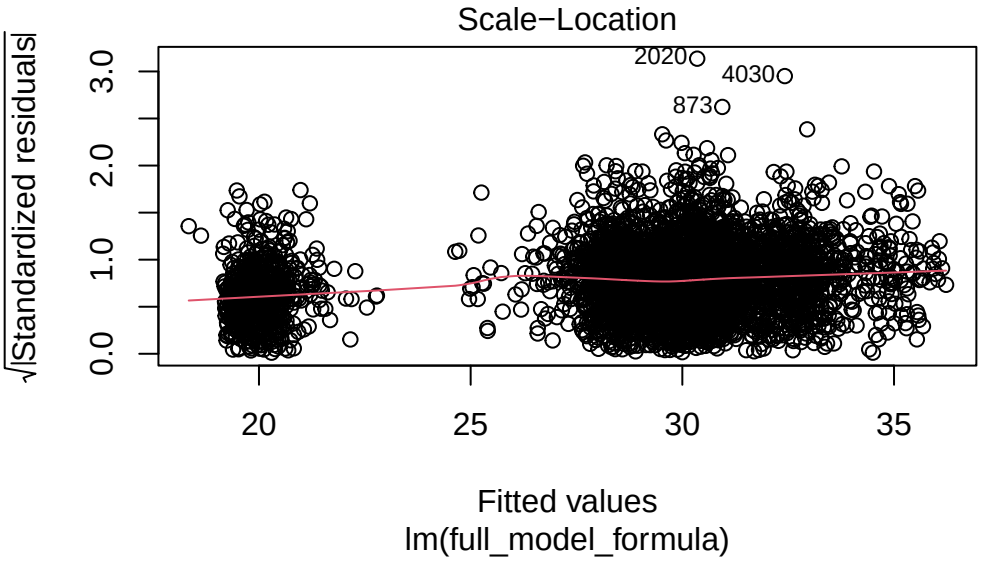
```
                    GVIF Df GVIF^(1/(2*Df))
gender          1.026488  1        1.013157
age             2.962727  1        1.721257
hypertension    1.115055  1        1.055962
heart_disease   1.112725  1        1.054858
```

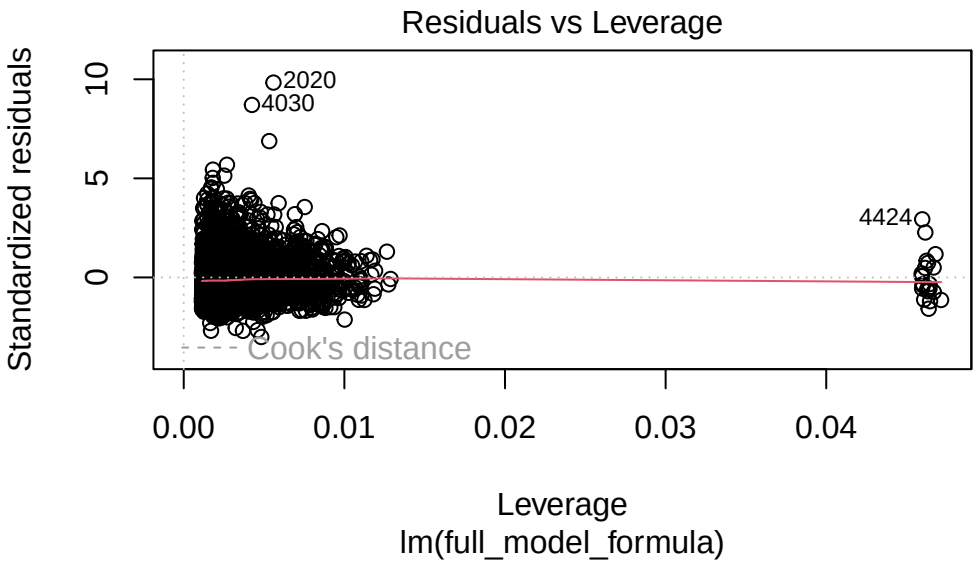
```
ever_married      1.980175  1        1.407187
work_type         2.365950  4        1.113655
Residence_type    1.002457  1        1.001228
avg_glucose_level 1.097308  1        1.047525
smoking_status    1.450265  3        1.063917
stroke            1.087775  1        1.042965
```

```r
plot(full_model)
```




Q−Q Residuals
Standardized residuals
10
6
2
−2
2020
4030
873
−4
−2
0
2
4
Theoretical Quantiles
lm(full_model_formula)


Scale−Location
√|Standardized residuals|
3.0
2.0
1.0
0.0
2020
4030
873
20
25
30
35
Fitted values
lm(full_model_formula)




```
hist(full_model$residuals)
```

The VIF for each variable is below our stated threshold indicating that muticollinearity between independent variables is likely not an issue. The residual plots do not show much of a discernible pattern. There is no funnel shape or curvature. Heteroskedasticity does not appear to be a problem. Several outliers are likely present and could be skewing results. The residual histogram appears to be skewed to the right and is not quite bell shaped. Lets prune the model by eliminating independent variables with insignificant t-tests before interpreting coefficients. We will interpret coefficients in our pruned models.

Pruned Model 1

```
pruned_model1_formula <- as.formula("bmi ~ age + hypertension + ever_married + work_type +
print(pruned_model1_formula)
```

```
bmi ~ age + hypertension + ever_married + work_type + avg_glucose_level +
    smoking_status
```

```
pruned_model1 <- lm(pruned_model1_formula, data = df_stroke)
summary(pruned_model1)
```

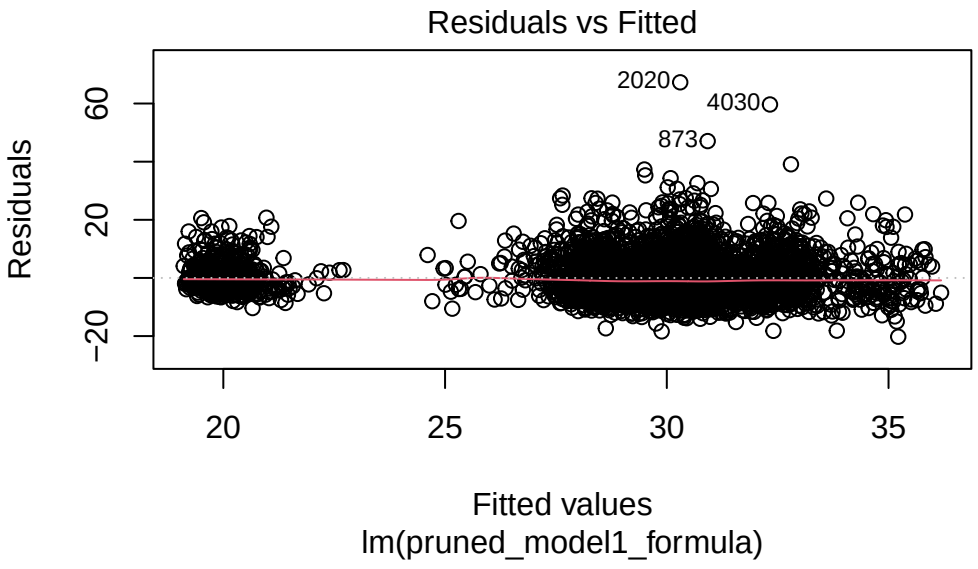
```
Call:
lm(formula = pruned_model1_formula, data = df_stroke)

Residuals:
    Min      1Q  Median      3Q     Max
-20.219  -4.545  -1.099   3.296  67.297

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          19.048103   0.469915  40.535  < 2e-16 ***
age                  -0.020058   0.007164  -2.800 0.005134 **
hypertensionYes       2.481696   0.356904   6.953 4.03e-12 ***
ever_marriedYes       2.178611   0.288252   7.558 4.86e-14 ***
work_typeGovt         8.670064   0.497573  17.425  < 2e-16 ***
work_typeNever        5.468937   1.494952   3.658 0.000257 ***
work_typePrivate      8.669800   0.411990  21.044  < 2e-16 ***
work_typeSelf         8.195878   0.509426  16.088  < 2e-16 ***
avg_glucose_level     0.019765   0.002292   8.625  < 2e-16 ***
smoking_statusFormer -0.414087   0.289661  -1.430 0.152908
```

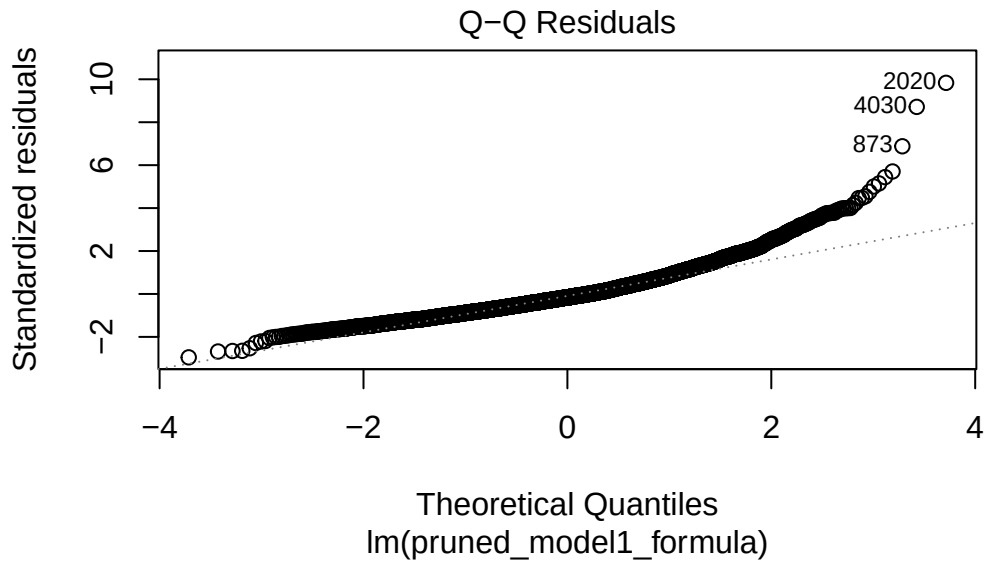
```
smoking_statusNever  -0.200421   0.350273  -0.572 0.567222
smoking_statusSmokes -0.774808   0.330108  -2.347 0.018958 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.86 on 4896 degrees of freedom
Multiple R-squared:  0.2389,    Adjusted R-squared:  0.2372
F-statistic: 139.7 on 11 and 4896 DF,  p-value: < 2.2e-16
```

