# Data Set name: Stroke Prediction Dataset

# About the Data Set:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This dataset is used to predict whether a patient is likely to get stroke based on the

input parameters like gender, age, various diseases, and smoking status. Each row

in the data provides relevant information about the patient.

### **Attribute Information:**

1. id: unique identifier  
   2) gender: "Male", "Female" or "Other"  
   3) age: age of the patient  
   4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension  
   5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease  
   6) ever\_married: "No" or "Yes"  
   7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"  
   8) Residence\_type: "Rural" or "Urban"  
   9) avg\_glucose\_level: average glucose level in blood  
   10) bmi: body mass index  
   11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*  
   12) stroke: 1 if the patient had a stroke or 0 if not

Here stroke is target attribute.

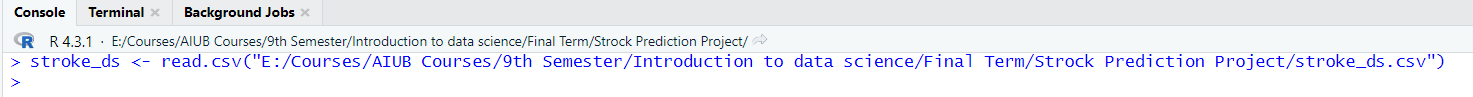
GitHub link of our project:

Importing the data set: Importing csv format of the data set in R studio.

Code:



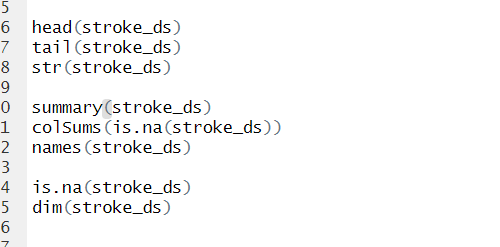
Output:



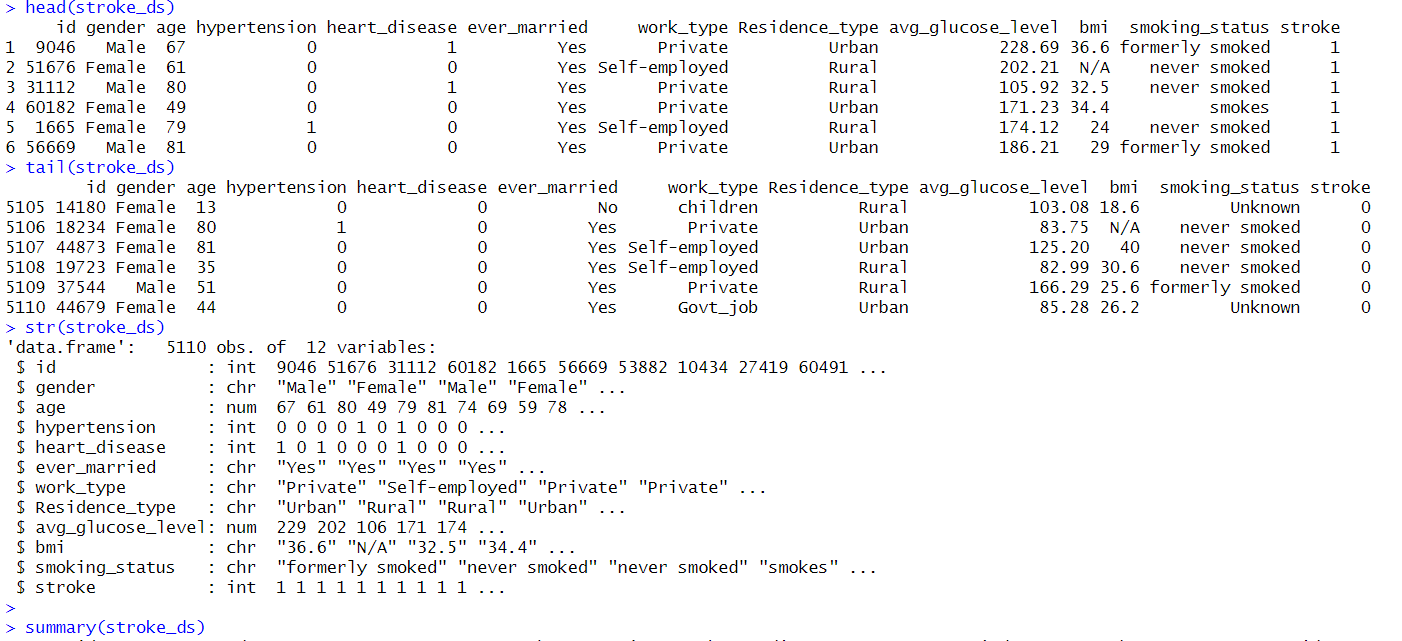
Inspecting the data set: In this code we are inspecting the data set using head(), tail(),

