Report for Healthcare Strokes Dataset

Team Details:

Written by:

Pratik Dhanraj Chavan

Other Team Members:

Harshini Kandimalla (1001960046)

Pratik Antoni Patekar (1001937948)

Exploratory Analysis of Healthcare strokes dataset using R:

Introduction:

For this assignment we have used R Language to perform our analysis.

We have used the healthcare_stroke_dataset.csv file which contained the dataset. This dataset has 5110 rows and 13 columns.

The dataset contains id

- date
- gender
- age
- hypertension
- heart_disease
- ever_married
- work_type
- Residence_type
- avg_glucose_level
- bmi
- smoking_status
- stroke

Retrieving The Data:

To get the data for analysis we have imported the data from the csv file into a dataframe.

We have imported the data into dataframe called **dataset.**

```
In [2]: # Read the file
        dataset <- read.csv("healthcare_stroke_dataset.csv")
```

Glimpse of Data:

Displaying the first 5 rows of data using the head command.

```
In [3]: #return the first 5 rows of the datasetResidence_type
        head(dataset,5)

        #Displaying the first 5 rows of the dataframe
```

id	date	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	12/30/2020	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	8/18/2020	Female	61	0	0	Yes	Self-employed	Rural	202.21	NA	never smoked	1
31112	3/5/2020	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	7/8/2020	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	6/5/2020	Female	79	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Data Exploration:

Task 1: Statistical Exploratory Data Analysis:

Task 1-a: Details of health_data date frame:

For the above task we have used the functions summary() and str().

```
str(dataset)
summary(dataset)
```

Output:

```
'data.frame':   5110 obs. of  13 variables:
 $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
 $ date             : Factor w/ 366 levels "1/1/2020","1/10/2020",..: 116 315 179 304 270 245 299 4 238 259 ...
 $ gender           : Factor w/ 3 levels "Female","Male",..: 2 1 2 1 1 2 2 1 1 1 ...
 $ 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     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 2 2 ...
 $ work_type        : Factor w/ 5 levels "children","Govt_job",..: 4 5 4 4 5 4 4 4 4 4 ...
 $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 1 2 1 2 1 2 1 2 ...
 $ 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   : Factor w/ 4 levels "-","formerly smoked",..: 2 3 3 4 3 2 3 3 1 1 ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...

       id                date         gender          age        
 Min.   :   67   6/9/2020  :  27   Female:2994   Min.   : 0.08  
 1st Qu.:17741   5/9/2020  :  26   Male  :2115   1st Qu.:25.00  
 Median :36932   5/3/2020  :  25   Other :   1   Median :45.00  
 Mean   :36518   12/21/2020:  23                 Mean   :43.23  
 3rd Qu.:54682   4/7/2020  :  23                 3rd Qu.:61.00  
 Max.   :72940   8/14/2020 :  23                 Max.   :82.00  
                 (Other)   :4963                                
  hypertension     heart_disease     ever_married         work_type    
 Min.   :0.00000   Min.   :0.00000   No :1757     children      : 687  
 1st Qu.:0.00000   1st Qu.:0.00000   Yes:3353     Govt_job      : 657  
 Median :0.00000   Median :0.00000                Never_worked :  22  
 Mean   :0.09746   Mean   :0.05401                Private       :2925  
 3rd Qu.:0.00000   3rd Qu.:0.00000                Self-employed: 819  
 Max.   :1.00000   Max.   :1.00000                                    

 Residence_type avg_glucose_level      bmi               smoking_status 
 Rural:2514     Min.   : 55.12    Min.   :10.30   -              :1544  
 Urban:2596     1st Qu.: 77.25    1st Qu.:23.50   formerly smoked: 885  
                Median : 91.89    Median :28.10   never smoked   :1892  
                Mean   :106.15    Mean   :28.89   smokes         : 789  
                3rd Qu.:114.09    3rd Qu.:33.10                         
                Max.   :271.74    Max.   :97.60                         
```

Task #1-b Find the number of rows and columns in dataset.