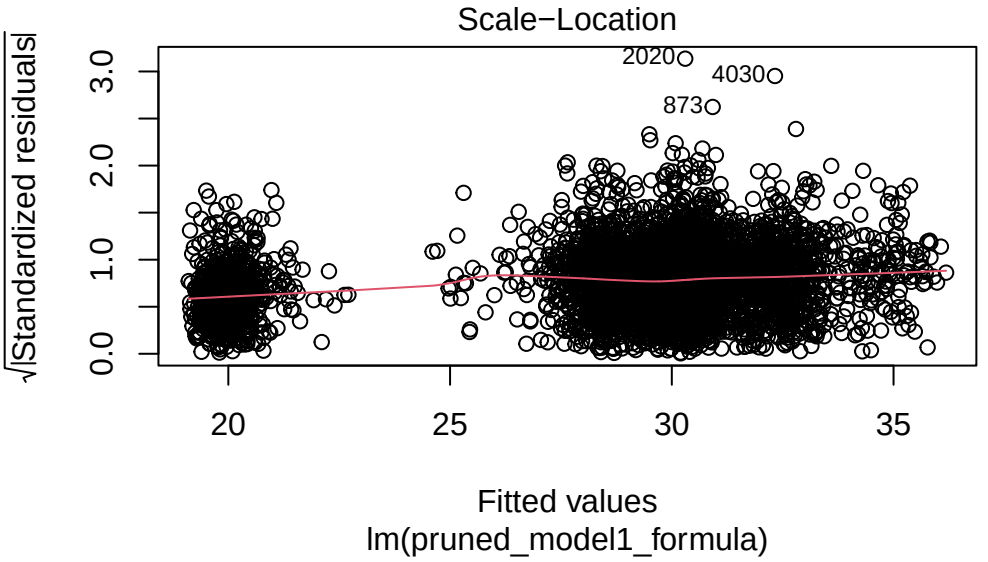
```
vif(pruned_model1)
```

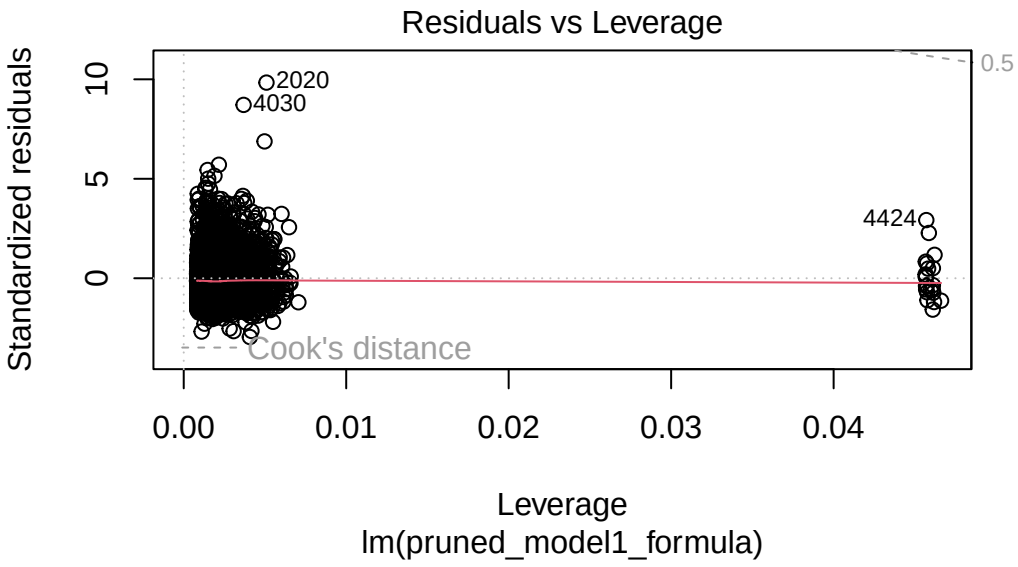
```
                    GVIF Df GVIF^(1/(2*Df))
age                2.723151  1        1.650197
hypertension       1.108678  1        1.052938
ever_married       1.964216  1        1.401505
work_type          2.325282  4        1.111244
avg_glucose_level 1.080856  1        1.039642
smoking_status     1.435372  3        1.062089
```

```
plot(pruned_model1)
```

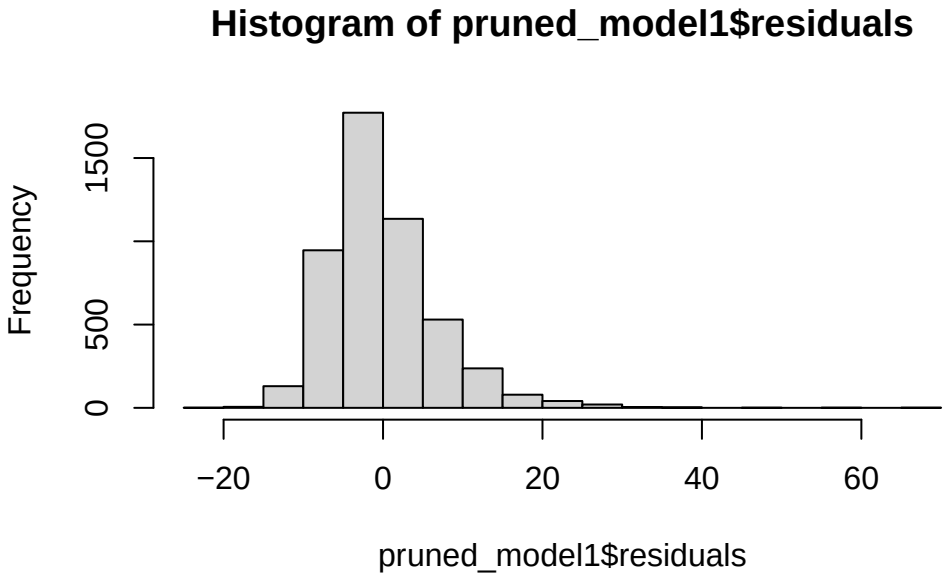



Q−Q Residuals
Standardized residuals
10
6
2
−2
2020
4030
873
−4
−2
0
2
4
Theoretical Quantiles
lm(pruned_model1_formula)


Scale−Location
√|Standardized residuals|
3.0
2.0
1.0
0.0
2020
4030
873
20
25
30
35
Fitted values
lm(pruned_model1_formula)




```r
hist(pruned_model1$residuals)
```



The same situation as before can be described regarding assumptions. There are clearly outliers present when examining the residual plots. Lets eliminate all observations greater than Q3 + 1.5 * IQR, and all observations less than Q1 - 1.5 * IQR with respect to continuous independent variables age, avg_glucose_level, and bmi.

```
outlier_range <- function(column){
  x <- length(column) * .25
  y <- length(column) *.75
  Q1 <- sort(column)[x]
  Q3 <- sort(column)[y]
  IQR <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
  return (c(lower,upper))
}
age_range <- outlier_range(df_stroke$age)
df_stroke_no_outlier <- df_stroke |>
  filter(age > age_range[1] & age < age_range[2])
glucose_range <- outlier_range(df_stroke$avg_glucose_level)
df_stroke_no_outlier <- df_stroke_no_outlier |>
  filter(avg_glucose_level > glucose_range[1] & avg_glucose_level < glucose_range[2])
bmi_range <- outlier_range(df_stroke$bmi)
df_stroke_no_outlier <- df_stroke_no_outlier |>
  filter(bmi > bmi_range[1] & bmi < bmi_range[2])
```

Pruned Model 1 No Outliers

```
pruned_model2_formula <- as.formula("bmi ~ age + hypertension + ever_married + work_type +
print(pruned_model2_formula)
```

```
bmi ~ age + hypertension + ever_married + work_type + avg_glucose_level +
    smoking_status
```

```
pruned_model2 <- lm(pruned_model2_formula,
                    data = df_stroke_no_outlier)
summary(pruned_model2)
```

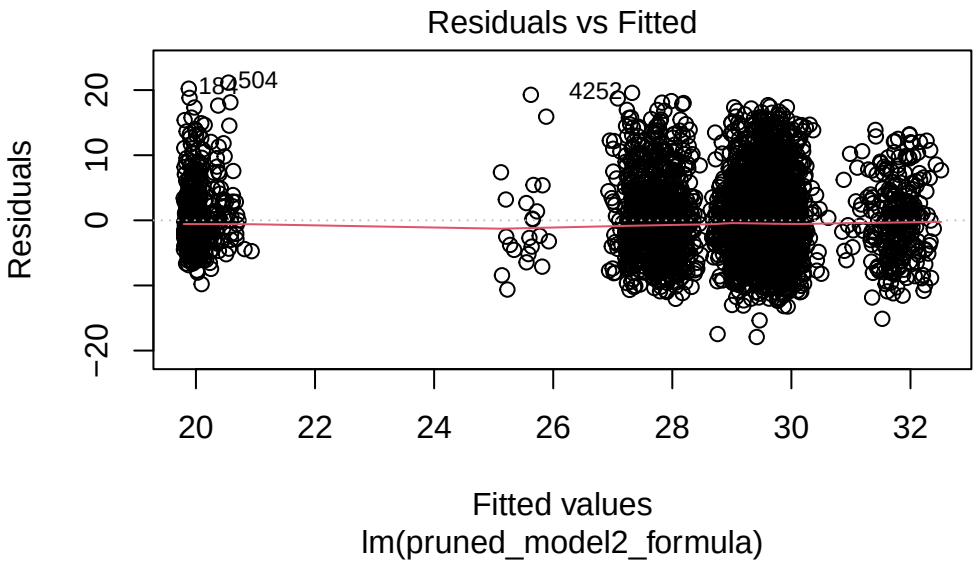
```
Call:
```

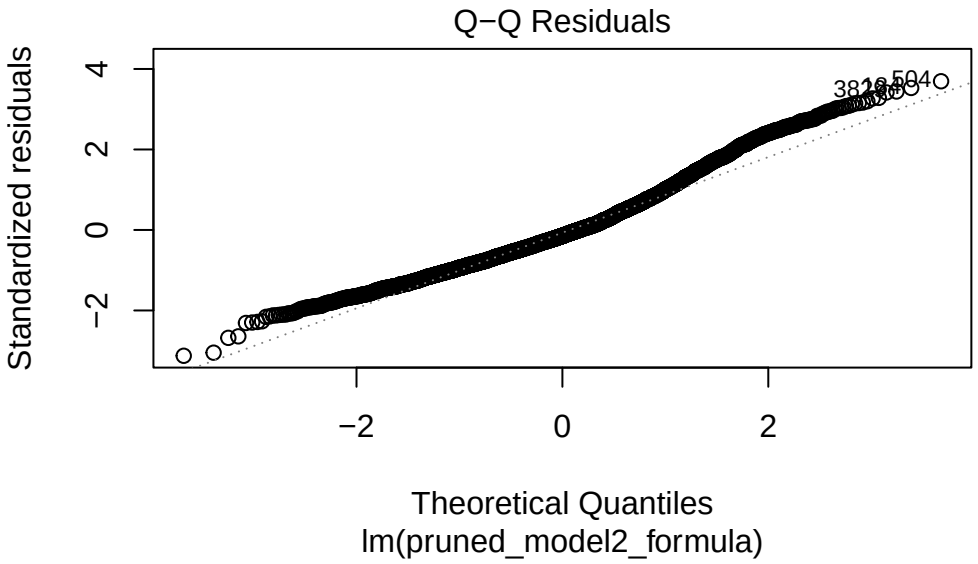
```
lm(formula = pruned_model2_formula, data = df_stroke_no_outlier)

Residuals:
     Min       1Q   Median       3Q      Max
-17.9178  -4.0169  -0.8341   3.2421  21.1532

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          20.342075   0.512797  39.669  < 2e-16 ***
age                  -0.001519   0.006450  -0.235    0.814
hypertensionYes       2.175677   0.360954   6.028 1.81e-09 ***
ever_marriedYes       1.790200   0.256155   6.989 3.21e-12 ***
work_typeGovt         7.689344   0.436241  17.626  < 2e-16 ***
work_typeNever        5.309918   1.251535   4.243 2.26e-05 ***
work_typePrivate      7.420691   0.355364  20.882  < 2e-16 ***
work_typeSelf         7.175992   0.449676  15.958  < 2e-16 ***
avg_glucose_level     0.004237   0.003880   1.092    0.275
smoking_statusFormer -0.332767   0.268135  -1.241    0.215
smoking_statusNever  -0.011296   0.320212  -0.035    0.972
smoking_statusSmokes -0.762238   0.299739  -2.543    0.011 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.736 on 4245 degrees of freedom
Multiple R-squared:  0.2726,    Adjusted R-squared:  0.2707
F-statistic: 144.6 on 11 and 4245 DF,  p-value: < 2.2e-16
```