Summary(), names(), dim(), is.na(), colSums() amd str() method.

Code:



Output:

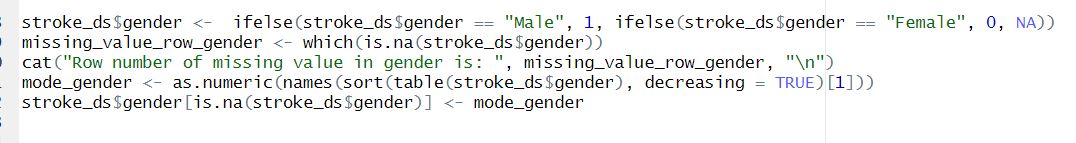


Preparing the data set:

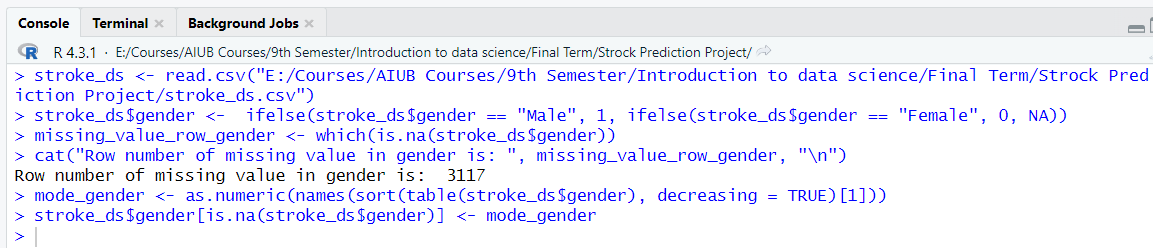
Gender Column:

In this part of the code we convert “Male” and “Female” with 1 and 0 to find the missing value and then we find the row of the missing value and recover that row of missing value with mode value.

Code:



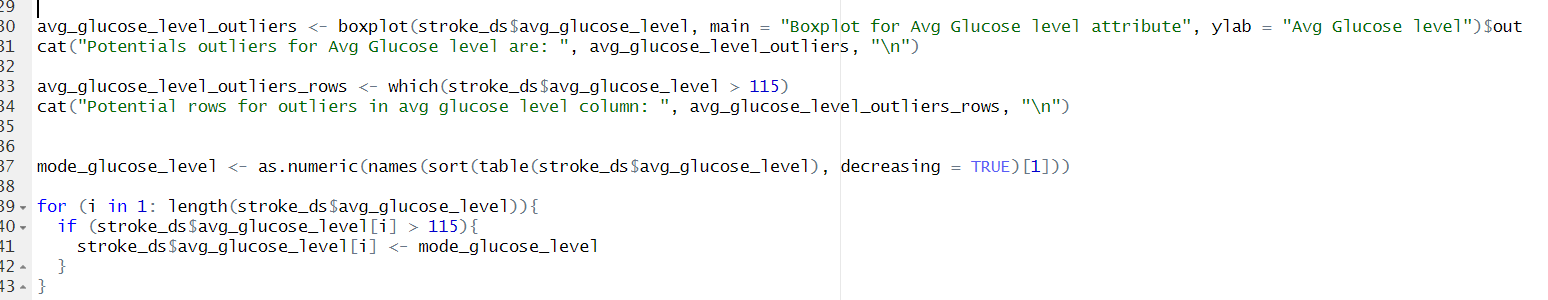
Output:

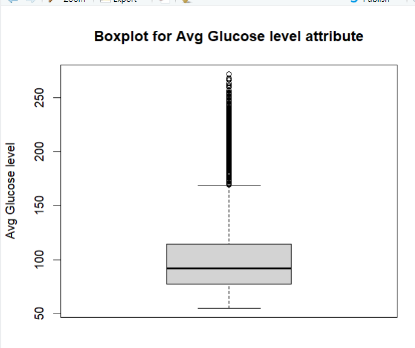


Glucose level column:

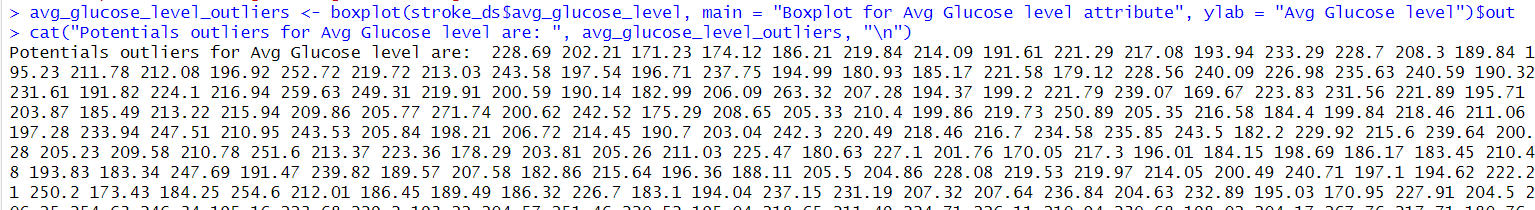
In this code we find the avg\_glucose\_level column’s outliers and their rows and also find mode value and recover those outliers with mode value.

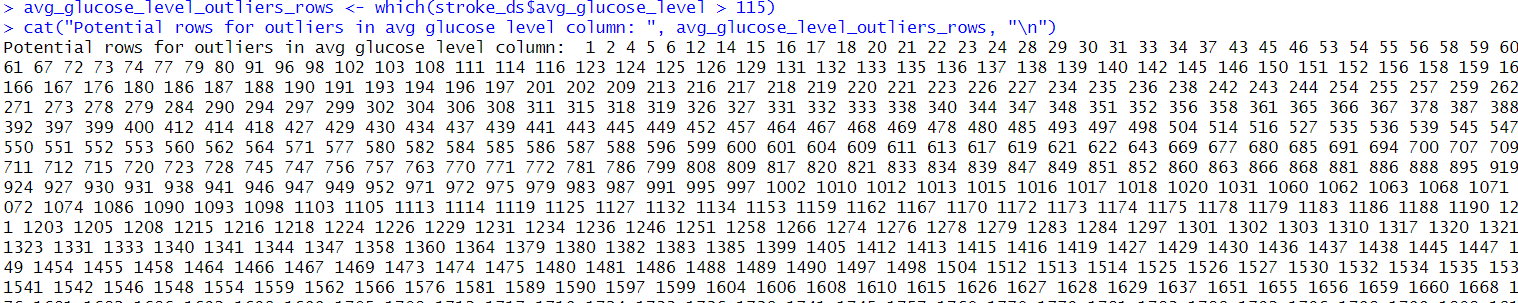
Code:

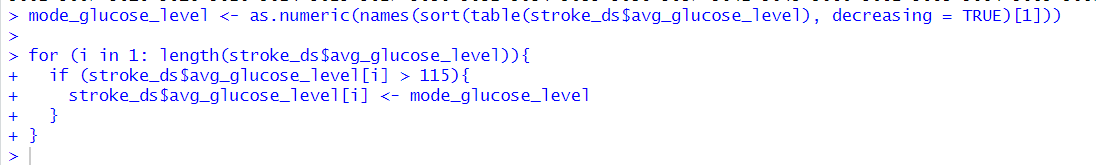




Output:



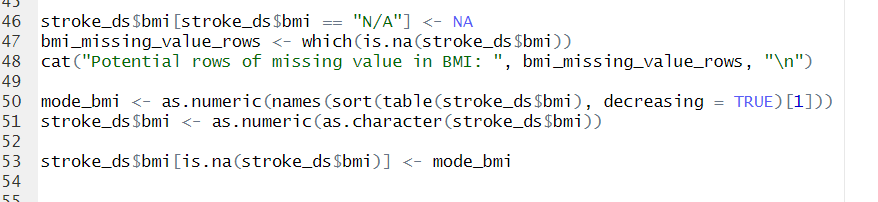


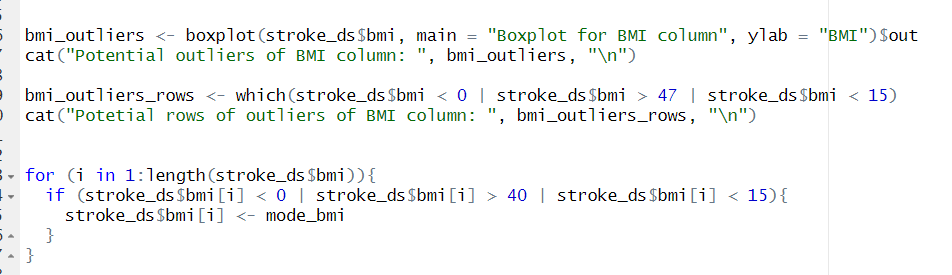


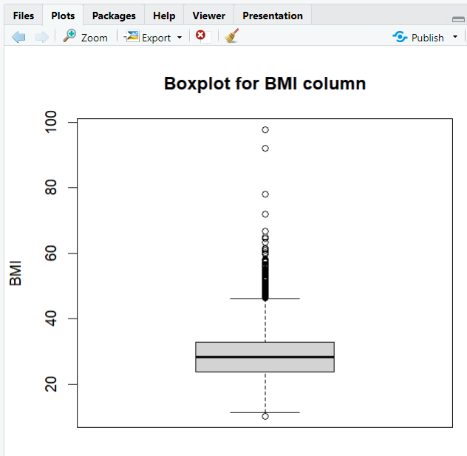
Bmi Column:

In this part of code we convert N/A into NA and find missing value’s row and also find mode value and recover those missing value rows with mode. Then we find outliers and also recover those outliers with mode value.

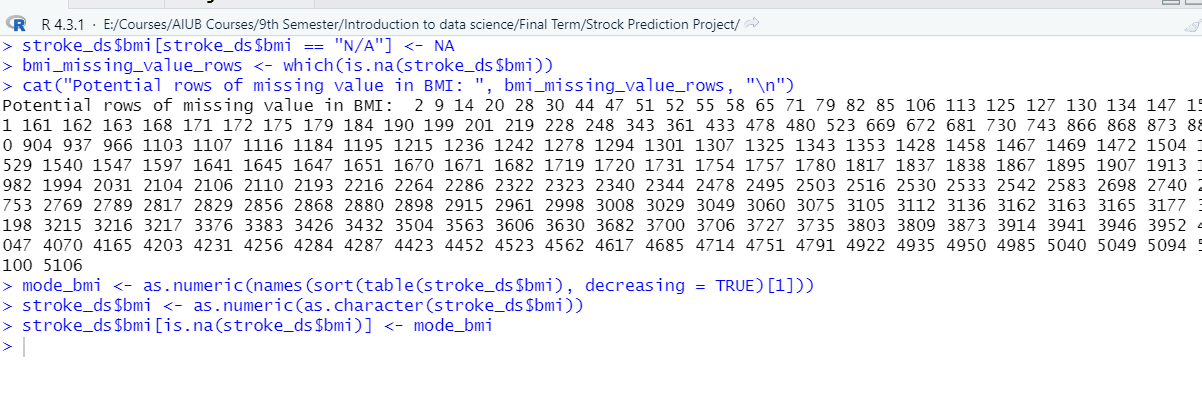
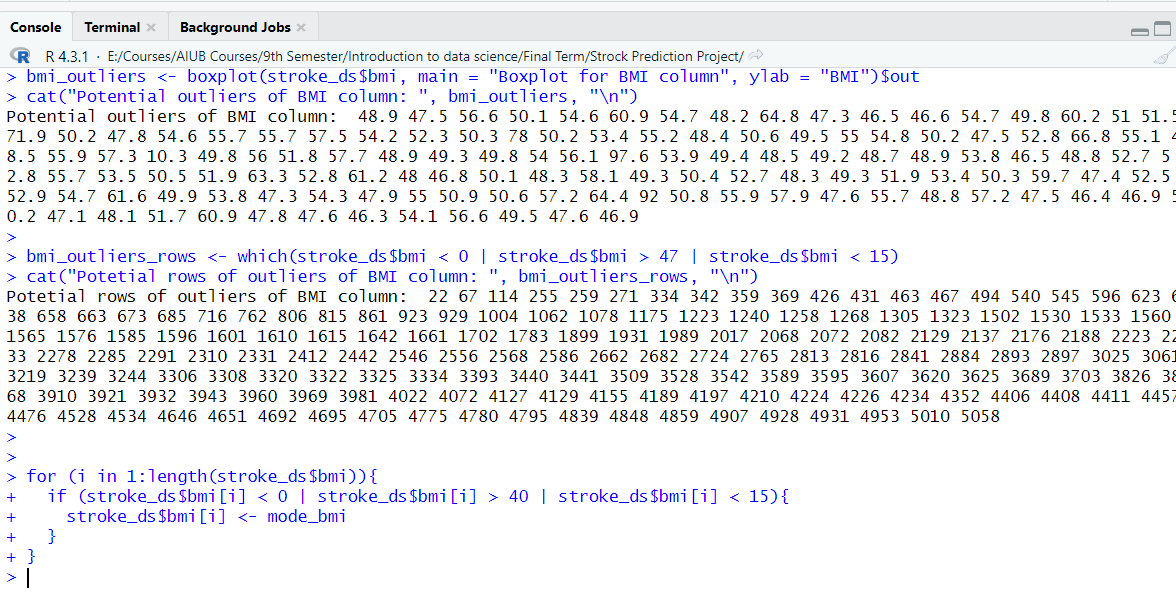
Code:







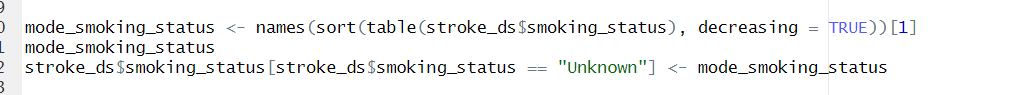
Output:



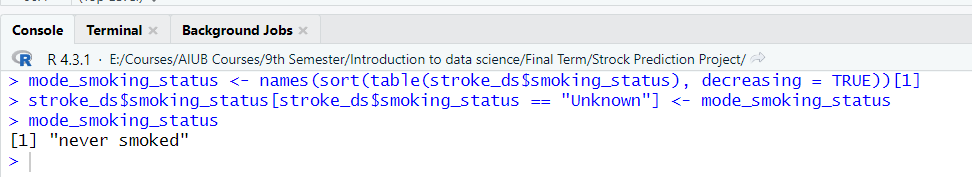
Smoking Status column:

In this part of code we find the mode value of smoking status column and recover the Unknown rows with mode value.

Code:



Output:



Converting into categorical value:

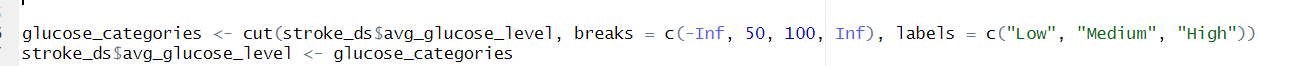
Avg glucose level column:

In this code we convert the avg\_glucose\_level column into categorical attribute.

Glucose levels are categorized as follows:

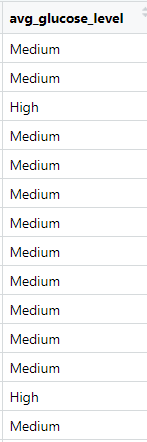
1. "Low" for values less than or equal to 50
2. "Medium" for values between 51 and 100
3. "High" for values greater than 100

Code:



Output:





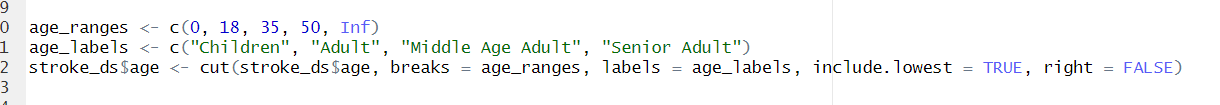
Age column:

In this code we convert the age column value into categorical value.

