Stroke Factors: Classification & Predictive Analytics

Danyang Liu (500936348). Supervisor: Dr. Ceni Babaoglu

3/7/2022

Dataset: removed NAs, removed outliers

Converted cat variables into num

With PCA analysis and Normalization

Read dataset

```
stroke <- read.csv(file="stroke_1_raw.csv",header=T, sep=",")
```

Exploratory Analytics and Data Cleaning

```
# Descriptive analysis
str(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 ...
```

```
summary(stroke)
```

```
##        id           gender               age         hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000
```

```
##  Max.   :72940                      Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  avg_glucose_level     bmi            smoking_status         stroke
##  Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
##  Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
##  Mean   :106.15                                          Mean   :0.04873
##  3rd Qu.:114.09                                          3rd Qu.:0.00000
##  Max.   :271.74                                          Max.   :1.00000
```

```
# Convert 'N/A's (strings) in dataset to NA
is.na(stroke) <- stroke == "N/A"
# Count number of NAs in dataset
sum(is.na(stroke))
```

```
## [1] 201
```

```
# Count number of NAs in all columns
colSums(is.na(stroke))
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0               201                 0                 0
```

```
# Count number of 'Unknown's in all columns
colSums(stroke == "Unknown")
```

```
##                id            gender               age      hypertension
##                 0                 0                 0                 0
##     heart_disease      ever_married         work_type    Residence_type
##                 0                 0                 0                 0
## avg_glucose_level               bmi    smoking_status            stroke
##                 0                NA              1544                 0
```

```
# Remove first column 'id'; irrelevant to data analysis
stroke <- stroke[2:12]

# Check attribute levels and convert data types to numeric
# For binary "Yes"/"No" values, "Yes" = 1 and "No" = 2
str(stroke)
```

```
## 'data.frame':    5110 obs. of  11 variables:
```

```
##  $ 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" NA "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 ...
```

```
unique(stroke$gender)
```

```
## [1] "Male"   "Female" "Other"
```

```
stroke$gender <- gsub("Male", 1, stroke$gender)
stroke$gender <- gsub("Female", 2, stroke$gender)
stroke$gender <- gsub("Other", 3, stroke$gender)
stroke$gender <- as.numeric(stroke$gender)
unique(stroke$gender)
```

```
## [1] 1 2 3
```

```
unique(stroke$ever_married)
```

```
## [1] "Yes" "No"
```

```
stroke$ever_married <- gsub("Yes", 1, stroke$ever_married)
stroke$ever_married <- gsub("No", 0, stroke$ever_married)
stroke$ever_married <- as.numeric(stroke$ever_married)
unique(stroke$ever_married)
```

```
## [1] 1 0
```

```
unique(stroke$work_type)
```

```
## [1] "Private"       "Self-employed" "Govt_job"      "children"
## [5] "Never_worked"
```

```
stroke$work_type <- gsub("Private", 1, stroke$work_type)
stroke$work_type <- gsub("Self-employed", 2, stroke$work_type)
stroke$work_type <- gsub("Govt_job", 3, stroke$work_type)
stroke$work_type <- gsub("children", 4, stroke$work_type)
stroke$work_type <- gsub("Never_worked", 5, stroke$work_type)
stroke$work_type <- as.numeric(stroke$work_type)
unique(stroke$work_type)
```

```
## [1] 1 2 3 4 5
```

```r
unique(stroke$Residence_type)
```

```
## [1] "Urban" "Rural"
```

```r
stroke$Residence_type <- gsub("Urban", 1, stroke$Residence_type)
stroke$Residence_type <- gsub("Rural", 2, stroke$Residence_type)
stroke$Residence_type <- as.numeric(stroke$Residence_type)
unique(stroke$Residence_type)
```

```
## [1] 1 2
```

```r
stroke$bmi <- as.numeric(stroke$bmi)

unique(stroke$smoking_status)
```

```
## [1] "formerly smoked" "never smoked"    "smokes"          "Unknown"
```

```r
stroke$smoking_status <- gsub("formerly smoked", 1, stroke$smoking_status)
stroke$smoking_status <- gsub("never smoked", 2, stroke$smoking_status)
stroke$smoking_status <- gsub("smokes", 3, stroke$smoking_status)
stroke$smoking_status <- gsub("Unknown", 4, stroke$smoking_status)
stroke$smoking_status <- as.numeric(stroke$smoking_status)
unique(stroke$smoking_status)
```