Used the nrows() and ncol() function to get the number of rows and columns.

Output:

```
In [7]: #1-b Find the number of rows and columns in dataset
        #
        print("Number of Rows in dataset are")
        nrow(dataset)

        print("Number of Columns in dataset are")
        ncol(dataset)
```

```
[1] "Number of Rows in dataset are"

5110

[1] "Number of Columns in dataset are"

13
```

Task #1-c Print descriptive detail of a column in dataset.

Used the summary() to get the descriptive details of column gender in dataset.

```
In [4]: #1-c Print descriptive detail of a column in dataset
        summary(dataset$gender)
        #printing the details for Column Gender from the dataset
```

Female	2994
Male	2115
Other	1

#1-d Find all the count of unique values for 'avg_glucose_level' column in dataset.

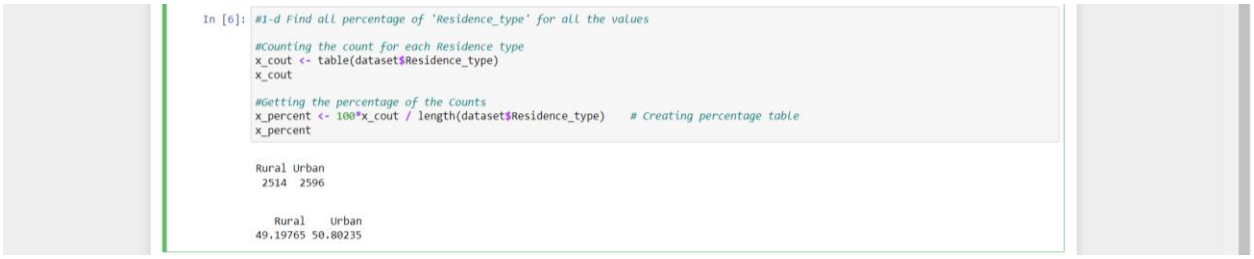
Used the unique command to get all unique values in the column

Output:

```
In [5]: #1-d Find all the count of unique values for a 'avg_glucose_level' column in dataset

        print("Count of all Unique values for Column avg_glucose_level is")
        length(unique(dataset$avg_glucose_level))

        [1] "Count of all Unique values for Column avg_glucose_level is"

        3979
```

Task #1-d Find all percentage of 'Residence_type' for all the values

Used the table command to get the count of residence type and used it to calculate percentage for urban and rural residence type.



Task 2: Aggregation & Filtering & Rank

Task 2-a: Find out the gender with largest number of records

Used the table command then compared it with other gender to get the max.

```
In [7]: #Task 2-a: Find out the gender with largest number of records

        #Creating Table with Counts of gender values
        gender_count <- table(dataset$gender)
        gender_count
        #Finding the gender type with max number of records
        names(gender_count[gender_count==max(gender_count)])
```

```
Female   Male  Other
  2994   2115      1

'Female'
```

Task 2-b: Find out the total number of Residence type "Urban" who are Male

Used subset command to sort the data based on residence type urban and are male.

The counted the number of rows

```
In [8]: #Task 2-b: Find out the total number of Residence_type "Urban" who are Male

        Male_Urban_resident <- subset(dataset, Residence_type=="Urban" & gender =="Male")

        nrow(Male_Urban_resident)
```

```
1067
```

Group by function for dataframe in R using pipe operator

2-c 1 question #Find the top 10 ages with highest avg_glucose_level

Selected the columns age and highest avg_glucose_level and arranged them in descending order

Printed the top 10 records.

Output:

In [9]:
```
# Group by function for dataframe in R using pipe operator
#2-c 1 question #Find the top 10 ages with highest av_glucose_level
ans2c <- dataset  %>% select(age,avg_glucose_level)  %>% arrange(desc(avg_glucose_level))

head(ans2c,10)
```

age	avg_glucose_level
68	271.74
49	267.76
76	267.61
76	267.60
60	266.59
67	263.56
71	263.32
62	261.67
67	260.85
80	259.63

#2-d 2nd question top 10 ages with more number of strokes

Used the pipe operator and filter the data with stroke data.

