2021 Stroke Prediction

Marriah Lewis and Glory Onyeugbo
IST 707 Data Mining Project

Background

- When a blood vessel carrying oxygen and nutrients to the brain is either blocked by a clot or bursts, a stroke occurs.
- Stroke is the No. 5 cause of death and leading cause of disability in the United States.
- Up to 80% of second clot-related strokes can be avoided. There is strong evidence that high glucose levels can contribute to stroke.
- Symptoms of stroke include trouble speaking and understanding, paralysis or numbness, lack of sight, headache, trouble walking.

Motivation

- Goal: Create a model to predict whether target patients had a stroke based on attribute subset.

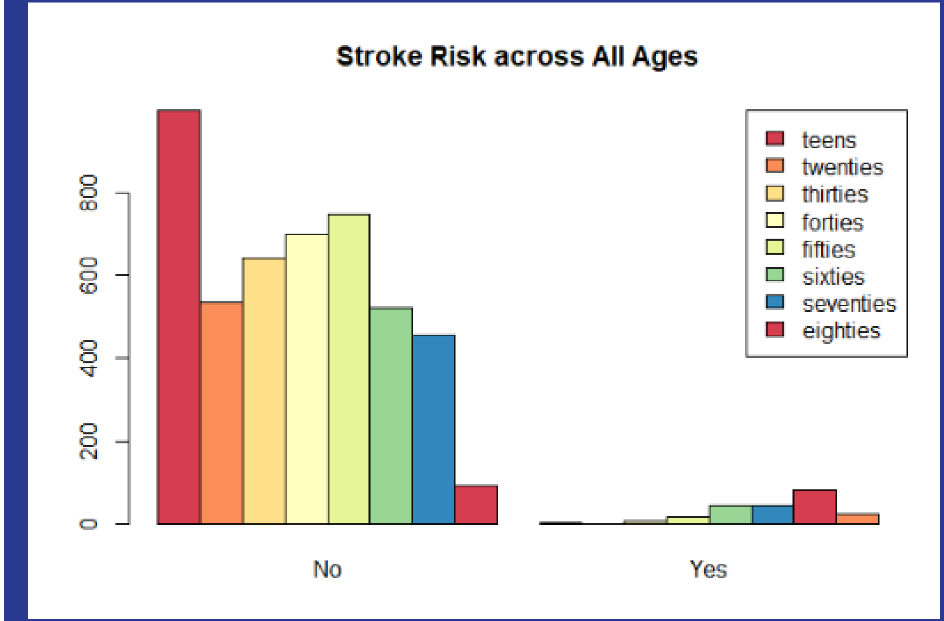
- Some stroke risk factors are gender, age, and family history.

- Lifestyle factors that can increase the risk of stroke are smoking, high glucose levels, poor diet, and lack of exercise.

Problem Statement

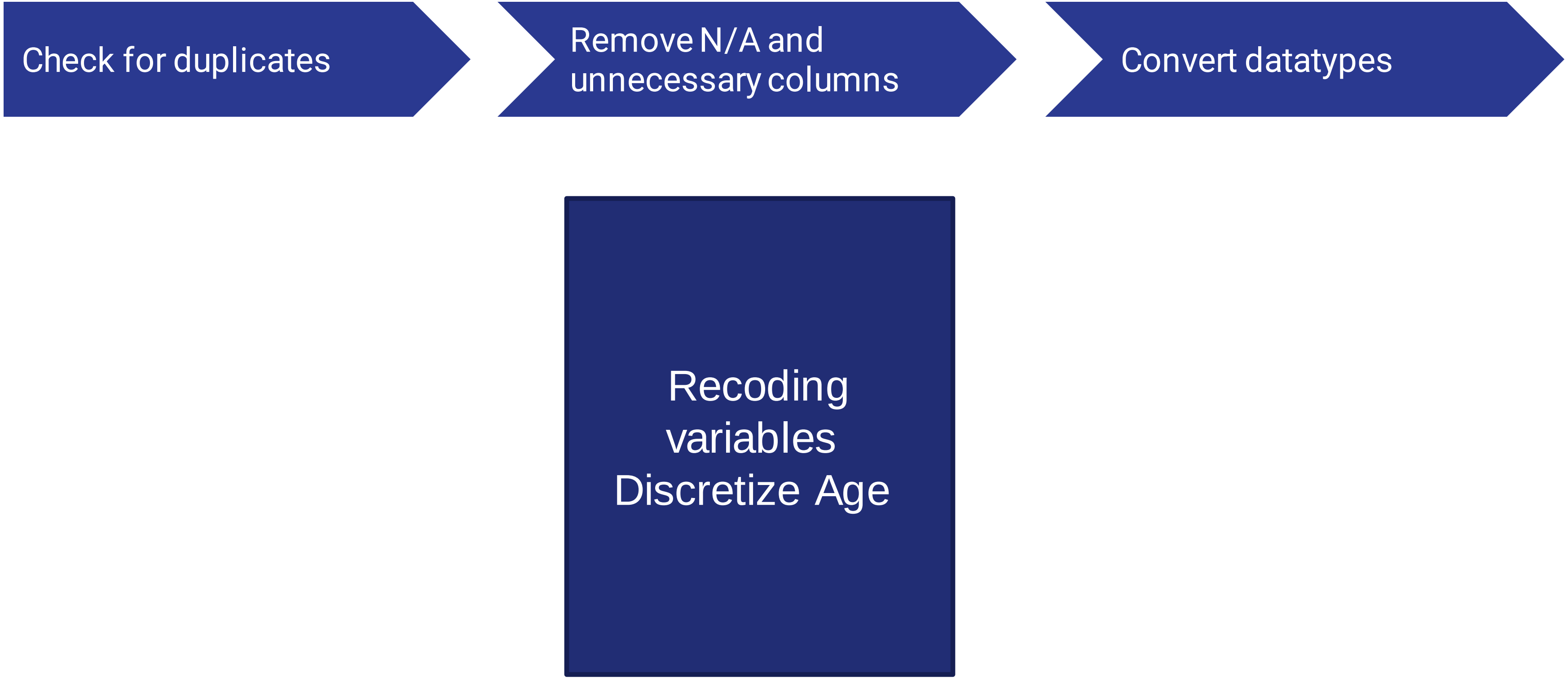
❑ The problem can be defined as follows: Determine whether a patient will have a stroke using historical medical information. This problem can be seen as a binary classification problem.