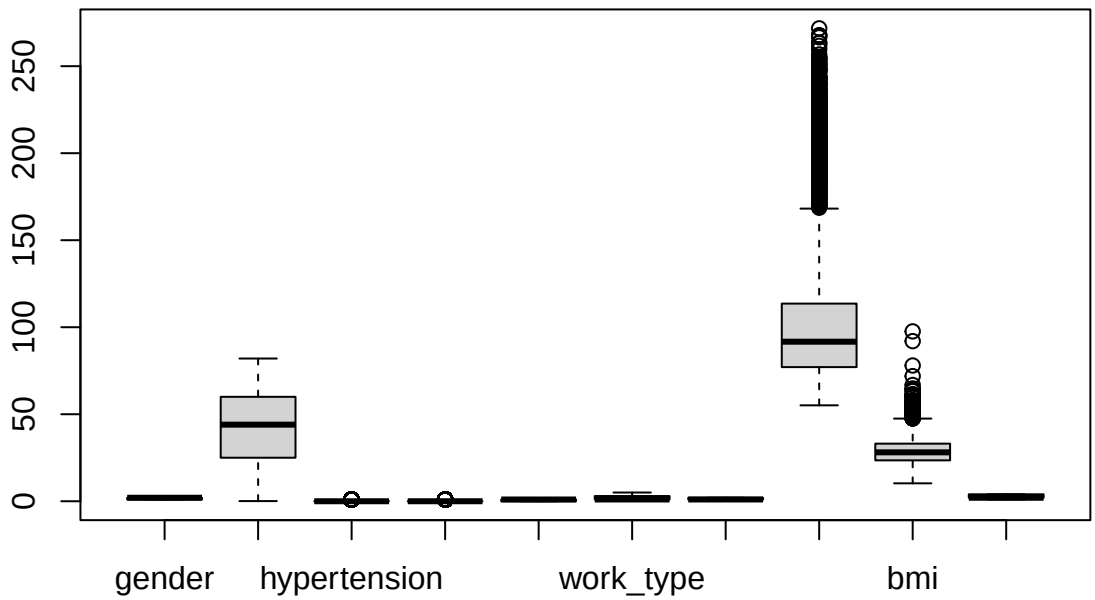
```
## [1] 1 2 3 4
```

```r
# Check that all attributes are now numeric data types
str(stroke)
```

```
## 'data.frame':    5110 obs. of  11 variables:
##  $ gender           : num  1 2 1 2 2 1 1 2 2 2 ...
##  $ 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     : num  1 1 1 1 1 1 1 0 1 1 ...
##  $ work_type        : num  1 2 1 1 2 1 1 1 1 1 ...
##  $ Residence_type   : num  1 2 2 1 2 1 2 1 2 1 ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : num  1 2 2 3 2 1 2 2 4 4 ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Deal with NAs
# Method 1: remove NAs
stroke_noNAs <- stroke[complete.cases(stroke), ]

# Deal with outliers
# Box plot to visualize outliers
boxplot(as.matrix(stroke_noNAs[1:10]))
```




```
# Excluding categorical variables, avg_glucose_level
# And bmi have several outliers
# Remove outliers using interquartile range values
agl_outliers <- boxplot(stroke$avg_glucose_level, plot = FALSE)$out
bmi_outliers <- boxplot(stroke$bmi, plot = FALSE)$out
stroke_noNAs_noOL <- stroke_noNAs
stroke_noNAs_noOL <- stroke_noNAs_noOL[-which(stroke_noNAs_noOL$avg_glucose_level %in% agl_outliers),]
stroke_noNAs_noOL <- stroke_noNAs_noOL[-which(stroke_noNAs_noOL$bmi %in% bmi_outliers),]

# Examine correlations between all Independent Variables
cor(stroke_noNAs_noOL[1:10])
```

```
##                       gender          age  hypertension heart_disease
## gender             1.0000000000  0.047163661 -0.0181116010  -0.087077746
## age                0.0471636606  1.000000000  0.2492046322   0.236193434
## hypertension      -0.0181116010  0.249204632  1.0000000000   0.106065206
## heart_disease     -0.0870777462  0.236193434  0.1060652062   1.000000000
## ever_married       0.0508315382  0.687498881  0.1488340141   0.105364898
## work_type         -0.0758616230 -0.439614390 -0.0721676438  -0.041084225
## Residence_type    -0.0003523739 -0.009598891  0.0038834139   0.014064422
## avg_glucose_level -0.0305248091 -0.023924488 -0.0009078475   0.004947325
## bmi                0.0054726191  0.378683833  0.1515384482   0.054618944
## smoking_status    -0.0590370218 -0.385509590 -0.1155068859  -0.057584935
##                    ever_married   work_type Residence_type avg_glucose_level
## gender             0.0508315382 -0.07586162  -0.0003523739     -0.0305248091
## age                0.6874988811 -0.43961439  -0.0095988913     -0.0239244877
```

```
## hypertension       0.1488340141 -0.07216764   0.0038834139     -0.0009078475
## heart_disease      0.1053648985 -0.04108422   0.0140644220      0.0049473252
## ever_married       1.0000000000 -0.39116104   0.0004186879     -0.0083602287
## work_type         -0.3911610411  1.00000000  -0.0155902656      0.0109823333
## Residence_type     0.0004186879 -0.01559027   1.0000000000      0.0145557947
## avg_glucose_level -0.0083602287  0.01098233   0.0145557947      1.0000000000
## bmi                0.3756328526 -0.38386175  -0.0110487374      0.0017839920
## smoking_status    -0.3177225122  0.33765019  -0.0042874498      0.0178261247
##                           bmi smoking_status
## gender             0.005472619    -0.05903702
## age                0.378683833    -0.38550959
## hypertension       0.151538448    -0.11550689
## heart_disease      0.054618944    -0.05758493
## ever_married       0.375632853    -0.31772251
## work_type         -0.383861746     0.33765019
## Residence_type    -0.011048737    -0.00428745
## avg_glucose_level  0.001783992     0.01782612
## bmi                1.000000000    -0.26338455
## smoking_status    -0.263384550     1.00000000
```

```
# PCA and normalization
stroke.pca.normdata <- prcomp(stroke_noNAs_noOL, scale = TRUE, center = TRUE)
stroke.pca.normdata$rotation
```

