Assignment-1 Report
Data Mining
CSE-5334-04

Student Name:

Mehul Ganjude – 1001990551
Gagan Ujjini Mallikarjuna – 1001851247
Venkata Nagendar Shantiswaroop Adibhatla-1001862413

Professor: Dr Elizabeth D Diaz
Teaching Assistant: Pralobh Lokhande

Language: R

Introduction

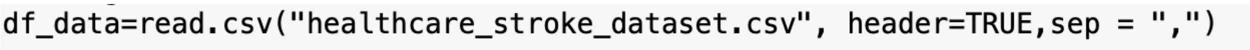
Data mining is a process of extracting and discovering patterns in large data sets that includes methods in the interface of machine learning, statistics and database systems. In the present assignment, we will be analyzing the data of healthcare that consists of vital signs and general bio data. The csv file used in the following assignment is healthcare_stroke_dataset.csv. The columns present in the dataset are mentioned below:

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

Retrieving the Data

Information retrieval is the process of identifying and extracting information from a document repositories particularly textual information. To retrieve and analyze the information provided we are using R programming language.

To retrieve information from given dataset, we are implementing the following code mentioned below:



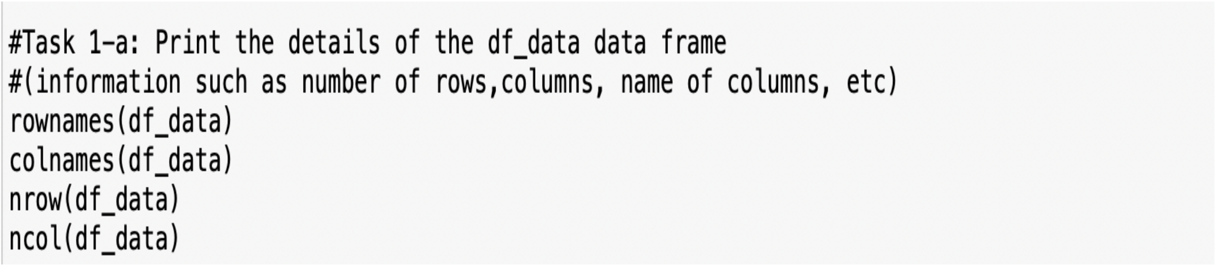
```
df_data=read.csv("healthcare_stroke_dataset.csv", header=TRUE,sep = ",")
```

From the above code, we are trying to read data from the csv file and then saving it into DataFrame. DataFrame is tightly coupled collections of variables which share many of properties of matrices and lists, used as fundamental data structure by most of R’s modelling software.

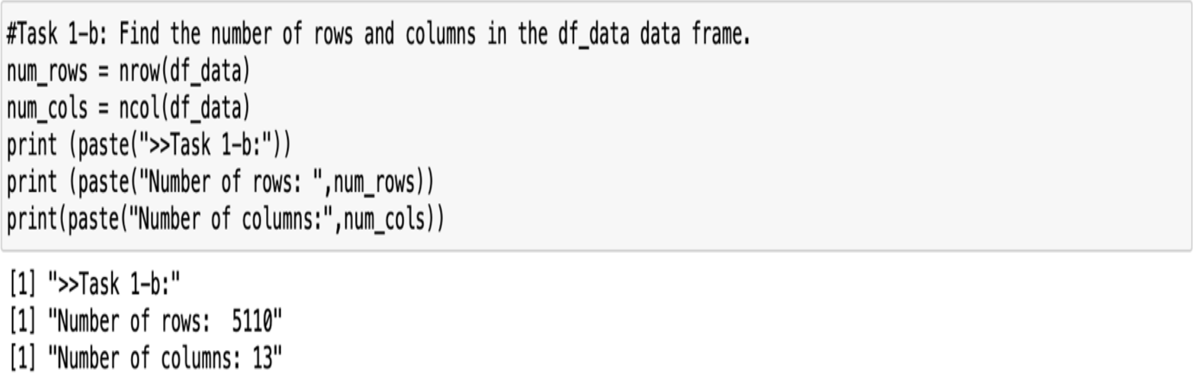
Glimpse of Data

Glimpse of Data is showing the underlying data to see the columns of the dataset and display some portion of the data with respect to each attribute that can be fit on a single line.

