Data Analysis Report on Beers

Table of Contents

[1. Introduction 1](#_Toc21912)

[1.1 Purpose and Use 1](#_Toc21913)

[1.2 Methodology 1](#_Toc21916)

[2. Data Setup 2](#_Toc21917)

[2.1 Cleaning in Microsoft Excel 2](#_Toc21918)

[2.2 Setting up the data in R 2](#_Toc21920)

[3. Exploratory Data Analyses 3](#_Toc21921)

[3.1 One-variable analyses 3](#_Toc21922)

[3.2 Two-variable analyses 9](#_Toc21925)

[4. Advanced Analyses 1](#_Toc21928)5

[4.1 K-means Clustering 15](#_Toc21929)

[4.1.1 Explanation 15](#_Toc21930)

[4.1.2 Clustering Beers 1](#_Toc21931)5

[4.2 Linear Regression 1](#_Toc21932)7

[4.2.1 Linear Regression Explanation 1](#_Toc21933)7

[4.2.2 Linear Regression forABV 17](#_Toc21934)

[5. Conclusion 19](#_Toc21936)

[6. Reflections 20](#_Toc21937)

# 1. Introduction

### 1.1 Purpose and Use

The report analyses beers and visualizes the data in an understandable and meaningful form. When it boils down to is, the purpose of the report is to find how to make good beer.

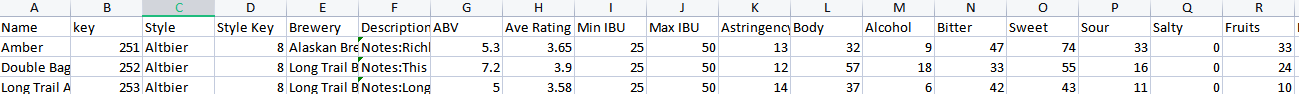
### 1.2 Methodology

The report will make use of one and two-variable analysis, clustering with K-means, and linear regression data analyses and a range of data visualization techniques, and will be performed in R.

# 2. Data Setup

### 2.1 Arrangement in Microsoft Excel

The data set is stored as “beer\_data\_set.csv”. The columns represent attributes of beers.



### 

### 2.2 Setting up the data in R

Prior to analyzing the data, the data is needed to be properly set-up in a data frame with the following R code. And arrange data in ascending order of rating for analysis.

Also, remove unnecessary items for analysis.

########### **Code** ###########

# Load data from ‘.csv’ file

data <- read.csv(file = 'beer\_data\_set.csv')

# Remove unnecessary column

data <- subset(data, select = -Description)

# 3. Exploratory Data Analyses

### 3.1 One-variable analyses

In this section, we will analyze what attributes top beers have.

For this, display graph that shows the relation between average rating and each style.

The following images show the relation between

########### **Code** ###########

# ABV

plot( arranged\_data$ABV, main="ABV attribute", ylab="ABV", xlab = "Rating", col="red", cex=.08)

# Body

plot( arranged\_data$Body, main="Body attribute", xlab = "Rating",ylab="Body", col="green", cex=.08)

#Alcohol

plot( arranged\_data$Alcohol, main="Alcohol attribute", xlab = "Rating", ylab="Alcohol", col="red", cex=.08)

#Bitter

plot( arranged\_data$Bitter, main="Bitter attribute", xlab = "Rating", ylab="Bitter", col="blue", cex=.08)

#Sweet

plot( arranged\_data$Sweet, main="Sweet attribute", xlab = "Rating", ylab="Sweet", col="green", cex=.08)

# Sour

plot( arranged\_data$Sour, main="Sour attribute", xlab = "Rating", ylab="Sour", col="red", cex=.08)

#Salty

plot( arranged\_data$Salty, main="Salty attribute", xlab = "Rating", ylab="Salty", col="blue", cex=.08)

# Fruits

plot( arranged\_data$Fruits, main="Fruits attribute", xlab = "Rating", ylab="Fruits", col="green", cex=.08)

