1. This question is worth 50 marks.

The dataset tips is located in the reshape2 package in R. It contains information on the tips received by a waiter while working in a restaurant for several months. Each row/record in the dataset corresponds to a different table of customers.

a. Install the reshape2 package, load the library, and access the dataset tips. Include the commands needed to look at both the structure of the dataset and its help file. Summarise the dataset. [8 marks]

install.packages('reshape2')

library(reshape2)

data("tips")

?tips

str(tips)

'data.frame': 244 obs. of 7 variables:

$ total\_bill: num 17 10.3 21 23.7 24.6 ...

$ tip : num 1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...

$ sex : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...

$ smoker : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...

$ day : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...

$ time : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...

$ size : int 2 3 3 2 4 4 2 4 2 2 ...

The dataset contains records on 244 tips, with 7 variables. They show the amount of the tip (dollars) received, the amount of the bill (dollars), the bill payer’s sex, whether there were smokers in the table, the day of the week, time of the day and size of the table.

b. The smoke variable records if there were any smokers at the table. How many tables had smokers present? [4 marks]

table(tips$smoker)

No Yes

151 93

We can see that there were 93 tables with smokers present.

c. What is the size of the largest group of diners? How many groups of this size dined at the restaurant? What was the largest bill amongst these groups? [8 marks]

tips[ which.max(tips$size) , ]

total\_bill tip sex smoker day time size

157 48.17 5 Male No Sun Dinner 6

tips[ which.max(tips$size) , ]$size

[1] 6

size6 <- tips[ tips$size == tips[ which.max(tips$size) , ]$size, ]

nrow(size6)

[1] 4

size6[ which.max(size6$total\_bill) , ]

total\_bill tip sex smoker day time size

157 48.17 5 Male No Sun Dinner 6

size6[ which.max(size6$total\_bill) , ]$total\_bill

[1] 48.17

The size of the largest group of diners is 6 people. There are 4 groups of this size and the largest bill amongst these groups is $48.17

d. It’s difficult to analyse the amount left as a tip, without taking the size of the corresponding bill into consideration. In order to do this, form a new column called percentage.tip which contains the percentage of the bill which the tip constitutes (e.g., if a bill is €50 and the tip is €5, then this new column would record that the percentage.tip is 10 - i.e., 10%). What is the average percentage tip? [6 marks]

tips$percentage.tip <- (tips$tip\*100)/tips$total\_bill

head(tips)

total\_bill tip sex smoker day time size percentage.tip

1 16.99 1.01 Female No Sun Dinner 2 5.944673

2 10.34 1.66 Male No Sun Dinner 3 16.054159

3 21.01 3.50 Male No Sun Dinner 3 16.658734

4 23.68 3.31 Male No Sun Dinner 2 13.978041

5 24.59 3.61 Female No Sun Dinner 4 14.680765

6 25.29 4.71 Male No Sun Dinner 4 18.623962

mean(tips$percentage.tip)

[1] 16.08026

The average percentage tip is 16.08%

e. Which sex/day combination left the smallest mean percentage.tip? [8 marks]

df <- aggregate(tips[ , 8], list(tips$sex, tips$day), mean)

Group.1 Group.2 x

1 Female Fri 19.93884

2 Male Fri 14.33852

3 Female Sat 15.64702

4 Male Sat 15.15768

5 Female Sun 18.15688

6 Male Sun 16.23441

7 Female Thur 15.75248

8 Male Thur 16.52765

df[which.min(df$x), ]

Group.1 Group.2 x

2 Male Fri 14.33852

The smallest mean percentage.tip was left by a male on a Friday.

f. Create a new column called rating which converts percentage.tip to an ordered factor using the cut() function. Use bins of 0 - 10% (“Normal”), 10 - 20% (“Generous”), and 20 - 50% (“Very generous”). [8 marks]

tips$rating <- cut(tips$percentage.tip,

breaks = c(0, 10, 20, 50),

labels = c("Normal", "Generous", "Very generous"),

ordered\_result = TRUE)

head(tips)

total\_bill tip sex smoker day time size percentage.tip rating

1 16.99 1.01 Female No Sun Dinner 2 5.944673 Normal

2 10.34 1.66 Male No Sun Dinner 3 16.054159 Generous

3 21.01 3.50 Male No Sun Dinner 3 16.658734 Generous

4 23.68 3.31 Male No Sun Dinner 2 13.978041 Generous

5 24.59 3.61 Female No Sun Dinner 4 14.680765 Generous

6 25.29 4.71 Male No Sun Dinner 4 18.623962 Generous

g. The time variable records whether a table of diners sat at Dinner or Lunch. Make a two-way table of time vs. rating. How many tables at lunch are considered generous? [8 marks]

table(tips$time, tips$rating)

Normal Generous Very generous

Dinner 23 126 26

Lunch 4 52 12

We can see that 52 tables are considered generous at lunch time.