Task 1: Statistical Exploratory Data Analysis

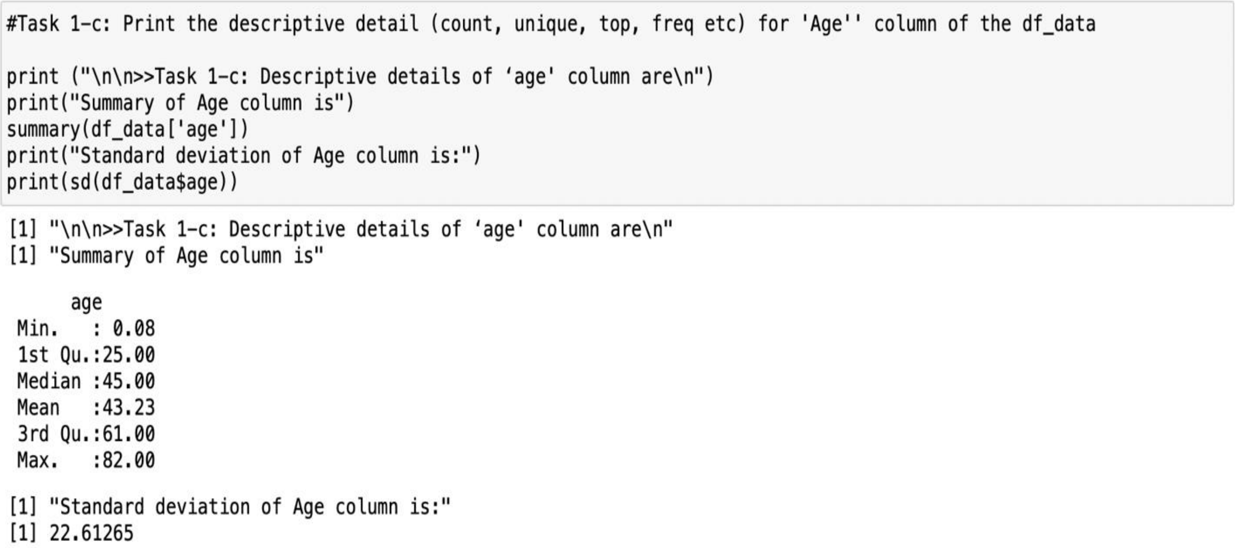


```
#Task 1-a: Print the details of the df_data data frame
#(information such as number of rows,columns, name of columns, etc)
rownames(df_data)
colnames(df_data)
nrow(df_data)
ncol(df_data)
```



```
#Task 1-b: Find the number of rows and columns in the df_data data frame.
num_rows = nrow(df_data)
num_cols = ncol(df_data)
print (paste(">>Task 1-b:"))
print (paste("Number of rows: ",num_rows))
print(paste("Number of columns:",num_cols))
```

```
[1] ">>Task 1-b:"
[1] "Number of rows:  5110"
[1] "Number of columns: 13"
```



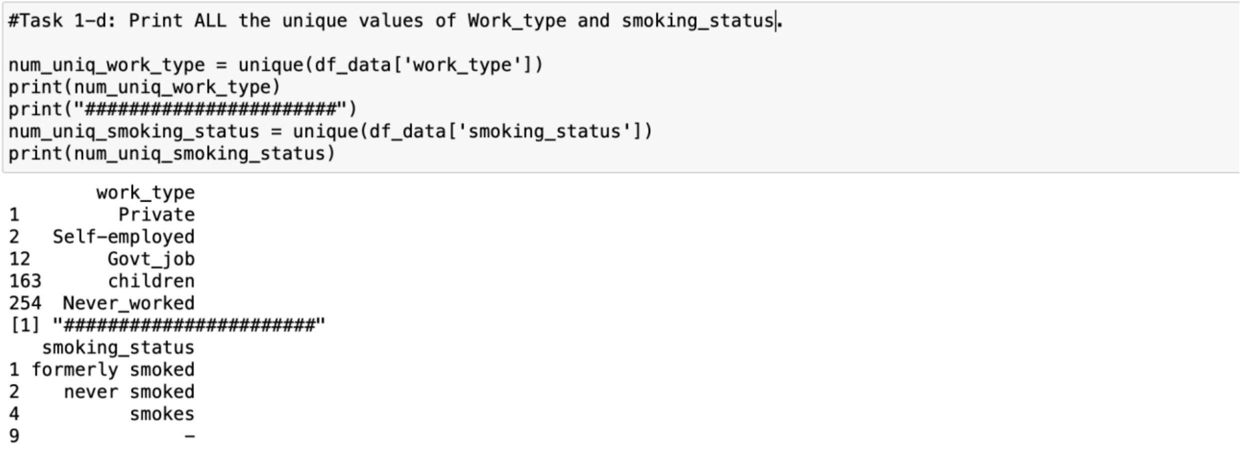
```
#Task 1-c: Print the descriptive detail (count, unique, top, freq etc) for 'Age'' column of the df_data

print ("\n\n>>Task 1-c: Descriptive details of ‘age' column are\n")
print("Summary of Age column is")
summary(df_data['age'])
print("Standard deviation of Age column is:")
print(sd(df_data$age))
```

```
[1] "\n\n>>Task 1-c: Descriptive details of ‘age' column are\n"
[1] "Summary of Age column is"

      age
 Min.   : 0.08
 1st Qu.:25.00
 Median :45.00
 Mean   :43.23
 3rd Qu.:61.00
 Max.   :82.00

[1] "Standard deviation of Age column is:"
[1] 22.61265
```



