Stoke Data

Miles Tweed

5/11/2021

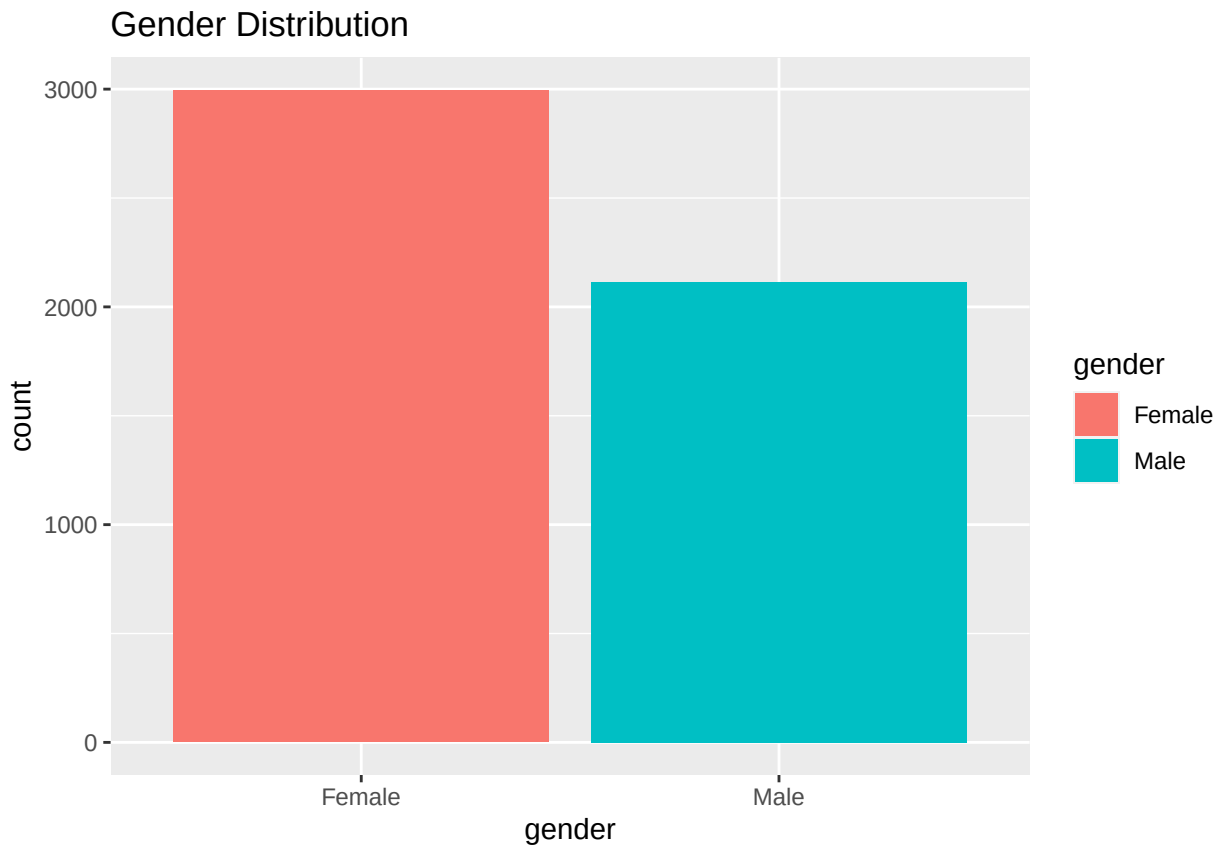
```
Stroke <- read_csv('healthcare-dataset-stroke-data.csv')
Stroke$bmi <- Stroke$bmi %>% as.numeric()
```

```
## Warning in Stroke$bmi %>% as.numeric(): NAs introduced by coercion
```

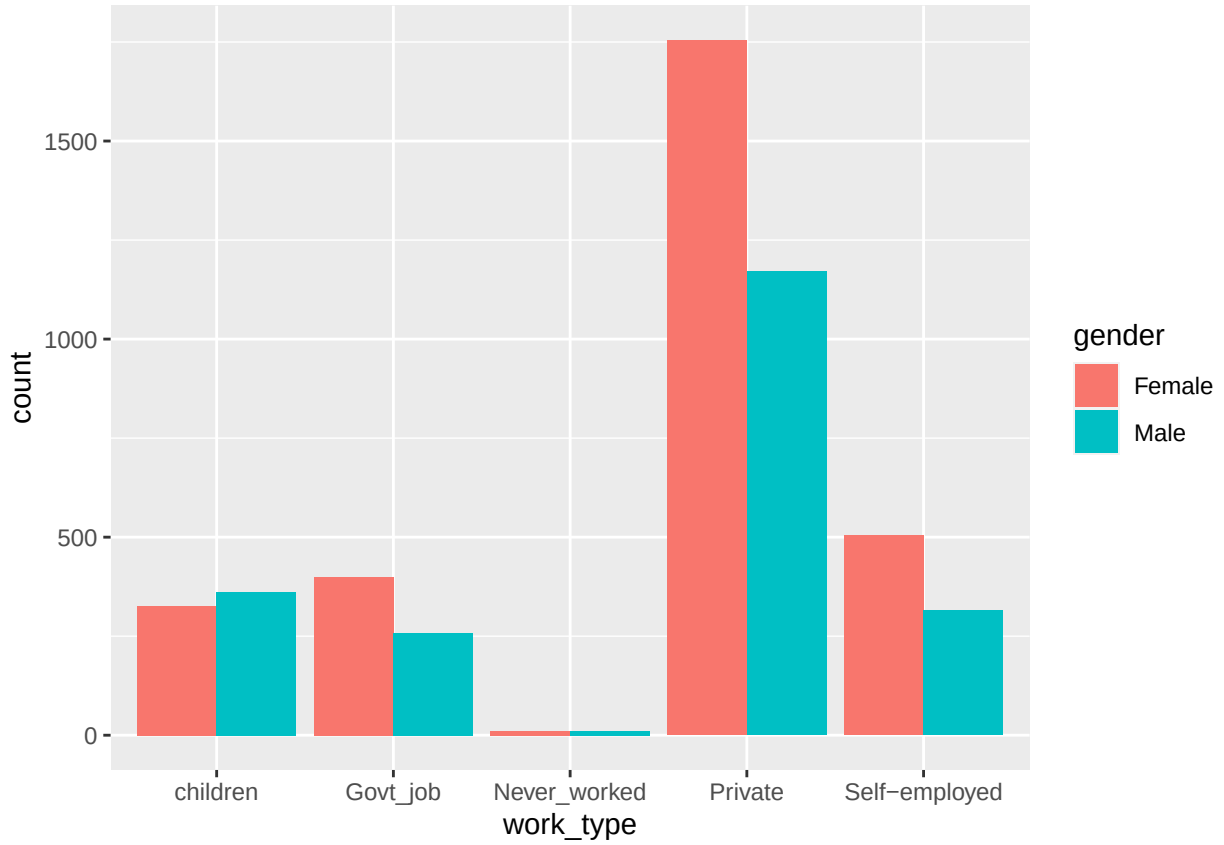
```
Stroke <- Stroke %>% mutate(bmi2 = ifelse(is.na(bmi), median(bmi, na.rm = TRUE), bmi)) %>% select(-bmi,
Stroke$stroke <- factor(Stroke$stroke, levels=c(0,1), labels = c("No","Yes"))
```