Then took sum of the stroke values and printed top 10 values.

In [10]:
```
#2-d 2nd question top 10 ages with more number of strokes
#summary(dataset$stroke)

ans3d1 <- dataset %>% select(age,stroke) %>% filter(stroke>=1) %>% group_by(age)

ans3d2 <- aggregate(stroke ~ age, ans3d1, sum)  %>% arrange(desc(stroke))
head(ans3d2,10)
```

age	stroke
78	21
79	17
80	17
81	14
57	11
76	10
63	9
68	9
74	9
82	9

TASK 3: VISUALIZATION

Task 3-a

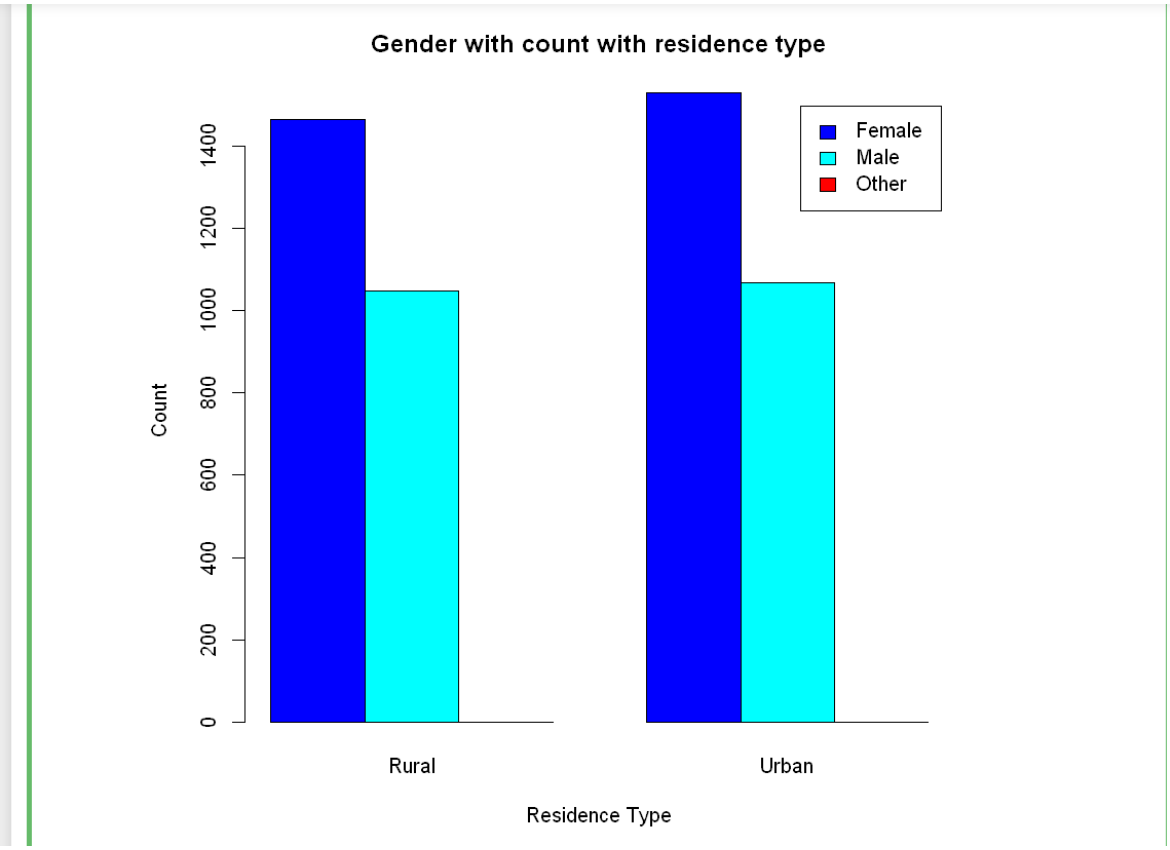
Create barplot showing gender with count with residence type

```
In [11]: #task 3-a
         #Create barplot showing gender with count with residence type
         x = table(dataset$gender,dataset$Residence_type)
         x
         barplot(x, main="Gender with count with residence type",
            ylab="Count",xlab = "Residence Type",col = c("blue","cyan","red"),legend.text = rownames(x),beside = TRUE)
```

```
        Rural Urban
Female   1465  1529
Male     1048  1067
Other       1     0
```