Data


Stroke Risk across All Ages
0
200
400
600
800
No
Yes
teens
twenties
thirties
forties
fifties
sixties
seventies
eighties

Variable	Data Type	Description
gender	factor	Patient's gender: "Male", "Female", or "Other"
age	numeric	Age of the Patient
hypertension	factor	0: patient does not have hypertension, 1:the patient has hypertension
heart_disease	factor	0: patient does not have any heart diseases, 1: patient has a heart disease
ever_married	factor	"No" or "Yes"
work_type	factor	"children", "Govt_job", "Never_worked", "Private", or "Self-employed"
residence_type	factor	"Rural" or "Urban"
avg_glucose_level	numeric	Patient's average glucose level
smoking_status	factor	"formerly smoked", "never smoked", "smokes", or "Unknown"
stroke	factor	1: patient had a stroke or 0: patient has not had a stroke
age_group	ordinal factor	"child", "teens", "twenties", "thirties", "forties", "fifties", "sixties","seventies", "eighties"

Data Cleaning



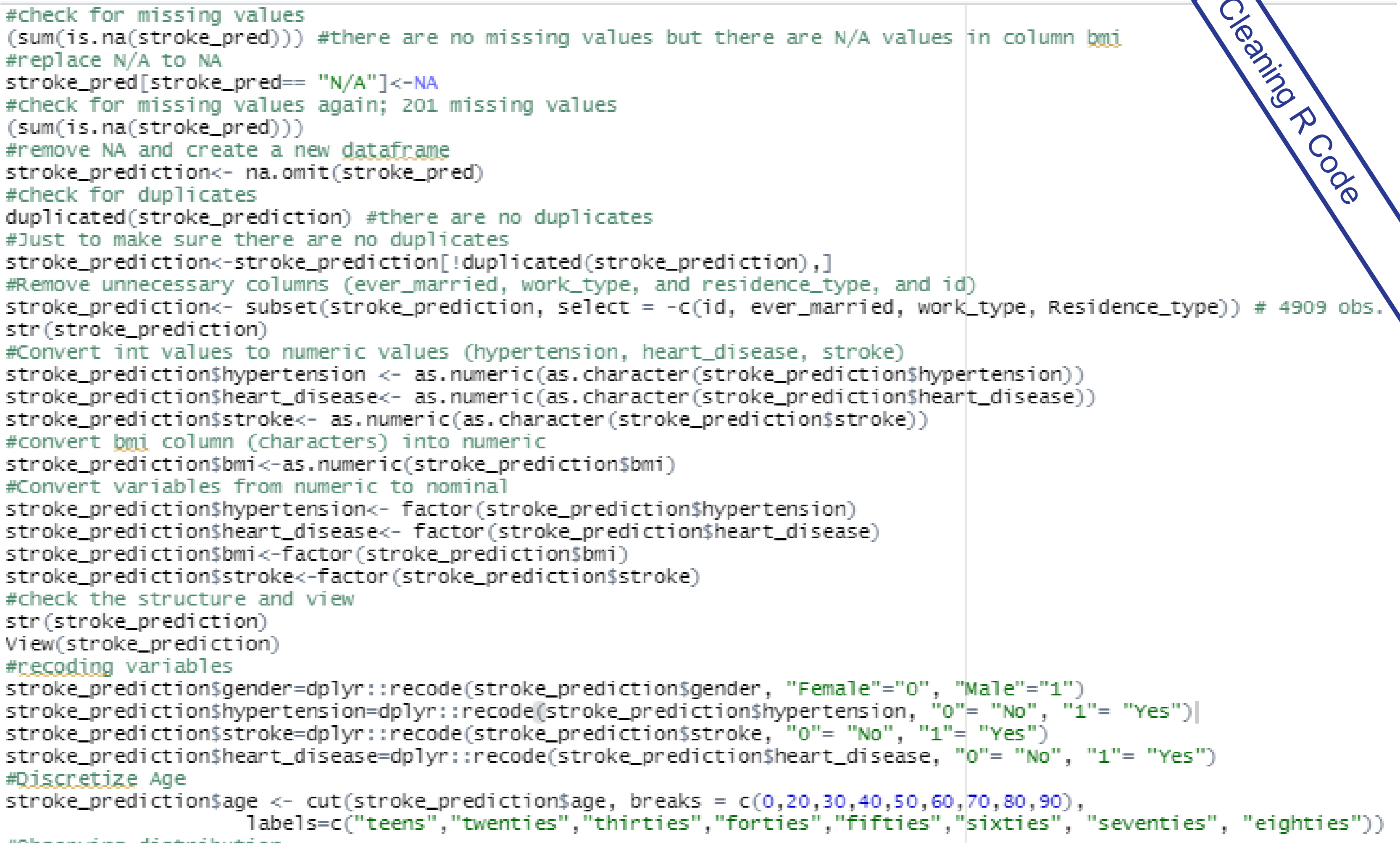
Check for duplicates

Remove N/A and unnecessary columns

Convert datatypes

Recoding variables
Discretize Age

Cleaning R Code

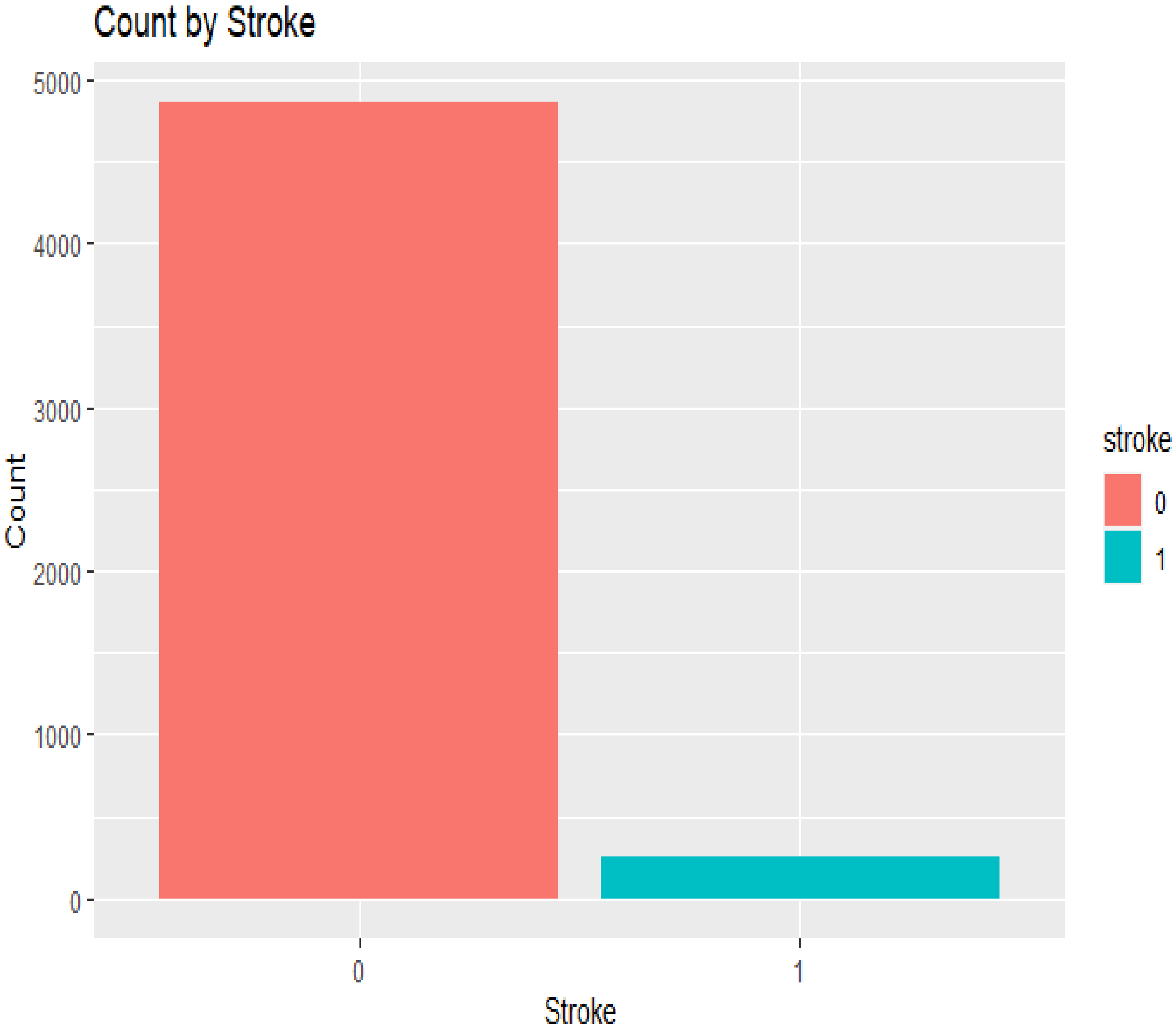


```
#check for missing values
(sum(is.na(stroke_pred))) #there are no missing values but there are N/A values in column bmi
#replace N/A to NA
stroke_pred[stroke_pred== "N/A"]<-NA
#check for missing values again; 201 missing values
(sum(is.na(stroke_pred)))
#remove NA and create a new dataframe
stroke_prediction<- na.omit(stroke_pred)
#check for duplicates
duplicated(stroke_prediction) #there are no duplicates
#Just to make sure there are no duplicates
stroke_prediction<-stroke_prediction[!duplicated(stroke_prediction),]