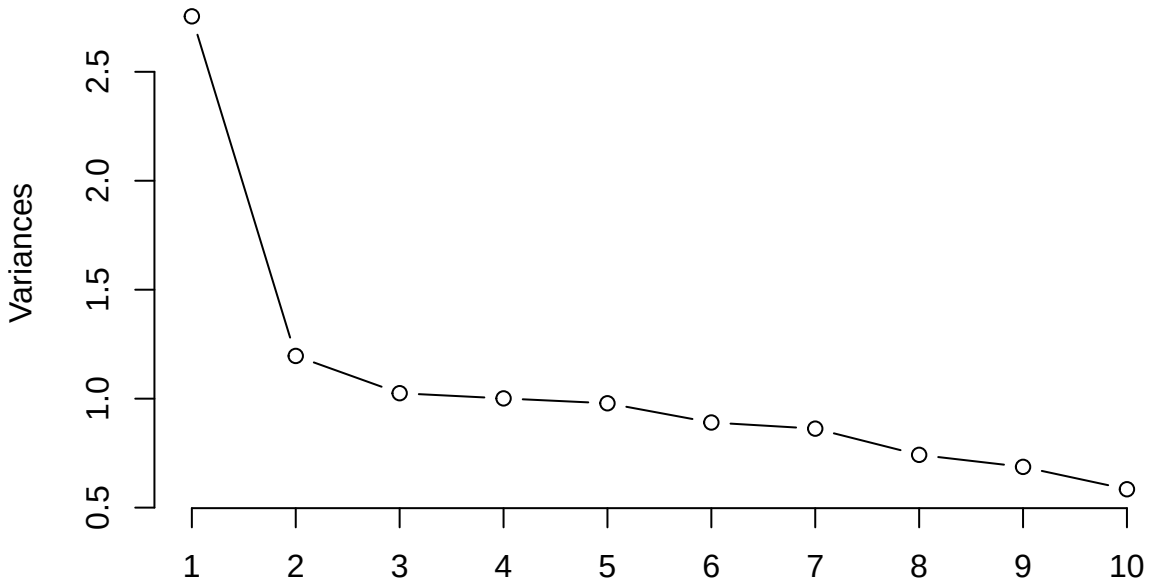
```
##                           PC1         PC2         PC3          PC4
## gender             0.0490864170 -0.44419923  0.42363430 -0.431526253
## age                0.5118693890  0.10031489  0.05238915 -0.007236396
## hypertension       0.2031465059  0.38573878  0.15993034 -0.076824727
## heart_disease      0.1472145917  0.56255222 -0.02783112  0.045596061
## ever_married       0.4704298966 -0.05773920 -0.03613982  0.032416974
## work_type         -0.4023363643  0.26853414  0.11922666 -0.018371044
## Residence_type     0.0008712861  0.03571302 -0.40367660 -0.831243305
## avg_glucose_level -0.0150753794  0.11536113 -0.66240218 -0.006347625
## bmi                0.3775435371 -0.14936062 -0.17580593  0.157269566
## smoking_status    -0.3570976388  0.16135348  0.01860592  0.018094077
## stroke             0.1384734760  0.43473327  0.37806169 -0.297147264
##                           PC5         PC6          PC7          PC8
## gender             0.375455637 -0.16177978  0.490589254 -0.07850954
## age                0.006022925 -0.10578801  0.025572370 -0.12230496
## hypertension       0.098130058  0.78023046  0.341662106  0.07724541
## heart_disease     -0.218757040 -0.52086365  0.497257450  0.06398503
## ever_married      -0.013916744 -0.08810412  0.001350278 -0.28090877
## work_type          0.037450781  0.07308880  0.120432388  0.02270044
## Residence_type    -0.364656259  0.06909450 -0.057251417 -0.03825000
## avg_glucose_level  0.727932484 -0.06013280  0.109058129  0.03468614
## bmi               -0.060474692  0.20098152 -0.050984963 -0.33029554
## smoking_status     0.011359942  0.01663442  0.079502560 -0.88055966
## stroke             0.365062017 -0.14287443 -0.596718497 -0.04738266
##                           PC9         PC10         PC11
## gender             0.121634957 -0.075842304  0.008265948
## age               -0.341209008 -0.005302509 -0.762783147
## hypertension       0.005593231  0.172686340  0.086698099
## heart_disease      0.268850401  0.009016145  0.127022939
## ever_married      -0.546317537 -0.135040278  0.607814650
```

```
## work_type         -0.313100597 -0.789638498 -0.090208803
## Residence_type    -0.014962089 -0.042902096 -0.015814985
## avg_glucose_level -0.027463824  0.006345799 -0.015956675
## bmi                0.601821434 -0.510945282 -0.022612977
## smoking_status    -0.013686996  0.241662036 -0.071226123
## stroke             0.191364035 -0.034736884  0.103810616
```

```
# Values of normalized data after transformation
head(stroke.pca.normdata$x)
```

```
##         PC1      PC2       PC3       PC4        PC5        PC6        PC7
## 3  3.441946 5.558467 0.3938259 -1.580346  0.4167229 -3.6309256 -1.5299321
## 7  3.846391 6.989832 2.1897170 -1.991781 -0.2974448 -0.5725982 -0.3177422
## 8  0.996406 1.805902 2.8499995 -1.326315  2.8221179 -1.3427154 -3.0699948
## 10 1.619917 1.809621 3.8372304 -1.186544  1.6575666 -1.3996745 -3.0938477
## 11 3.441993 3.114412 2.8552370 -3.064537  1.9468186  1.8884097 -1.9351728
## 12 2.320312 5.158517 0.8979079 -2.370695  1.6771036 -3.6392073 -0.2360772
##            PC8        PC9       PC10        PC11
## 3  -0.04066612 1.89671178 -0.1506647  0.39677169
## 7   0.49369646 1.59666867  0.9118594  0.98604783
## 8   0.63228342 1.17671979  0.7477191 -1.07123493
## 10 -1.73056225 0.05765966  0.7931528 -0.23440349
## 11 -0.14313150 0.49538000  0.5358033  0.07165724
## 12 -1.05263441 2.24826190 -1.7630739  0.81384944
```

