ASSIGNMENT 1

GSI Intro to Big Data and Data Mining

The University of Texas at Austin

Zhaowen Fan
Rafael Ignacio Gonzalez Chong

Table of Contents

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

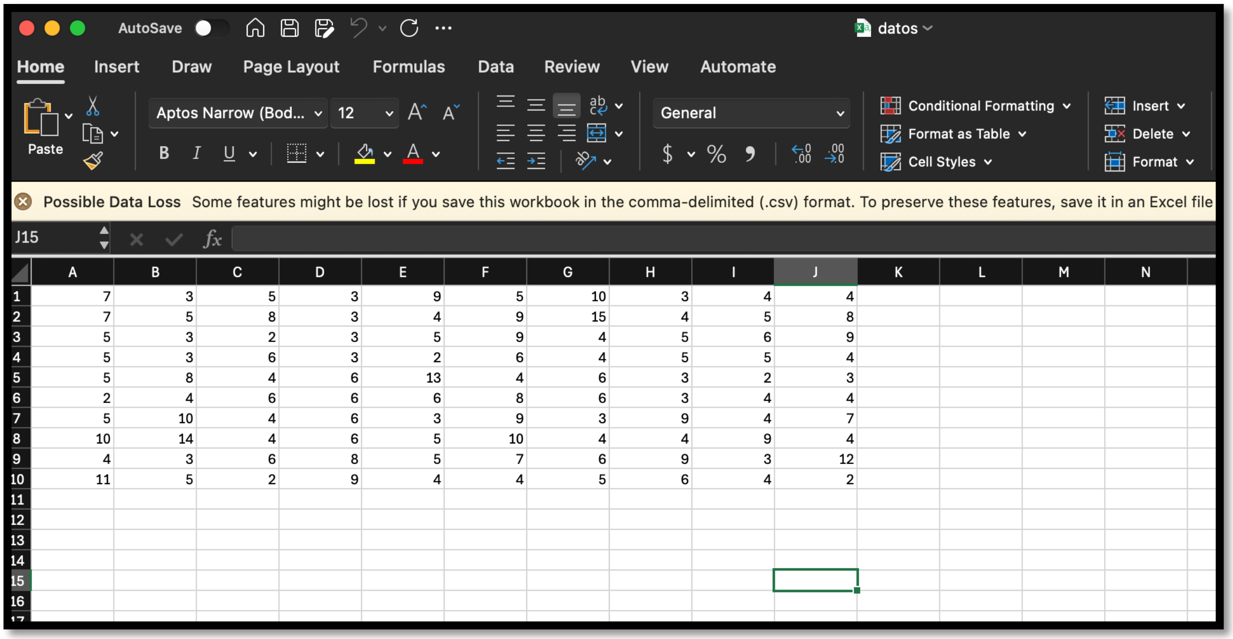
Task - 2 Make a histogram of the duration of days of hospital stays. Ensure the histogram is labelled appropriately. Use a width of 1 day. Describe the shape center and spread of the data. Are there any outliers? (5 points) *3*

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*

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

Appendices (Code) *5*

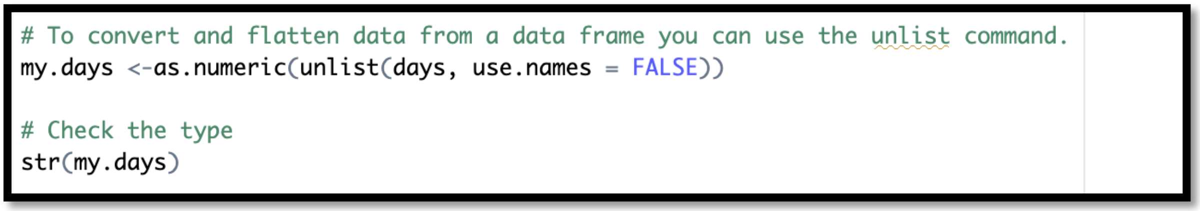
Task -1 Save the data to a CSV file and read into R for analysis. (4 points)



Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file

	A	B	C	D	E	F	G	H	I	J
1	7	3	5	3	9	5	10	3	4	4
2	7	5	8	3	4	9	15	4	5	8
3	5	3	2	3	5	9	4	5	6	9
4	5	3	6	3	2	6	4	5	5	4
5	5	8	4	6	13	4	6	3	2	3
6	2	4	6	6	6	8	6	3	4	4
7	5	10	4	6	3	9	3	9	4	7
8	10	14	4	6	5	10	4	4	9	4
9	4	3	6	8	5	7	6	9	3	12
10	11	5	2	9	4	4	5	6	4	2

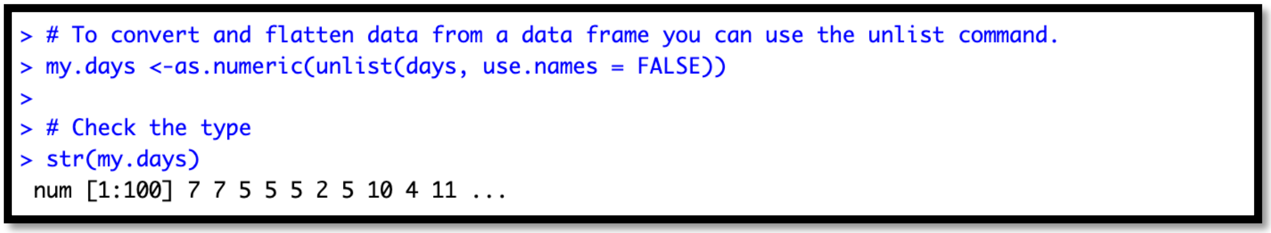
Fig 1. Data in a csv file



```
# 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)
```

