DATA ANALYTICS WITH R, EXCEL AND TABLEU

Assignment 2 (Session 6-10)

Name – Abhilash Singh

Task 1

1. Import the Titanic Dataset from the link **=> Titanic Data Set**.
Perform the following:
a. Is there any difference in fares by a different class of tickets?
Note - Show a boxplot displaying the distribution of fares by class.

Answer:
Yes, the Fare is highest for first class, moderate for second class and lowest for the third class.

```
boxplot(fare~pclass,data= titanic_dataset, main="Fare vs Pclass",xlab="Class",ylab="Fare",col=topo.colors(3))
```

b. Is there any association with Passenger class and gender?
Note – Show a stacked bar chart

Answer:
No, there is no association with Pclass and Gender.

```
library(ggplot2)
ggplot(titanic_dataset, aes(x = pclass, fill = factor(sex))) + geom_bar(stat='Count', position='stack') + labs(x = 'Pclass')
```

Task 2:

1. Create a box and whisker plot by class using mtcars dataset.
Answer:

Task 3:

1. A recent national study showed that approximately 44.7% of college students have used Wikipedia as a source in at least one of their term papers. Let X equal the number of students in a random sample of size n = 31 who have used Wikipedia as a source.
Perform the below functions
a. Find the probability that X is equal to 17

```
> dbinom(17, size = 31, prob = 0.447)
[1] 0.07532248
```

b. Find the probability that X is at most 13

```
> pbinom(13, size = 31, prob = 0.447)
[1] 0.451357
```

c. Find the probability that X is bigger than 11.

```
> pbinom(11, size = 31, prob = 0.447, lower.tail = FALSE)
[1] 0.8020339
```

d. Find the probability that X is at least 15.

```
> pbinom(14, size = 31, prob = 0.447, lower.tail = FALSE)
[1] 0.406024
```

e. Find the probability that X is between 16 and 19, inclusive

```
> sum(dbinom(16:19, size = 31, prob = θ.447))
[1] 0.2544758
> diff(pbinom(c(19, 15), size = 31, prob = θ.447, lower.tail = FALSE))
[1] 0.2544758
```

Task 4:

1. If Z is norm (mean = 0, sd = 1)
Find P(Z > 2.64)
Answer:

```
#We need to take the whole of the right hand side (area 0.5)
#and subtract the area from z = 0 to z = 2.64, which we get from the z-table.
#the probability value of z =2.64 in table is 0.4959
#so P(Z > 2.64)=0.5-P( 0 < z < 2.64)=0.5-0.4959=0.0041


#or we can do like this
1 - pnorm(2.64, mean=0, sd=1)
#0.004145301
```

Find P(|Z| > 1.39)
Answer:

```
#we can find by pnorm function too
pnorm(1.39)
#0.9177356
pnorm(-1.39)
#0.08226444
#1-(pnorm(1.39)-pnorm(-1.39))
#1-(0.9177356-0.08226444)
#1-0.8354712
#0.1645288
```

2. Suppose p = the proportion of students who are admitted to the graduate school of the University of California at Berkeley, and suppose that a public relation officer boasts that UCB has historically had a 40% acceptance rate for its graduate school. Consider the data stored in the table UCB Admissions from 1973. Assuming these observations constituted a simple random sample, are they consistent with the officers claim, or do they provide evidence that the acceptance rate was significantly less than 40%? Use an Î± = 0.01 significance level.
Answer:

```
#to check for wheather there is consistency with the officers claim or do they provide evidence
#that the acceptance rate was significantly less than 40%
#thus defining the null hypo as Ho:p is equal to 0.40
#and Ha:p less than 0.40
#Ho :  p = 0.4
#Ha :  p < 0.4
#alpha = 0.01
```

```
#Thus to find we use qnorm() function


-qnorm(0.99)
#-2.326348
#Now to find out our test statistic
newucb_data <- as.data.frame(UCBAdmissions)
View(newucb_data)
dim(newucb_data)
summary(newucb_data$Admit)
phat <- 12/(24)
t <- (phat-0.4)/sqrt(0.4*0.6/(24))
t
#1
#by calculations it is clear that our test statistic is not less than -2.326348
#So we accept our null hypothesis Ho
#hence we say that the observed data are consistent with the officer's claim at alpha = 0.01(Level of Significance)
```

Task 5:

Import dataset from the following link: **AirQuality Data Set**
Perform the following written operations:
1. Read the file in Zip format and get it into R.
Answer:

1. mydata<-read_csv("AirqualityUCI.zip") library(readr) AirQualityUCI <- read_delim("AirQualityUCI.zip", ";", escape_double = FALSE, trim_ws = TRUE) View(AirQualityUCI)

Multiple files in zip: reading 'AirQualityUCI.csv' Parsed with column specification: cols(Date;Time;CO(GT);PT08.S1(CO);NMHC(GT);C6H6(GT);PT08.S2(NMHC);NOx(GT);PT08.S3(NOx);NO2(GT);PT08.S4(NO2);PT08.S5(O3);T;RH;AH;; = col_character()) number of columns of result is not a multiple of vector length (arg 1)9357 parsing failures. row # A tibble: 5 x 5 col row col expected actual file expected actual 1 1 NA 1 columns 6 columns 'AirqualityUCI.zip' file 2 2 NA 1 columns 5 columns 'AirqualityUCI.zip' row 3 3 NA 1 columns 6 columns 'AirqualityUCI.zip' col 4 4 NA 1 columns 6 columns 'AirqualityUCI.zip' expected 5 5 NA 1 columns 6 columns 'AirqualityUCI.zip' ... .
....
....
....
....
....
....
.... See problems(...) for more details.
Multiple files in zip: reading 'AirQualityUCI.csv' Missing column names filled in: 'X16' [16],

```
'X17' [17]Parsed with column specification: cols( Date = col_character(), Time =
col_character(), CO(GT) = col_character(), PT08.S1(CO) = col_integer(),NMHC(GT) =
col_integer(), C6H6(GT) = col_character(), PT08.S2(NMHC) = col_integer(), NOx(GT) =
col_integer(),PT08.S3(NOx) = col_integer(), NO2(GT) = col_integer(), PT08.S4(NO2) =
col_integer(), PT08.S5(O3) = col_integer(), T = col_number(), RH = col_number(), AH =
col_character(), X16 = col_character(), X17 = col_character() ) Other method
```

2. Create Univariate for all the columns.
Answer:

```
mydata<-read_csv("AirqualityUCI.zip") library(readr) AirQualityUCI <-
read_delim("AirQualityUCI.zip", ";", escape_double = FALSE, trim_ws = TRUE)
View(AirQualityUCI)
```

```
Multiple files in zip: reading 'AirQualityUCI.csv' Parsed with column specification:
cols(Date;Time;CO(GT);PT08.S1(CO);NMHC(GT);C6H6(GT);PT08.S2(NMHC);NOx(GT);PT08.S3(NOx);NO2(GT);P
T08.S4(NO2);PT08.S5(O3);T;RH;AH;; = col_character() ) number of columns of result is not a
multiple of vector length (arg 1)9357 parsing failures. row # A tibble: 5 x 5 col row col
expected actual file expected actual 1 1 NA 1 columns 6 columns 'AirqualityUCI.zip' file 2 2
NA 1 columns 5 columns 'AirqualityUCI.zip' row 3 3 NA 1 columns 6 columns
'AirqualityUCI.zip' col 4 4 NA 1 columns 6 columns 'AirqualityUCI.zip' expected 5 5 NA 1
columns 6 columns 'AirqualityUCI.zip' ... ................................ ...
................................................ ..........................................................................
........ ........................................................................
Multiple files in zip: reading 'AirQualityUCI.csv' Missing column names filled in: 'X16' [16],
'X17' [17]Parsed with column specification: cols( Date = col_character(), Time =
col_character(), CO(GT) = col_character(), PT08.S1(CO) = col_integer(),NMHC(GT) =
col_integer(), C6H6(GT) = col_character(), PT08.S2(NMHC) = col_integer(), NOx(GT) =
col_integer(),PT08.S3(NOx) = col_integer(), NO2(GT) = col_integer(), PT08.S4(NO2) =
col_integer(), PT08.S5(O3) = col_integer(), T = col_number(), RH = col_number(), AH =
col_character(), X16 = col_character(), X17 = col_character() ) Other method
```

3. Check for missing values in all columns.
Answer:

```
colSums(is.na(AirQualityUCI)) # Number of missing per column/variable Date Time CO(GT)
PT08.S1(CO) NMHC(GT) C6H6(GT) 114 114 114 114 114 114 PT08.S2(NMHC) NOx(GT) PT08.S3(NOx)
NO2(GT) PT08.S4(NO2) PT08.S5(O3) 114 114 114 114 114 114 T RH AH X16 X17 114 114 114 9471
9471
```