Expectation 1

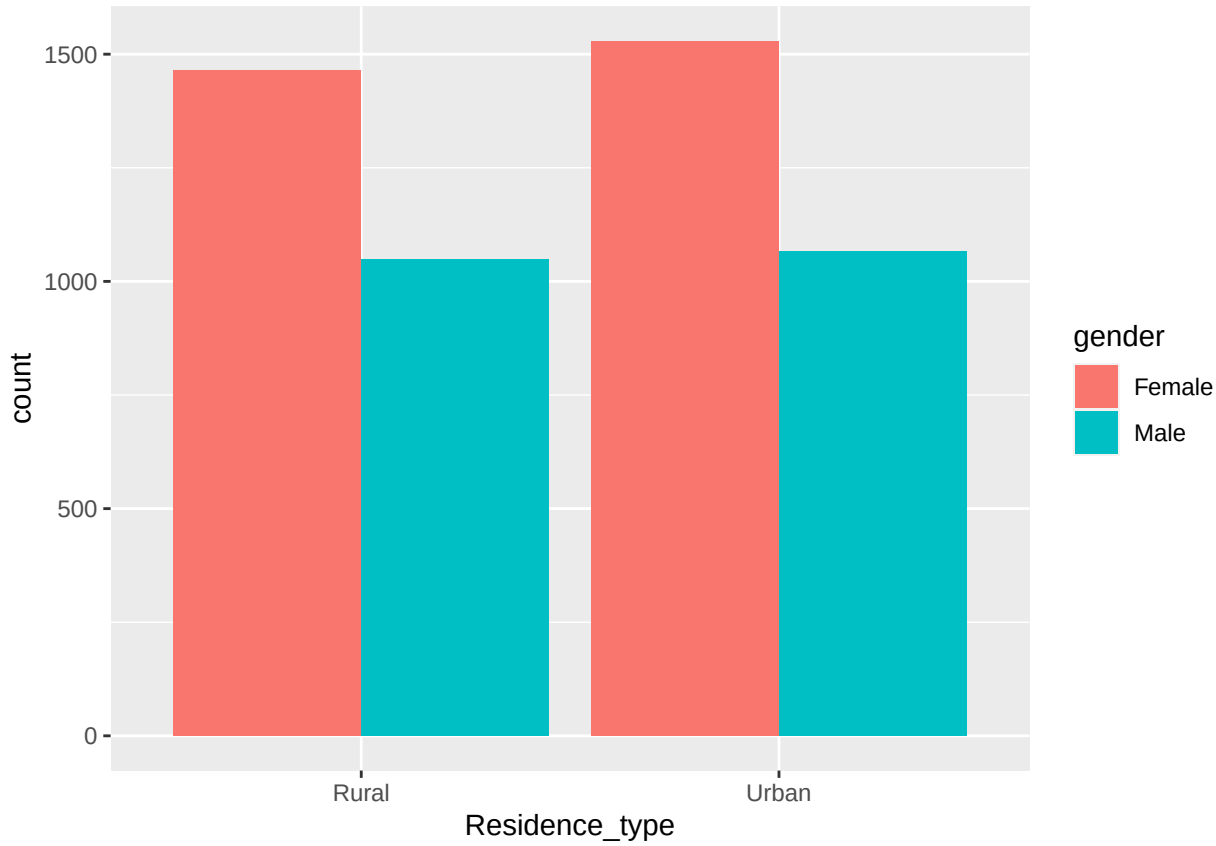
```
ggplot(Stroke, aes(x = gender, fill =  gender)) +
  geom_bar(position = 'dodge') +
  labs(title = "Gender Distribution")
```



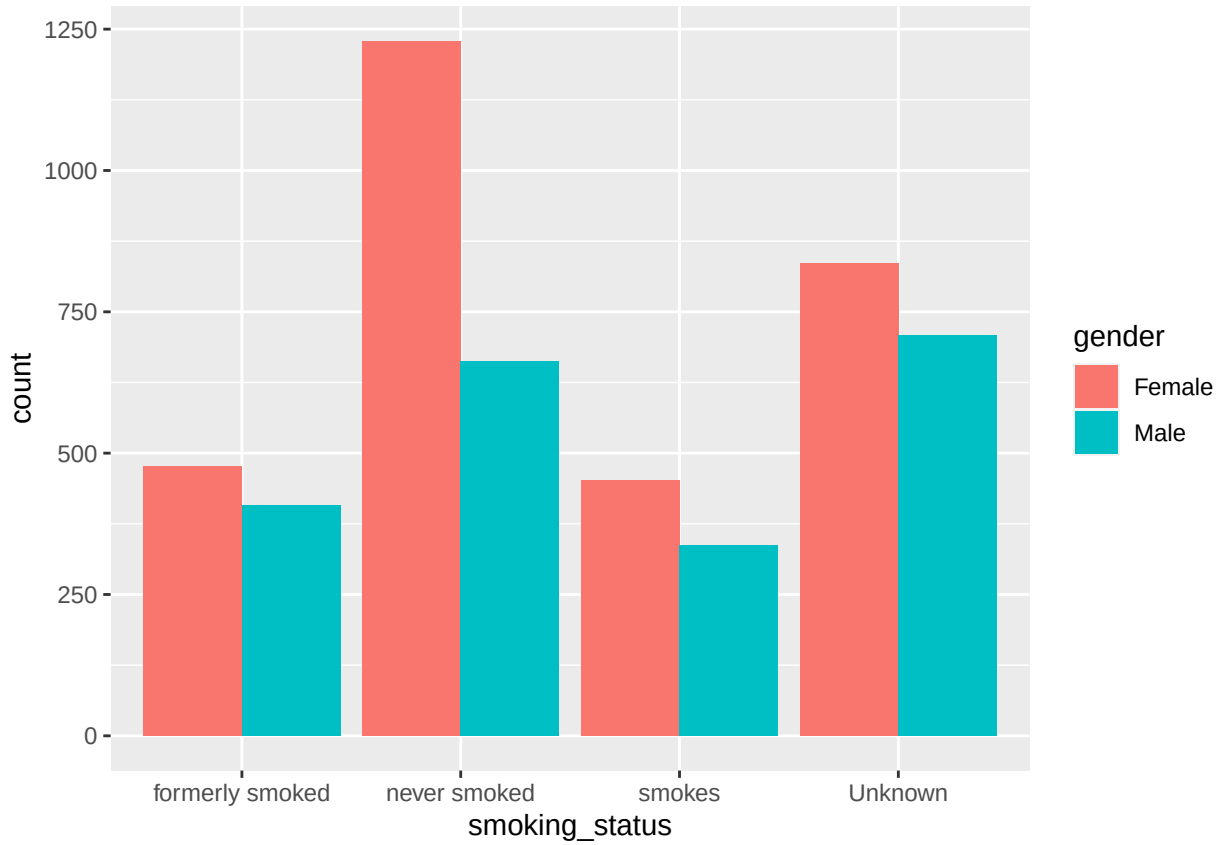
```
ggplot(Stroke, aes(x = work_type, fill =  gender)) +
  geom_bar(position = 'dodge')
```