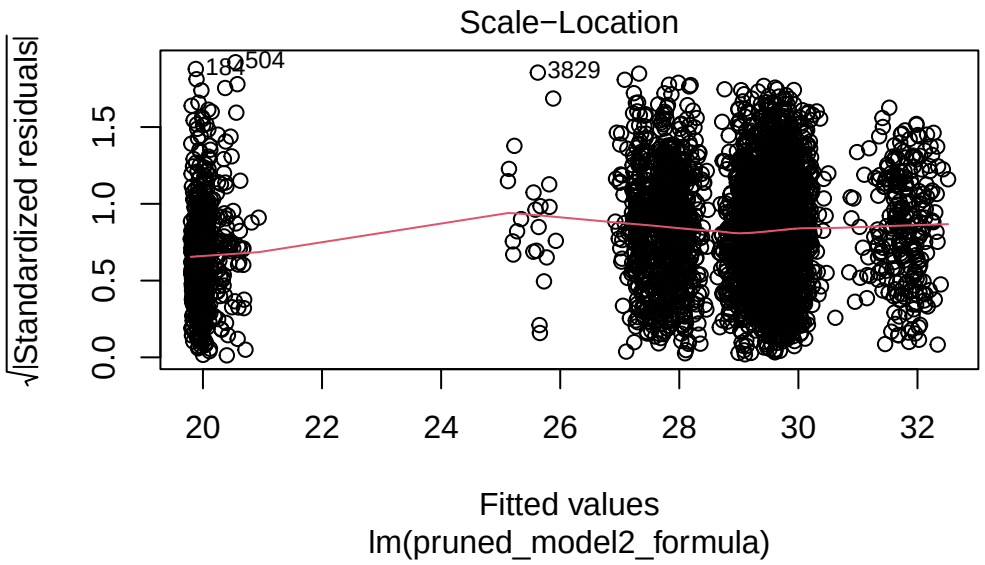
```r
vif(pruned_model2)
```

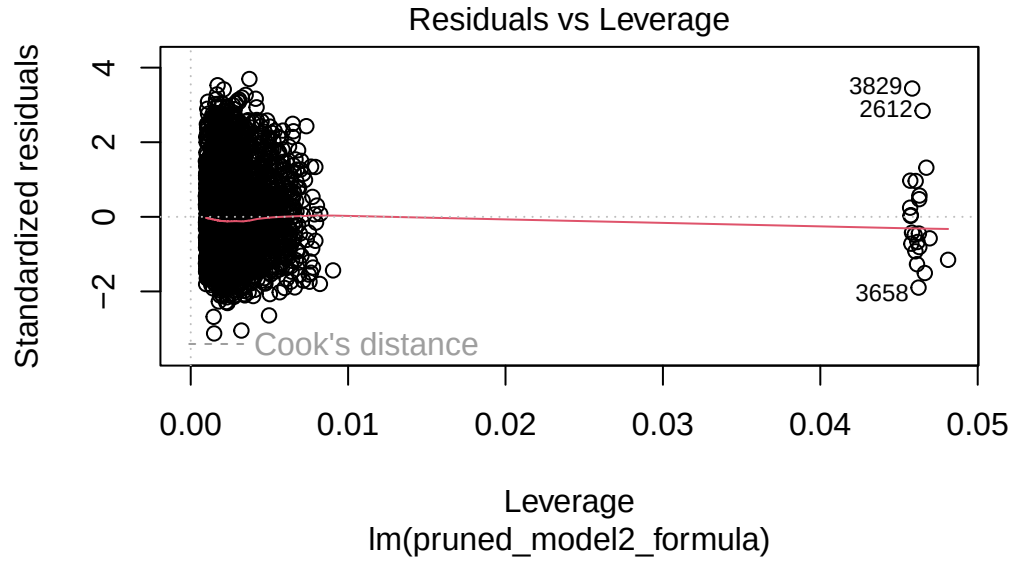
```
                     GVIF Df GVIF^(1/(2*Df))
age              2.712543  1        1.646980
hypertension     1.077061  1        1.037815
ever_married     1.998064  1        1.413529
work_type        2.391418  4        1.115147
avg_glucose_level 1.001575  1        1.000787
smoking_status   1.456998  3        1.064739
```

```r
plot(pruned_model2)
```

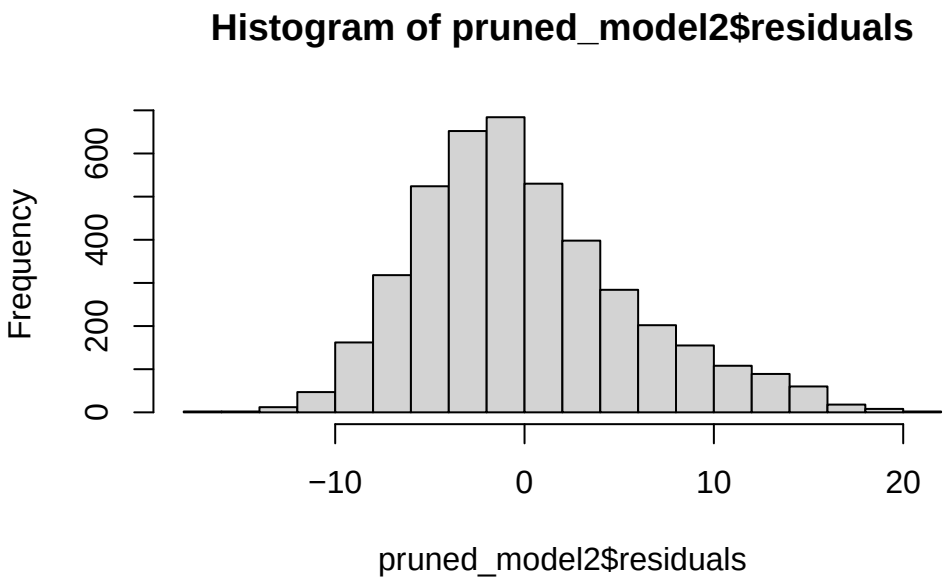

Residuals vs Fitted
Residuals
Fitted values
lm(pruned_model2_formula)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(pruned_model2_formula)






```
hist(pruned_model2$residuals)
```



The situation is the same as before regarding assumptions. Multicollinearity does not appear to be present and heteroskedasticity is likely not a problem. The residual plot is skewed right, and this problem should be investigated further. The adjusted R-squared has increased from .2372 in the first pruned model to .2707 in the second. While this is an improvement, this model still does not carry much explanatory power. Below are explanations of the significant coefficients.

Significant Coefficients

1. hypertensionYes - when a subject goes from not having hypertension to having hypertension, on average their bmi increases 2.1757 points, provided all other explanatory variables remain constant
2. ever_marriedYes - when a subject goes from being never married to currently or formerly married, on average their bmi increases 1.7902 points, provided all other explanatory variables remain constant
3. work_typeGovt - when a subject goes from being a full time parent to doing government work, on average their bmi increases 7.6893 points, provided all other explanatory variables remain constant
4. work_typeNever - when a subject goes from being a full time parent to having never worked, on average their bmi increases 5.3099 points, provided all other explanatory variables remain constant
5. work_typePrivate - when a subject goes from being a full time parent to doing private work, on average their bmi increases 7.4207 points, provided all other explanatory

variables remain constant

6. work_typeSelf - when a subject goes from being a full time parent to being self-employed, on average their bmi increases 7.176 points, provided all other explanatory variables remain constant
7. smoking_statusSmokes - when a subjects smoking status goes from unknown to smoking, on average bmi decreases .7622 points, provided all other explanatory variables remain constant

Pruned Model 2 No Outliers

```
print(as.formula("bmi ~ ever_married + work_type"))
```

```
bmi ~ ever_married + work_type
```

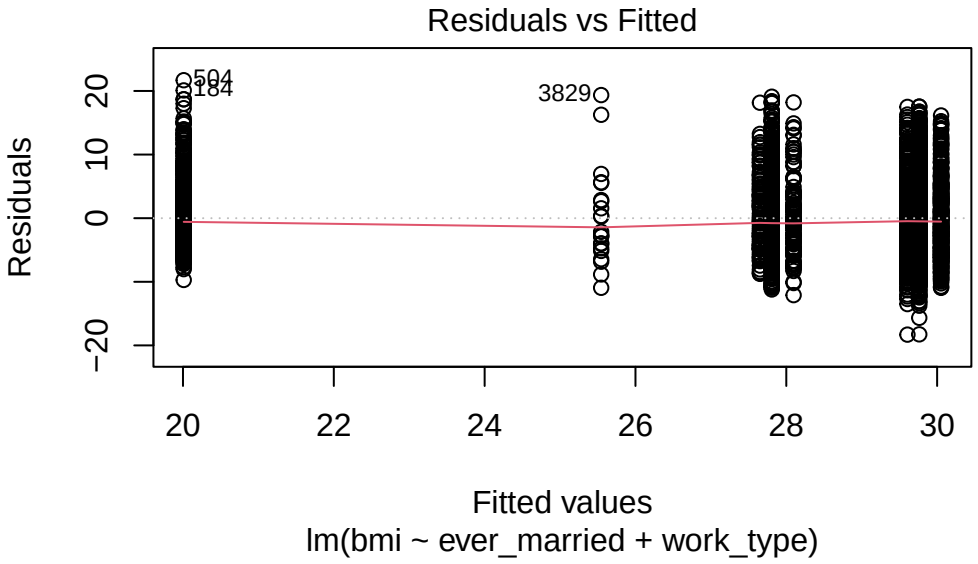
```
pruned_model3 <- lm(bmi ~ ever_married + work_type,
                    data = df_stroke_no_outlier)
summary(pruned_model3)
```