# Hoppy

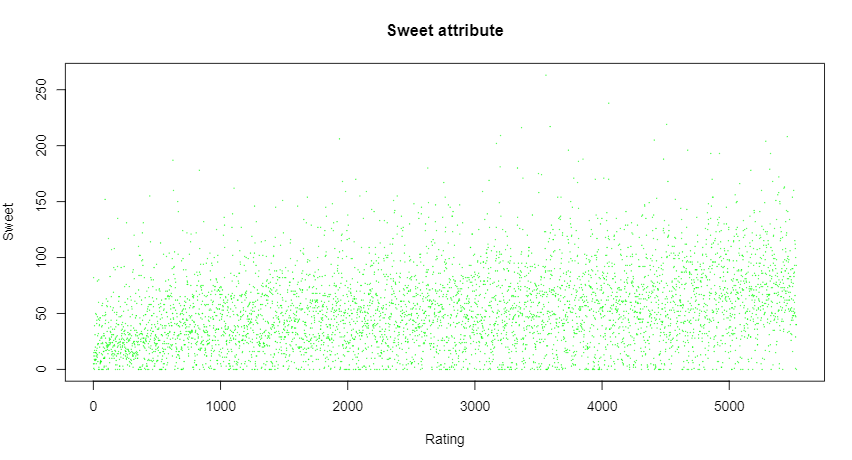
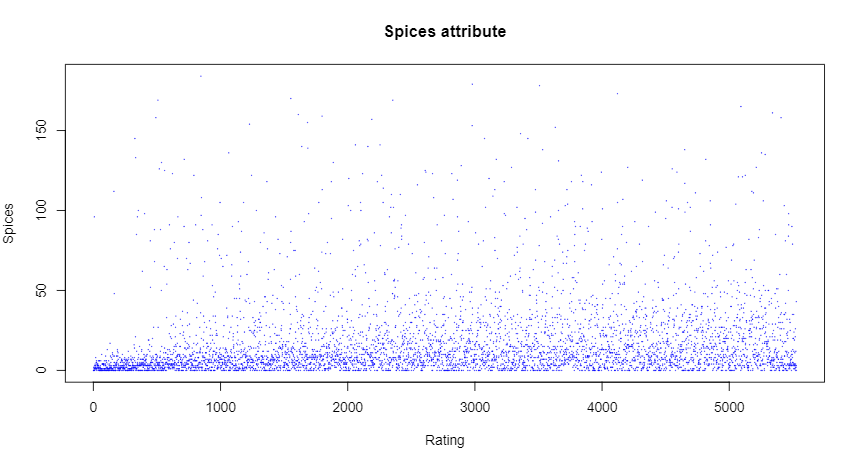
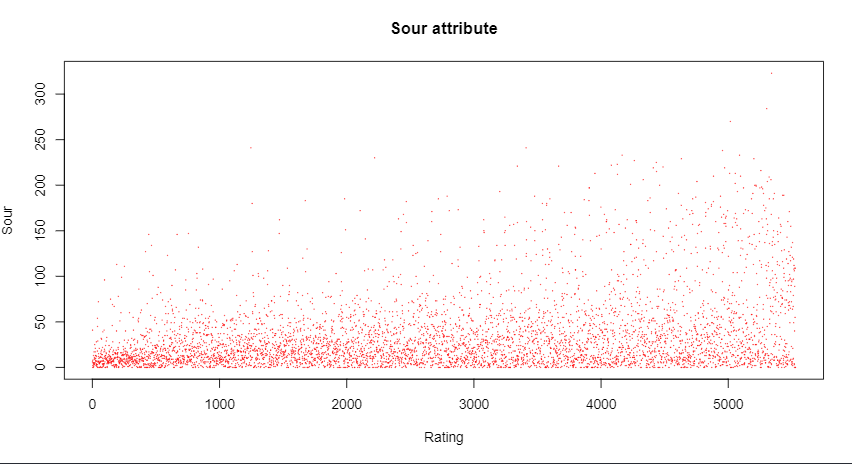
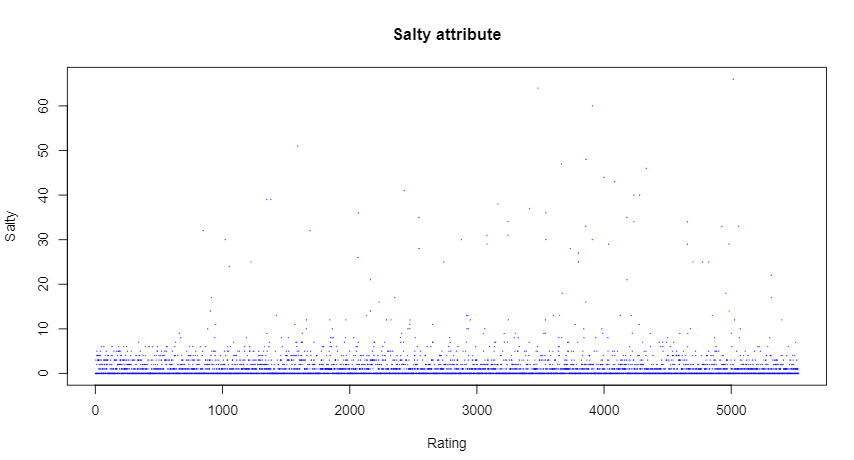