Fig 2. Code line for reading csv file



```
> # 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)
 num [1:100] 7 7 5 5 5 2 5 10 4 11 ...
```

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)

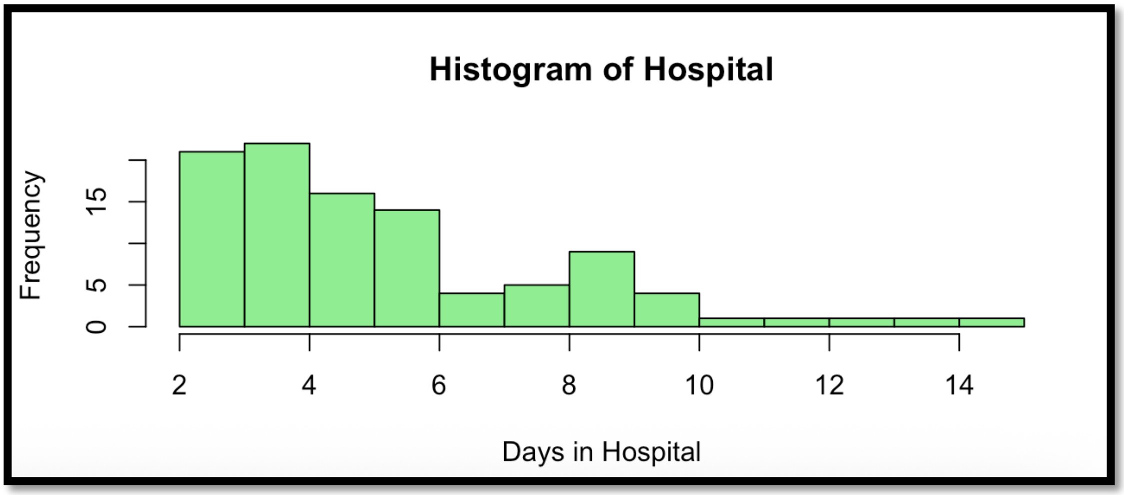


Fig 4. Histogram

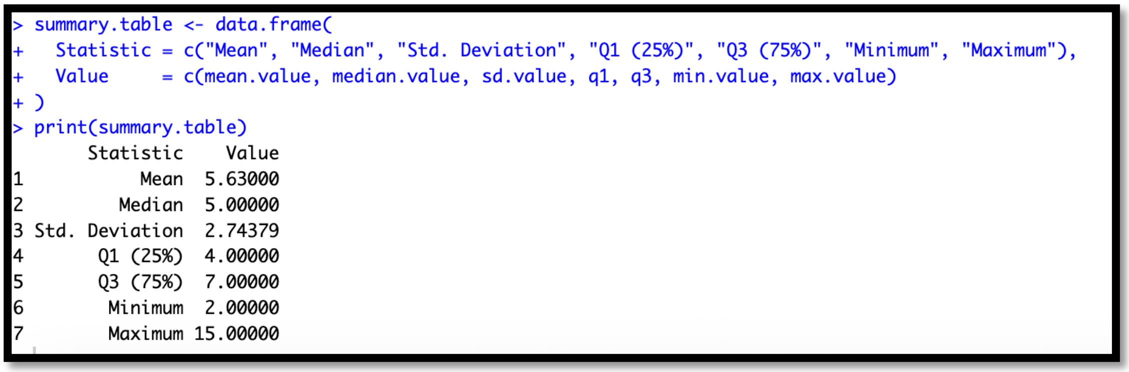
The histogram shows that most patients stay only a few days in the hospital, with the highest frequency at the shortest stays. As the number of days increases, the frequency drops sharply, creating a long right tail. This indicates a positively skewed distribution, where a small number of patients have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



```
> 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)
       Statistic    Value
1           Mean  5.63000
2         Median  5.00000
3 Std. Deviation  2.74379
4       Q1 (25%)  4.00000
5       Q3 (75%)  7.00000
6        Minimum  2.00000
7        Maximum 15.00000
```

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)

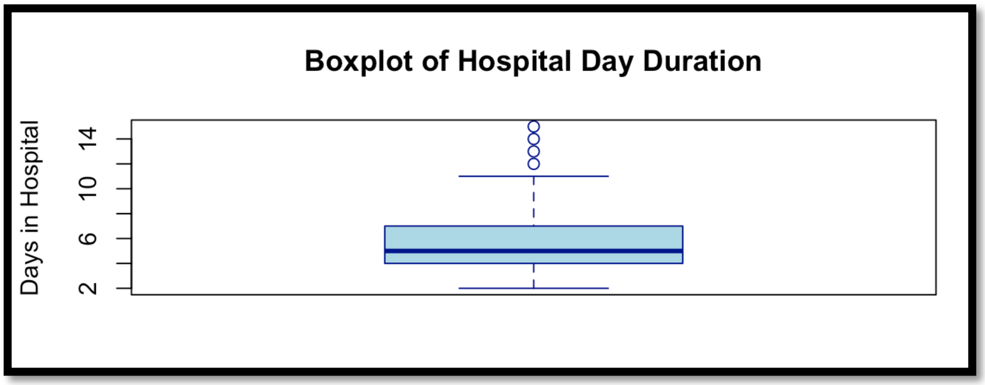


Fig 6. Boxplot of Hospital Day Duration

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")
```