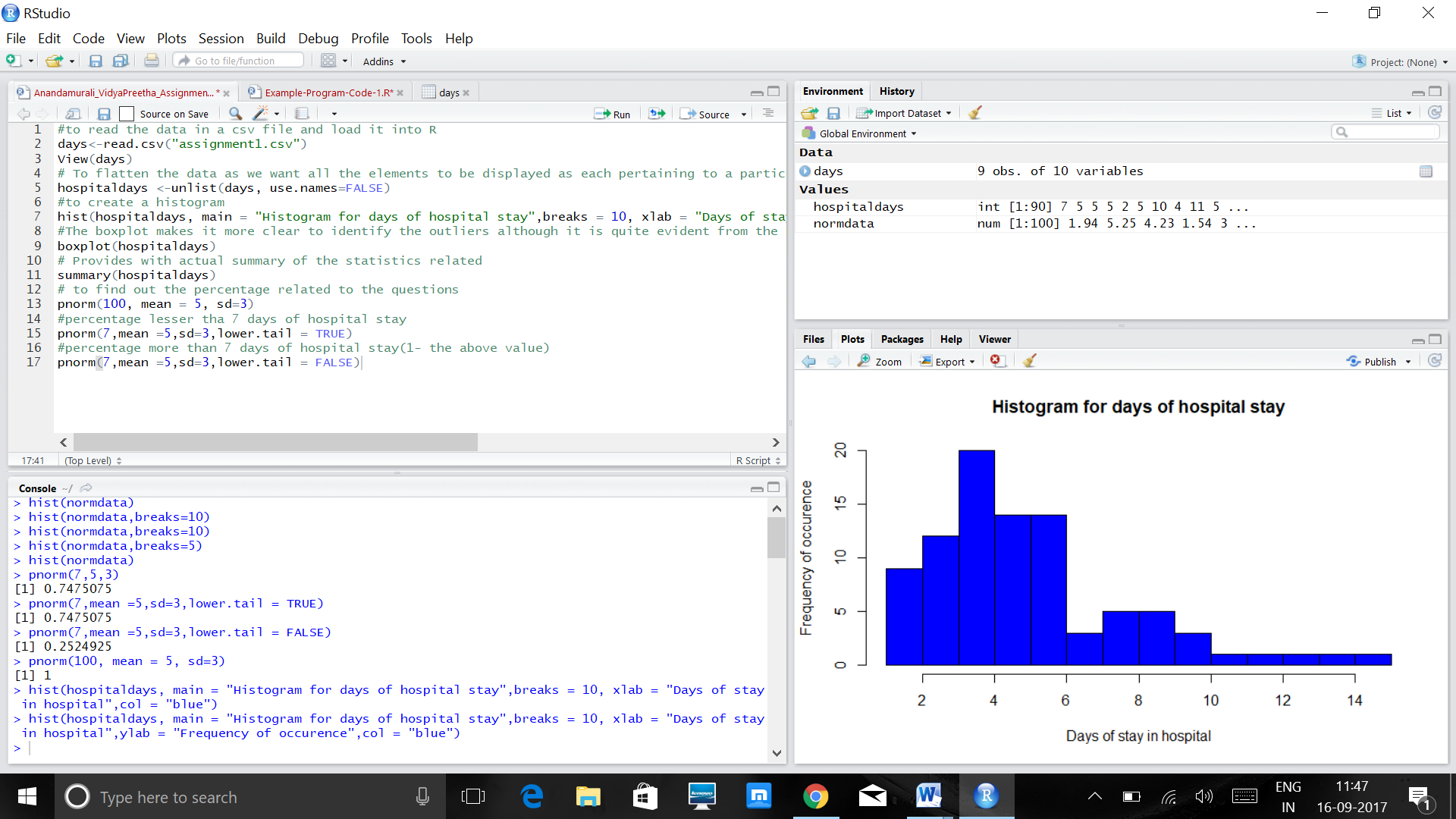
Assignment 1 – Vidya Preetha Anandamurali

1. *To save the data we will just have to copy it to an excel sheet and then convert it to a csv file.*
2. *Histogram created –*



It is quite evident from the above histogram that the histogram is labelled and has a width of 1 day. The following are the observations from the graph :-  
(i) The **range** of the data goes from 1 day to 15 days which shows the variability in the data points.   
(ii) The **mean** of the data is around 5 to 6 days of range from the graph. The **centre** is not very clear with this graph as it has skewed values present in it.  
(iii) The highest occurrence of stay in the hospital is 3-4 days of gap, which has a **frequency** of 20 occurences, this is therefore the peak of the graph and is **unimodal** in nature as one evident **peak** is only present.  
(iv) The graph is **skewed to the right** as fewer observations are found more to the right of the graph, which shows that the graph is not symmetric in nature.  
(v) There are no evident **gaps** present in this graph.   
(vi) There are **outliers** that are present in the graph which include days ranging from 10 days to 15 days. These range of days provide inadequate results to the graph summaries with respect to the mean, as these skewed values will make the mean look bigger than usual, therefore at instances like this the median is a more reliable parameter to describe the middle value of the data.

1. *Summary Table:-*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mean | Median | Standard deviation | 1st quartile | 2nd quartile | Min Value | Max Value |
| 5.4 | 5 | 2.827633 | 4 | 6 | 1 | 15 |

(i) The best single number for the center will be the median of the data set which is 5 days as this will actually tell us the middle data point that keeps occurring more often in the center of the data. The mean value is a little out of track as it includes the extreme values also which contribute very less to the major trend that the data set follows and can end up misleading. (ii) The quartiles, variance and standard deviation are the number statistics usually used for determining the nature of spread of the data. The best single number statistic for the spread of the data is the quartiles as it is less affected by the outliers which provide skewed values in our dataset. The interquartile range tells us about the difference between the first and the third quartile which is 2 days , which tells us more about the spread of the central part of the dataset.

(4) *Working with Normal Distribution:-*(i) Given all the details of the published literature we use the standard and deviation given and then find out the percentage using the R code. **Ans: 0.7475705 which is almost 75%**  
(ii) The sample mean will be closer to the population mean for a particular data. Here, we take a sample size of 10 which is ‘n’, so the mean remains 5, but the standard deviation is the (population sd/root(n)) = (3/root(10)) = 0.948683 . Now we should use this in our R code as the standard deviation with mean as 5. Here, since we want the probability for more than 7 as the mean for the sample, we assign lower.tail as FALSE. Hence, The **Ans : 0.01750746 which is almost 1.75%**

**The R Code:**

#to read the data in a csv file and load it into R  
days<-read.csv("assignment1.csv")  
View(days)

# To flatten the data as we want all the elements to be displayed as each pertaining to a particular column  
hospitaldays <-unlist(days, use.names=FALSE)

#to create a histogram   
hist(hospitaldays, main = "Histogram for days of hospital stay",breaks = 10, xlab = "Days of stay in hospital",ylab = "Frequency of occurence",col = "blue")

#The boxplot makes it more clear to identify the outliers although it is quite evident from the histogram  
boxplot(hospitaldays)

# Provides with actual summary of the statistics related   
summary(hospitaldays)  
sd(hospitaldays)

#more detailed analysis  
library(psych)  
describe(hospitaldays)

# to find out the percentage related to the questions  
pnorm(100, mean = 5, sd=3)

#percentage lesser than 7 days of hospital stay  
pnorm(7,mean =5,sd=3,lower.tail = TRUE)

#percentage more than 7 days of hospital stay for sample of 10 from the normal distribution data given  
pnorm(7,mean =5,sd=0.948683,lower.tail = FALSE)