Sorted the data into variable and then Used it barplot to generate graph

Output:

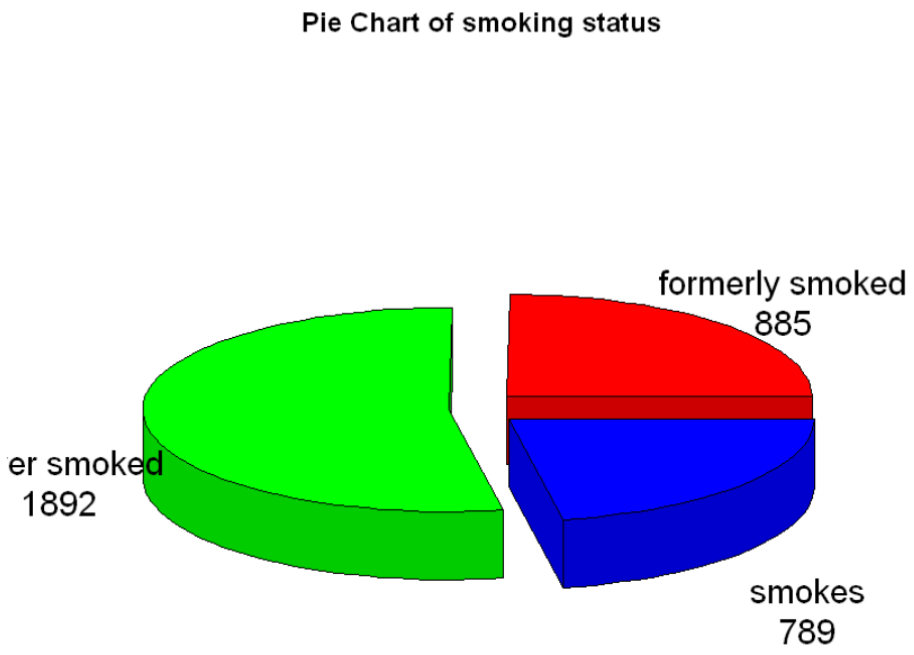


Task 3-b

Display pie chart for the smoking status data

```
In [12]: #task 3-b
         #Display pie chart for the smoking status data

         library(plotrix)
         allstatustable <- table(dataset$smoking_status)
         statustable <- allstatustable[c(0,2:4)]
         lbls <- paste("\n",names(statustable), "\n", statustable, sep="")

         pie3D(statustable, labels = lbls,
            main="Pie Chart of smoking status",radius=0.9,explode=0.1)
```

Output:



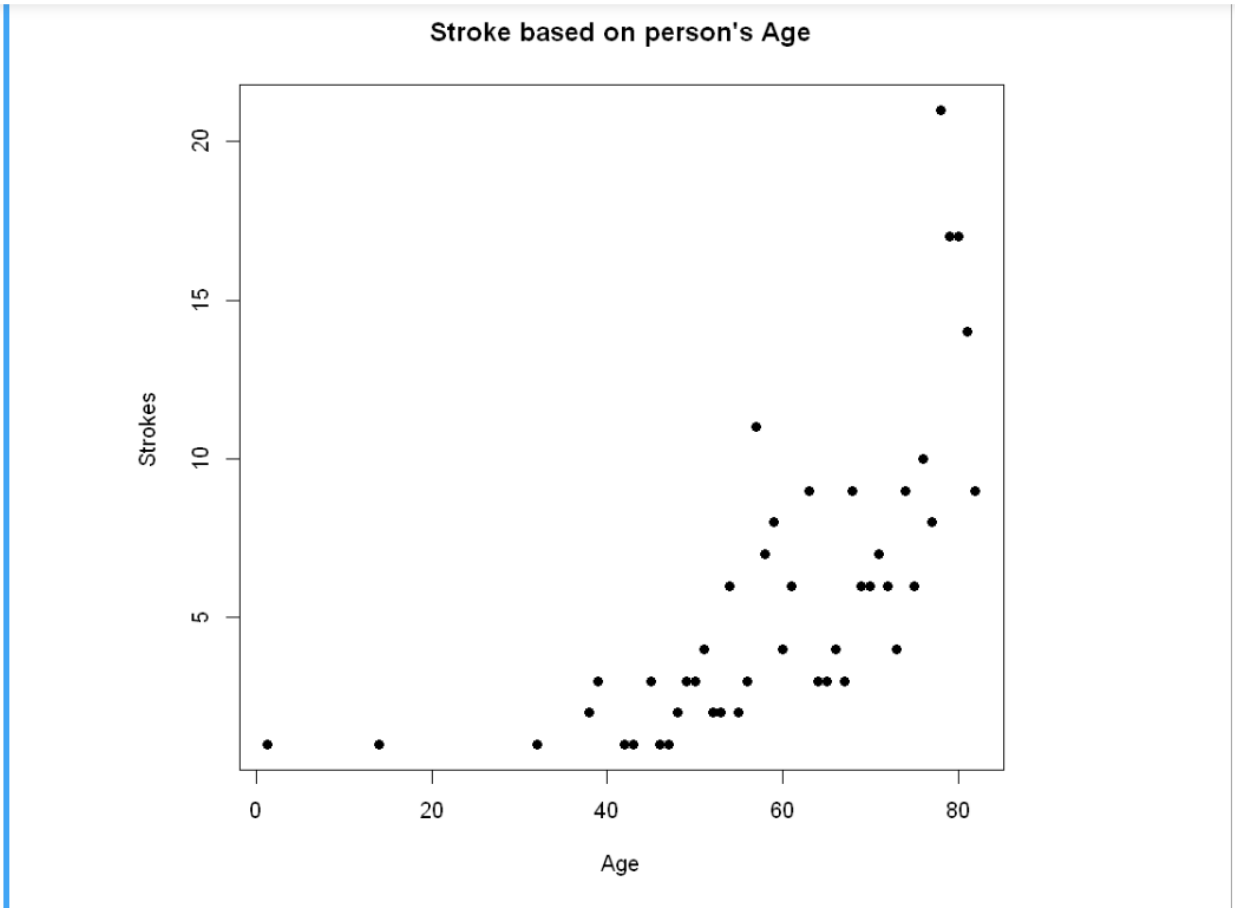
Task4 finding an interesting pattern

#atleast two visualization with explanation.

```
In [13]: #Task4 finding an interesting pattern
         # atleast two visualization with explanation

         #Stroke based on persons Age

         data1 <- dataset %>% select(age,stroke) %>% filter(stroke>=1) %>% group_by(age)

         plotdata <- aggregate(stroke ~ age, data1, sum)  %>% arrange(desc(stroke))


         plot(plotdata, main="Stroke based on person's Age",
            xlab="Age", ylab="Strokes ", pch=16)
```

Output:

Here as you can see in graph we can analyze that people with age above 40 as seen to have strokes more often.