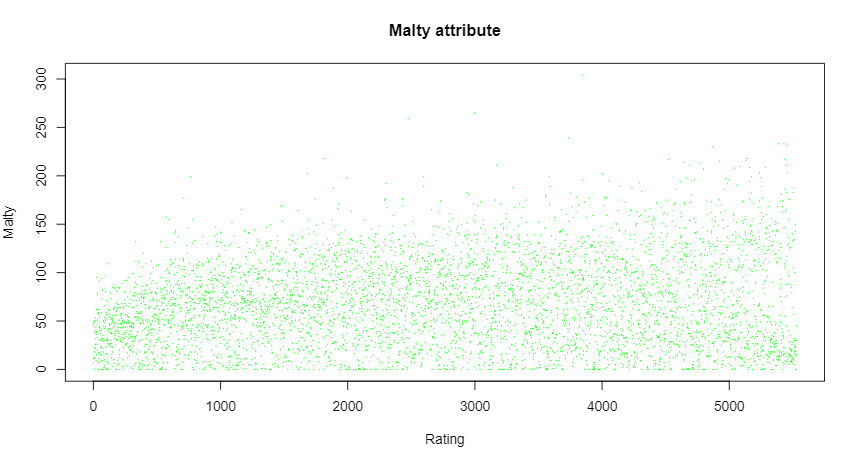
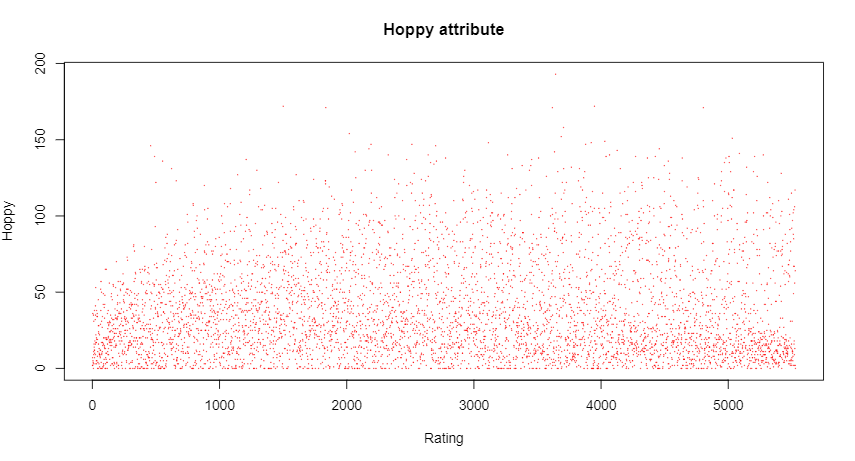
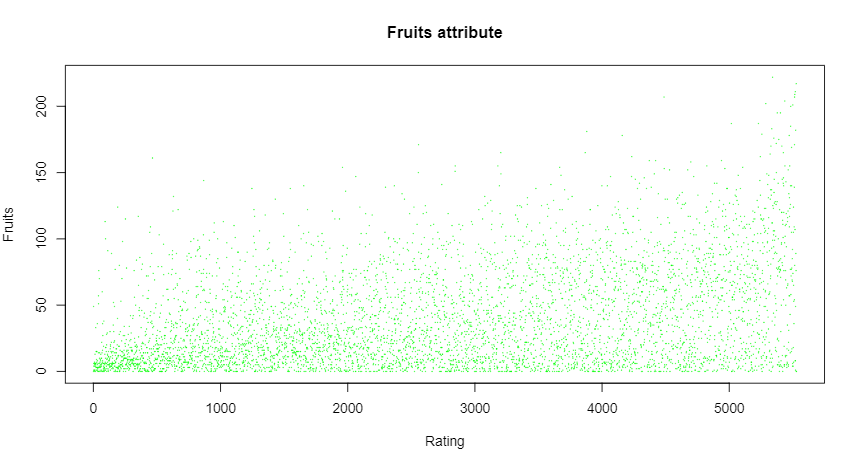
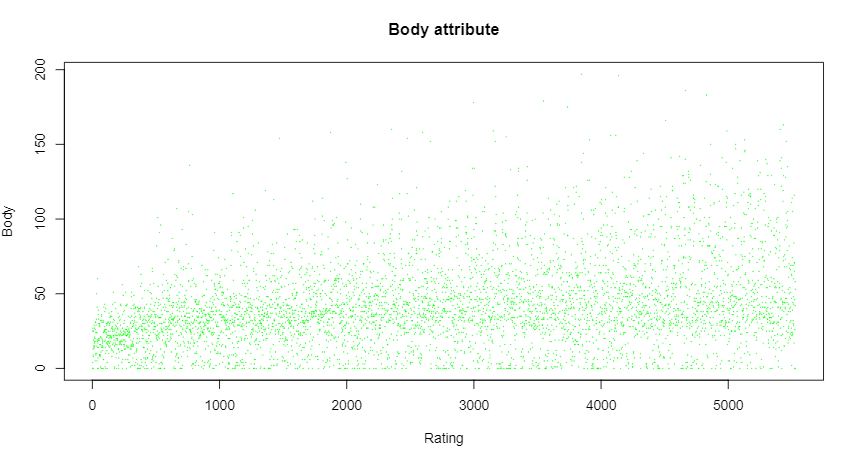
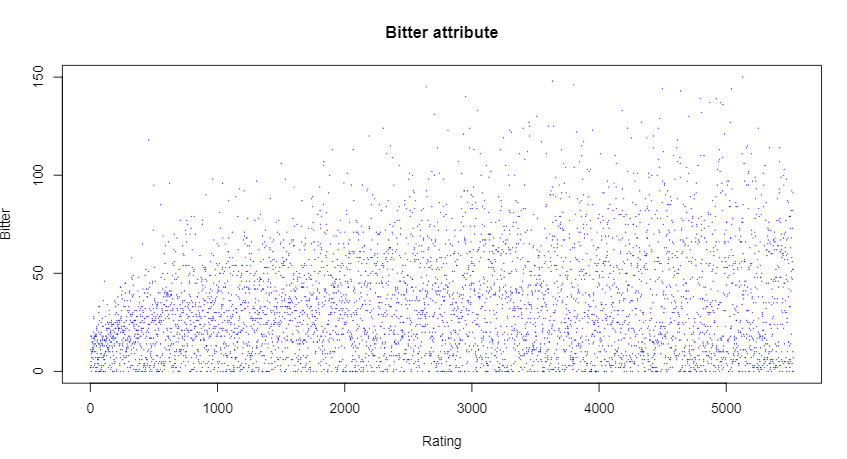