4. Impute the missing values using appropriate methods.
Answer:

```
colSums(is.na(AirQualityUCI)) # Number of missing per column/variable #filling the missing values by
NA library(plyr) AirQualityUCI[AirQualityUCI==-200.0]<-NA #Replacing the NA by mean of each
columns for(i in 1:ncol(AirQualityUCI)){ AirQualityUCI[is.na(AirQualityUCI[,i]),i] <-
mean(AirQualityUCI[,i], na.rm = TRUE)} summary(AirQualityUCI) Mode :character Mode :character
Mode :character Median :1063
Mean :1100
```

```
3rd Qu.:1231
Max. :2040
NA's :480
NMHC(GT) C6H6(GT) PT08.S2(NMHC) NOx(GT)
Min. : 7.0 Length:9471 Min. : 383.0 Min. : 2.0
1st Qu.: 67.0 Class :character 1st Qu.: 734.5 1st Qu.: 98.0
Median : 150.0 Mode :character Median : 909.0 Median : 180.0
Mean : 218.8 Mean : 939.2 Mean : 246.9
3rd Qu.: 297.0 3rd Qu.:1116.0 3rd Qu.: 326.0
Max. :1189.0 Max. :2214.0 Max. :1479.0
NA's :8557 NA's :480 NA's :1753
PT08.S3(NOx) NO2(GT) PT08.S4(NO2) PT08.S5(O3) T
Min. : 322.0 Min. : 2.0 Min. : 551 Min. : 221.0 Min. :-19.0
1st Qu.: 658.0 1st Qu.: 78.0 1st Qu.:1227 1st Qu.: 731.5 1st Qu.:118.0
Median : 806.0 Median :109.0 Median :1463 Median : 963.0 Median :178.0
Mean : 835.5 Mean :113.1 Mean :1456 Mean :1022.9 Mean :183.2
3rd Qu.: 969.5 3rd Qu.:142.0 3rd Qu.:1674 3rd Qu.:1273.5 3rd Qu.:244.0
Max. :2683.0 Max. :340.0 Max. :2775 Max. :2523.0 Max. :446.0
NA's :480 NA's :1756 NA's :480 NA's :480 NA's :480
RH AH X16 X17
Min. : 92.0 Length:9471 Length:9471 Length:9471
1st Qu.:358.0 Class :character Class :character Class :character
Median :496.0 Mode :character Mode :character Mode :character
Mean :492.3
3rd Qu.:625.0
Max. :887.0
NA's :480
```

5. Create bivariate analysis for all relationships.
Answer:

```
summary(AirQualityUCI) plot(AirQualityUCI$NOx(GT)~AirQualityUCI$PT08.S2(NMHC))
plot(AirQualityUCI$PT08.S1(CO)~AirQualityUCI$PT08.S3(NOx))
plot(AirQualityUCI$NO2(GT)~AirQualityUCI$PT08.S4(NO2))
plot(AirQualityUCI$PT08.S5(O3)~AirQualityUCI$T)
```

6. Test relevant hypothesis for valid relations.
Answer:

```
plot(AirQualityUCI$PT08.S1(CO),AirQualityUCI$T)
lm(formula=AirQualityUCI$PT08.S3(NOx)~AirQualityUCI$NOx(GT)) lm(formula =
AirQualityUCI$PT08.S1(CO)~AirQualityUCI$T) lm(formula =
AirQualityUCI$NMHC(GT)~AirQualityUCI$PT08.S2(NMHC))
plot(AirQualityUCI$PT08.S5(O3),AirQualityUCI$NOx(GT)) lm(formula
=AirQualityUCI$PT08.S5(O3)~AirQualityUCI$NOx(GT) ) pnorm(1.49) pnorm(1.097)
qnorm(0.9318879) qnorm(0.8636793)
Call: lm(formula = AirQualityUCI$PT08.S3(NOx) ~ AirQualityUCI$NOx(GT))
Coefficients: (Intercept) AirQualityUCI$NOx(GT)
1022.2737 -0.8165
Call: lm(formula = AirQualityUCI$PT08.S1(CO) ~ AirQualityUCI$T)
Coefficients: (Intercept) AirQualityUCI$T
1077.9402 0.1195
```

```
Call: lm(formula = AirQualityUCI$NMHC(GT) ~ AirQualityUCI$PT08.S2(NMHC))
Coefficients: (Intercept) AirQualityUCI$PT08.S2(NMHC)
-410.0522 0.6663
Call: lm(formula = AirQualityUCI$PT08.S5(O3) ~ AirQualityUCI$NOx(GT))
Coefficients: (Intercept) AirQualityUCI$NOx(GT)
670.796 1.548
library(car) mod=lm(AirQualityUCI$PT08.S5(O3) ~ AirQualityUCI$NOx(GT)) summary(mod)
predict(mod) Call: lm(formula = AirQualityUCI$PT08.S5(O3) ~ AirQualityUCI$NOx(GT))
Residuals: Min 1Q Median 3Q Max -978.34 -172.18 -16.95 143.35 1324.95
```