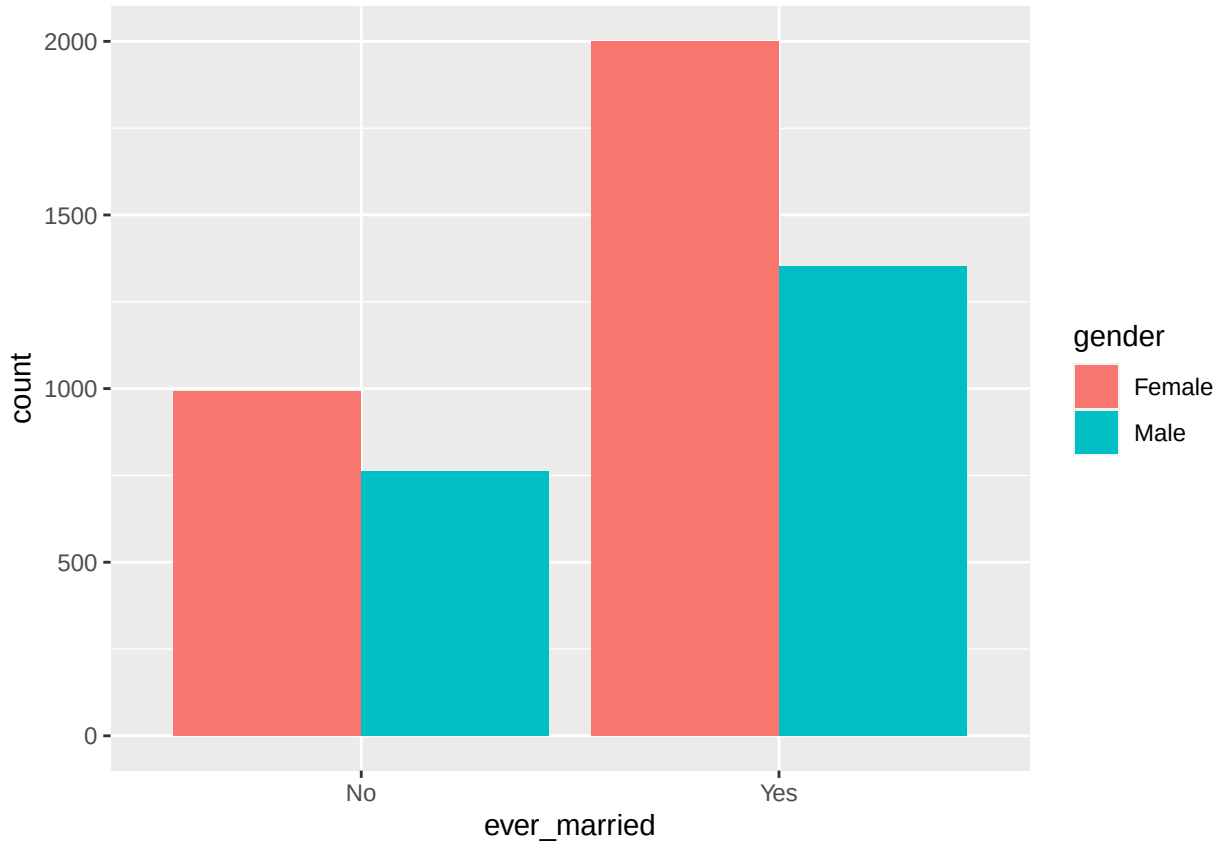

```
ggplot(Stroke, aes(x = Residence_type, fill =  gender)) +
  geom_bar(position = 'dodge')
```



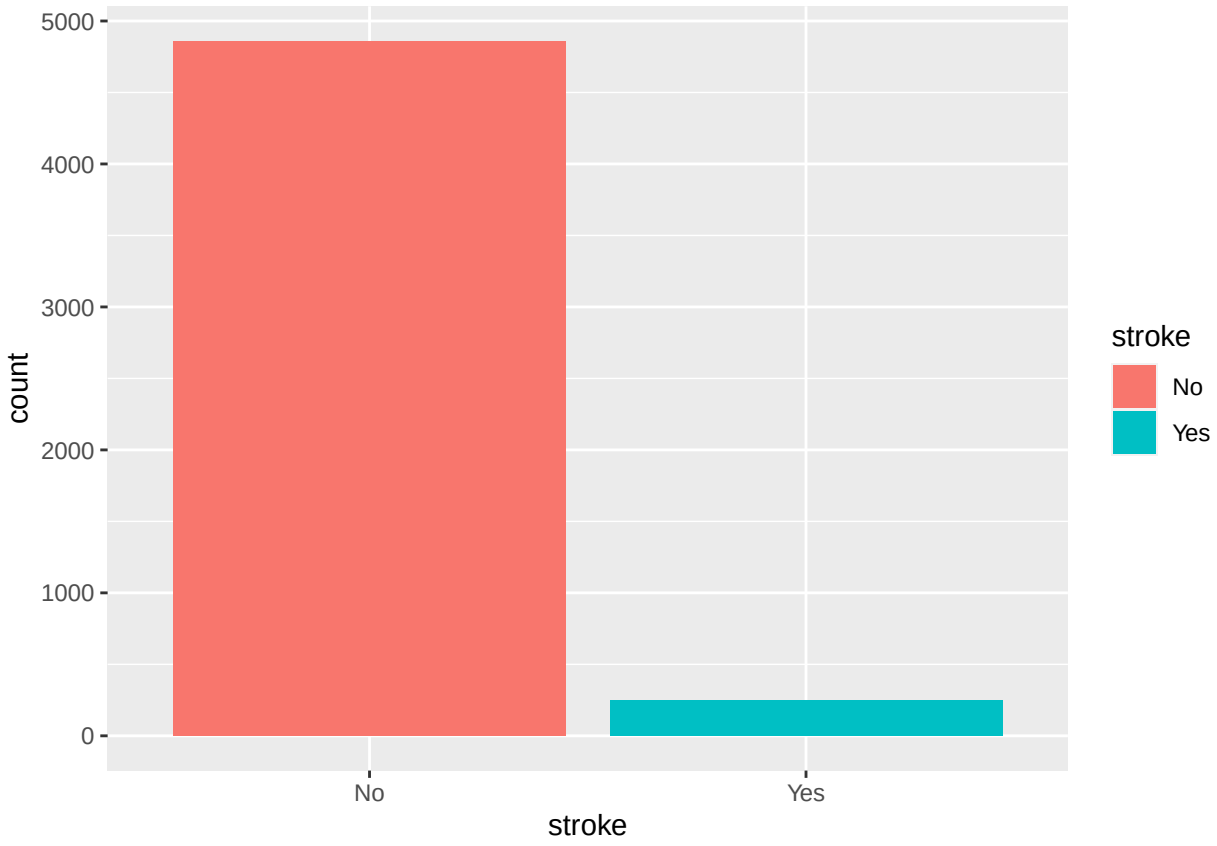

```
ggplot(Stroke, aes(x = smoking_status, fill =  gender)) +
  geom_bar(position = 'dodge')
```




```
ggplot(Stroke, aes(x = ever_married, fill =  gender)) +
  geom_bar(position = 'dodge')
```




```
ggplot(Stroke, aes(x = stroke, fill = stroke)) +
  geom_bar(position = 'dodge')
```