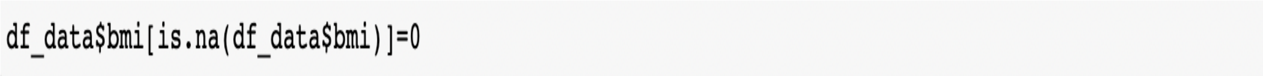
```
#Task 1-d: Print ALL the unique values of Work_type and smoking_status.

num_uniq_work_type = unique(df_data['work_type'])
print(num_uniq_work_type)
print("#######################")
num_uniq_smoking_status = unique(df_data['smoking_status'])
print(num_uniq_smoking_status)
```

```
        work_type
1         Private
2   Self-employed
12       Govt_job
163      children
254  Never_worked
[1] "#######################"
   smoking_status
1 formerly smoked
2    never smoked
4          smokes
9               -
```

Check for Missing Data

At times when we are trying to import data from the csv file, we encounter missing data in some columns. We need to check this data cautiously because missing data can change the results during data analysis. In R missing values are represented by the symbol NA (not available). We need to fill the data in the column according to type of data saved in the column.



```
df_data$bmi[is.na(df_data$bmi)]=0
```

In the above code snippet, we are filling the value 0 in the cells which have N.A as a value that is the cells are empty.

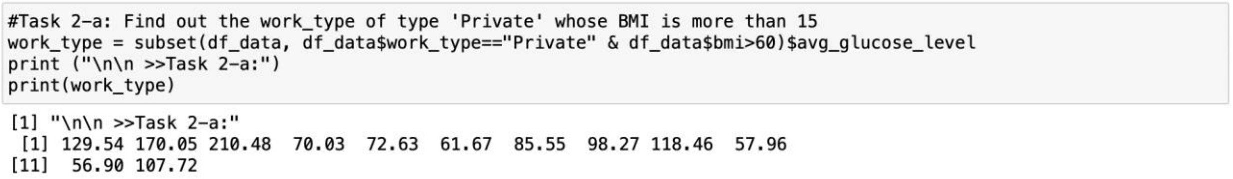
Data Exploration

Data Exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe each dataset characterizations, such as size, quantity and accuracy in order to understand the nature of the data.