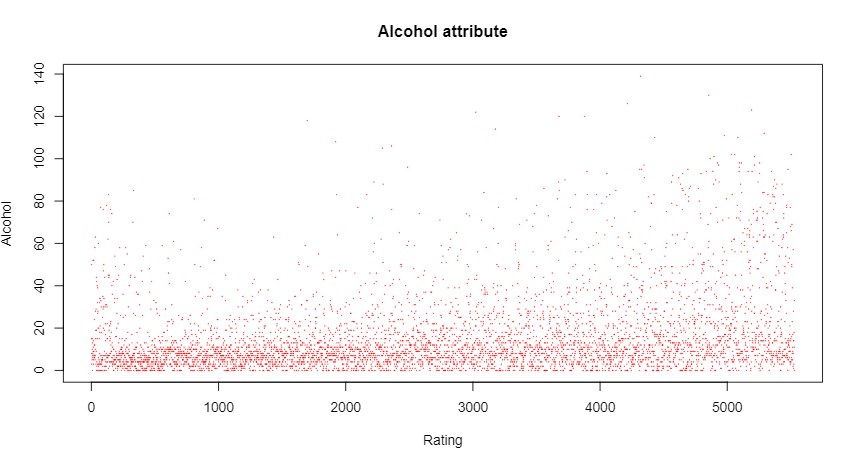
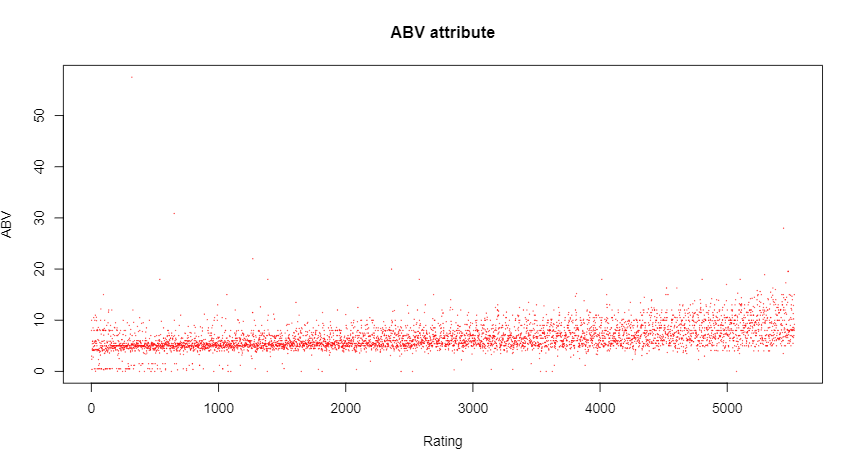
plot( arranged\_data$Hoppy, main="Hoppy attribute", xlab = "Rating",ylab="Hoppy", col="red", cex=.08)