#Remove unnecessary columns (ever_married, work_type, and residence_type, and id)
stroke_prediction<- subset(stroke_prediction, select = -c(id, ever_married, work_type, Residence_type)) # 4909 obs.
str(stroke_prediction)
#Convert int values to numeric values (hypertension, heart_disease, stroke)
stroke_prediction$hypertension <- as.numeric(as.character(stroke_prediction$hypertension))
stroke_prediction$heart_disease<- as.numeric(as.character(stroke_prediction$heart_disease))
stroke_prediction$stroke<- as.numeric(as.character(stroke_prediction$stroke))
#convert bmi column (characters) into numeric
stroke_prediction$bmi<-as.numeric(stroke_prediction$bmi)
#Convert variables from numeric to nominal
stroke_prediction$hypertension<- factor(stroke_prediction$hypertension)
stroke_prediction$heart_disease<- factor(stroke_prediction$heart_disease)
stroke_prediction$bmi<-factor(stroke_prediction$bmi)
stroke_prediction$stroke<-factor(stroke_prediction$stroke)
#check the structure and view
str(stroke_prediction)
View(stroke_prediction)
#recoding variables
stroke_prediction$gender=dplyr::recode(stroke_prediction$gender, "Female"="0", "Male"="1")
stroke_prediction$hypertension=dplyr::recode(stroke_prediction$hypertension, "0"= "No", "1"= "Yes")
stroke_prediction$stroke=dplyr::recode(stroke_prediction$stroke, "0"= "No", "1"= "Yes")
stroke_prediction$heart_disease=dplyr::recode(stroke_prediction$heart_disease, "0"= "No", "1"= "Yes")
#Discretize Age
stroke_prediction$age <- cut(stroke_prediction$age, breaks = c(0,20,30,40,50,60,70,80,90),
                    labels=c("teens","twenties","thirties","forties","fifties","sixties", "seventies", "eighties"))
#observing distribution
```

