Chapter 2

library(datasets)  
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --  
## v dplyr 1.1.2 v readr 2.1.4  
## v forcats 1.0.0 v stringr 1.5.0  
## v ggplot2 3.4.2 v tibble 3.2.1  
## v lubridate 1.9.2 v tidyr 1.3.0  
## v purrr 1.0.1   
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

head(airquality)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6

tail(airquality)

## Ozone Solar.R Wind Temp Month Day  
## 148 14 20 16.6 63 9 25  
## 149 30 193 6.9 70 9 26  
## 150 NA 145 13.2 77 9 27  
## 151 14 191 14.3 75 9 28  
## 152 18 131 8.0 76 9 29  
## 153 20 223 11.5 68 9 30

mean(airquality$Ozone, na.rm =TRUE) #remove missing value

## [1] 42.12931

## Subset

Month.Ozone = subset(airquality, select = c(Month,Day,Ozone)) #create subset of selected attribute  
Month.Ozone

## Month Day Ozone  
## 1 5 1 41  
## 2 5 2 36  
## 3 5 3 12  
## 4 5 4 18  
## 5 5 5 NA  
## 6 5 6 28  
## 7 5 7 23  
## 8 5 8 19  
## 9 5 9 8  
## 10 5 10 NA  
## 11 5 11 7  
## 12 5 12 16  
## 13 5 13 11  
## 14 5 14 14  
## 15 5 15 18  
## 16 5 16 14  
## 17 5 17 34  
## 18 5 18 6  
## 19 5 19 30  
## 20 5 20 11  
## 21 5 21 1  
## 22 5 22 11  
## 23 5 23 4  
## 24 5 24 32  
## 25 5 25 NA  
## 26 5 26 NA  
## 27 5 27 NA  
## 28 5 28 23  
## 29 5 29 45  
## 30 5 30 115  
## 31 5 31 37  
## 32 6 1 NA  
## 33 6 2 NA  
## 34 6 3 NA  
## 35 6 4 NA  
## 36 6 5 NA  
## 37 6 6 NA  
## 38 6 7 29  
## 39 6 8 NA  
## 40 6 9 71  
## 41 6 10 39  
## 42 6 11 NA  
## 43 6 12 NA  
## 44 6 13 23  
## 45 6 14 NA  
## 46 6 15 NA  
## 47 6 16 21  
## 48 6 17 37  
## 49 6 18 20  
## 50 6 19 12  
## 51 6 20 13  
## 52 6 21 NA  
## 53 6 22 NA  
## 54 6 23 NA  
## 55 6 24 NA  
## 56 6 25 NA  
## 57 6 26 NA  
## 58 6 27 NA  
## 59 6 28 NA  
## 60 6 29 NA  
## 61 6 30 NA  
## 62 7 1 135  
## 63 7 2 49  
## 64 7 3 32  
## 65 7 4 NA  
## 66 7 5 64  
## 67 7 6 40  
## 68 7 7 77  
## 69 7 8 97  
## 70 7 9 97  
## 71 7 10 85  
## 72 7 11 NA  
## 73 7 12 10  
## 74 7 13 27  
## 75 7 14 NA  
## 76 7 15 7  
## 77 7 16 48  
## 78 7 17 35  
## 79 7 18 61  
## 80 7 19 79  
## 81 7 20 63  
## 82 7 21 16  
## 83 7 22 NA  
## 84 7 23 NA  
## 85 7 24 80  
## 86 7 25 108  
## 87 7 26 20  
## 88 7 27 52  
## 89 7 28 82  
## 90 7 29 50  
## 91 7 30 64  
## 92 7 31 59  
## 93 8 1 39  
## 94 8 2 9  
## 95 8 3 16  
## 96 8 4 78  
## 97 8 5 35  
## 98 8 6 66  
## 99 8 7 122  
## 100 8 8 89  
## 101 8 9 110  
## 102 8 10 NA  
## 103 8 11 NA  
## 104 8 12 44  
## 105 8 13 28  
## 106 8 14 65  
## 107 8 15 NA  
## 108 8 16 22  
## 109 8 17 59  
## 110 8 18 23  
## 111 8 19 31  
## 112 8 20 44  
## 113 8 21 21  
## 114 8 22 9  
## 115 8 23 NA  
## 116 8 24 45  
## 117 8 25 168  
## 118 8 26 73  
## 119 8 27 NA  
## 120 8 28 76  
## 121 8 29 118  
## 122 8 30 84  
## 123 8 31 85  
## 124 9 1 96  
## 125 9 2 78  
## 126 9 3 73  
## 127 9 4 91  
## 128 9 5 47  
## 129 9 6 32  
## 130 9 7 20  
## 131 9 8 23  
## 132 9 9 21  
## 133 9 10 24  
## 134 9 11 44  
## 135 9 12 21  
## 136 9 13 28  
## 137 9 14 9  
## 138 9 15 13  
## 139 9 16 46  
## 140 9 17 18  
## 141 9 18 13  
## 142 9 19 24  
## 143 9 20 16  
## 144 9 21 13  
## 145 9 22 23  
## 146 9 23 36  
## 147 9 24 7  
## 148 9 25 14  
## 149 9 26 30  
## 150 9 27 NA  
## 151 9 28 14  
## 152 9 29 18  
## 153 9 30 20

August.Ozone = subset(airquality, Month == 8, select = c(Month,Day,Ozone))  
August.Ozone

## Month Day Ozone  
## 93 8 1 39  
## 94 8 2 9  
## 95 8 3 16  
## 96 8 4 78  
## 97 8 5 35  
## 98 8 6 66  
## 99 8 7 122  
## 100 8 8 89  
## 101 8 9 110  
## 102 8 10 NA  
## 103 8 11 NA  
## 104 8 12 44  
## 105 8 13 28  
## 106 8 14 65  
## 107 8 15 NA  
## 108 8 16 22  
## 109 8 17 59  
## 110 8 18 23  
## 111 8 19 31  
## 112 8 20 44  
## 113 8 21 21  
## 114 8 22 9  
## 115 8 23 NA  
## 116 8 24 45  
## 117 8 25 168  
## 118 8 26 73  
## 119 8 27 NA  
## 120 8 28 76  
## 121 8 29 118  
## 122 8 30 84  
## 123 8 31 85

analysis = lm(Temp~Month, data = airquality)  
summary(analysis)

##   
## Call:  
## lm(formula = Temp ~ Month, data = airquality)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.5263 -6.2752 0.9121 6.2865 17.9121   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 58.2112 3.5191 16.541 < 2e-16 \*\*\*  
## Month 2.8128 0.4933 5.703 6.03e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.614 on 151 degrees of freedom  
## Multiple R-squared: 0.1772, Adjusted R-squared: 0.1717   
## F-statistic: 32.52 on 1 and 151 DF, p-value: 6.026e-08

namiss = drop\_na(airquality) #data frame with removed missing value  
head(namiss)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 23 299 8.6 65 5 7  
## 6 19 99 13.8 59 5 8

set1 = read.csv("set1.csv", header =T, na.strings = "?")  
set1

## X X1990 X1991 X1992 X1993 X1994 X1995  
## 1 Xsar 1 2 3 4 5 6  
## 2 Tesla 5 6 7 8 9 10  
## 3 Motor Inc. 9 10 11 12 9 10  
## 4 Time Warner 13 14 15 16 9 10

## tidyverse

set1.yum = rownames\_to\_column(set1, var="industry") #assign column name to row number

set1.yum

## industry X X1990 X1991 X1992 X1993 X1994 X1995  
## 1 1 Xsar 1 2 3 4 5 6  
## 2 2 Tesla 5 6 7 8 9 10  
## 3 3 Motor Inc. 9 10 11 12 9 10  
## 4 4 Time Warner 13 14 15 16 9 10

longset1 = gather(set1.yum,Year,Millions\_Dollars,3:8) #rearrange data frame so it reduce dimension  
longset1

## industry X Year Millions\_Dollars  
## 1 1 Xsar X1990 1  
## 2 2 Tesla X1990 5  
## 3 3 Motor Inc. X1990 9  
## 4 4 Time Warner X1990 13  
## 5 1 Xsar X1991 2  
## 6 2 Tesla X1991 6  
## 7 3 Motor Inc. X1991 10  
## 8 4 Time Warner X1991 14  
## 9 1 Xsar X1992 3  
## 10 2 Tesla X1992 7  
## 11 3 Motor Inc. X1992 11  
## 12 4 Time Warner X1992 15  
## 13 1 Xsar X1993 4  
## 14 2 Tesla X1993 8  
## 15 3 Motor Inc. X1993 12  
## 16 4 Time Warner X1993 16  
## 17 1 Xsar X1994 5  
## 18 2 Tesla X1994 9  
## 19 3 Motor Inc. X1994 9  
## 20 4 Time Warner X1994 9  
## 21 1 Xsar X1995 6  
## 22 2 Tesla X1995 10  
## 23 3 Motor Inc. X1995 10  
## 24 4 Time Warner X1995 10

spread(longset1, Year, Millions\_Dollars) #revert gather

## industry X X1990 X1991 X1992 X1993 X1994 X1995  
## 1 1 Xsar 1 2 3 4 5 6  
## 2 2 Tesla 5 6 7 8 9 10  
## 3 3 Motor Inc. 9 10 11 12 9 10  
## 4 4 Time Warner 13 14 15 16 9 10

filter(longset1, X == "Tesla") #filter data frame

## industry X Year Millions\_Dollars  
## 1 2 Tesla X1990 5  
## 2 2 Tesla X1991 6  
## 3 2 Tesla X1992 7  
## 4 2 Tesla X1993 8  
## 5 2 Tesla X1994 9  
## 6 2 Tesla X1995 10