2. This question is worth 25 marks.

Load the dublin.Rdata object, which you will find in the Exam folder on Moodle. It includes a dataframe with daily measurements of the following variables: mean temperature (meantp); maximum temperature (maxtp); minimum temperature (mintp); precipitation (rain); mean windspeed (wdsp); and sunshine duration (sun). Temperature is measured in degrees Celsius; precipitation in millimetres (mm); mean windspeed in knots; and sunshine duration is measured in hours.

a. A data analyst wants to study the daily mean windspeed (wdsp) variable. Produce a histogram of this variable, and make the graph look neat and presentable (paying particular attention to labels, colour, titles etc.).

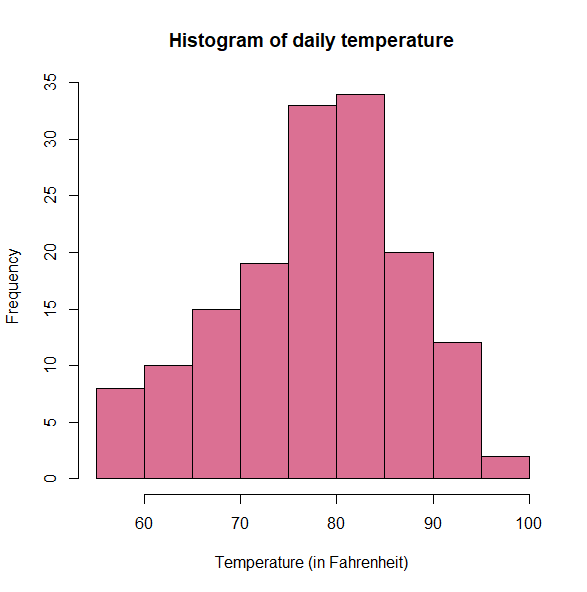
Comment on the resulting histogram. [10 marks]

hist(airquality$Temp, breaks = 10,

xlab = "Temperature (in Fahrenheit)",

main = "Histogram of daily temperature",

col = "palevioletred")



b. The data analyst is also interested in learning more about the relationship between the mean temperature (meantp) variable and the month variable. Produce a plot showing boxplots of the mean temperature variable grouped by month.

In addition to producing this plot, you should:

■ colour the boxplots

■ include x- and y-axis labels and a main title

■ make the plot look neat (e.g., change the numbers on the y-axis to appear horizontal rather than vertical)

Comment on the resulting plot. [15 marks]

boxplot(airquality$Temp ~ airquality$Month,

xlab = "Month",

xaxt = "n",

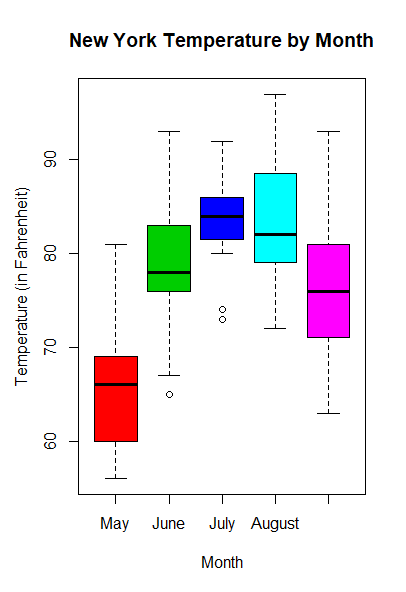
col = 2:6,

main = "New York Temperature by Month", ylab = "")

axis(1, at = 1:5,

labels = c("May", "June", "July", "August", "September"))

mtext("Temperature (in Fahrenheit)", side = 2, line = 2.5, las = 0)



3. This question is worth 25 marks.

This questions uses the dublin.RData object from Question 2. Reload this object. Using the class() function, assign the class weather to the list dublin.

a. Write a summary method for an object of class weather which includes the following information in its output:

■ The beginning and end years in the dataset

■ The minimum of the mintp variable

■ The maximum of the maxtp variable

■ One other piece of information from the dataset

Your summary method should be neat and clear and easy to read.

Test your summary method on the object dublin. [17 marks]

air <- airquality

class(air)

[1] "data.frame"

class(air) <- "weather"

class(air)

[1] "weather"

summary\_weather <- function( weather\_list ){

# beginning and end years in the dataset

beginning <- min(weather\_list$Month)

end <- max(weather\_list$Month)

print( paste('beginning month', beginning, 'end month', end))

# The minimum of the mintp variable

mintp <- min(weather\_list$Temp)

print(paste('Min temperature in the dateset', mintp))

# The maximum of the maxtp variable

maxtp <- max(weather\_list$Temp)

print(paste('Max temperature in the dataset', maxtp))

# One other piece of information from the dataset

mean\_wind <- mean(weather\_list$Wind)

print(paste('Average wind speed in the dataset', mean\_wind))

}

summary\_weather(air)

[1] "beginning month 5 end month 9"

[1] "Min temperature in the dateset 56"

[1] "Max temperature in the dataset 97"

[1] "Average wind speed in the dataset 9.95751633986928"

b. Explain in your own words what the following object-oriented programming (OOP) terms mean in the R language:

■ Polymorphism

■ Inheritance [8 marks]

End of Exam