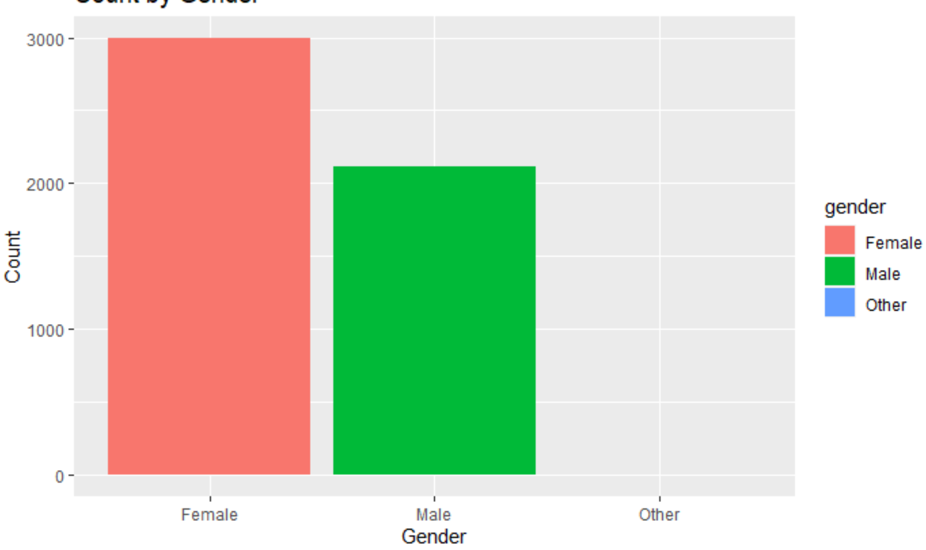
Initial Results from EDA

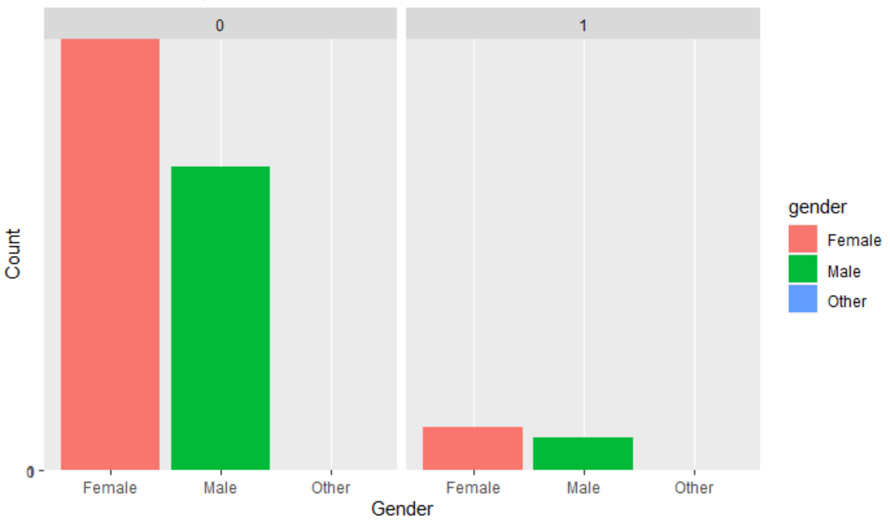
Target Variable

- There is a huge class imbalance in this dataset. We will certainly address the imbalance for future models.
- There are 5110 patients in this data set and only 249 suffered a stroke; the other 4,861 have not.