Expectation 2

```
# Making ever_married and smoking_status factors
Stroke$ever_married <- factor(Stroke$ever_married)
Stroke$smoking_status <- factor(Stroke$smoking_status)

# Contingency Table
con.table <- table(Stroke$ever_married, Stroke$smoking_status)
con.table
```

```
##
##       formerly smoked never smoked smokes Unknown
##   No              146          530    179     901
##   Yes             738         1362    610     643
```

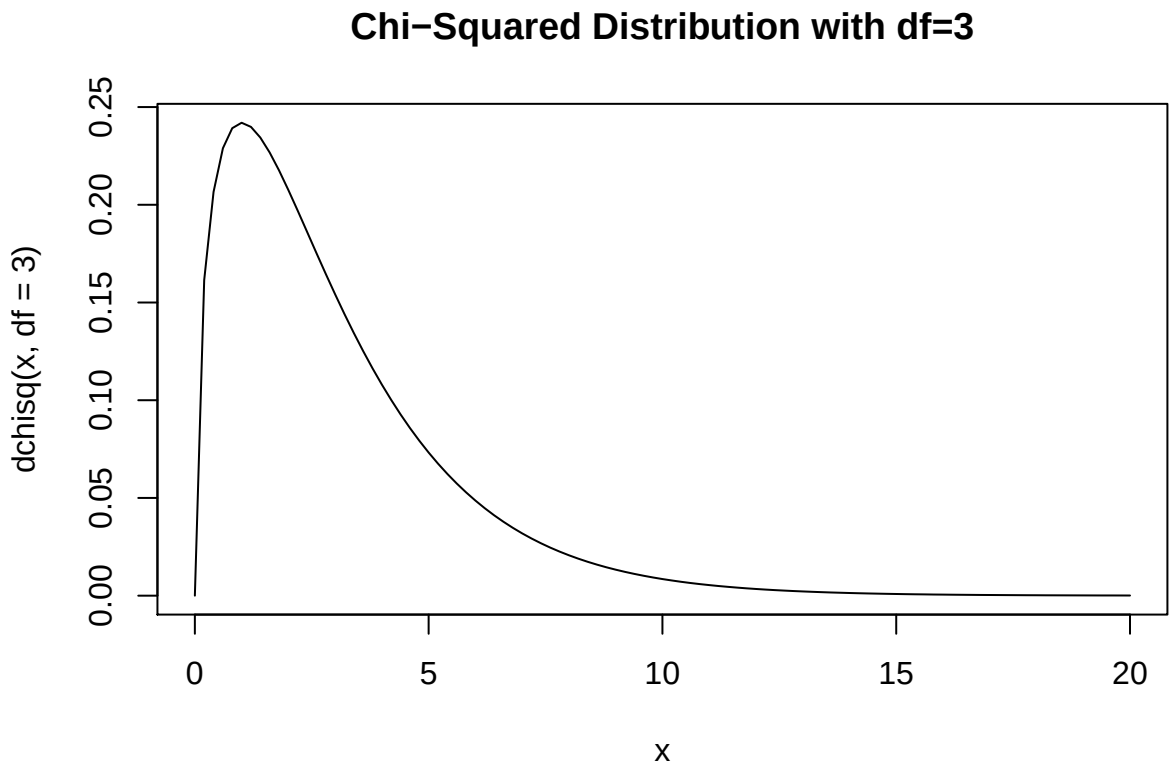
```
# Chi-Squared Test
chisq.test(con.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  con.table
## X-squared = 600.33, df = 3, p-value < 2.2e-16
```

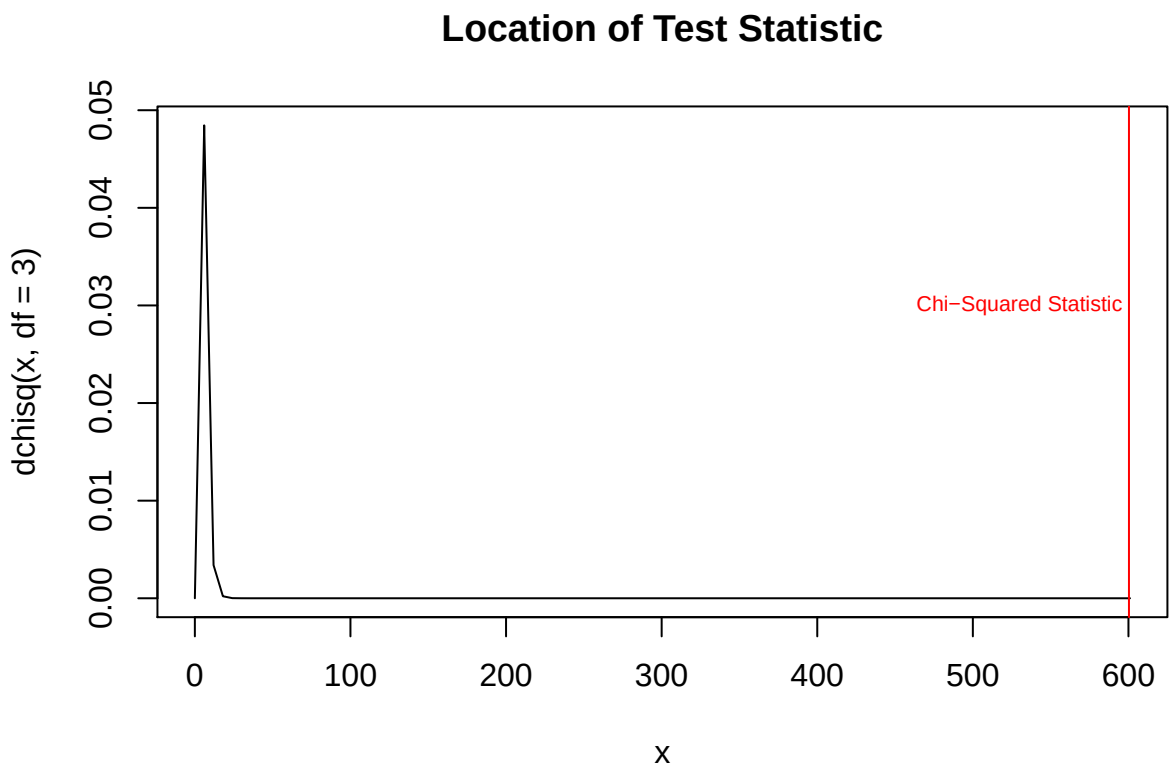
```
# Probability that marriage and smoking status are independent
pchisq(600.33, df=3, lower.tail=F)
```

```
## [1] 8.547784e-130
```

```
x11()
curve(dchisq(x, df = 3), from = 0, to = 20, main = "Chi-Squared Distribution with df=3")
```




```
savePlot("chi_sq.png")
curve(dchisq(x, df = 3), from = 0, to = 601, main = "Location of Test Statistic")
abline(v = 600.33, col='red')
text(x=530, y = 0.03, labels = "Chi-Squared Statistic", col = 'red', cex = 0.7)
```




```
savePlot("chi_sq_stat.png")
```