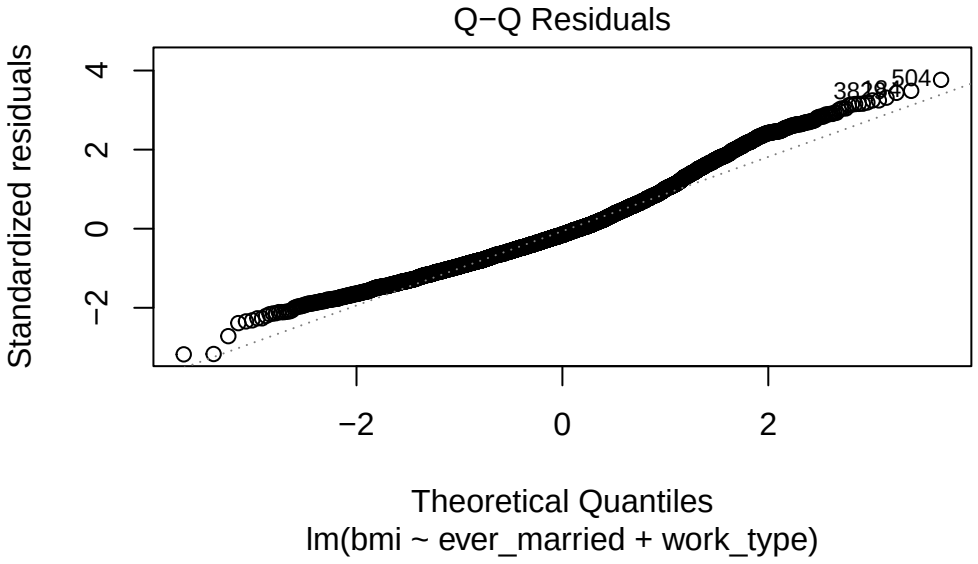
```
Call:
lm(formula = bmi ~ ever_married + work_type, data = df_stroke_no_outlier)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3063  -4.0076  -0.8631   3.2937  21.6900

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       20.0100     0.2241  89.298  < 2e-16 ***
ever_marriedYes    1.9555     0.2224   8.792  < 2e-16 ***
work_typeGovt      8.0869     0.3812  21.214  < 2e-16 ***
work_typeNever     5.5355     1.2495   4.430 9.65e-06 ***
work_typePrivate   7.7976     0.2964  26.304  < 2e-16 ***
work_typeSelf      7.6408     0.3728  20.496  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.765 on 4251 degrees of freedom
Multiple R-squared:  0.264, Adjusted R-squared:  0.2632
F-statistic:   305 on 5 and 4251 DF,  p-value: < 2.2e-16
```

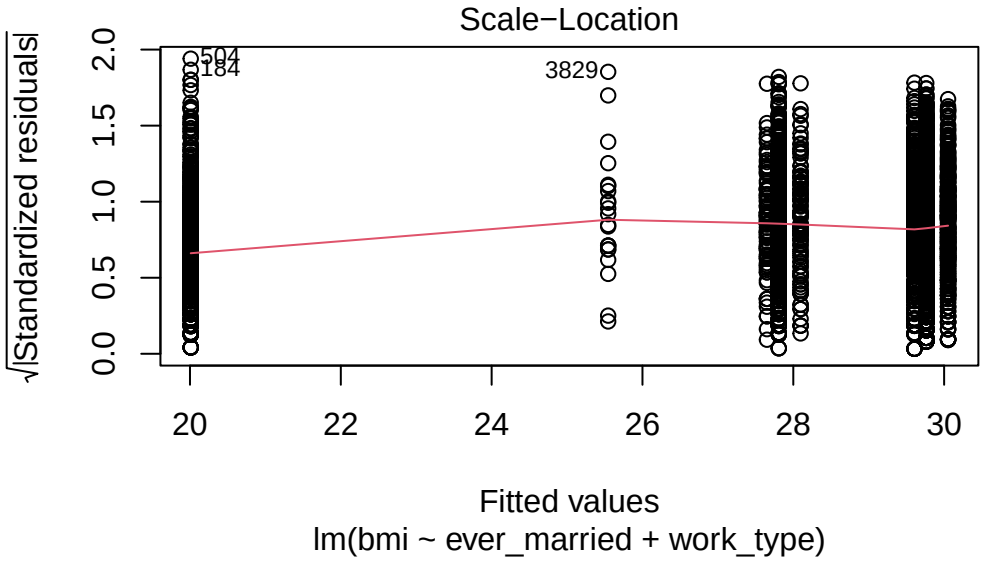
```r
vif(pruned_model3)
```

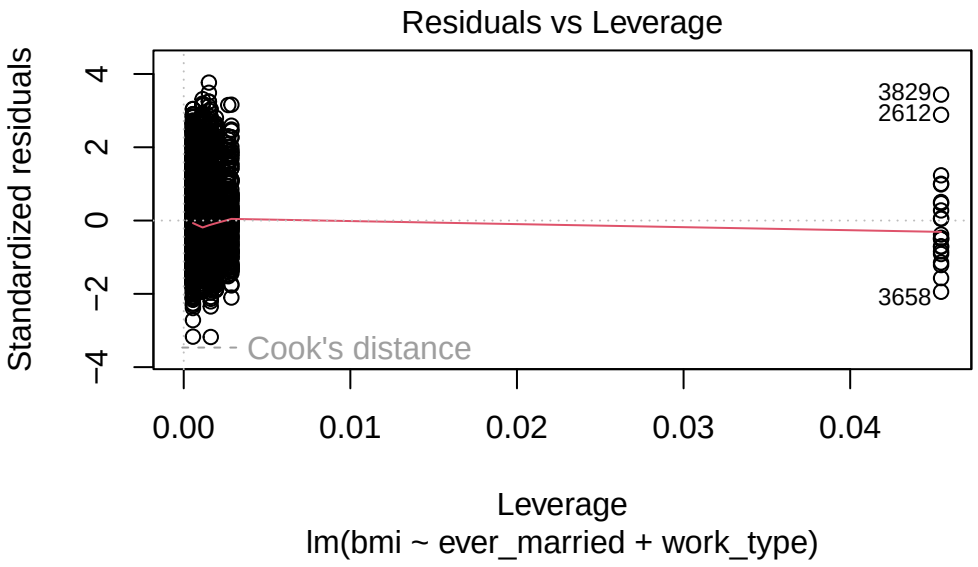
```
                GVIF Df GVIF^(1/(2*Df))
ever_married 1.491103  1        1.221107
work_type    1.491103  4        1.051208
```

```r
plot(pruned_model3)
```




Q−Q Residuals
Standardized residuals
−2
0
2
4
504
−2
0
2
Theoretical Quantiles
lm(bmi ~ ever_married + work_type)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
504
184
3829
20
22
24
26
28
30
Fitted values
lm(bmi ~ ever_married + work_type)




```r
hist(pruned_model3$residuals)
```

Residual patterns can be seen to form four distinct groups and show approximately even spread around a mean of zero. The Q-Q residual plot shows the straightest line yet, and the residual histogram has become less skewed to the right, better approximating a bell shaped curve. Assumptions do not appear to be violated in a meaningful way. Below are explanations for significant coefficients.

Significant Coefficients

1. ever_marriedYes - when a subject goes from being never married to currently or formerly married, on average their bmi increases 1.9555 points, provided all other explanatory variables remain constant
2. work_typeGovt - when a subject goes from being a full time parent to doing government work, on average their bmi increases 8.0869 points, provided all other explanatory variables remain constant
3. work_typeNever - when a subject goes from being a full time parent to having never worked, on average their bmi increases 5.5355 points, provided all other explanatory variables remain constant
4. work_typePrivate - when a subject goes from being a full time parent to doing private work, on average their bmi increases 7.7976 points, provided all other explanatory variables remain constant
5. work_typeSelf - when a subject goes from being a full time parent to being self-employed, on average their bmi increases 7.6408 points, provided all other explanatory variables remain constant

Eliminating smoking_status, hypertension, and avg_glucose level from the model decreased the adjusted R-squared, but not by much. The final pruned model may be the best due to the fact that not much more explanatory power is gained by adding in extra independent variables.

Interpretation

While the best performing multivariate regression models had little ability to explain the variation in bmi, it is still worth commenting on some of the discoveries made so that future studies can be better informed. The strongest associations between predictor and response were found in marital history and work type. This is interesting because these are not what one would normally think of as medical factors. On the contrary, these factors fall squarely within the behavioral category. This investigation has highlighted the need for medical questionnaires to inqure about aspects of a patients life that may seem wholly unrelated to the the diagnosis at hand. It also bears mentioning that a currently or previously married full time parent scores on average much less on the bmi scale than all other groups. Because so many health issues are linked with being overweight or obese, it appears that married full time parents are one of the healthiest groups of people, despite all the stress of raising children.

Logistic Regression Analysis - stroke

Full Model

```
glm_stroke_formula <- as.formula("stroke ~ gender + age + hypertension + heart_disease + e
print(glm_stroke_formula)
```

```
stroke ~ gender + age + hypertension + heart_disease + ever_married +
    work_type + Residence_type + avg_glucose_level + smoking_status +
    bmi
```

```
glm_stroke <- glm(glm_stroke_formula,
                  family = "binomial",
                  data = df_stroke_no_outlier)
summary(glm_stroke)
```

```
Call:
glm(formula = glm_stroke_formula, family = "binomial", data = df_stroke_no_outlier)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.948833   1.154809  -6.017 1.77e-09 ***
genderMale            0.008873   0.188896   0.047   0.9625
age                   0.072620   0.007100  10.227  < 2e-16 ***
hypertensionYes       0.589023   0.226481   2.601   0.0093 **
heart_diseaseYes      0.153903   0.287802   0.535   0.5928
ever_marriedYes      -0.300654   0.275769  -1.090   0.2756
work_typeGovt        -0.408050   1.139724  -0.358   0.7203
work_typeNever      -10.784261 509.558011  -0.021   0.9831
work_typePrivate     -0.305426   1.118473  -0.273   0.7848
work_typeSelf        -0.489751   1.141854  -0.429   0.6680
Residence_typeUrban  -0.034892   0.182487  -0.191   0.8484
avg_glucose_level     0.003259   0.003898   0.836   0.4031
smoking_statusFormer -0.134052   0.230031  -0.583   0.5601
smoking_statusNever   0.071873   0.283293   0.254   0.7997
smoking_statusSmokes -0.330387   0.289293  -1.142   0.2534
bmi                  -0.004505   0.016900  -0.267   0.7898
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1204.29  on 4256  degrees of freedom
Residual deviance:  974.62  on 4241  degrees of freedom
AIC: 1006.6

Number of Fisher Scoring iterations: 15
```

```
suppressPackageStartupMessages(library(pROC))
auc(glm_stroke$y, glm_stroke$fitted.values)
```

```
Setting levels: control = 0, case = 1
```

```
Setting direction: controls < cases
```

```
Area under the curve: 0.8467
```

Most of the parameter estimates for the above model have insignificant p-values. The area under the ROC curve is quite good, but lets prune the model and see if the AIC goes down without doing much to anything else.