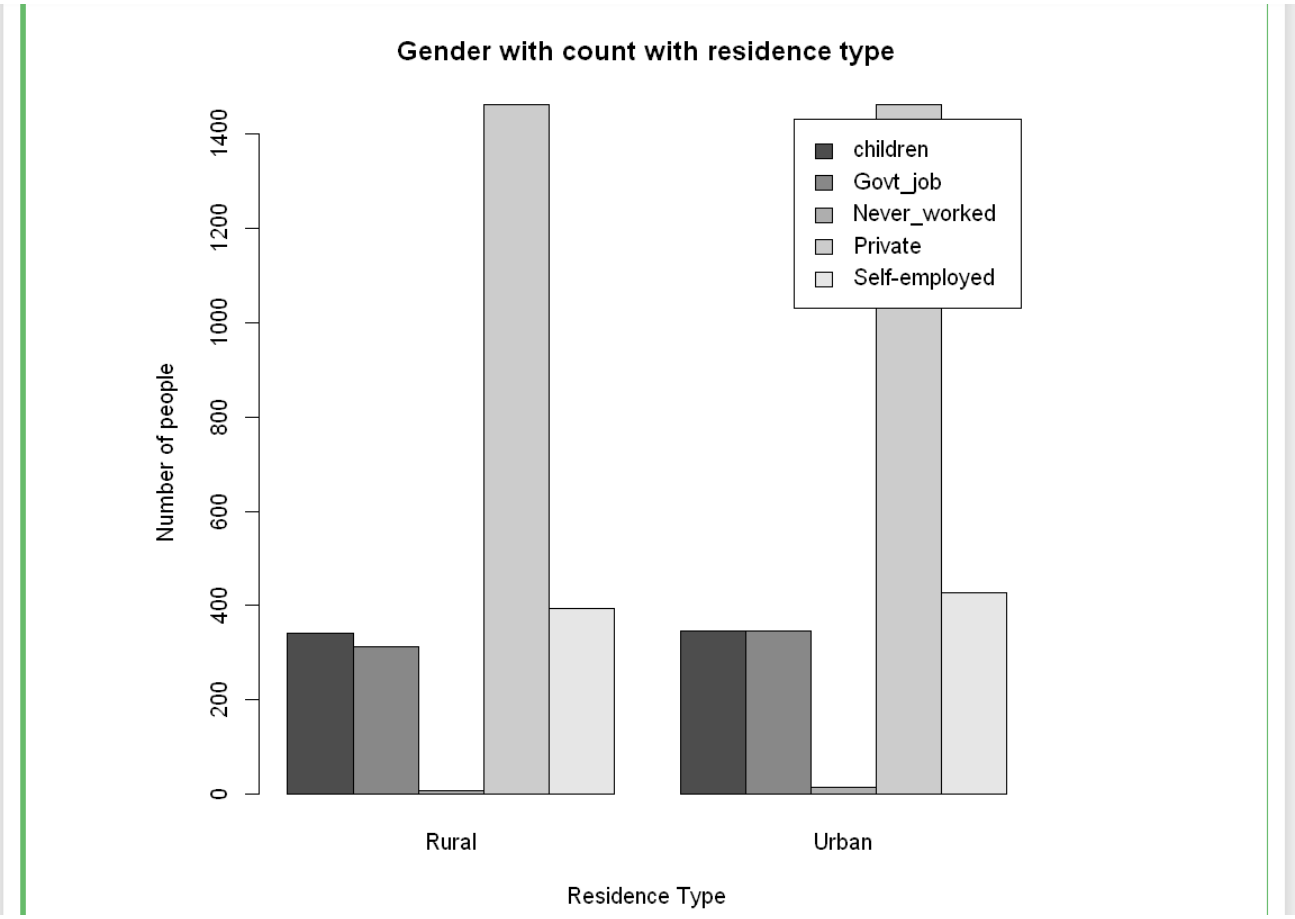
work_type and Residence type

```
In [14]: # work_type and Residence type
         workdata <- table(dataset$work_type,dataset$Residence_type)

         barplot(workdata, main="Gender with count with residence type",
            ylab="Number of people",xlab = "Residence Type",legend.text = rownames(workdata),beside = TRUE)
```

Output:

People are working more in Private Sector than any other sector

We can also analyze that people working in each sector are almost the same in all the sector.



Observation by other Team members

We found that R Language is great to work on datasets.

It has functions inbuild and therefore analyzing the data is very easy.

Plus, there are libraries available that are powerful to perform tasks.

Compared to other Languages R is very easy to work and understand.