7. Create cross tabulations with derived variables.
Answer:

```
mydata<-AirQualityUCI View(mydata)
attach(mydata) mytable <- table(A,B) # A will be rows, B will be columns mytable # print table margin.table(mytable, 1) # A frequencies (summed over B) margin.table(mytable, 2) # B frequencies (summed over A) prop.table(mytable) # cell percentages prop.table(mytable, 1) # row percentages prop.table(mytable, 2) # column percentages Chi-squared approximation may be incorrect Pearson's Chi-squared test

data: mytable X-squared = 2450, df = 2401, p-value = 0.2382
```

8. Check for trends and patterns in time series.
Answer:

```
ts (AirQualityUCI, frequency = 4, start = c(1959, 2)) # frequency 4 => Quarterly Data ts (1:10, frequency = 12, start = 1990) # freq 12 => Monthly data. ts (AirQualityUCI, start=c(2009), end=c(2014), frequency=1) # Yearly Data ts (1:1000, frequency = 365, start = 1990)# freq 365 => daily data. tsAirqualityUCI <- EuStockMarkets[, 1] # ts data copied some time series data as below copie[326] 326 327 328 329 330 331 332 333 334 335 336 337 338 NAs introduced by coercionNAs introduced by coercionNAs introduced by coercionNAs introduced by coercionNAs introduced by coercion Date Time CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC) 1959 Q2 NA NA NA 1360 150 NA 1046 1959 Q3 NA NA 2 1292 112 NA 955 1959 Q4 NA NA NA 1402 88 NA 939 1960 Q1 NA NA NA 1376 80 NA 948 1960 Q2 NA NA NA 1272 51 NA 836 1960 Q3 NA NA NA 1197 38 NA 750 1960 Q4 NA NA NA 1185 31 NA 690 1961 Q1 NA NA 1 1136 31 NA 672 1961 Q2 NA NA NA 1094 24 NA 609 1961 Q3 NA NA NA 1010 19 NA 561 1961 Q4 NA NA NA 1011 14 NA 527 1962 Q1 NA NA NA 1066 8 NA 512 1962 Q2 NA NA NA 1052 16 NA 553 1962 Q3 NA NA NA 1144 29 NA 667 1962 Q4 NA NA 2 1333 64 NA 900 1963 Q1 NA NA NA 1351 87 NA 960 1963 Q2 NA NA NA 1233 77 NA 827 1963 Q3 NA NA NA 1179 43 NA 762 1963 Q4 NA NA NA 1236 61 NA 774 1964 Q1 NA NA NA 1286 63 NA 869 1964 Q2 NA NA NA 1371 164 NA 1034 1964 Q3 NA NA NA 1310 79 NA 933 1964 Q4 NA NA NA 1292 95 NA 912 1965 Q1 NA NA NA 1383 150 NA 1020 1965 Q2 NA NA NA 1581 307 NA 1319 1965 Q3 NA NA NA 1776 461 NA 1488 1965 Q4 NA NA NA 1640 401 NA 1404 1966 Q1 NA NA NA 1313 197 NA 1076 1966 Q2 NA NA NA 965 61 NA 749 1966 Q3 NA NA 1 913 26 NA 629 1966 Q4 NA NA NA 1080 55 NA 805 1967 Q1 NA
#plot time series tsAirqualityUCI <- EuStockMarkets[, 1] # ts data decomposedRes <- decompose(tsAirqualityUCI, type="mult") # use type = "additive" for additive components plot (decomposedRes)
```

