SLIIT UNI
THE KNOWLEDGE UNIVERSITY

Faculty of Computing

Year 2 Semester 1 (2025)

IT2120 - Probability and Statistics — Lab Sheet 04

Lab Exercise 4 (Descriptive Statistics)

Major League Baseball is known as "America's pastime." The role of Major League Baseball has been ingrained into American culture. The heroic figures and memorable moments of Major League Baseball reflect the type of attitude that American culture is built on. Given below are some measurements observed in this significant sport during the 1998 league.

X1 = Team Attendance
(Average number of spectators for a match that the team play)

X2 = Team Salary
(Earning of the team)

X3 = Years
(Years since the team has owned a stadium)

Before starting the lab sheet, you need to create a folder in your desktop and save all your working inside the folder. Set the working directory to that folder using the following command:

```
setwd("paste the path of the folder")
```

Eg:- `setwd("D:\\2025 - Sem 2\\IT2120\\Lab Sessions\\Lab 04")`

1. Identify the variables and enter the given data set into R.

```
##Part 1
##Setting Directory
setwd("D:\\2025 - Sem 2\\IT2120 - New\\Lab Sessions\\Lab 04")

##Importing the data set
data<-read.table("DATA 4.txt",header=TRUE,sep = " ")

##view the file in a separate window
fix(data)

##Attach the file into R. So, you can call the variables by their names.
attach(data)
```

2. Obtain the following for each variable

   a. Box-Plot, Histogram and Stem-Leaf Plot.

```
##Part 2
##Part (a)
##Obtaining Box Plots
boxplot(X1,main="Box plot for Team Attendence",outline=TRUE,outpch=8,horizontal=TRUE)
boxplot(X2,main="Box plot for Team Salary",outline=TRUE,outpch=8,horizontal=TRUE)
boxplot(X3,main="Box plot for Years",outline=TRUE,outpch=8,horizontal=TRUE)

##Obtaining Histogram
hist(X1,ylab="Frequency",xlab="Team Attendence",main="Histogram for Team Attendence")
hist(X2,ylab="Frequency",xlab="Team Salary",main="Histogram for Team Salary")
hist(X3,ylab="Frequency",xlab="Years",main="Histogram for Years")

#Stem & Leaf Plot
stem(X1)
stem(X2)
stem(X3)
```

   b. Mean, Median and Standard Deviation.

```
##Part (b)
##Mean
mean(X1)
mean(X2)
mean(X3)

##Median
median(X1)
median(X2)
median(X3)

##Standard Deviation
sd(X1)
sd(X2)
sd(X3)
```

c. First and Third Quartile.

```
##Part (c)
##Getting five number summary along with mean value
summary(X1)
summary(X2)
summary(X3)

##Getting only five number summary for X1 variable
quantile(X1)

##Calling first Quartile of X1 using index value
quantile(X1)[2]

##Calling third Quartile of X1 using index value
quantile(X1)[4]
```

d. Interquartile Range.

```
##Part (d)
##Obtaining Inter Quartile Range (IQR) of each variable
IQR(X1)
IQR(X2)
IQR(X3)
```

3. Write a function to find the modes of a given set of values. Check the function by finding the mode of the variable "Years".

```
##Part 3
##Function to get the mode of a data set
get.mode<-function(y){
  counts<-table(X3)
  names(counts[counts == max(counts)])
}

##Obtaining the mode of a variable using the function defined above
get.mode(X3)

###Explanation on how each command inside the function works
##Following command is to get the frequency table for the variable
table(X3)

##Following command will give the maximum frequency in the frequency table
max(counts)

##Following command will check whether frequencies in the frequency table equals
##to the maximum frequency obtained
counts == max(counts)

##This extracts both value and the frequency which gives "TRUE" in earlier logical function
counts[counts == max(counts)]

##This extracts the value which gives maximum frequency (mode) in earlier logical function
names(counts[counts == max(counts)])
```

4. Write a function that would produce the outliers when the values are given. Check the function with the 3 variables in the dataset.

```r
##Part 4
##Function to check the existence of outliers of a data set
get.outliers<-function(z){
  q1 <- quantile(z)[2]
  q3 <- quantile(z)[4]
  iqr <- q3 - q1

  ub <- q3 + 1.5*iqr
  lb <- q1 - 1.5*iqr

  print(paste("Upper Bound = ", ub))
  print(paste("Lower Bound = ", lb))
  print(paste("Outliers:", paste(sort(z[z<lb | z>ub]), collapse = ",")))
}

##Checking the outliers of a variable using the function defined above
get.outliers(X1)
get.outliers(X2)
get.outliers(X3)
```

```r
###Explanation on how each command inside the function works
##Following command is to calculate the interval for outliers
get.outliers<-function(z){
  q1 <- quantile(z)[2]
  q3 <- quantile(z)[4]
  iqr <- q3 - q1

  ub <- q3 + 1.5*iqr
  lb <- q1 - 1.5*iqr

##Following command is to display the upper boundry and lower boundry of the interval
  print(paste("Upper Bound = ", ub))
  print(paste("Lower Bound = ", lb))

##Checking the existence of outliers and Display the outliers if exists
  print(paste("Outliers:", paste(sort(z[z<lb | z>ub]), collapse = ",")))
```

DataSet

Team	Team Attendance (X_1)	Team Salary (X_2)	Years (X_3)
Atlanta Braves	3.361	59.536	3
New York Mets	2.288	49.518	35
Philadelphia Phillies	1.716	34.370	28
Florida Marlins	0.914	9.162	23
Houston Astros	1.750	33.434	12
Chicago Cubs	2.450	40.629	34
St. Louis Cardinals	2.623	49.433	85
Cincinnati Reds	3.195	52.575	33
Milwaukee Brewers	1.794	21.995	29
Pittsburgh Pirates	1.812	32.393	46
San Diego Padres	1.561	13.352	29
San Francisco Giants	2.556	45.368	32
Colorado Rockies	1.926	40.571	39
Arizona Diamondbacks	3.089	47.970	37
New York Yankees	3.789	47.435	4
Boston Red Sox	3.603	30.572	1
Toronto Blue jays	2.950	63.461	76
Baltimore Orioles	2.344	51.647	87
Tampa Bay Devil	2.454	48.666	10
Cleveland Indians	3.685	68.988	7
Chicago White Sox	2.506	25.318	9
Kansas City Royals	3.467	59.584	5
Minnesota Twins	1.391	36.840	8
Detroit Tigers	1.495	32.963	26
Texas Rangers	1.166	26.183	17
Anaheim Angels	1.409	22.725	87
Seattle Mariners	2.927	55.305	5
Oakland Athletics	2.519	38.702	33
Montreal Expos	2.644	52.027	23
Los Angeles Dodgers	1.232	20.063	33

Exercise

Instructions: Create a folder in your desktop with your registration number (Eg: "IT....."). You need to save the R script file and take screenshots of the command prompt with answers and save it in a word document inside the folder. Save both R script file and word document with your registration number (Eg: "IT....."). After you finish the exercise, zip the folder and upload the zip file to the submission link.

1. Import the dataset ('Exercise.txt') into R and store it in a data frame called "branch_data".
2. Identify the variable type and scale of measurement for each variable.
3. Obtain boxplot for sales and interpret the shape of the sales distribution.
4. Calculate the five number summary and IQR for advertising variable.
5. Write an R function to find the outliers in a numeric vector and check for outliers in years variables.