*The University of Texas at Austin*

*Zhaowen Fan*

*Rafael Ignacio Gonzalez Chong*

*Assignment 1*

GSI Intro to Big Data and Data Mining

Table of Contents

[Task -1  Save the data to a CSV file and read into R for analysis. (4 points) 2](#_Toc203556990)

[Task - 2 Make a histogram of the duration of days of hospital stays.  Ensure the histogram is labelled appropriately.  Use a width of 1 day.  Describe the shape center and spread of the data.  Are there any outliers? (5 points) 3](#_Toc203556991)

[Task - 3 Find the mean, median, standard deviation, first and third quartiles, minimum and maximum of the durations of hospital stay in the sample. 3](#_Toc203556992)

[Task  - 4:  Make a Boxplot of the duration of days of hospital stays.  Ensure the Boxplot is labelled appropriately. (5 points) 4](#_Toc203556993)

[Appendices (Code) 5](#_Toc203556994)

# **Task -1  Save the data to a CSV file and read into R for analysis. (4 points)**

A screenshot of a computer

AI-generated content may be incorrect.

Fig 1. Data in a csv file

A close-up of a computer screen

AI-generated content may be incorrect.

Fig 2. Code line for reading csv file

A close-up of a computer screen

AI-generated content may be incorrect.

Fig 3. Reading line for csv file

# **Task - 2 Make a histogram of the duration of days of hospital stays.  Ensure the histogram is labelled appropriately.  Use a width of 1 day.** **Describe the shape center and spread of the data.  Are there any outliers? (5 points)**

**A graph of a patient

AI-generated content may be incorrect.**

Fig 4. Histogram

The histogram shows that most patients stay only a few days in the hospital, with the highest frequency at the shortest stays. As the number of days increases, the frequency drops sharply, creating a long right tail. This indicates a positively skewed distribution, where a small number of patients have much longer stays, raising the average, but most hospitalizations are short.

# **Task - 3 Find the mean, median, standard deviation, first and third quartiles, minimum and maximum of the durations of hospital stay in the sample.**

**Summarize these values in a table that you create in your report. Describe the values.**

**- Given the shape of the distribution, what is the best single number summary of the center of the distribution?**

* **What is the best single number summary of the spread of the distribution? (6 points)**

**A computer screen shot of a computer

AI-generated content may be incorrect.**

Fig 5. Table with results

Description of Values

* Mean (5.63 days): This is the average stay
* Median (5.00 days): The middle value, showing that half the patients stay 5 days or less.
* Standard Deviation (2.74 days): most stays fall within about 2-3 days of the mean.
* First Quartile (Q1, 4.00 days): 25% of patients stay less than 4 days.
* Third Quartile (Q3, 7.00 days): 75% of patients stay less than 7 days.
* Minimum (2.00 days): Shortest observed stay in the dataset.
* Maximum (15.00 days): Longest observed stay

Best Single Number Summary for Center

The best measure of central tendency is the median (5.00 days).

Brief Interpretation

Most hospital stays are short, with the majority clustered between 4 and 7 days. A small number of patients have significantly longer stays, which affects the mean more than the median.

# **Task  - 4:  Make a Boxplot of the duration of days of hospital stays.  Ensure the Boxplot is labelled appropriately. (5 points)**

A diagram of a box plot

AI-generated content may be incorrect.

Fig 6. Boxplot of Hospital Day Duration

# **Appendices (Code)**

#ASSIGNMENT 1

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong

#Reading data

days <- read.csv("~/Desktop/work1/datos.csv", header = FALSE)

# To convert and flatten data from a data frame you can use the unlist command.

my.days <-as.numeric(unlist(days, use.names = FALSE))

# Check the type

str(my.days)

# Create a histogram

hist(my.days,

main = "Histogram of Hospital",

xlab = "Days in Hospital",

col = "lightgreen",

border = "black",

breaks = 10)

# Print basic statistics

mean.value <- mean (my.days)

median.value <- median (my.days )

sd.value <- sd( my.days )

q1 <- quantile(my.days, 0.25)

q3 <- quantile(my.days, 0.75)

min.value <- min (my.days )

max.value <- max ( my.days )

iqr.value <- IQR(my.days)

# Summary

summary(my.days)

# Print a summary table

summary.table <- data.frame(

Statistic = c("Mean", "Median", "Std. Deviation", "Q1 (25%)", "Q3 (75%)", "Minimum", "Maximum"),

Value = c(mean.value, median.value, sd.value, q1, q3, min.value, max.value)

)

print(summary.table)

# Create a boxplot

boxplot(my.days,

main = "Boxplot of Hospital Day Duration",

ylab = "Days in Hospital",

col = "lightblue",

border = "darkblue")