Task 2 Aggregation & Filtering & Rank

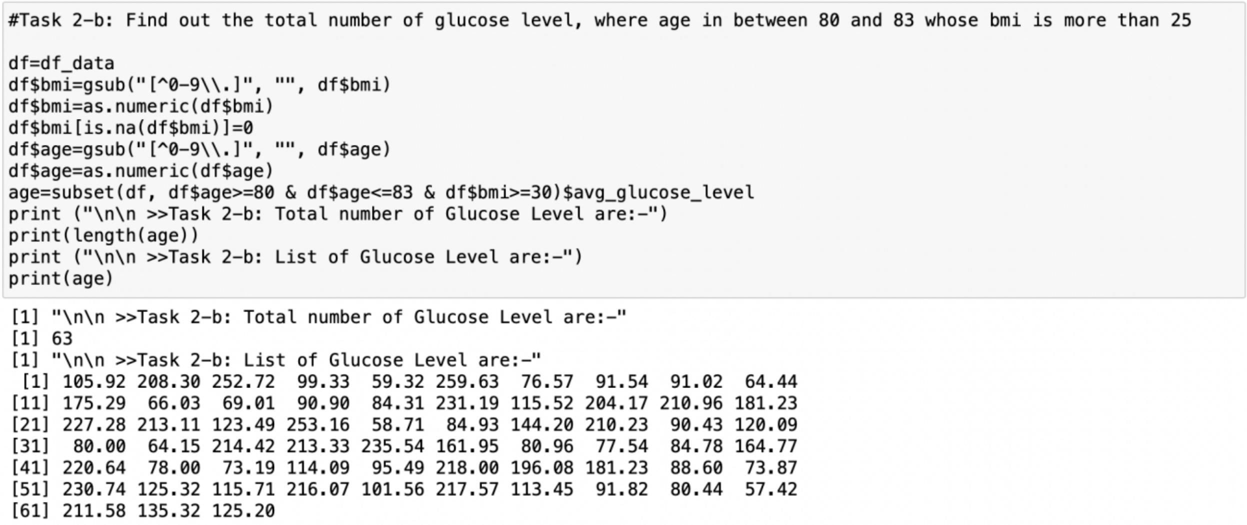
Aggregation:

The process of gathering data and presenting it in a summarized format is called aggregation. We have implemented aggregation on following dataset and generated outcomes mentioned below.



```
#Task 2-a: Find out the work_type of type 'Private' whose BMI is more than 15
work_type = subset(df_data, df_data$work_type=="Private" & df_data$bmi>60)$avg_glucose_level
print ("\n\n >>Task 2-a:")
print(work_type)
```

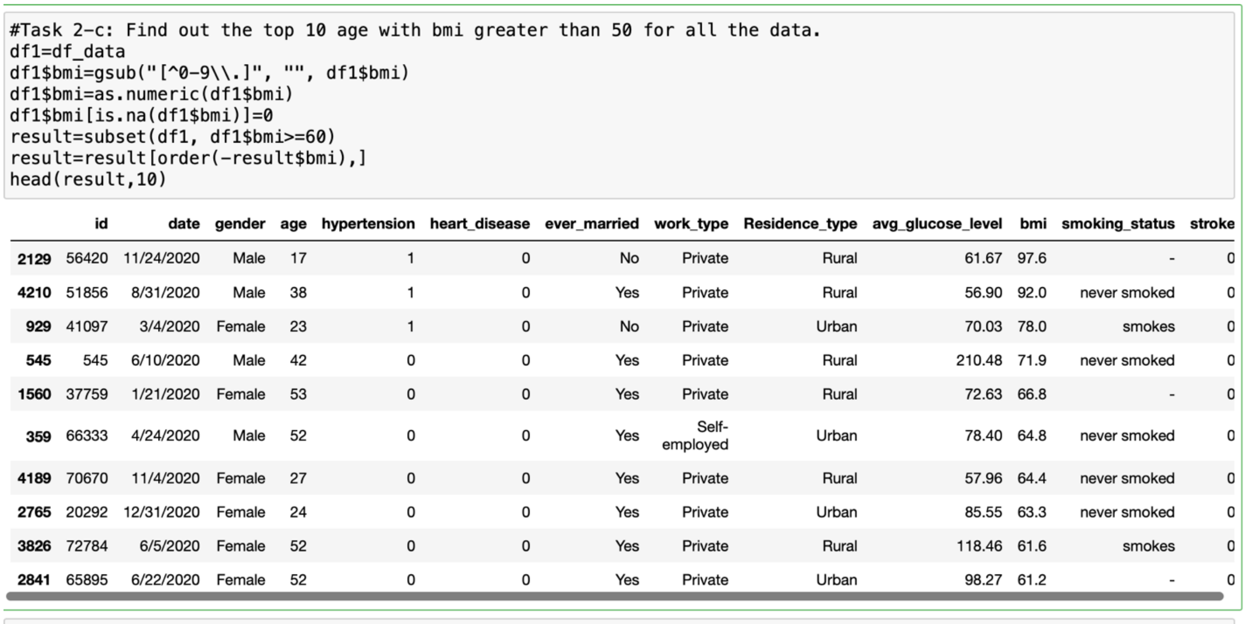
```
[1] "\n\n >>Task 2-a:"
 [1] 129.54 170.05 210.48  70.03  72.63  61.67  85.55  98.27 118.46  57.96
[11]  56.90 107.72
```



```
#Task 2-b: Find out the total number of glucose level, where age in between 80 and 83 whose bmi is more than 25

df=df_data
df$bmi=gsub("[^0-9\\.]", "", df$bmi)
df$bmi=as.numeric(df$bmi)
df$bmi[is.na(df$bmi)]=0
df$age=gsub("[^0-9\\.]", "", df$age)
df$age=as.numeric(df$age)
age=subset(df, df$age>=80 & df$age<=83 & df$bmi>=30)$avg_glucose_level
print ("\n\n >>Task 2-b: Total number of Glucose Level are:-")
print(length(age))
print ("\n\n >>Task 2-b: List of Glucose Level are:-")
print(age)
```

```
[1] "\n\n >>Task 2-b: Total number of Glucose Level are:-"
[1] 63
[1] "\n\n >>Task 2-b: List of Glucose Level are:-"
 [1] 105.92 208.30 252.72  99.33  59.32 259.63  76.57  91.54  91.02  64.44
[11] 175.29  66.03  69.01  90.90  84.31 231.19 115.52 204.17 210.96 181.23
[21] 227.28 213.11 123.49 253.16  58.71  84.93 144.20 210.23  90.43 120.09
[31]  80.00  64.15 214.42 213.33 235.54 161.95  80.96  77.54  84.78 164.77
[41] 220.64  78.00  73.19 114.09  95.49 218.00 196.08 181.23  88.60  73.87
[51] 230.74 125.32 115.71 216.07 101.56 217.57 113.45  91.82  80.44  57.42
[61] 211.58 135.32 125.20
```