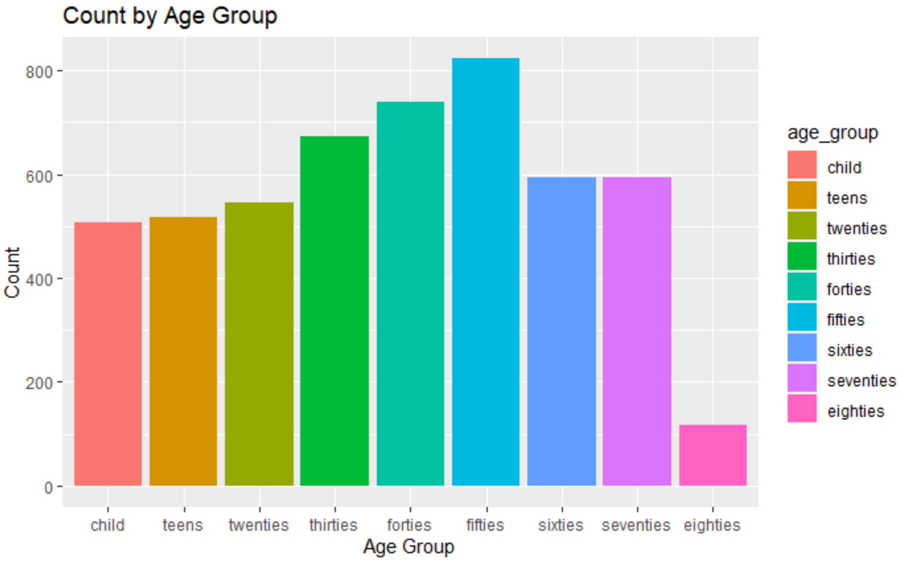
Results from EDA

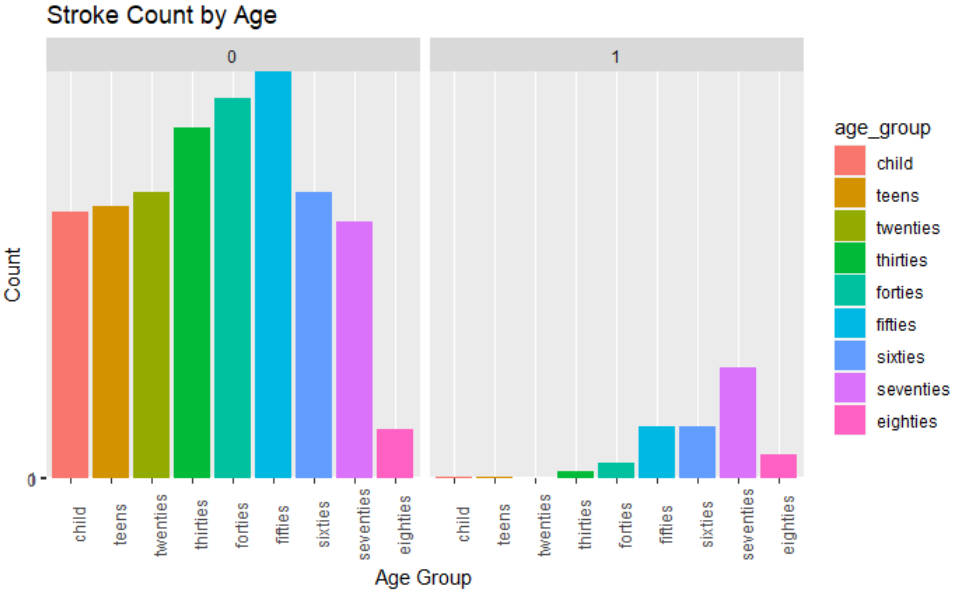

Count by Gender


Stroke Count by Gender

There are 2994 Females in the data. 141 have suffered from a stroke.
There are 2115 Males in the data. 108 have suffered from a stroke.
The is only one patient who classifies as other but never had a stroke.

Results from EDA

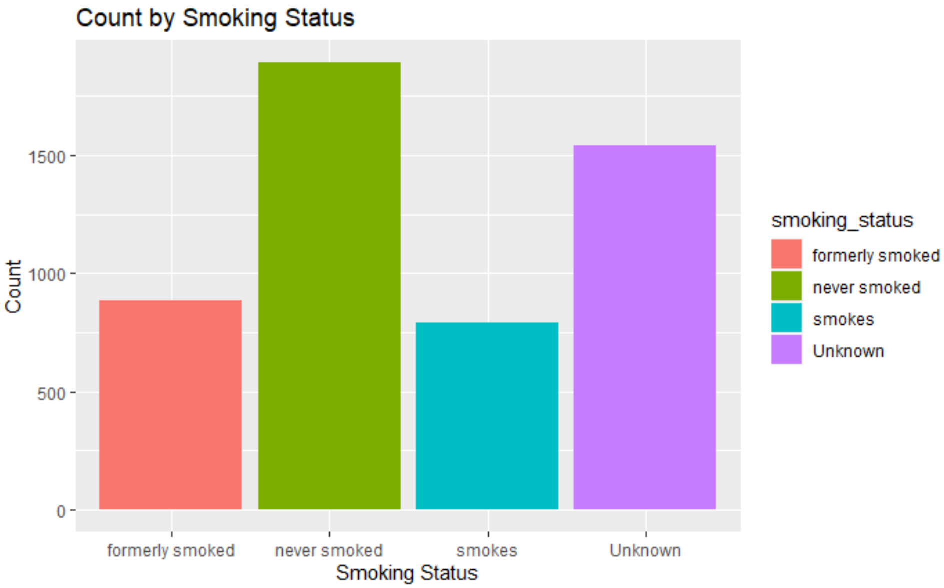


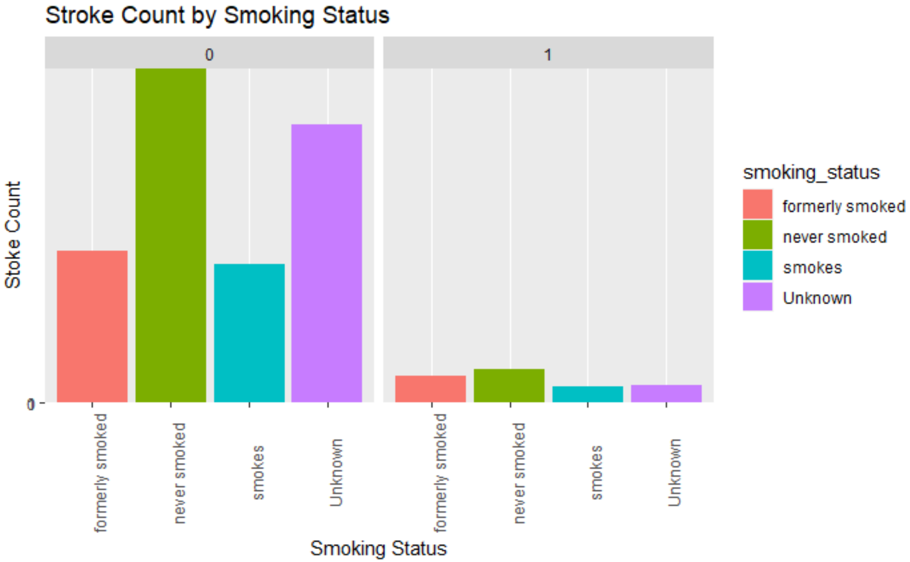


507 children; 1 suffered a stroke
518 teens; 1 suffered a stroke
545 twenties; 0 had a stroke
674 thirties; 6 suffered a stroke
739 forties; 15 suffered a stroke

823 fifties; 49 suffered a stroke
594 sixties; 49 suffered a stroke
594 seventies; 105 suffered a stroke
116 eighties; 23 suffered a stroke

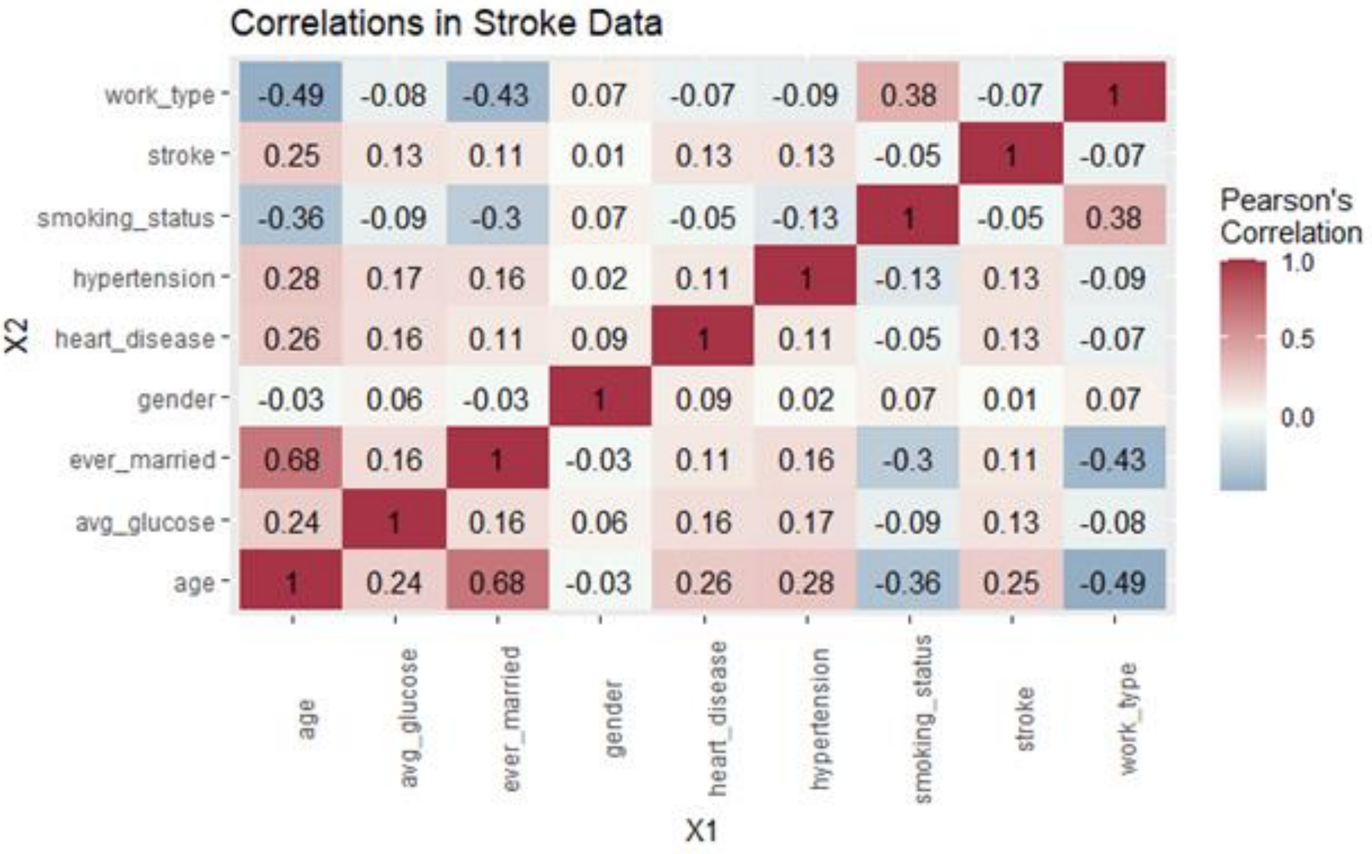
Results from EDA





885 former smokers: 70 strokes
1892 never smoked; 90 strokes
789 smokers; 42 strokes
1544 unknown; 47 strokes