9. Find out the most polluted time of the day and the name of the chemical compound.
Answer:

```
tsAirqualityUCI <- EuStockMarkets[, 1] # ts data decomposedRes <-
decompose(tsAirqualityUCI, type="mult") # use type = "additive" for additive components
plot (decomposedRes) # see plot below stlRes <- stl(tsAirqualityUCI, s.window = "periodic")
plot(AirQualityUCI$T, type = "l") 118 119 120 121 122 123 124 125 126 127 128 129 130
[131
PT08.S4(NO2) is the highest pollution at 18.00 hrs PTO*s4
132 133 134 135 136 137 138 139 140 141 142 143 [144] 144 145 146 147 148 149 150 151
152 153 154 155 156 [157] 157 158 159 160 161 162 163 164 165 166 167 168 169 [1 Date
Time
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3) 6/8/2004 8:00:00 376 525 125 2746 1708 6/9/2004 8:00:00 357 507 151 2691
2147 10/26/2004 18:00:00 952 325 180 2775 2372 max 1479.0 2682.8 339.7 2775.0 2522.8
70] 170 171 172 173 174 175 176 177 178 179 180 181 182 [183] 183 184 185 186 187 188
189 190 191 192 193 194 195 [196] 196 197 198 199 200 201 202 203 204 205 206 207 208
[209] 209 210 211 212 213 214 215 216 217 218 219 220 221 [222] 222 223 224 225 226
227 228 229 230 231 232 233 234 [235] 235 236 237 238 239 240 241 242 243 244 245 246
247 [248] 248 249 250 251 252 253 254 255 256 257 258 259 260 [261] 261 262 263 264
265 266 267 268 269 270 271 272 273 [274] 274 275 276 277 278 279 280 281 282 283 284
285 286 [287] 287 288 289 290 291 292 293 294 295 296 297 298 299 [300] 300 301 302
303 304 305 306 307 308 309 310 311 312 [313] 313 314 315 316 317 318 319 320 321 322
323 324 325 [326] 326 327 328 329 330 331 332 333 334 335 336 337 338 NAs introduced
by coercionNAs introduced by coercionNAs introduced by coercionNAs introduced by
coercionNAs introduced by coercion Date Time CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT)
PT08.S2(NMHC) 1959 Q2 NA NA NA 1360 150 NA 1046 1959 Q3 NA NA 2 1292 112 NA 955
1959 Q4 NA NA NA 1402 88 NA 939 1960 Q1 NA NA NA 1376 80 NA 948 1960 Q2 NA NA NA
1272 51 NA 836 1960 Q3 NA NA NA 1197 38 NA 750 1960 Q4 NA NA NA 1185 31 NA 690
1961 Q1 NA NA 1 1136 31 NA 672 1961 Q2 NA NA NA 1094 24 NA 609 1961 Q3 NA NA NA
1010 19 NA 561 1961 Q4 NA NA NA 1011 14 NA 527 1962 Q1 NA NA NA 1066 8 NA 512
1962 Q2 NA NA NA 1052 16 NA 553 1962 Q3 NA NA NA 1144 29 NA 667 1962 Q4 NA NA 2
1333 64 NA 900 1963 Q1 NA NA NA 1351 87 NA 960 1963 Q2 NA NA NA 1233 77 NA 827
1963 Q3 NA NA NA 1179 43 NA 762 1963 Q4 NA NA NA 1236 61 NA 774 1964 Q1 NA NA NA
1286 63 NA 869 1964 Q2 NA NA NA 1371 164 NA 1034 1964 Q3 NA NA NA 1310 79 NA 933
1964 Q4 NA NA NA 1292 95 NA 912 1965 Q1 NA NA NA 1383 150 NA 1020 1965 Q2 NA NA
NA 1581 307 NA 1319 1965 Q3 NA NA NA 1776 461 NA 1488 1965 Q4 NA NA NA 1640 401
NA 1404 1966 Q1 NA NA NA 1313 197 NA 1076 1966 Q2 NA NA NA 965 61 NA 749 1966 Q3
NA NA 1 913 26 NA 629 1966 Q4 NA NA NA 1080 55 NA 805 1967 Q1 NA
Date Time CO(GT) PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC) 6/8/2004 8:00:00 5.8
1377 -200 36.1 1688 6/9/2004 8:00:00 6.4 1496 -200 36.9 1705 10/26/2004 18:00:00 9.5
1908 -200 52.1 2007 max 11.9 2039.8 1189.0 63.7 2214.0 Date Time
NOx(GT)
PT08.S3(NOx)
NO2(GT)
```

PT08.S4(NO2)
PT08.S5(O3) 6/8/2004 8:00:00 376 525 125 2746 1708 6/9/2004 8:00:00 357 507 151 2691
2147 10/26/2004 18:00:00 952 325 180 2775 2372 max 1479.0 2682.8 339.7 2775.0 2522.8
[989] 989 990 991 992 993 994 995 996 997 998 999 1000