```
# Visualize PCA to see most important principal components
plot(stroke.pca.normdata, type = "l", main = "With data normalization")
```




```
# Elbow point occurs around PC2 (if 1.0 as threshold) or PC8 (if 0.75 as threshold)
```

```
# Check correlations of original vs normalized transformed data
# Original
cor(stroke_noNAs_noOL)
```

```
##                        gender          age  hypertension heart_disease
## gender             1.0000000000  0.047163661 -0.0181116010  -0.087077746
## age                0.0471636606  1.000000000  0.2492046322   0.236193434
## hypertension      -0.0181116010  0.249204632  1.0000000000   0.106065206
## heart_disease     -0.0870777462  0.236193434  0.1060652062   1.000000000
## ever_married       0.0508315382  0.687498881  0.1488340141   0.105364898
## work_type         -0.0758616230 -0.439614390 -0.0721676438  -0.041084225
## Residence_type    -0.0003523739 -0.009598891  0.0038834139   0.014064422
## avg_glucose_level -0.0305248091 -0.023924488 -0.0009078475   0.004947325
## bmi                0.0054726191  0.378683833  0.1515384482   0.054618944
## smoking_status    -0.0590370218 -0.385509590 -0.1155068859  -0.057584935
## stroke             0.0031328665  0.209844238  0.1198515115   0.093080480
##                    ever_married   work_type Residence_type avg_glucose_level
## gender             0.0508315382 -0.07586162  -3.523739e-04     -0.0305248091
## age                0.6874988811 -0.43961439  -9.598891e-03     -0.0239244877
## hypertension       0.1488340141 -0.07216764   3.883414e-03     -0.0009078475
## heart_disease      0.1053648985 -0.04108422   1.406442e-02      0.0049473252
## ever_married       1.0000000000 -0.39116104   4.186879e-04     -0.0083602287
## work_type         -0.3911610411  1.00000000  -1.559027e-02      0.0109823333
```

```
## Residence_type     0.0004186879 -0.01559027   1.000000e+00      0.0145557947
## avg_glucose_level -0.0083602287  0.01098233   1.455579e-02      1.0000000000
## bmi                0.3756328526 -0.38386175  -1.104874e-02      0.0017839920
## smoking_status    -0.3177225122  0.33765019  -4.287450e-03      0.0178261247
## stroke             0.0896453377 -0.04859507  -7.082066e-05      0.0056519121
##                             bmi smoking_status        stroke
## gender             0.005472619    -0.05903702  3.132866e-03
## age                0.378683833    -0.38550959  2.098442e-01
## hypertension       0.151538448    -0.11550689  1.198515e-01
## heart_disease      0.054618944    -0.05758493  9.308048e-02
## ever_married       0.375632853    -0.31772251  8.964534e-02
## work_type         -0.383861746     0.33765019 -4.859507e-02
## Residence_type    -0.011048737    -0.00428745 -7.082066e-05
## avg_glucose_level  0.001783992     0.01782612  5.651912e-03
## bmi                1.000000000    -0.26338455  3.092483e-02
## smoking_status    -0.263384550     1.00000000 -6.724810e-02
## stroke             0.030924826    -0.06724810  1.000000e+00
```