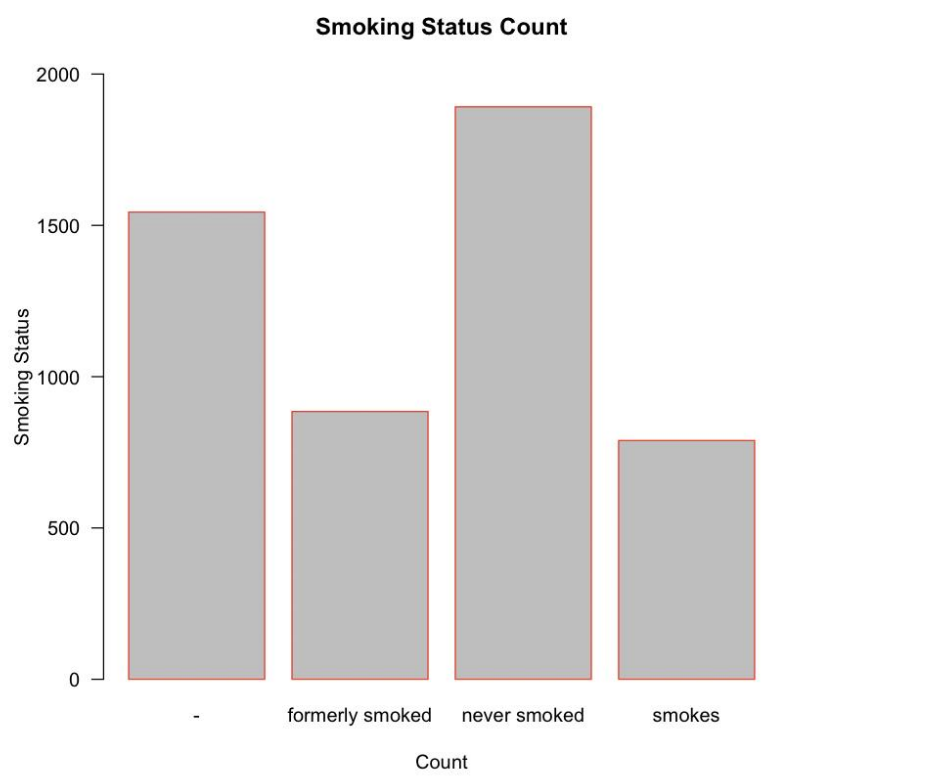
```
#Task 2-c: Find out the top 10 age with bmi greater than 50 for all the data.
df1=df_data
df1$bmi=gsub("[^0-9\\.]", "", df1$bmi)
df1$bmi=as.numeric(df1$bmi)
df1$bmi[is.na(df1$bmi)]=0
result=subset(df1, df1$bmi>=60)
result=result[order(-result$bmi),]
head(result,10)
```