Expectation 3

```
lm.obj <- lm(bmi2~., Stroke)

summary(lm.obj)
```

```
##
## Call:
## lm(formula = bmi2 ~ ., data = Stroke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.940  -4.389  -1.170   3.215  67.568
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                19.279828   0.473568  40.712  < 2e-16 ***
## genderMale                  0.067688   0.194807   0.347 0.728259
## age                        -0.014734   0.007227  -2.039 0.041534 *
## hypertension                2.207150   0.336475   6.560 5.93e-11 ***
## heart_disease              -0.891137   0.442456  -2.014 0.044055 *
## ever_marriedYes             2.048699   0.279911   7.319 2.89e-13 ***
## work_typeGovt_job           8.372414   0.485973  17.228  < 2e-16 ***
## work_typeNever_worked       5.204431   1.474000   3.531 0.000418 ***
## work_typePrivate            8.374553   0.403623  20.748  < 2e-16 ***
## work_typeSelf-employed      7.911740   0.497444  15.905  < 2e-16 ***
## Residence_typeUrban         0.010257   0.189566   0.054 0.956852
```

```
## avg_glucose_level           0.018547   0.002189   8.472  < 2e-16 ***
## smoking_statusnever smoked -0.369075   0.279917  -1.319 0.187389
## smoking_statussmokes       -0.258365   0.334997  -0.771 0.440597
## smoking_statusUnknown      -0.790955   0.316362  -2.500 0.012445 *
## strokeYes                  -0.771554   0.459276  -1.680 0.093032 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.765 on 5093 degrees of freedom
## Multiple R-squared:  0.2304, Adjusted R-squared:  0.2281
## F-statistic: 101.6 on 15 and 5093 DF,  p-value: < 2.2e-16
```

```
step(lm.obj, trace = 0)
```

```
##
## Call:
## lm(formula = bmi2 ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + smoking_status + stroke,
##     data = Stroke)
##
## Coefficients:
##                (Intercept)                        age
##                   19.32014                    -0.01473
##               hypertension              heart_disease
##                    2.20932                    -0.87894
##           ever_marriedYes           work_typeGovt_job
##                    2.04949                     8.36268
##      work_typeNever_worked            work_typePrivate
##                    5.20717                     8.36521
##     work_typeSelf-employed           avg_glucose_level
##                    7.90031                     0.01858
## smoking_statusnever smoked        smoking_statussmokes
##                   -0.37634                    -0.26007
##      smoking_statusUnknown                   strokeYes
##                   -0.79347                    -0.77187
```

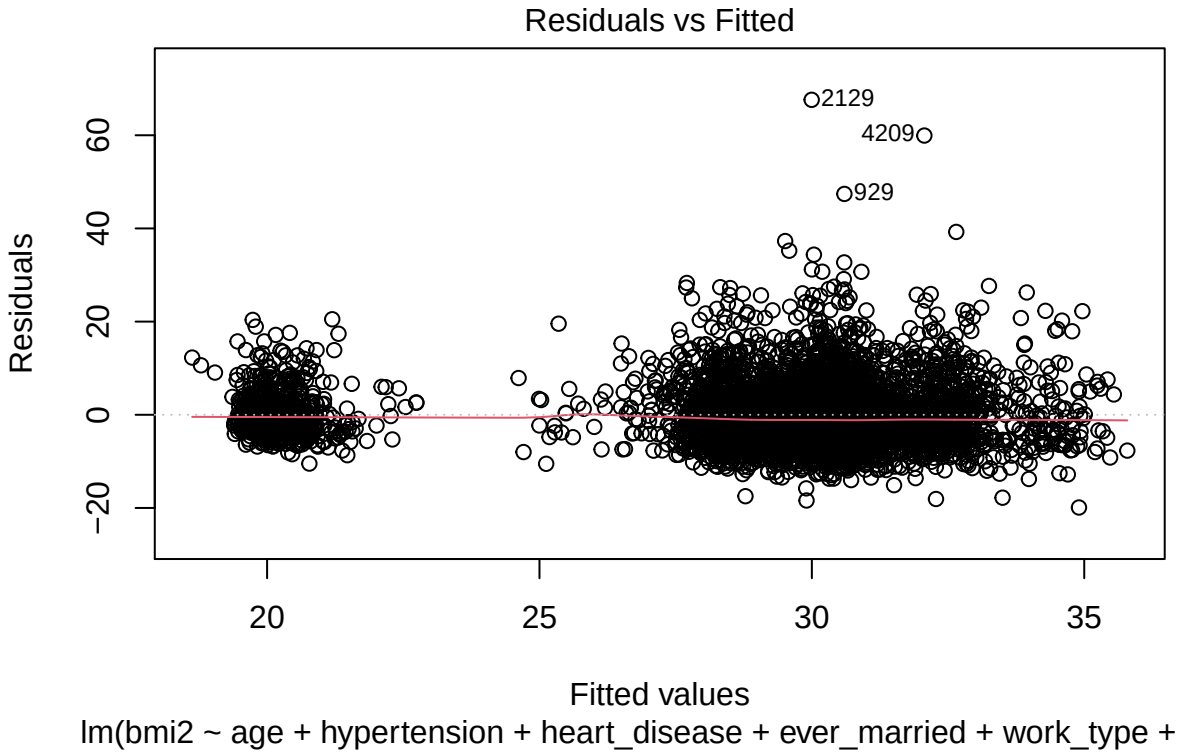
```
lm.reduced <- lm(bmi2 ~ age + hypertension + heart_disease + ever_married +
    work_type + avg_glucose_level + smoking_status + stroke, data = Stroke)

summary(lm.reduced)
```

```
##
## Call:
## lm(formula = bmi2 ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + smoking_status + stroke,
##     data = Stroke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.903  -4.402  -1.169   3.210  67.603
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               19.320144   0.451918  42.751  < 2e-16 ***
## age                       -0.014734   0.007225  -2.039 0.041478 *
```