```
# Normalized transformed
cor(stroke.pca.normdata$x)
```

```
##                PC1           PC2           PC3           PC4           PC5
## PC1   1.000000e+00 -1.303699e-15 -1.824337e-15  1.337965e-15 -1.415912e-15
## PC2  -1.303699e-15  1.000000e+00  8.146696e-15 -9.281104e-15  7.735391e-15
## PC3  -1.824337e-15  8.146696e-15  1.000000e+00  9.535617e-15 -9.564393e-15
## PC4   1.337965e-15 -9.281104e-15  9.535617e-15  1.000000e+00  8.258918e-15
## PC5  -1.415912e-15  7.735391e-15 -9.564393e-15  8.258918e-15  1.000000e+00
## PC6  -3.375997e-15 -8.679475e-15  1.715411e-15 -2.412462e-15  2.784916e-15
## PC7  -2.400832e-15  1.028706e-14 -1.161268e-14  1.257343e-14 -1.053726e-14
## PC8   2.434771e-15 -1.345733e-15  9.663284e-16 -1.722727e-15  4.889136e-16
## PC9   1.940548e-15  5.476395e-15 -3.000821e-15  3.770428e-15 -4.247938e-15
## PC10 -1.244897e-15 -3.261725e-15  2.949046e-15 -2.763150e-15  1.787171e-15
## PC11 -8.155478e-15 -4.843047e-16 -9.157315e-16  3.682951e-16 -7.114039e-16
##                PC6           PC7           PC8           PC9           PC10
## PC1  -3.375997e-15 -2.400832e-15  2.434771e-15  1.940548e-15 -1.244897e-15
## PC2  -8.679475e-15  1.028706e-14 -1.345733e-15  5.476395e-15 -3.261725e-15
## PC3   1.715411e-15 -1.161268e-14  9.663284e-16 -3.000821e-15  2.949046e-15
## PC4  -2.412462e-15  1.257343e-14 -1.722727e-15  3.770428e-15 -2.763150e-15
## PC5   2.784916e-15 -1.053726e-14  4.889136e-16 -4.247938e-15  1.787171e-15
## PC6   1.000000e+00 -4.254377e-16 -1.311467e-15  1.356848e-16 -2.882571e-15
## PC7  -4.254377e-16  1.000000e+00  1.642074e-15 -2.147479e-15  2.378676e-15
## PC8  -1.311467e-15  1.642074e-15  1.000000e+00  2.568080e-16 -1.449092e-15
## PC9   1.356848e-16 -2.147479e-15  2.568080e-16  1.000000e+00 -5.955315e-16
## PC10 -2.882571e-15  2.378676e-15 -1.449092e-15 -5.955315e-16  1.000000e+00
## PC11  4.843115e-16  1.200636e-15  1.417276e-15  3.483670e-15 -9.372143e-16
##               PC11
## PC1  -8.155478e-15
## PC2  -4.843047e-16
## PC3  -9.157315e-16
## PC4   3.682951e-16
## PC5  -7.114039e-16
## PC6   4.843115e-16
## PC7   1.200636e-15
## PC8   1.417276e-15
```

```
## PC9   3.483670e-15
## PC10 -9.372143e-16
## PC11  1.000000e+00
```

```
# Correlations between PCs in normalized transformed data are almost 0 - these PCs are now orthogonal
```

```
# Normalize continuous numeric variables
# Such as age, avg_blood_glucose, and bmi
# Using z-score methods
stroke_noNAs_noOL$age <- (stroke_noNAs_noOL$age - mean(stroke_noNAs_noOL$age))/sd(stroke_noNAs_noOL$age)
stroke_noNAs_noOL$avg_glucose_level <- (stroke_noNAs_noOL$avg_glucose_level - mean(stroke_noNAs_noOL$avg
stroke_noNAs_noOL$bmi <- (stroke_noNAs_noOL$bmi - mean(stroke_noNAs_noOL$bmi))/sd(stroke_noNAs_noOL$bmi)
```

Classification

Predictive Analytics: Logistic Regression

```
# Split dataset into 70% training, 30% testing sets
stroke_index1 <- sample(1:nrow(stroke_noNAs_noOL), 0.7 * nrow(stroke_noNAs_noOL))

# Assign selected sample as training set
# Assign leftover dataset as test set
train.set1 <- stroke_noNAs_noOL[stroke_index1,]
test.set1 <- stroke_noNAs_noOL[-stroke_index1,]

# Logistic regression model for prediction
glm_model1 <- glm(formula = stroke~., data = train.set1, family = "binomial")
summary(glm_model1)
```

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = train.set1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0987  -0.2374  -0.1283  -0.0722   3.2982
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.315813   0.685178  -6.299    3e-10 ***
## gender            -0.066869   0.229814  -0.291 0.771075
## age                1.663418   0.179908   9.246  < 2e-16 ***
## hypertension       0.878286   0.259329   3.387 0.000707 ***
## heart_disease      0.174166   0.352491   0.494 0.621235
## ever_married      -0.472407   0.316686  -1.492 0.135773
## work_type         -0.007895   0.149501  -0.053 0.957884
## Residence_type     0.258132   0.222560   1.160 0.246117
## avg_glucose_level  0.055587   0.106026   0.524 0.600089
## bmi               -0.119108   0.135910  -0.876 0.380828
## smoking_status    -0.103036   0.110076  -0.936 0.349251
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 834.93  on 2981  degrees of freedom
## Residual deviance: 653.47  on 2971  degrees of freedom
## AIC: 675.47
##
## Number of Fisher Scoring iterations: 8
```

Evaluation Metrics

```
predicted1 <- predict(glm_model1, test.set1, type = "response")
# Setting 0.5 as threshold - binary prediction
predicted_class1 <- ifelse(predicted1 >= 0.5, "Stroke", "No Stroke")
ConfusionMatrix1 <- table(actual = test.set1$stroke, predicted = predicted_class1)
ConfusionMatrix1
```

```
##       predicted
## actual No Stroke
##      0      1237
##      1        42
```

Abysmal predictions using only logistic regression applied to dataset with NAs removed (from BMI column) and outliers removed. No strokes are predicted at all