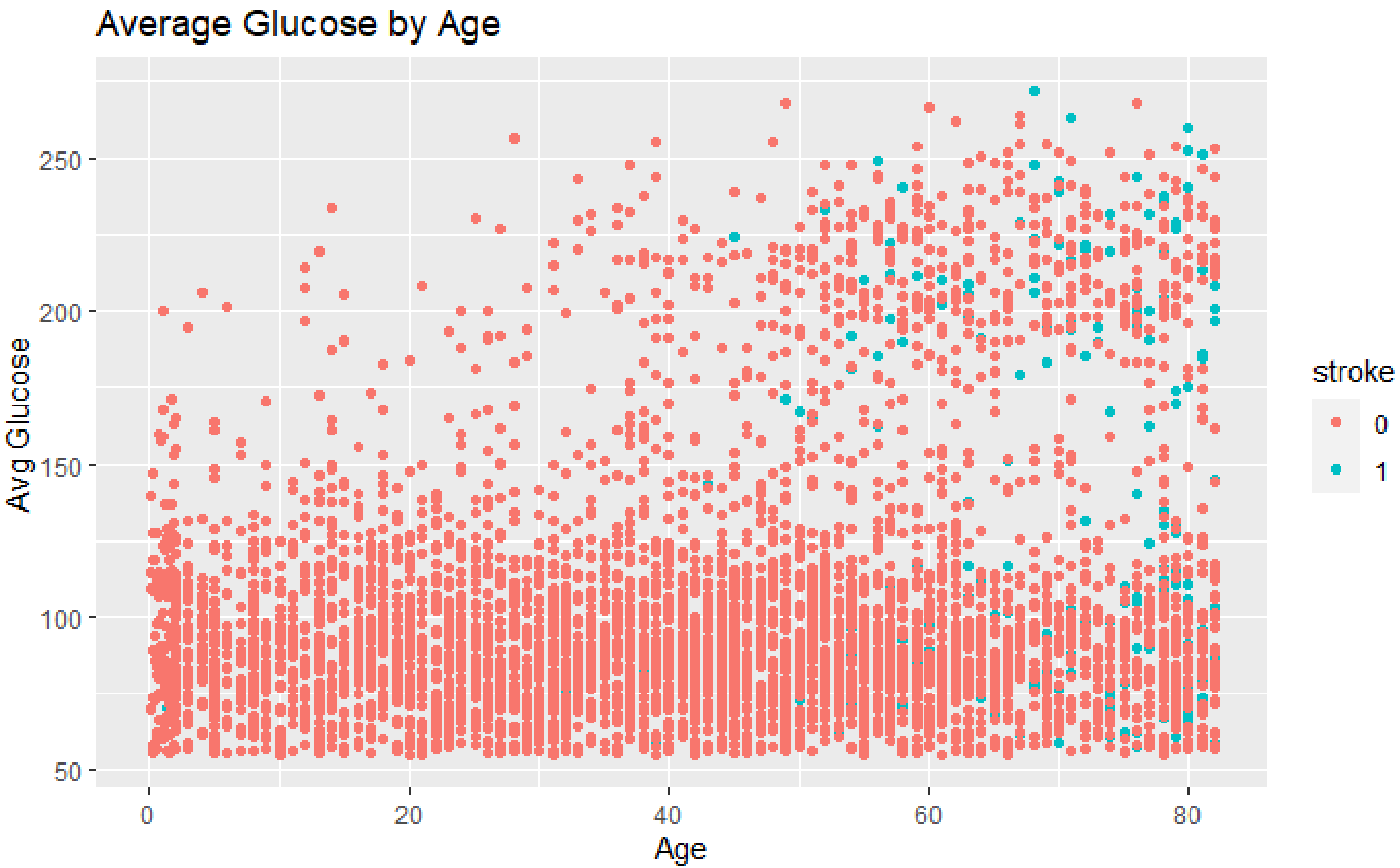
Correlation Analysis



Correlations in Stroke Data

	age	avg_glucose	ever_married	gender	heart_disease	hypertension	smoking_status	stroke	work_type
work_type	-0.49	-0.08	-0.43	0.07	-0.07	-0.09	0.38	-0.07	1
stroke	0.25	0.13	0.11	0.01	0.13	0.13	-0.05	1	-0.07
smoking_status	-0.36	-0.09	-0.3	0.07	-0.05	-0.13	1	-0.05	0.38
hypertension	0.28	0.17	0.16	0.02	0.11	1	-0.13	0.13	-0.09
heart_disease	0.26	0.16	0.11	0.09	1	0.11	-0.05	0.13	-0.07
gender	-0.03	0.06	-0.03	1	0.09	0.02	0.07	0.01	0.07
ever_married	0.68	0.16	1	-0.03	0.11	0.16	-0.3	0.11	-0.43
avg_glucose	0.24	1	0.16	0.06	0.16	0.17	-0.09	0.13	-0.08
age	1	0.24	0.68	-0.03	0.26	0.28	-0.36	0.25	-0.49

Average Glucose



Possible Analytical Methods

- Association Rules Mining: Find patterns that lead to the possibility of a patient having a stroke.
- K-means and HAC
- Decision Tree Model: Predict the classification of entries in the data frame.