Pruned Model

```
print(as.formula("stroke ~ age + hypertension"))
```

```
stroke ~ age + hypertension
```

```
glm_stroke_pruned <- glm(stroke ~ age + hypertension,
                         family = "binomial",
                         data = df_stroke_no_outlier)
summary(glm_stroke_pruned)
```

```
Call:
glm(formula = stroke ~ age + hypertension, family = "binomial",
    data = df_stroke_no_outlier)
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.349605   0.412904 -17.800  < 2e-16 ***
age              0.069960   0.006214  11.258  < 2e-16 ***
hypertensionYes  0.630426   0.222023   2.839  0.00452 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1204.29  on 4256  degrees of freedom
Residual deviance:  979.82  on 4254  degrees of freedom
AIC: 985.82

Number of Fisher Scoring iterations: 8
```

```r
exp(cbind(OR = coef(glm_stroke_pruned),
          confint(glm_stroke_pruned)))
```

```
Waiting for profiling to be done...

                          OR        2.5 %      97.5 %
(Intercept)     0.000642846 0.0002744663 0.001387925
age             1.072465271 1.0599060015 1.086076879
hypertensionYes 1.878410965 1.1999860138 2.871972930
```

```r
auc(glm_stroke_pruned$y, glm_stroke_pruned$fitted.values)
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Area under the curve: 0.8435
```

This pruned model has a lower AIC which indicates that the balance between complexity and explanatory power is better with this model. The area under the ROC curve remaines relatively unchanged. The pruned model classifies nearly just as good without the added complexity. Below are the odds-ratio interpretations for the pruned model parameter estimates.

Odds-Ratio Interperetations

exp(age) - for every one unit increase in age the odds ratio for stroke goes up by a factor of 1.0725, provided all other explanatory variables remain constant

exp(hypertensionYes) - when hypertension is changed from baseline No to Yes, the odds ratio for stroke increases by a factor of 1.8784, provided all other explanatory variables remain constant.

Post-Estimation

```
my_row <- df_stroke[72,]
predictions <- predict(glm_stroke_pruned, my_row, type="response")
predictions
```

```
       72
0.01693178
```

Automation

Data Cleaning

```
df_stroke_clean <- function(df_stroke) {
  suppressPackageStartupMessages(library(dplyr))
  df_stroke$id <- NULL
  df_stroke <- df_stroke |>
    filter(gender != "Other")
  df_stroke$gender <- as.factor(df_stroke$gender)
  df_stroke$hypertension <- as.factor(df_stroke$hypertension)
  levels(df_stroke$hypertension) <- c("No", "Yes")
  df_stroke$heart_disease <- as.factor(df_stroke$heart_disease)
  levels(df_stroke$heart_disease) <- c("No", "Yes")
  df_stroke$ever_married <- as.factor(df_stroke$ever_married)
  df_stroke$work_type <- as.factor(df_stroke$work_type)
  levels(df_stroke$work_type) <- c("Children", "Govt", "Never", "Private", "Self")
  df_stroke$Residence_type <- as.factor(df_stroke$Residence_type)
  df_stroke <- df_stroke |>
    filter(bmi != "N/A")
  df_stroke$bmi <- as.numeric(df_stroke$bmi)
  df_stroke$smoking_status <- factor(df_stroke$smoking_status)
```

```
  levels(df_stroke$smoking_status) = c("Unknown", "Former", "Never", "Smokes")
  df_stroke$stroke <- as.factor(df_stroke$stroke)
  levels(df_stroke$stroke) <- c("No", "Yes")
  return(df_stroke)
}
```

Summary Statistics

```
factor_summary <- function(Factors, title, color = "black", las_value = 1, make_mean = TRU
  counts <- table(Factors)
  if (make_mean == TRUE){
    MEAN <- mean(as.numeric(Factors))
  } else {
    MEAN <- "N/A"
  }
  mean_mode <- data.frame(MEAN = MEAN, MODE = names(which.max(counts)))
  column_plot <- barplot(counts, col = color, main = title, ylab = "Frequency", las = las_
  return(list(counts = counts, mean_mode = mean_mode, column_plot = column_plot))
}

numeric_summary <- function(column, title, color) {
  column_summary <- summary(column)
  column_box <- boxplot(column,
                  main = title,
                  ylab = "Values",
                  col = color)
  return(list(summary = column_summary, plot = column_box))
}

column_summary <- function(df_stroke_column) {
  if (is.numeric(df_stroke_column)) {
    numeric_summary(df_stroke_column)
  } else if (is.factor(df_stroke_column)) {
    num_levels <- nlevels(df_stroke_column)
    full_name <- deparse(substitute(df_stroke_column))
    column_name <- gsub(".*\\$", "", full_name)
    if(num_levels > 2) {
      factor_summary(df_stroke_column, title = column_name, las_value = 2, make_mean = FAL
    } else {
      factor_summary(df_stroke_column, title = column_name)
```

```
    }
  }
}

df_summary <- function(df) {
  my_list <- list()
  for (col_name in names(df)) {
    my_list[[col_name]] <- column_summary(df$col_name)
  }
  return (my_list)
}
```

One Sample Hypothesis Test: bmi

```
proportions_test_upper <- function(df, greater_than_age, greater_than_bmi, p0) {
  po <- p0
  n <- df |>
    filter(age >= greater_than_age) |>
    summarise(count = n())
  n <- n[1,1]
  x <- df |>
    filter(age >= greater_than_age & bmi > greater_than_bmi) |>
    summarise(count = n())
  x <- x[1,1]
  result <- prop.test(x, n, p = po, alternative = "greater", conf.level = .99, correct = F
  return (result)
}
```

Two Sample Hypothesis Test: bmi

```
two_sample_test_bmi <- function(df,column){
  if (!is.factor(column)) {
    stop("COLUMN MUST BE FACTOR!!")
  }
  if (nlevels(column) > 2) {
    stop("COLUMN CAN ONLY HAVE TWO LEVELS!!")
  }
  full_name <- deparse(substitute(column))
```

```
  column_name <- gsub(".*\\$", "", full_name)
  column_sym <- sym(column_name)
  x <- df |>
    filter(!!column_sym == levels(column)[1])
  y <- df |>
    filter(!!column_sym == levels(column)[2])
    test <- t.test(x$bmi, y$bmi, mu=0, paired=FALSE, var.equal=TRUE, conf.level=0.99)
    return(test)
}
```

Outlier Reduction

```
outlier_range <- function(column){
  x <- length(column) * .25
  y <- length(column) *.75
  Q1 <- sort(column)[x]
  Q3 <- sort(column)[y]
  IQR <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
  return (c(lower,upper))
}
df_no_outliers <- function(df_stroke) {
  age_range <- outlier_range(df_stroke$age)
  df_stroke_no_outlier <- df_stroke |>
    filter(age > age_range[1] & age < age_range[2])
  glucose_range <- outlier_range(df_stroke$avg_glucose_level)
  df_stroke_no_outlier <- df_stroke_no_outlier |>
    filter(avg_glucose_level > glucose_range[1] &      avg_glucose_level < glucose_range[2
  bmi_range <- outlier_range(df_stroke$bmi)
  df_stroke_no_outlier <- df_stroke_no_outlier |>
    filter(bmi > bmi_range[1] & bmi < bmi_range[2])
  return(df_stroke_no_outlier)
}
```

Multivariate Regression Analysis

```
multi_regression <- function(model_formula, df){
  suppressPackageStartupMessages(library(car))
  x <- lm(model_formula, data = df)
  my_list <- list()
  my_list[["summary"]] <- summary(x)
  my_list[["vif"]] <- vif(x)
  plot(x)
  hist(x$residuals)
  return (my_list)
}
```

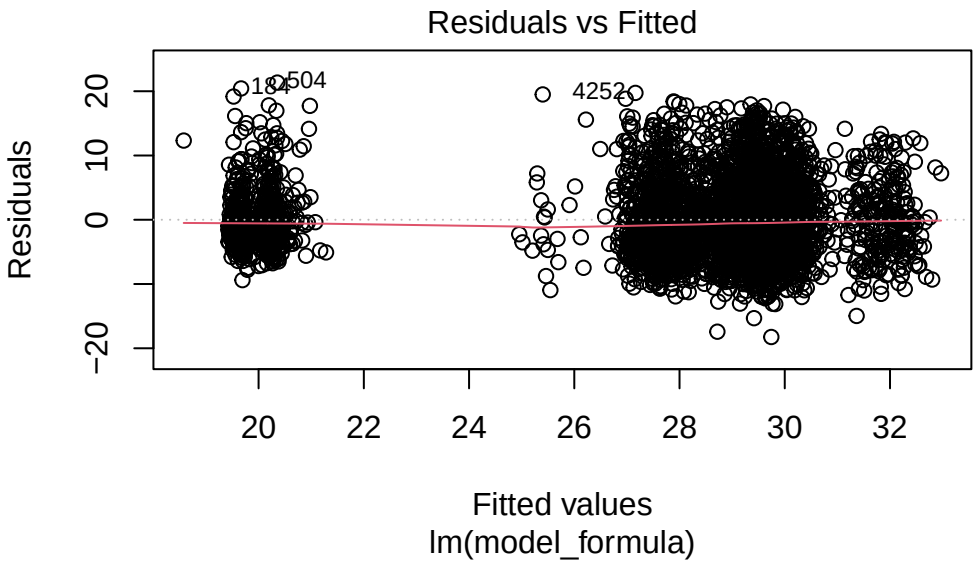
Logistic Regression

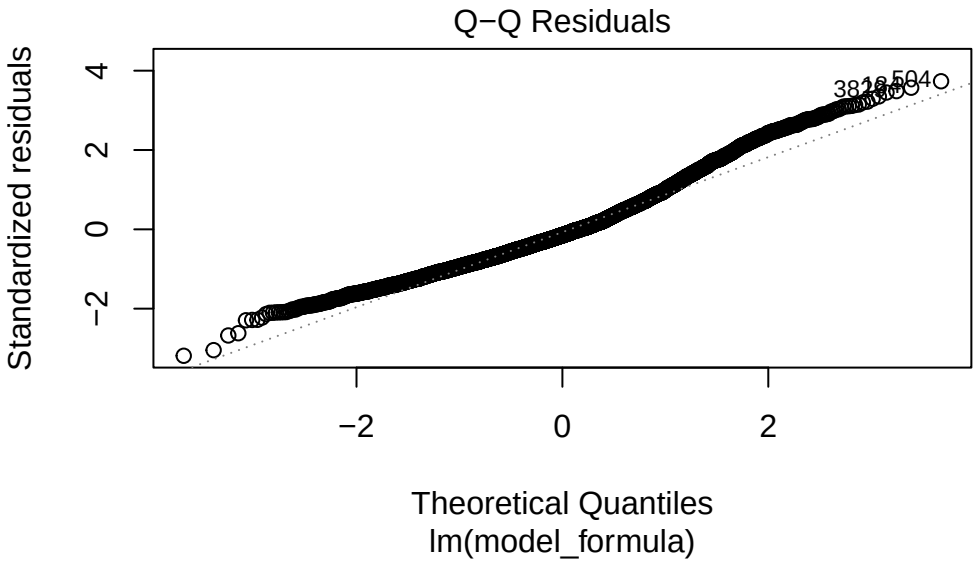
```
logi_regression <- function(model_formula, df) {
  x <- glm(model_formula, family = "binomial", data = df)
  my_list <- list()
  my_list[["summary"]] <- summary(x)
  odds_ratios <- exp(cbind(OR = coef(x),
          confint(x)))
  my_list[["odds_ratios"]] <- odds_ratios
  suppressPackageStartupMessages(library(pROC))
  my_list[["auc"]] <- auc(x$y, x$fitted.values)
  my_row <- df[sample(1:nrow(df),1),]
  predictions <- predict(x, my_row, type="response")
  my_list[["predicion"]] <- predictions
  return(my_list)
}
```