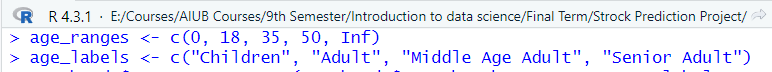
Ages are categorized as follows:

1. “0-18” to Children
2. “19-35” to Adults
3. “36-50” to Middle age adult
4. 51++ to Senior adult

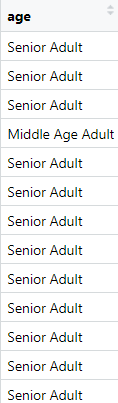
Code:



Output:







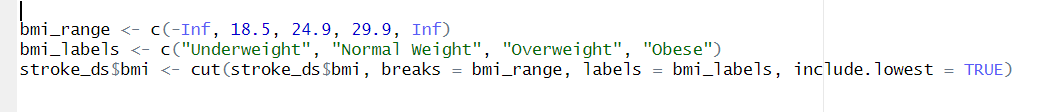
Bmi Column:

In this code we convert the bmi column value into categorical value.

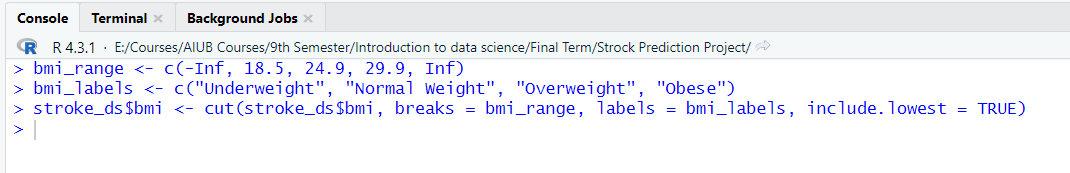
Bmi are categorized as follows:

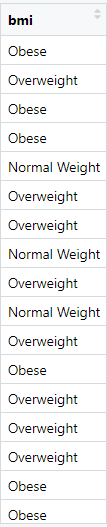
1. “0-18.5” to Underweight
2. “18.6-24.9” to Normal weight
3. “25-29.9” to Overweight
4. “30++” to Obese

Code:



Output:

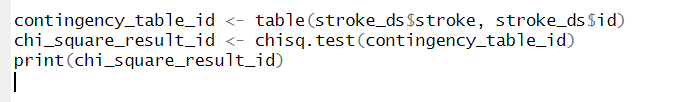




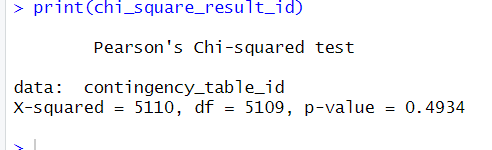
**Applying correlation technique** (**Pearson's Chi-squared test**):

**Id column:** In this part we apply **Pearson's Chi-squared test** onId column and find that the value of p is 0.4934 which is greater than 0.05. So, this column is insignificant and later we will delete this column.

Code:

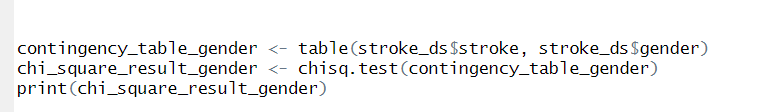


Output:

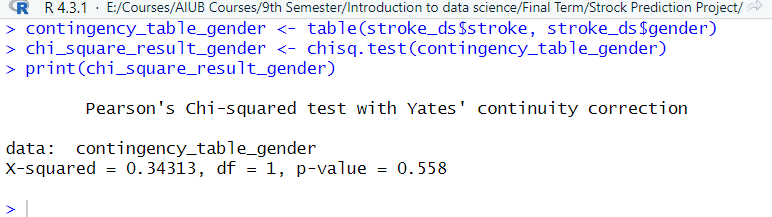


**Gender Column:** In this part we apply **Pearson's Chi-squared test on** gender column and find that the value of p is 0.558 which is greater than 0.05. So, this column is insignificant and later we will delete this column.