# Spices

plot( arranged\_data$Spices, main="Spices attribute", xlab = "Rating", ylab="Spices", col="blue", cex=.08)

# Malty

plot( arranged\_data$Malty, main="Malty attribute", xlab = "Rating", ylab="Malty", col="green", cex=.08)



As you can see in the pictures, ABV, Alcohol, Fruits, Sour and Spices attributes has significant increase as the rating increases. From this, we can say that top beers have attributes of ABV, Alcohol, Fruits, Sour and Spices. Also, almost attribute has increasing value as the rating increases.

### 3.2 Two-variable analyses

In this part, we are looking for attributes of top beers that could be put together to create a distinctive yet tasty beer. For this, we are going to analyze correlation between attributes mentioned above and find out a few attributes that have less correlation. The code is the same as follows.

########### **Code** ###########

plot(arranged\_data$ABV, arranged\_data$Alcohol, xlab = "ABV", ylab="Alcohol", cex=.08)

plot(arranged\_data$ABV, arranged\_data$Fruits, xlab = "ABV", ylab="Fruits", cex=.08)

plot(arranged\_data$ABV, arranged\_data$Sour, xlab = "ABV", ylab="Sour", cex=.08)

plot(arranged\_data$ABV, arranged\_data$Spices, xlab = "ABV", ylab="Spices", cex=.08)

plot(arranged\_data$Alcohol, arranged\_data$Fruits, xlab = "Alcohol", ylab="Fruits", cex=.08)

plot(arranged\_data$Alcohol, arranged\_data$Sour, xlab = "Alcohol", ylab="Sour", cex=.08)

plot(arranged\_data$Alcohol, arranged\_data$Spices, xlab = "Alcohol", ylab="Spices", cex=.08)

plot(arranged\_data$Fruits, arranged\_data$Sour, xlab = "Fruits", ylab="Sour", cex=.08)

plot(arranged\_data$Fruits, arranged\_data$Spices, xlab = "Fruits", ylab="Spices", cex=.08)

plot(arranged\_data$Sour, arranged\_data$Spices, xlab = "Sour", ylab="Spices", cex=.08)

##### ABV-AlcoholABV-FruitsABV-SourABV-SpicesAlcohol-FruitsAlcohol-SourAlcohol-SpicesFruits-SourFruits-SpicesSour-Spices

As we can see in pictures, there are correlations between ABV and Alcohol, Fruits and Soul.

So if we consider ABV which is the most important attribute, Fruits and Spices, we can make up good beer.

# 4. Advanced Analyses

### 4.1 K-means Clustering

#### 4.1.1 Explanation

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means clustering is a widely used approach for clustering. Generally, practitioners begin by learning about the architecture of the data set. K-means clusters data points into unique, non-overlapping groupings.