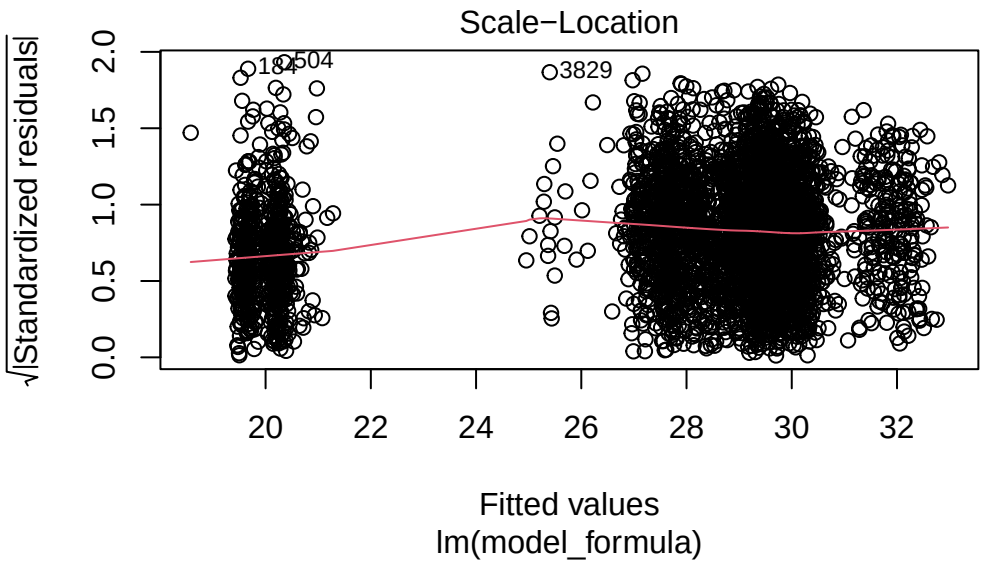
Automated Pipeline

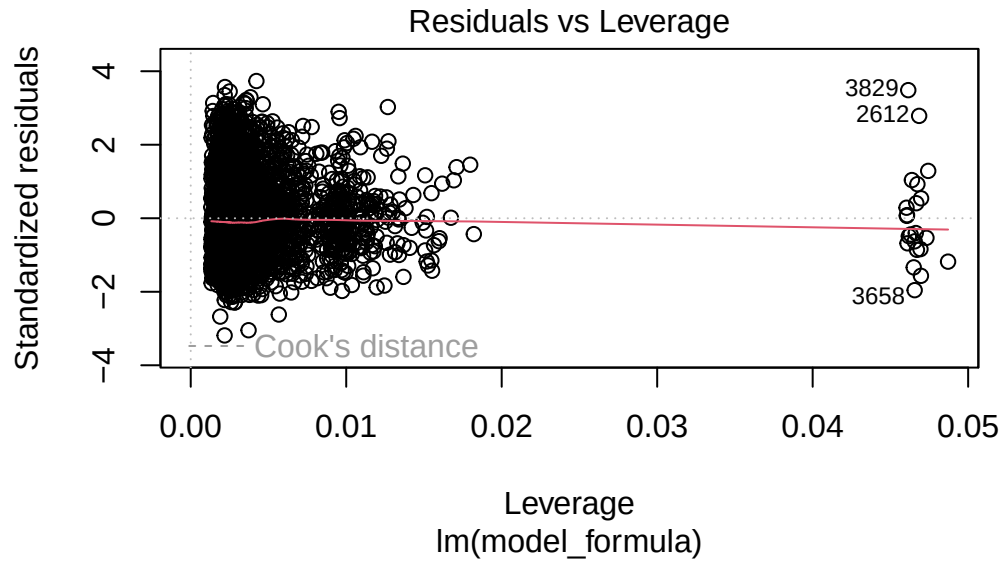
```
#import data
df_stroke <- read.csv("~/Desktop/healthcare-dataset-stroke-data.csv")

#clean data
df_stroke <- df_stroke_clean(df_stroke)
```

```r
#get summary statistics and charts
df_summary <- df_summary(df_stroke)

#one sample hypothesis test
results_one_sample <- proportions_test_upper(df_stroke,20,25,.64)

#two sample hypothesis test
results_two_sample <- two_sample_test_bmi(df_stroke, df_stroke$Residence_type)

#remove outliers
df_stroke <- df_no_outliers(df_stroke)

#multivariate regression
multi_model1 <- as.formula("bmi ~ gender + age + hypertension + heart_disease + ever_marri
multi_model2 <- as.formula("bmi ~ age + hypertension + ever_married + work_type + avg_gluc
multi_model3 <- as.formula("bmi ~ ever_married + work_type")
multi_model_list1 <- list(multi_model1, multi_model2, multi_model3)
multi_model_list2 <- list()
for (i in 1:length(multi_model_list1)) {
    multi_model_list2[[i]] <- multi_regression(multi_model_list1[[i]],df_stroke)
    print(multi_model_list1[[i]])
    print(multi_model_list2[[i]][1])
    print(multi_model_list2[[i]][2])
}
```


Residuals vs Fitted
Residuals
Fitted values
lm(model_formula)
504
4252

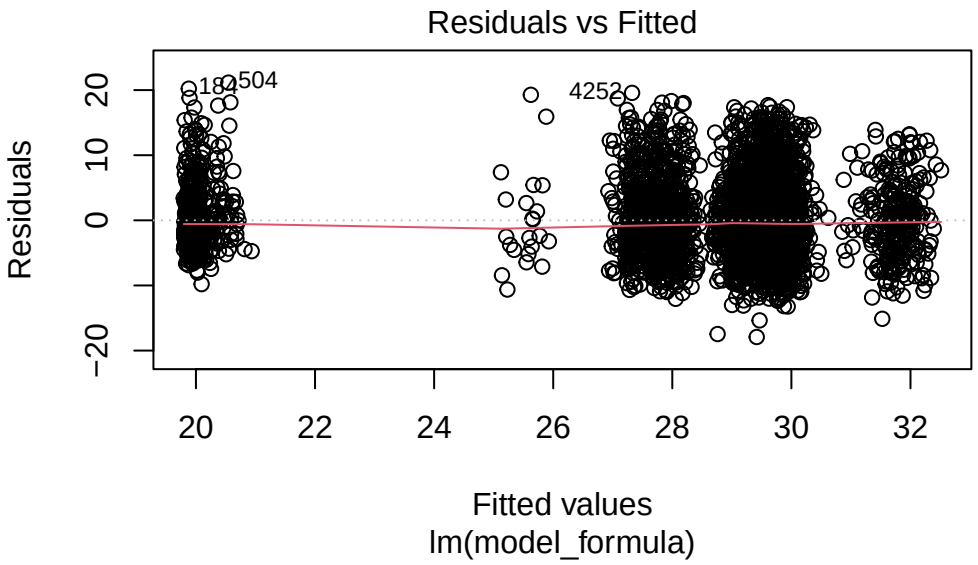

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(model_formula)
504

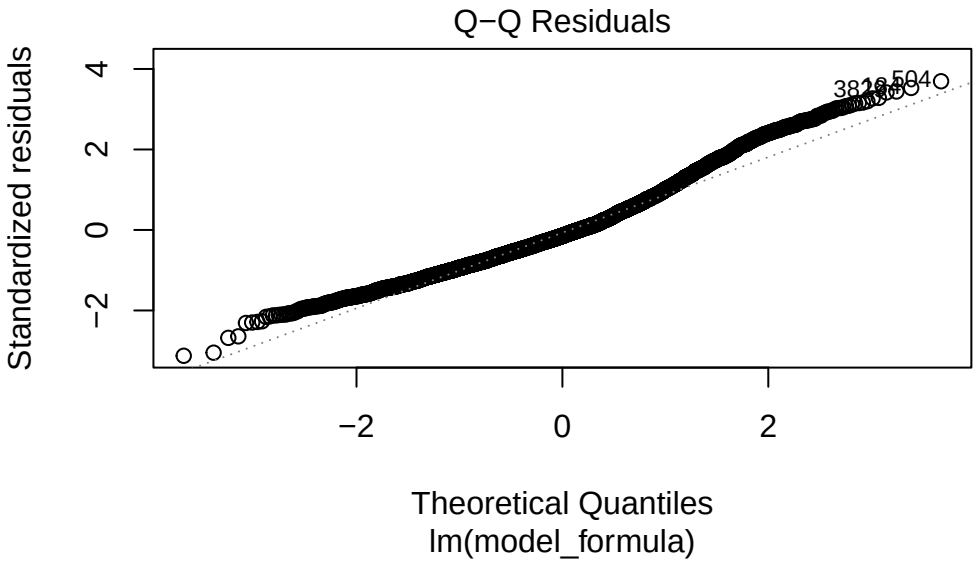

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
184
504
3829
20
22
24
26
28
30
32
Fitted values
lm(model_formula)

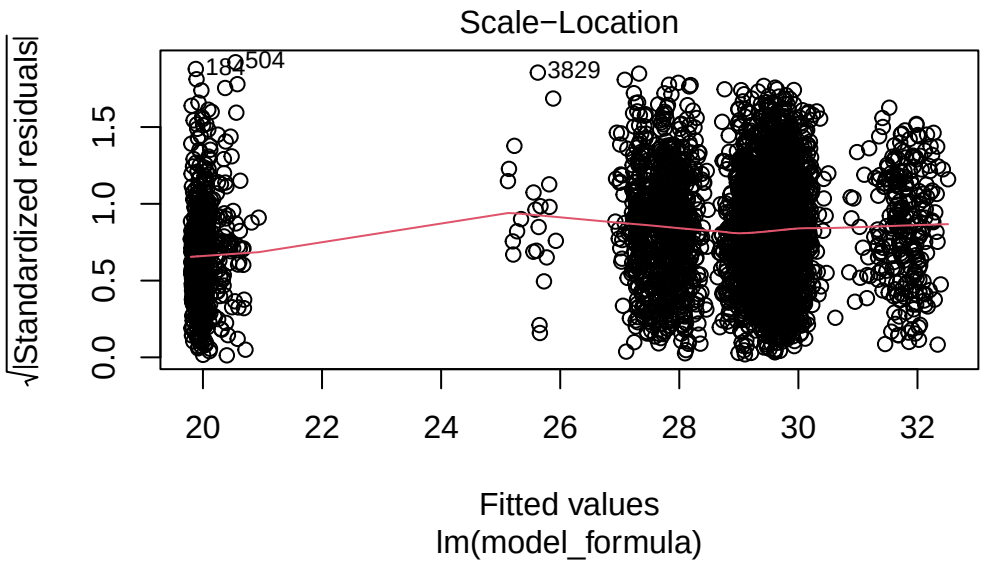

Residuals vs Leverage
Standardized residuals
−4
−2
0
2
4
3829
2612
3658
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(model_formula)

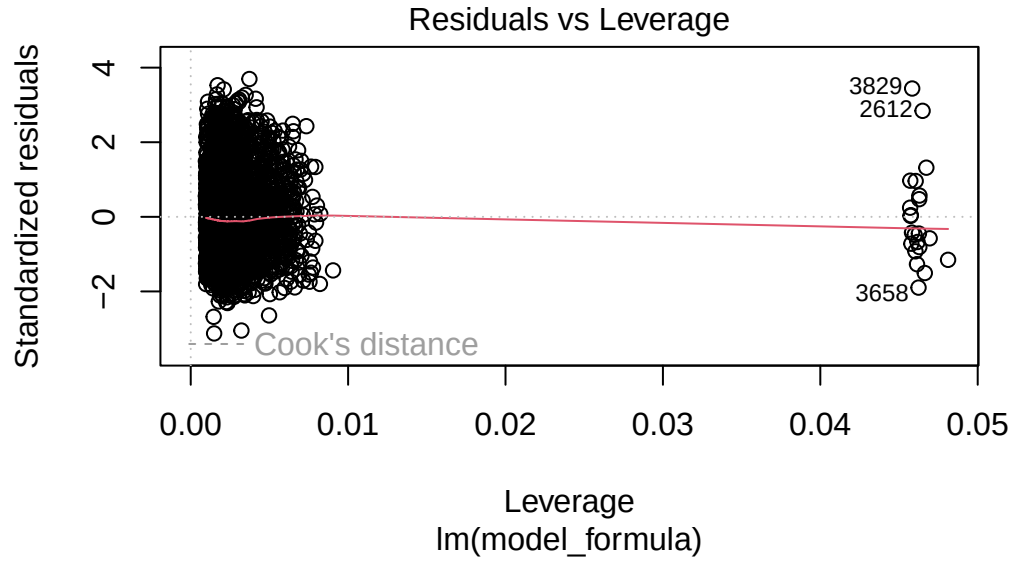

```
bmi ~ gender + age + hypertension + heart_disease + ever_married +
    work_type + Residence_type + avg_glucose_level + smoking_status +
    stroke
$summary

Call:
lm(formula = model_formula, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2444  -4.0572  -0.8534   3.2407  21.3440

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          19.9605611  0.5284058  37.775  < 2e-16 ***
genderMale            0.5825666  0.1811880   3.215  0.00131 **
age                   0.0005613  0.0067126   0.084  0.93337
hypertensionYes       2.1917567  0.3616707   6.060 1.48e-09 ***
heart_diseaseYes     -0.1803703  0.4893222  -0.369  0.71243
ever_marriedYes       1.7662493  0.2568554   6.876 7.03e-12 ***
work_typeGovt         7.7283751  0.4383945  17.629  < 2e-16 ***
work_typeNever        5.2563876  1.2505800   4.203 2.69e-05 ***
work_typePrivate      7.4618272  0.3576166  20.865  < 2e-16 ***
```

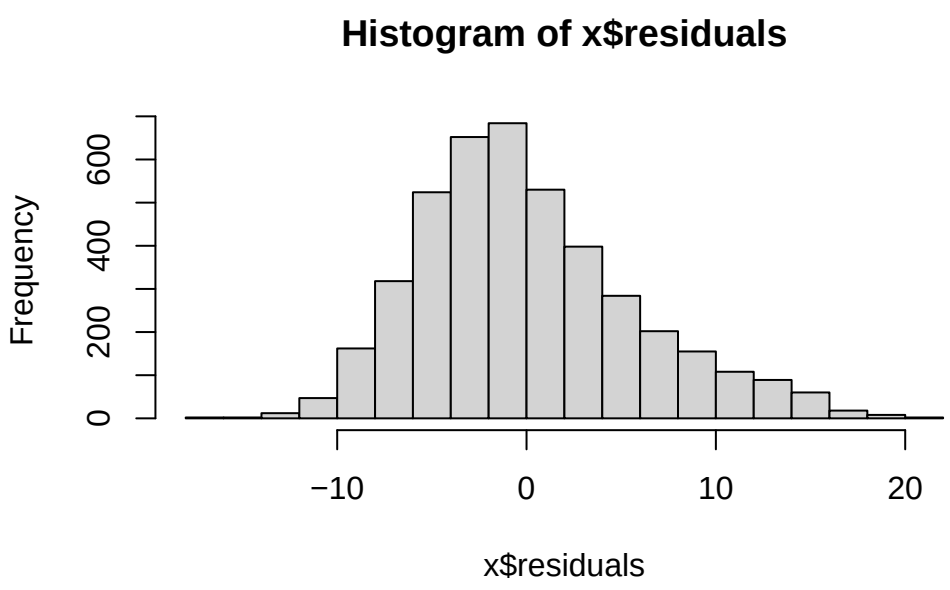
```
work_typeSelf          7.2340944  0.4517788  16.012  < 2e-16 ***
Residence_typeUrban    0.1542926  0.1759102   0.877  0.38048
avg_glucose_level      0.0040127  0.0038784   1.035  0.30090
smoking_statusFormer -0.2938988  0.2681527  -1.096  0.27314
smoking_statusNever  -0.0211077  0.3201457  -0.066  0.94744
smoking_statusSmokes -0.7574113  0.2994266  -2.530  0.01146 *
strokeYes            -0.8663443  0.5149692  -1.682  0.09258 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.729 on 4241 degrees of freedom
Multiple R-squared:  0.275, Adjusted R-squared:  0.2724
F-statistic: 107.2 on 15 and 4241 DF,  p-value: < 2.2e-16


$vif
                     GVIF Df GVIF^(1/(2*Df))
gender           1.026801  1        1.013312
age              2.944474  1        1.715947
hypertension     1.083896  1        1.041103
heart_disease    1.089562  1        1.043821
ever_married     2.013761  1        1.419071
work_type        2.434458  4        1.117636
Residence_type   1.003171  1        1.001584
avg_glucose_level 1.002900  1        1.001449
smoking_status   1.469562  3        1.066264
stroke           1.063748  1        1.031382
```


Residuals vs Fitted
Residuals
Fitted values
lm(model_formula)
4252


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(model_formula)

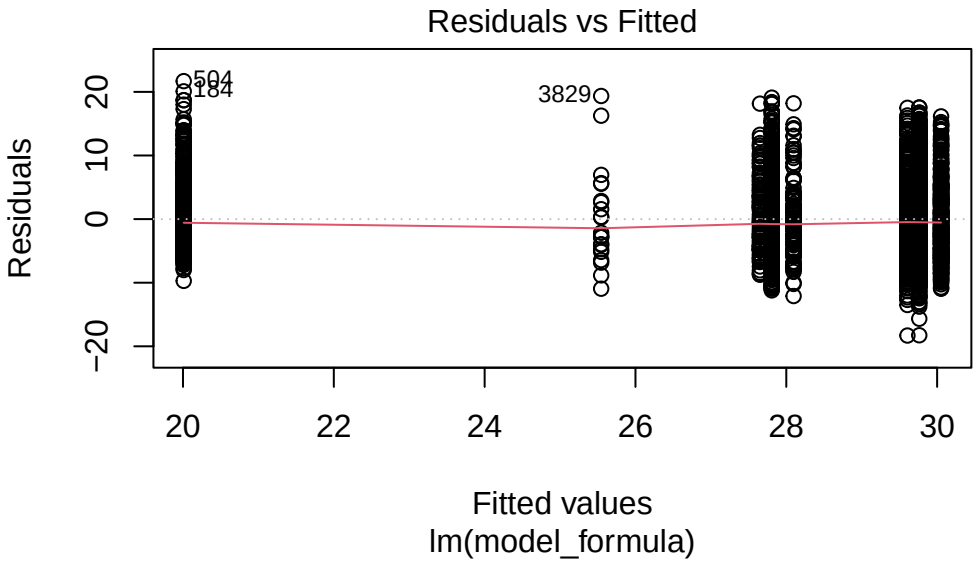

Scale−Location
√|Standardized residuals|
184
2504
3829
0.0
0.5
1.0
1.5
20
22
24
26
28
30
32
Fitted values
lm(model_formula)

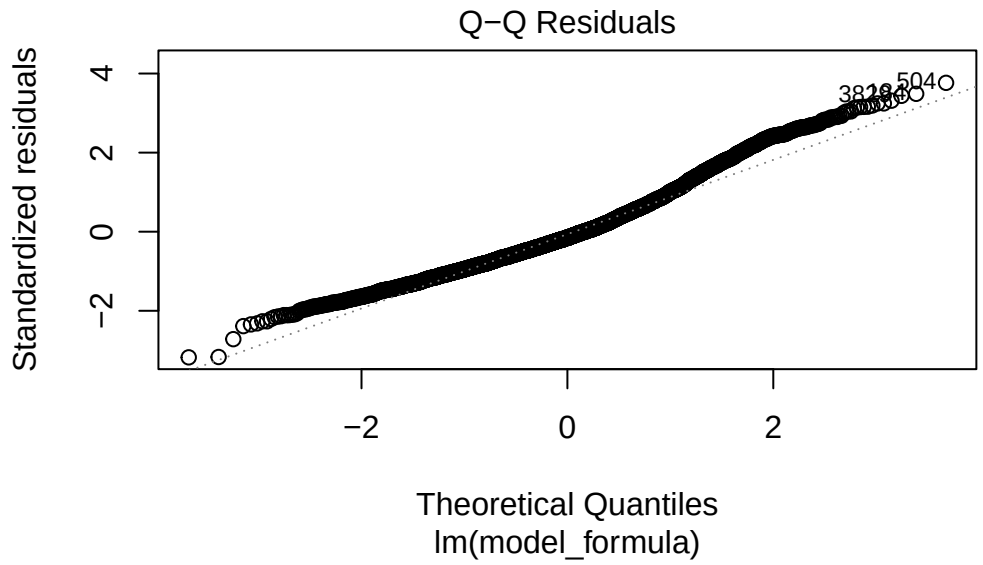

Residuals vs Leverage
Standardized residuals
3829
2612
3658
Cook's distance
−2
0
2
4
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(model_formula)

