Data Handling in R Subashree Bhaskaran

1. Read the input data file **“Bank500.csv”** in to R as input.data
   1. Write the code to read a .csv file  
      my\_data = read.csv("Bank500.csv”)
2. Use following commands to learn about the data you have read into **input.data** – Write code for each;
   1. Number of rows

str(my\_data) 500 rows

* 1. Number of columns  
     str(my\_data) 21 columns
  2. Display Top 6 rows   
     \_\_head(my\_data)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
  3. Display bottom 6 rows  
     \_\_tail(my\_data)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
  4. Display first 20 rows  
     \_head(my\_data, 20)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Enter code to understand structure of **input.data**  
   \_\_\_\_str(my\_data)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. Enter code to get a better understanding of the data with a summary  
   \_\_\_\_\_\_summary(my\_data)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. What is the datatype of variable “contact”? What are the different levels or values that the variable contact can assume?  
   \_str(my\_data) Factor w/ 3 levels\_\_\_\_\_\_\_\_\_\_\_\_  
     
   \_summary(my\_data) cellular, telephone, unknown\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. Write code to display only the “age” variable  
   \_\_my\_data$age\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
5. Write code to list the names of all the variables in input.data  
   \_\_\_\_names(my\_data)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
6. Create a new variable “age\_below\_40” which is a subset of all input.data where all the rows of age <= 40. Use subset function.  
   age\_below\_40 <- subset(my\_data, my\_data$age <=40)

age\_below\_40

1. What is the data type of age\_below\_40?  
   data.frame
2. How many rows are in age\_below\_40?

nrow(age\_below\_40)

[1] 248

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Order the rows in age\_below\_40 in descending order by age – Write code below  
   age\_below\_40[order(age\_below\_40$age, decreasing = TRUE),]
2. From the output above, What is the balance of the customer listed on the top row?  
   \_\_\_\_\_3571\_\_\_\_\_\_\_\_\_\_\_\_\_\_

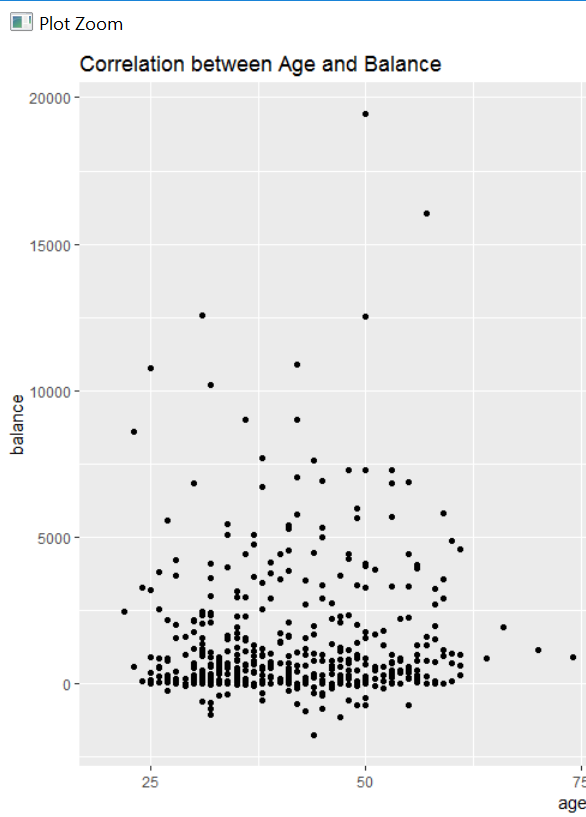
Basic Graphs in R

In this section, we will learn the following:

* Graphically understand data
* Generate different carts based on data type
* Interpret and gather insights from charts generated

1. Generate a Histogram for ‘**age’** field in **input.data** – Write code below:  
   \_\_\_\_\_\_\_hist(my\_data$age)\_\_
2. Do you observe anything unusual from the histogram generated above? If yes, state your observation  
   Data available for Age between 120 and 140, which is an outlier – Suspicious data
3. Generate a Barplot to understand the distribution of customers based on marital status – Write code below:  
   plot(my\_data$marital)
4. Generate a Scatter Plot to understand the correlation between **Duration** on x-axis and **Balance** on y-axis. Write code below:  
   plot(my\_data$duration, my\_data$balance, main = "Duration Vs Balance", xlab = " Duration in days", ylab = "Balance in Rs.")
5. Generate a Box Plot for the variable ‘balance’. Make sure to specify the title and labels for X and Y axis. Write the code below:  
   boxplot(my\_data$balance, main = "Balance of customers", xlab = "Customers", ylab = "Rs.")
6. Generate a Box Plot for balance versus job. Write code below:  
   boxplot(my\_data$balance ~ my\_data$job, main = "Balance of customers by Job Profile", xlab = "Job", ylab = "Balance")
7. From the Box Plot generated above, which job category has the highest median balance?  
   “unknown” job category
8. Using ggplot2 package, generate a point plot where x-axis is education and y-axis is balance. Write code below:  
   ggplot(my\_data, aes(x = education, y = balance)) + geom\_point() + ggtitle("Education Vs Bank Balance")
9. Using ggplot2 package, generate a point plot where x-axis is age and y-axis is balance. Write code below: - Is there a correlation between age and balance?  
   ggplot(my\_data, aes(x = age, y = balance)) + geom\_point() + ggtitle("Correlation between Age and Balance")

There is not a direct obvious correlation between age and balance, as most observations have balance around the min balance of 0. However, between age 35 and 60, balances spread between 0 & Rs8000 and not dense at Rs0. We can infer that this age group, people bank more often than age between 25 to 35.



1. Enhance the point plot above so that we can identify those who have housing loan based on colour. Write code below:  
   ggplot(my\_data, aes(x = age, y = balance, color = housing)) + geom\_point() + ggtitle("Correlation between Age and Balance")
2. Let us try to understand the overall Monthly Average **Balance**. Run the code below and answer relevant questions:  
     
   means = tapply(input.data$balance, input.data$numMonth, mean)  
     
   barplot(means, xlab = "Month", ylab = "Average Balance", main = "Monthly Average Balance")  
   1. Which month had the highest average Balance?  
      \_\_Month 12 - December\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   2. Which month had the lowest average Balance?  
      \_Month 7 - July\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
   3. How many months have average balance of less than 4000?  
      \_\_\_\_11\_\_\_\_\_\_\_