	id	date	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
2129	56420	11/24/2020	Male	17	1	0	No	Private	Rural	61.67	97.6	-	0
4210	51856	8/31/2020	Male	38	1	0	Yes	Private	Rural	56.90	92.0	never smoked	0
929	41097	3/4/2020	Female	23	1	0	No	Private	Urban	70.03	78.0	smokes	0
545	545	6/10/2020	Male	42	0	0	Yes	Private	Rural	210.48	71.9	never smoked	0
1560	37759	1/21/2020	Female	53	0	0	Yes	Private	Rural	72.63	66.8	-	0
359	66333	4/24/2020	Male	52	0	0	Yes	Self-employed	Urban	78.40	64.8	never smoked	0
4189	70670	11/4/2020	Female	27	0	0	Yes	Private	Rural	57.96	64.4	never smoked	0
2765	20292	12/31/2020	Female	24	0	0	Yes	Private	Urban	85.55	63.3	never smoked	0
3826	72784	6/5/2020	Female	52	0	0	Yes	Private	Rural	118.46	61.6	smokes	0
2841	65895	6/22/2020	Female	52	0	0	Yes	Private	Urban	98.27	61.2	-	0

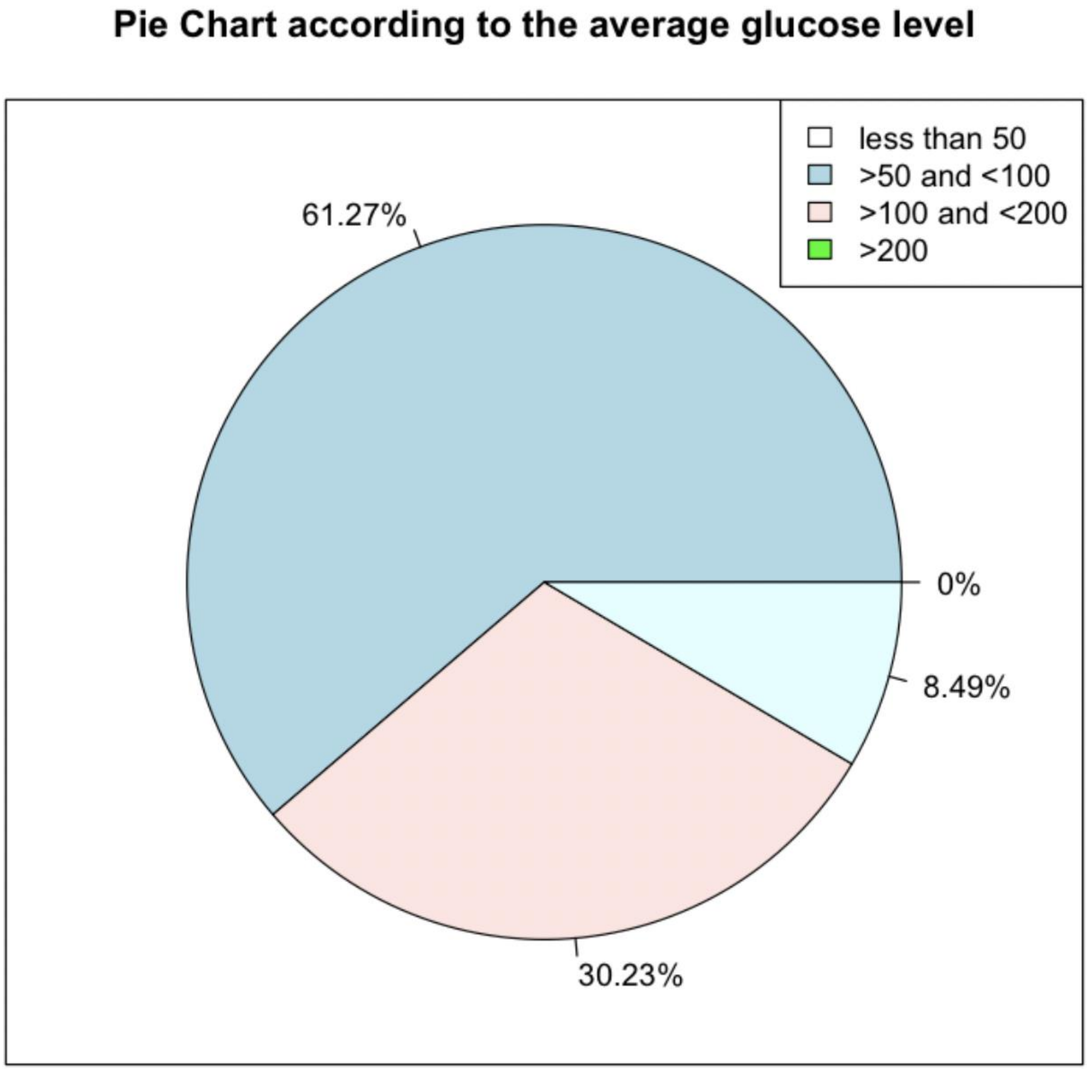
Task 3 Visualization:

The process of representing data in form of graphical representation is called visualization. Elements like charts, graphs and maps are used for visualizing data. Here we have visualized data of smokers count based on smoking status.

We can find histogram representing number of people based on each category regarding their smoking status

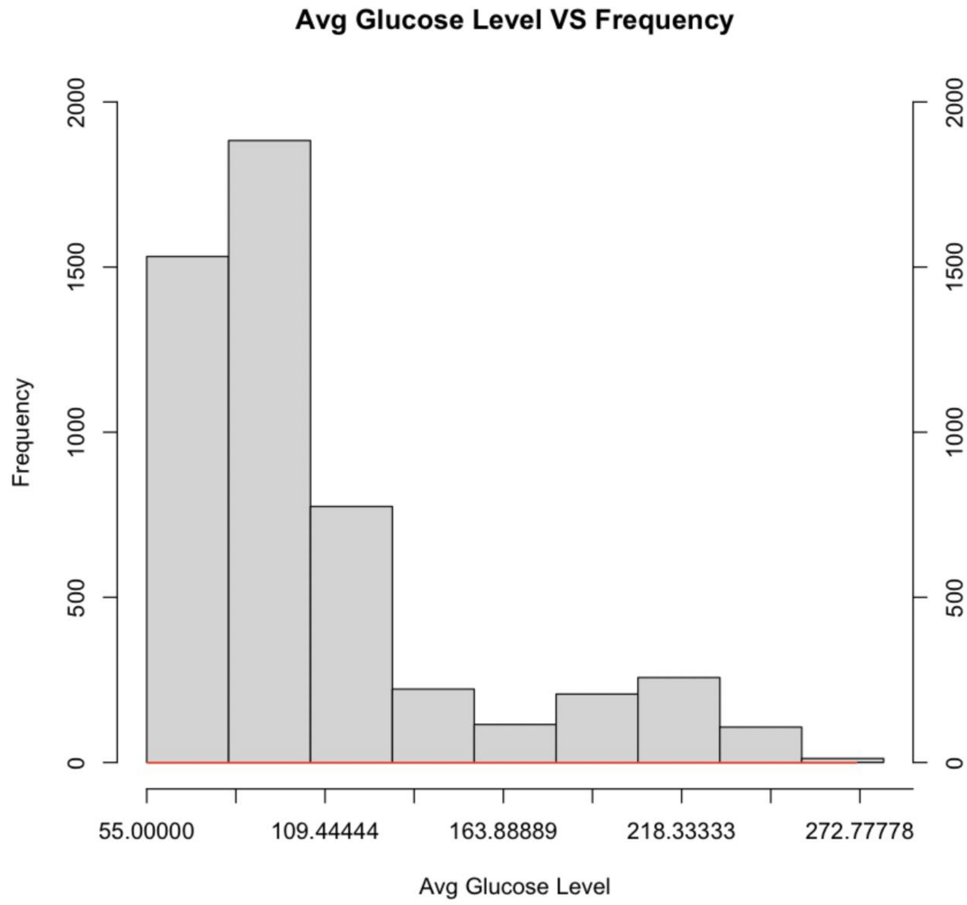


A pie chart is made representing smoking status count on basis of the given data.



Task 4:
Deep dive in dataset to extract unique information and visualizing it

Here we have tried to figure out correct column to visualize unique data and found average glucose level as convenient column and extracted the following output.



From the above graph we can see that Avg Glucose level from 82-100(approx.) is having highest level of glucose.

References:

- https://www.geeksforgeeks.org/r-programming-language-introduction/
- https://www.r-project.org/other-docs.html
- https://www.datamentor.io/r-programming/histogram/
- https://en.wikipedia.org/wiki/R_(programming_language