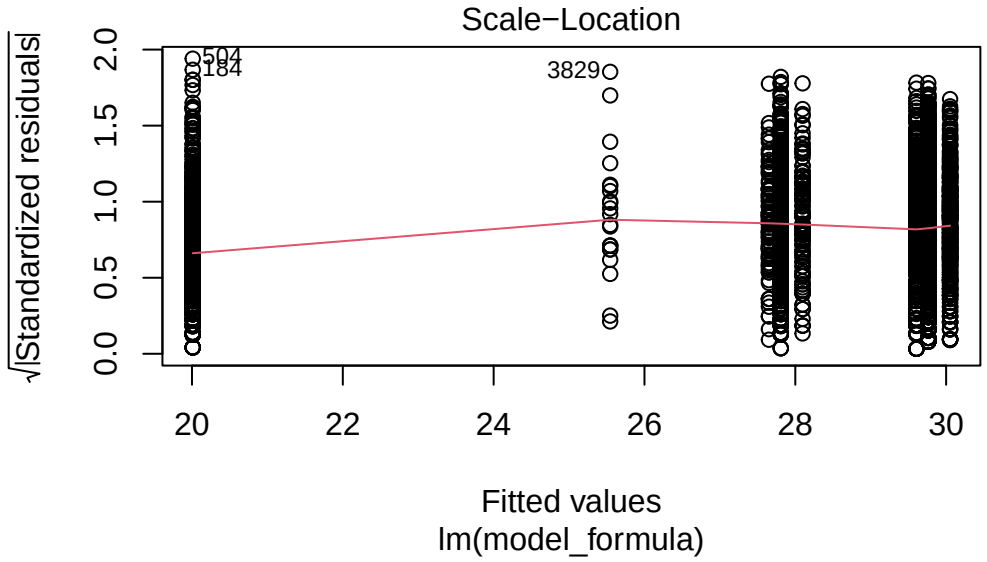


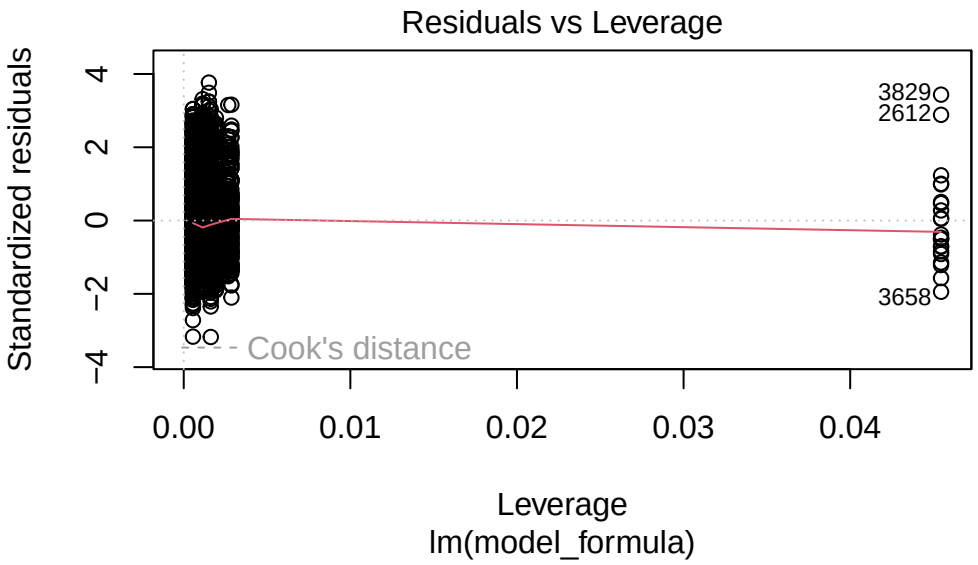

```
bmi ~ age + hypertension + ever_married + work_type + avg_glucose_level +
    smoking_status
$summary

Call:
lm(formula = model_formula, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-17.9178  -4.0169  -0.8341   3.2421  21.1532

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          20.342075   0.512797  39.669  < 2e-16 ***
age                  -0.001519   0.006450  -0.235    0.814
hypertensionYes       2.175677   0.360954   6.028 1.81e-09 ***
ever_marriedYes       1.790200   0.256155   6.989 3.21e-12 ***
work_typeGovt         7.689344   0.436241  17.626  < 2e-16 ***
work_typeNever        5.309918   1.251535   4.243 2.26e-05 ***
work_typePrivate      7.420691   0.355364  20.882  < 2e-16 ***
work_typeSelf         7.175992   0.449676  15.958  < 2e-16 ***
avg_glucose_level     0.004237   0.003880   1.092    0.275
smoking_statusFormer -0.332767   0.268135  -1.241    0.215
```

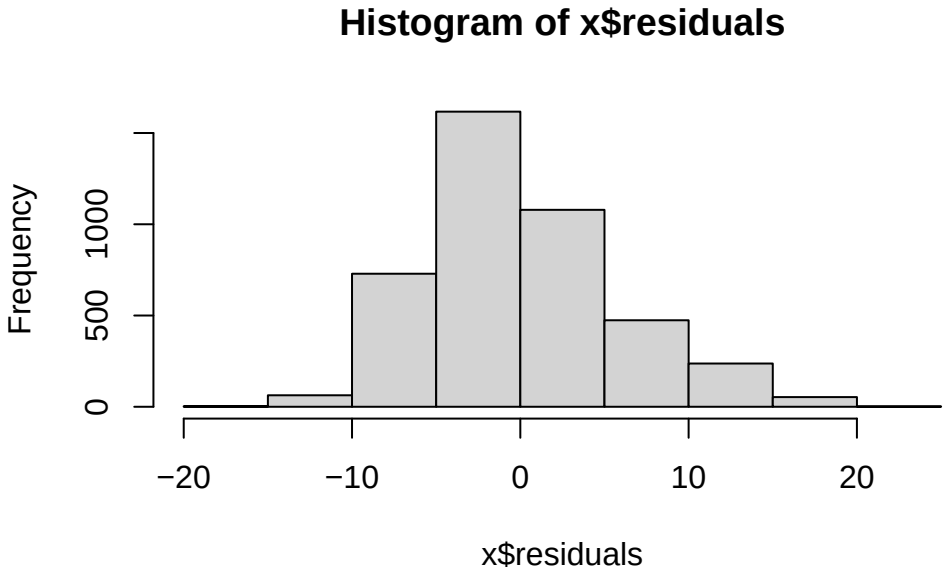
```
smoking_statusNever  -0.011296   0.320212  -0.035    0.972
smoking_statusSmokes -0.762238   0.299739  -2.543    0.011 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.736 on 4245 degrees of freedom
Multiple R-squared:  0.2726,    Adjusted R-squared:  0.2707
F-statistic: 144.6 on 11 and 4245 DF,  p-value: < 2.2e-16


$vif
                     GVIF Df GVIF^(1/(2*Df))
age                2.712543  1        1.646980
hypertension       1.077061  1        1.037815
ever_married       1.998064  1        1.413529
work_type          2.391418  4        1.115147
avg_glucose_level  1.001575  1        1.000787
smoking_status     1.456998  3        1.064739
```




Q−Q Residuals
Standardized residuals
−2
0
2
4
504
−2
0
2
Theoretical Quantiles
lm(model_formula)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
504
184
3829
20
22
24
26
28
30
Fitted values
lm(model_formula)






```
bmi ~ ever_married + work_type
$summary
```

```
Call:
lm(formula = model_formula, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3063  -4.0076  -0.8631   3.2937  21.6900

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       20.0100     0.2241  89.298  < 2e-16 ***
ever_marriedYes    1.9555     0.2224   8.792  < 2e-16 ***
work_typeGovt      8.0869     0.3812  21.214  < 2e-16 ***
work_typeNever     5.5355     1.2495   4.430 9.65e-06 ***
work_typePrivate   7.7976     0.2964  26.304  < 2e-16 ***
work_typeSelf      7.6408     0.3728  20.496  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.765 on 4251 degrees of freedom
Multiple R-squared:  0.264, Adjusted R-squared:  0.2632
F-statistic:   305 on 5 and 4251 DF,  p-value: < 2.2e-16


$vif
                 GVIF Df GVIF^(1/(2*Df))
ever_married 1.491103  1        1.221107
work_type    1.491103  4        1.051208
```

```r
#logistic regression
logi_model1 <- as.formula("stroke ~ gender + age + hypertension + heart_disease + ever_mar
logi_model2 <- as.formula("stroke ~ age + hypertension")
logi_model_list1 <- list(logi_model1, logi_model2)
logi_model_list2 <- list()
for (i in 1:length(logi_model_list1)) {
    logi_model_list2[[i]] <- logi_regression(logi_model_list1[[i]],df_stroke)
    print(logi_model_list1[[i]])
    print(logi_model_list2[[i]])
}
```

```
Waiting for profiling to be done...
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases

stroke ~ gender + age + hypertension + heart_disease + ever_married +
    work_type + Residence_type + avg_glucose_level + smoking_status +
    bmi
$summary

Call:
glm(formula = model_formula, family = "binomial", data = df)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.948833   1.154809  -6.017 1.77e-09 ***
genderMale            0.008873   0.188896   0.047   0.9625
age                   0.072620   0.007100  10.227  < 2e-16 ***
hypertensionYes       0.589023   0.226481   2.601   0.0093 **
heart_diseaseYes      0.153903   0.287802   0.535   0.5928
ever_marriedYes      -0.300654   0.275769  -1.090   0.2756
work_typeGovt        -0.408050   1.139724  -0.358   0.7203
work_typeNever      -10.784261 509.558011  -0.021   0.9831
work_typePrivate     -0.305426   1.118473  -0.273   0.7848
work_typeSelf        -0.489751   1.141854  -0.429   0.6680
Residence_typeUrban  -0.034892   0.182487  -0.191   0.8484
avg_glucose_level     0.003259   0.003898   0.836   0.4031
smoking_statusFormer -0.134052   0.230031  -0.583   0.5601
smoking_statusNever   0.071873   0.283293   0.254   0.7997
smoking_statusSmokes -0.330387   0.289293  -1.142   0.2534
bmi                  -0.004505   0.016900  -0.267   0.7898
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1204.29  on 4256  degrees of freedom
Residual deviance:  974.62  on 4241  degrees of freedom
AIC: 1006.6

Number of Fisher Scoring iterations: 15
```

```
$odds_ratios
                                OR        2.5 %       97.5 %
(Intercept)          9.597550e-04 4.678478e-05 6.609637e-03
genderMale           1.008912e+00 6.937131e-01 1.456861e+00
age                  1.075322e+00 1.060802e+00 1.090801e+00
hypertensionYes      1.802227e+00 1.141860e+00 2.781092e+00
heart_diseaseYes     1.166378e+00 6.462533e-01 2.007672e+00
ever_marriedYes      7.403341e-01 4.406573e-01 1.306734e+00
work_typeGovt        6.649459e-01 9.863895e-02 1.338461e+01
work_typeNever       2.072312e-05 1.256664e-95 5.657401e-64
work_typePrivate     7.368095e-01 1.157241e-01 1.450528e+01
work_typeSelf        6.127787e-01 9.045022e-02 1.236341e+01
Residence_typeUrban  9.657095e-01 6.749708e-01 1.382165e+00
avg_glucose_level    1.003264e+00 9.954772e-01 1.010825e+00
smoking_statusFormer 8.745446e-01 5.593354e-01 1.381839e+00
smoking_statusNever  1.074518e+00 6.106146e-01 1.862929e+00
smoking_statusSmokes 7.186456e-01 4.022406e-01 1.257217e+00
bmi                  9.955051e-01 9.626951e-01 1.028698e+00

$auc
Area under the curve: 0.8467

$predicion
       839
0.05629201


Waiting for profiling to be done...


Setting levels: control = 0, case = 1


Setting direction: controls < cases


stroke ~ age + hypertension
$summary

Call:
glm(formula = model_formula, family = "binomial", data = df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.349605   0.412904 -17.800  < 2e-16 ***
```

```
age              0.069960   0.006214  11.258  < 2e-16 ***
hypertensionYes  0.630426   0.222023   2.839  0.00452 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1204.29  on 4256  degrees of freedom
Residual deviance:  979.82  on 4254  degrees of freedom
AIC: 985.82

Number of Fisher Scoring iterations: 8


$odds_ratios
                          OR        2.5 %      97.5 %
(Intercept)     0.000642846 0.0002744663 0.001387925
age             1.072465271 1.0599060015 1.086076879
hypertensionYes 1.878410965 1.1999860138 2.871972930

$auc
Area under the curve: 0.8435

$predicion
       1915
0.005215943
```