Code:

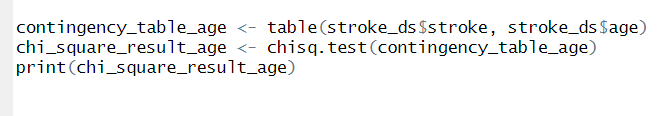


Output:

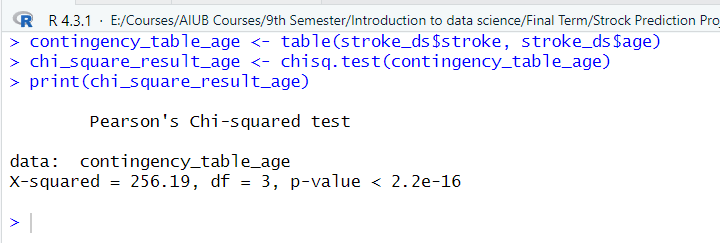


**Age column**: In this part we apply **Pearson's Chi-squared test** onage column and find that the value of p is 2.2e-16 which is less than 0.05. So, this column is significant.

Code:

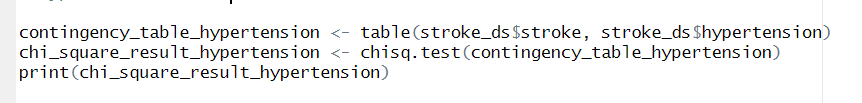


Output:

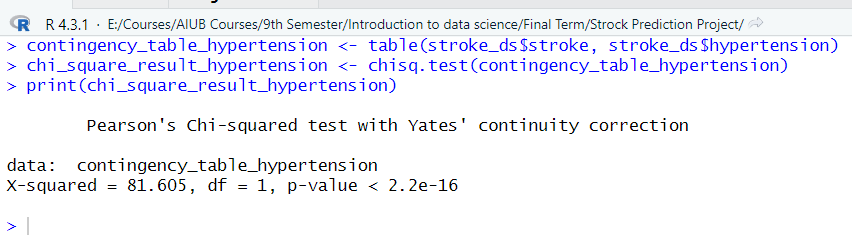


**Hypertension Column:** In this part we apply **Pearson's Chi-squared test** on hypertension column and find that the value of p is 2.2e-16 which is less than 0.05. So, this column is significant.

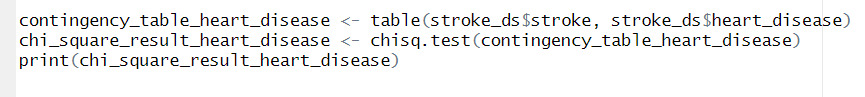
Code:



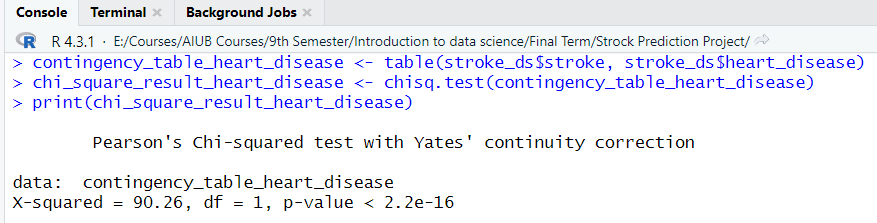
Output:



**Heart Disease column:** In this part we apply **Pearson's Chi-squared test** on heart disease column and find that the value of p is 2.2e-16 which is less than 0.05. So, this column is significant.  
Code:



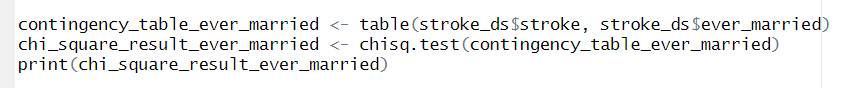
Output:



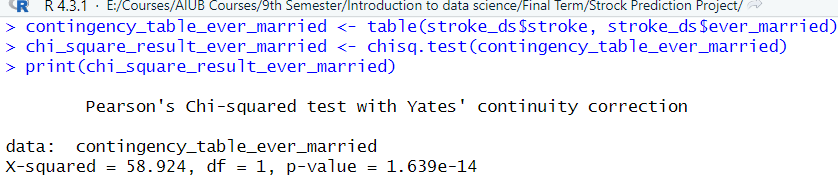
**Ever married column:**

In this part we apply **Pearson's Chi-squared test** on ever married column and find that the value of p is 1.639e-14 which is less than 0.05. So, this column is significant.