#### 4.1.2 Clustering Beers

Here, k-means clustering would be performed on the average rating of beers.

For clear visualization, we used only 20 data.

########### **Code** ###########

png(file="k\_mean.png")

plot(head(arranged\_data[, 7], n=20))

dev.off()

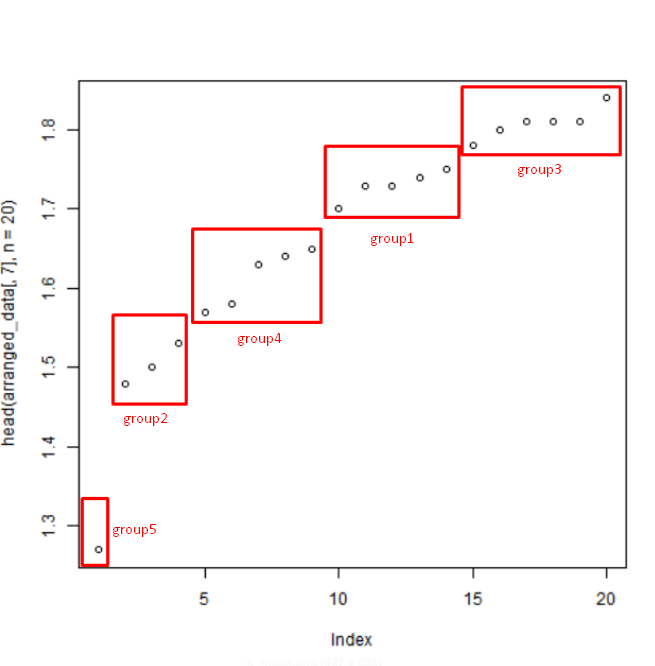
fit <- kmeans(head(arranged\_data[, 7], n = 20), 5, nstart=25)

plot(fit$cluster)

points(fit$centers, col = 1:5, pch = 8)

#Output

5 2 2 2 4 4 4 4 4 1 1 1 1 1 3 3 3 3 3 3



### 4.2 Linear Regression

##### 4.2.1 Linear Regression Explanation

Linear regression is used to predict the value of an outcome variable y on the basis of one or more input predictor variables x. In other words, linear regression is used to establish a linear relationship between the predictor and response variables.

In linear regression, predictor and response variables are related through an equation in which the exponent of both these variables is 1. Mathematically, a linear relationship denotes a straight line, when plotted as a graph.

##### **4.2.2 Linear Regression for ABV**

##### In this part, we are going to predict ABV according to Ave Rating.

########### **Code** ###########

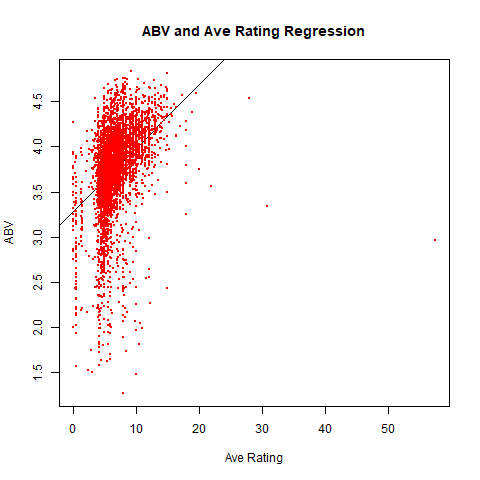
y = arranged\_data$ABV

x = arranged\_data$Ave\_Rating

linear\_regression\_model <- lm(y ~ x)

print(summary(linear\_regression\_model))

plot(y, x, col = "red", main = "ABV and Ave Rating Regression", abline(lm(x~y)),cex = 1.3,pch = 16, xlab = "Ave Rating", ylab = "ABV")



# Conclusion

I had more knowledge of data science and R programming language. Especially I had good knowledge of R and statistics. I have analyzed what attributes top beers have and found out how to do to make good beer. And by several mathematical methods, I have found out some interesting things.

# 6. Reflections

I have suffered from less knowledge of R language. In particular, I couldn’t code fast because I spent lots of time for finding out functions to process “.csv” files and to learn how to plot data on screen.

Also, I spent lots of time to fix bugs. Anyhow, I had good knowledge of R language through this assignment.

When I started assignment, I didn’t know k-mean clustering well. But I have learned by myself and fully understood.