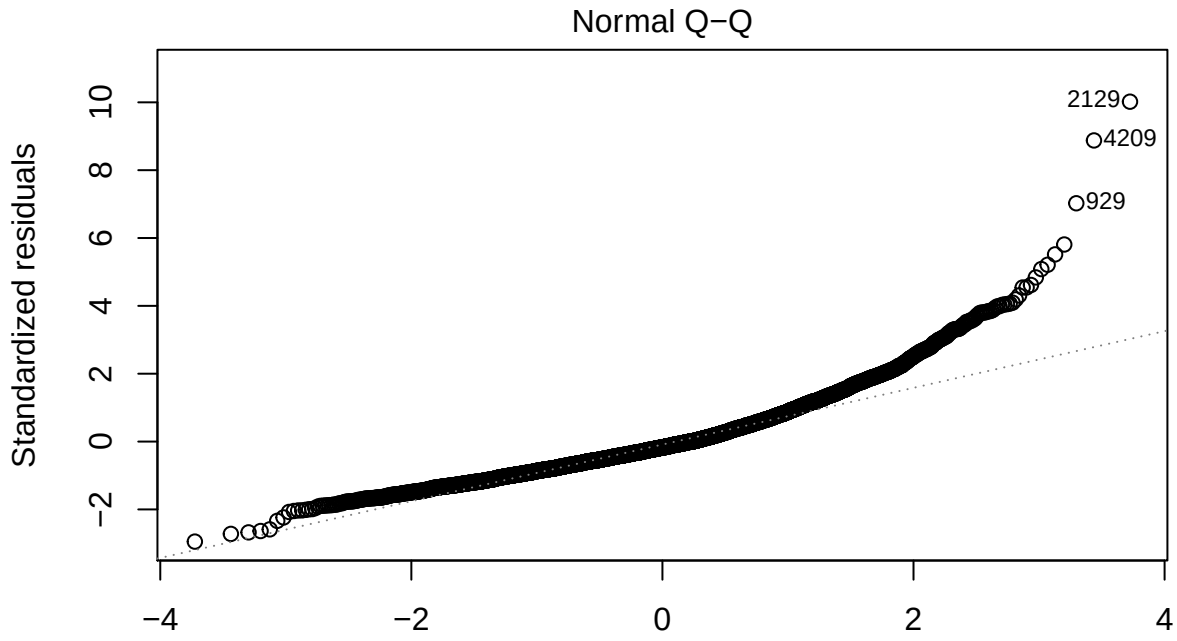
```
## hypertension                 2.209322   0.336316   6.569 5.56e-11 ***
## heart_disease               -0.878938   0.440962  -1.993 0.046290 *
## ever_marriedYes              2.049489   0.279848   7.324 2.79e-13 ***
## work_typeGovt_job            8.362684   0.485079  17.240  < 2e-16 ***
## work_typeNever_worked        5.207175   1.473332   3.534 0.000413 ***
## work_typePrivate             8.365211   0.402670  20.774  < 2e-16 ***
## work_typeSelf-employed       7.900314   0.496280  15.919  < 2e-16 ***
## avg_glucose_level            0.018581   0.002186   8.499  < 2e-16 ***
## smoking_statusnever smoked -0.376343   0.279097  -1.348 0.177580
## smoking_statussmokes       -0.260065   0.334843  -0.777 0.437385
## smoking_statusUnknown      -0.793474   0.316223  -2.509 0.012130 *
## strokeYes                  -0.771871   0.459151  -1.681 0.092808 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.763 on 5095 degrees of freedom
## Multiple R-squared:  0.2304, Adjusted R-squared:  0.2284
## F-statistic: 117.3 on 13 and 5095 DF,  p-value: < 2.2e-16
```

Expectation 4

```
plot(lm.reduced,which=1)
```



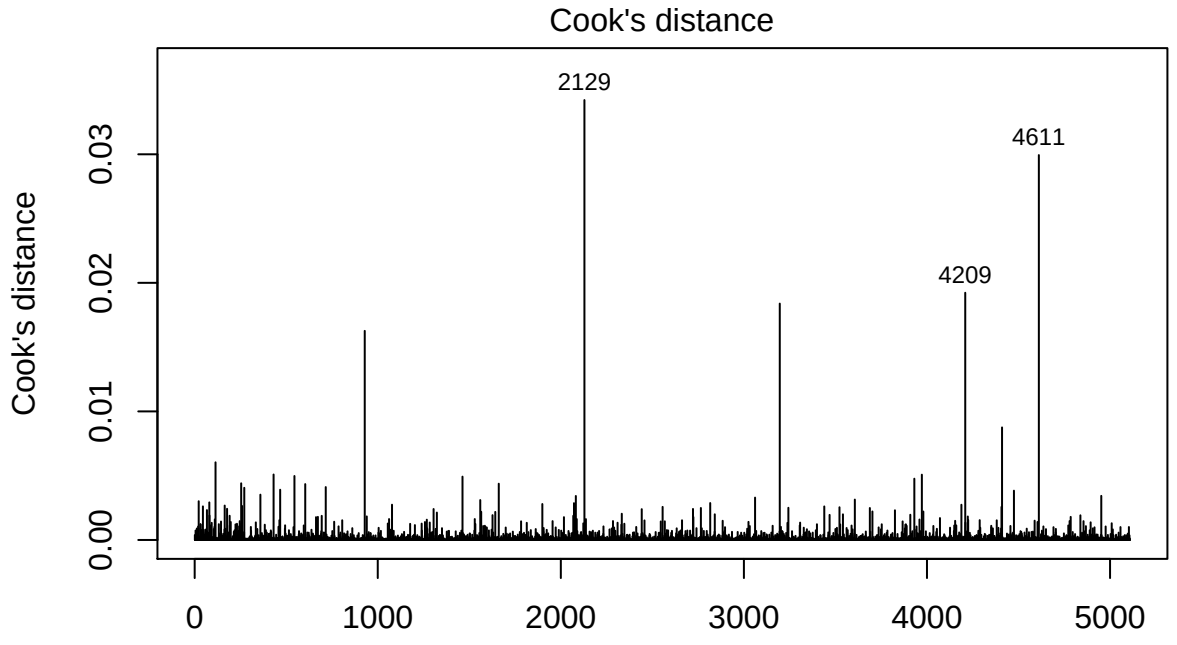

```
plot(lm.reduced,which=2)
```



lm(bmi2 ~ age + hypertension + heart_disease + ever_married + work_type + a ...

Expectation 5

```
plot(lm.reduced, which=4)
```



lm(bmi2 ~ age + hypertension + heart_disease + ever_married + work_type + a ...

```r
# remove outliers based on Cook's distance
Stroke.Out <- Stroke[-c(2129, 4209, 4611),]

# Refit Model
lm.outliers <- lm(bmi2~age + hypertension + heart_disease + ever_married +
    work_type + avg_glucose_level + smoking_status + stroke, data = Stroke.Out)
summary(lm.outliers)
```

```
##
## Call:
## lm(formula = bmi2 ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + smoking_status + stroke,
##     data = Stroke.Out)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.724  -4.353  -1.147   3.224  47.784
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  19.284011   0.443650  43.467  < 2e-16 ***
## age                          -0.012193   0.007095  -1.719  0.08575 .
## hypertension                  1.886813   0.330983   5.701 1.26e-08 ***
## heart_disease                -0.890029   0.432861  -2.056  0.03982 *
## ever_marriedYes               2.047347   0.274768   7.451 1.08e-13 ***
## work_typeGovt_job             8.241659   0.476339  17.302  < 2e-16 ***
## work_typeNever_worked         4.231717   1.478454   2.862  0.00422 **
## work_typePrivate              8.211052   0.395524  20.760  < 2e-16 ***
## work_typeSelf-employed        7.773723   0.487339  15.951  < 2e-16 ***
## avg_glucose_level             0.019313   0.002147   8.997  < 2e-16 ***
## smoking_statusnever smoked   -0.393230   0.273990  -1.435  0.15129
## smoking_statussmokes         -0.237890   0.328695  -0.724  0.46926
## smoking_statusUnknown        -0.848071   0.310477  -2.732  0.00633 **
## strokeYes                    -0.761370   0.450716  -1.689  0.09123 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.639 on 5092 degrees of freedom
## Multiple R-squared:  0.2362, Adjusted R-squared:  0.2343
## F-statistic: 121.1 on 13 and 5092 DF,  p-value: < 2.2e-16
```

Expectation 6

```r
glm.obj <- glm(stroke~., Stroke, family='binomial')

glm.null <- glm(stroke~1, Stroke, family = 'binomial')

# Test of overall model significance
# Likelihood Ratio Test
anova(glm.null, glm.obj, test = "LRT")
```

```
## Analysis of Deviance Table
##
```

```
## Model 1: stroke ~ 1
## Model 2: stroke ~ gender + age + hypertension + heart_disease + ever_married +
##     work_type + Residence_type + avg_glucose_level + smoking_status +
##     bmi2
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5108     1990.3
## 2      5093     1581.2 15   409.12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Forward Stepwise Selection
step(glm.obj, trace = 0)
```

```
##
## Call:  glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
##     family = "binomial", data = Stroke)
##
## Coefficients:
##       (Intercept)                age       hypertension      heart_disease
##         -7.488996           0.068920           0.381396           0.329972
## avg_glucose_level
##          0.004121
##
## Degrees of Freedom: 5108 Total (i.e. Null);  5104 Residual
## Null Deviance:       1990
## Residual Deviance: 1591  AIC: 1601
```

Let $p_i = p(stroke_i = 1 \mid age_i,\ hypertension_i,\ heart\_disease_i,\ avg\_glucose\_level_i)$

$$\begin{cases} stroke_i & \sim_{indep.} Bin(1, p_i), \\ \log\left(\frac{p_i}{1-p_i}\right) & = \beta_0 + \beta_1 age_i + \beta_2 D_{hypertension,i} + \\ & \beta_3 D_{heart_disease,i} + \beta_4 avg\_glucose\_level_i \end{cases}$$

```
# Fit the reduced model
glm.reduced <- glm(stroke ~ age + hypertension + heart_disease + avg_glucose_level, data = Stroke, famil
glm.reduced$coefficients
```

```
##       (Intercept)                age       hypertension      heart_disease
##      -7.488995909        0.068919711        0.381396493        0.329972246
## avg_glucose_level
##        0.004120979
```

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -7.489 + 0.0689 age_i + 0.381 D_{hypertension,i} + 0.330 D_{heart_disease,i} + 0.004 avg\_glucose\_level_i$$

```
anova(glm.null, glm.reduced, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: stroke ~ 1
## Model 2: stroke ~ age + hypertension + heart_disease + avg_glucose_level
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5108     1990.3
```

```
## 2      5104     1591.5  4   398.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#checking for collinearity
vif(glm.reduced)
```

```
##               age      hypertension     heart_disease avg_glucose_level
##          1.076504          1.044221          1.061891          1.049907
```

```
# Reduced model accuracy
glm.pred <- predict(glm.reduced, type='response')
stroke.pred <- ifelse(glm.pred > 0.50, "Yes","No")
stroke.labs <- Stroke$stroke

# Confusion matrix
conf.mat <- table(Pred=stroke.pred,
                  True=stroke.labs)
conf.mat
```

```
##     True
## Pred   No  Yes
##   No 4860  249
```

```
# Misclassification rate
mean(stroke.pred != stroke.labs)
```

```
## [1] 0.04873752
```

```
glm.pred <- predict(glm.obj, type='response')
stroke.pred <- ifelse(glm.pred > 0.50, "Yes","No")
stroke.labs <- Stroke$stroke

# Confusion matrix
conf.mat <- table(Pred=stroke.pred,
                  True=stroke.labs)
conf.mat
```

```
##      True
## Pred    No  Yes
##   No  4860  248
##   Yes    0    1
```

```
# Misclassification rate
mean(stroke.pred != stroke.labs)
```

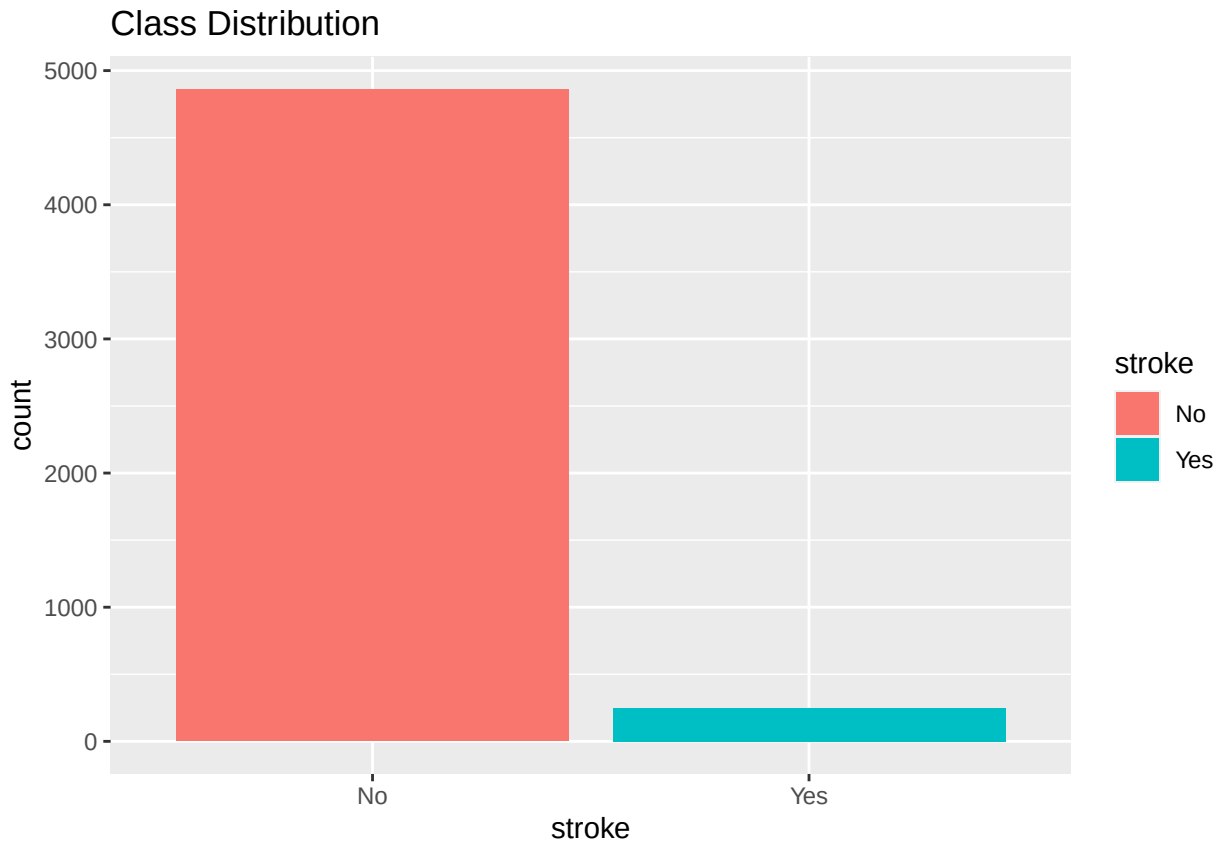
```
## [1] 0.04854179
```

```
Confint(glm.reduced)
```

```
##                      Estimate        2.5 %       97.5 %
## (Intercept)       -7.488995909 -8.216161681 -6.811974269
## age                0.068919711  0.059100995  0.079265708
## hypertension       0.381396493  0.057114291  0.695267479
## heart_disease      0.329972246 -0.046508263  0.690596302
## avg_glucose_level  0.004120979  0.001822614  0.006381289
```

```
ggplot(Stroke, aes(x = stroke, fill = stroke)) +
  geom_bar(position = 'dodge') +
```

```
  labs(title="Class Distribution") +
  ggsave("classDist.png", width = 100, height = 60, units = 'mm')
```




```
## ON A MORE BALANCED DATA SET
set.seed(42)
had.stroke <- Stroke %>%
  filter(stroke == "Yes") %>%
  sample_n(150)
no.stroke <- Stroke %>%
  filter(stroke == "No") %>%
  sample_n(150)

red.stroke.df <- had.stroke %>% union(no.stroke)


glm.obj.bal <- glm(stroke~., red.stroke.df, family='binomial')

glm.null.bal <- glm(stroke~1, red.stroke.df, family = 'binomial')

# Test of overall model significance
# Likelihood Ratio Test
anova(glm.null.bal, glm.obj.bal, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: stroke ~ 1
## Model 2: stroke ~ gender + age + hypertension + heart_disease + ever_married +
```

```
##     work_type + Residence_type + avg_glucose_level + smoking_status +
##     bmi2
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       299     415.89
## 2       285     290.81 14   125.08 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Forward Stepwise Selection
step(glm.obj.bal, trace = 0)
```

```
##
## Call:  glm(formula = stroke ~ gender + age + avg_glucose_level, family = "binomial",
##     data = red.stroke.df)
##
## Coefficients:
##       (Intercept)         genderMale                age  avg_glucose_level
##         -5.141017          -0.440937           0.076473           0.006764
##
## Degrees of Freedom: 299 Total (i.e. Null);  296 Residual
## Null Deviance:       415.9
## Residual Deviance: 295.4     AIC: 303.4
```

```
# Fit the reduced model
glm.reduced.bal <- glm(formula = stroke ~ gender + age + avg_glucose_level, family = "binomial",
    data = red.stroke.df)

glm.reduced.bal$coefficients
```

```
##       (Intercept)        genderMale               age avg_glucose_level
##      -5.141017133      -0.440937162       0.076472765       0.006764088
```

$$\log\left(\frac{p_i}{1-p_i}\right) = -5.141 - 0.441 gender Male_i + 0.0765 age_i + 0.0068 avg\_glucose\_level_i$$

```
anova(glm.null.bal, glm.reduced.bal, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: stroke ~ 1
## Model 2: stroke ~ gender + age + avg_glucose_level
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       299     415.89
## 2       296     295.41  3   120.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#checking for collinearity
vif(glm.reduced.bal)
```

```
##            gender               age avg_glucose_level
##          1.032516          1.026567          1.008227
```

```r
glm.pred.bal <- predict(glm.reduced.bal, type='response')
stroke.pred.bal <- ifelse(glm.pred.bal > 0.50, "Yes","No")
stroke.labs.bal <- red.stroke.df$stroke

# Confusion matrix
conf.mat.bal <- table(Pred=stroke.pred.bal,
                      True=stroke.labs.bal)
conf.mat.bal
```

```
##      True
## Pred   No Yes
##   No  107  32
##   Yes  43 118
```

```r
# Misclassification rate
mean(stroke.pred.bal != stroke.labs.bal)
```

```
## [1] 0.25
```

```r
Confint(glm.reduced.bal)
```

```
##                       Estimate       2.5 %      97.5 %
## (Intercept)       -5.141017133 -6.54967816 -3.89601608
## genderMale        -0.440937162 -1.02203249  0.12848492
## age                0.076472765  0.05835243  0.09674579
## avg_glucose_level  0.006764088  0.00173335  0.01202212
```