Code:



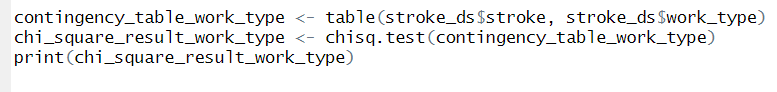
Output:



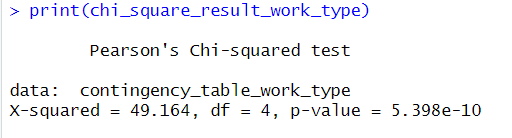
**Work Type column:**

In this part we apply **Pearson's Chi-squared test** on work type column and find that the value of p is 5.398e-10 which is less than 0.05. So, this column is significant.

Code:

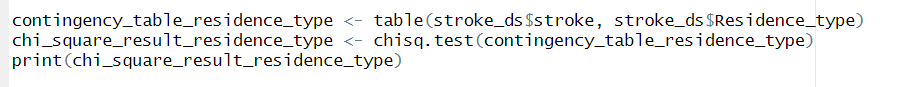


Output:

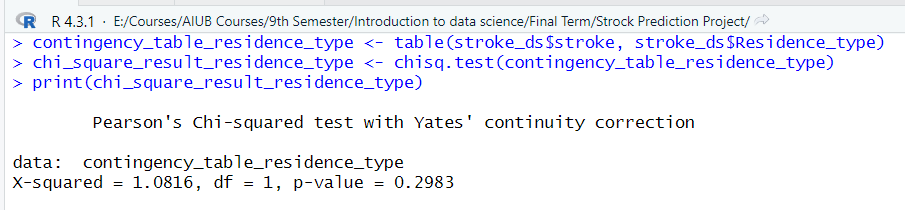


**Residence Type column:** In this part we apply **Pearson's Chi-squared test** on residence type column and find that the value of p is 0.2983 which is greater than 0.05. So, this column is insignificant and later we will delete this column.

**Code:**

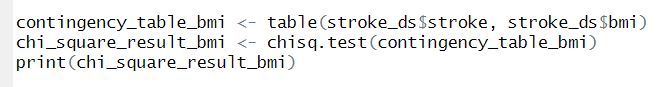


**Output:**

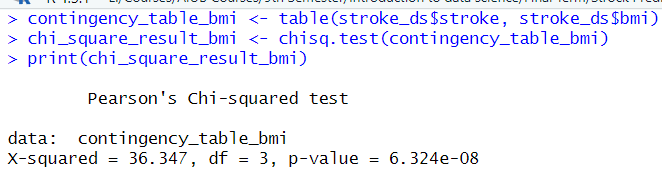


**Bmi Column:** In this part we apply **Pearson's Chi-squared test** on bmi column and find that the value of p is 6.324e-8 which is less than 0.05. So, this column is significant.

Code:

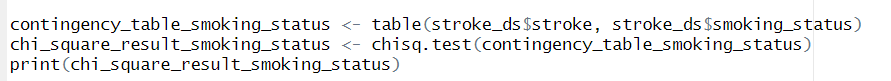


Output:

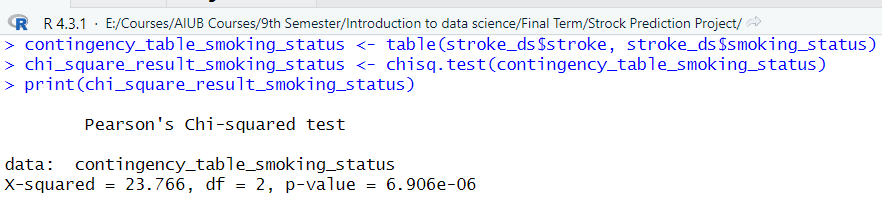


**Smoking Status Column:** In this part we apply **Pearson's Chi-squared test** on smoking status column and find that the value of p is 6.906e-06 which is less than 0.05. So, this column is significant.

Code:



Output:



**Deleting Insignificant column from the data set:**

In this part we delete the insignificant column such as Id, gender and avg\_glucose\_level from our dataset.

Code: