SAVEETHA SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

ITA 0443 - STATISTICS WITH R PROGRAMMING FOR REAL TIME PROBLEM

DAY 4 – LAB ASSESSMENT

Reg No:192124086

Name:L.Uthra

1.Randomly Sample the iris dataset such as 80% data for training and 20% for test and create Logistics regression with train data, use species as target and petals width and

length as feature variables , Predict the probability of the model using test data, Create Confusion matrix for above test model

Program:

2. (i)Write suitable R code to compute the mean, median ,mode of the following values

c(90, 50, 70, 80, 70, 60, 20, 30, 80, 90, 20)

Program:

```
values<-c(90,50,70,80,70,60,20,30,80,90,20)

mean(values)

median(values)

mode_num<-names(which.max(table(values)))

mode_num
```

Output:

```
> mean(values)

[1] 60

> median(values)

[1] 70
```

```
> mode_num<-names(which.max(table(values)))

> mode_num

[1] "20"
```

(ii) Write R code to find 2nd highest and 3rd Lowest value of above problem.

Program:

```
values<-c(90,50,70,80,70,60,20,30,80,90,20)

sort(unique(values), decreasing=TRUE)[2]

sort(unique(values))[3]
```

Output:

```
> sort(unique(values), decreasing=TRUE)[2]
[1] 80
> sort(unique(values))[3]
[1] 50
```

3. Explore the airquality dataset. It contains daily air quality measurements from New York during a period of five months:
• Ozone: mean ozone concentration (ppb), • Solar.R: solar radiation (Langley),
• Wind: average wind speed (mph), • Temp: maximum daily temperature in degrees Fahrenheit,
• Month: numeric month (May=5, June=6, and so on),• Day: numeric day of the month (1 -4).

i. Compute the mean temperature(don't use build in function)

Program:

```
data(airquality)
mean_temp<-sum(airquality$Temp)/nrow(airquality)
mean_temp
```

Output:

```
> mean_temp
[1] 77.88235
```

ii.Extract the first five rows from airquality.

Program:

```
data(airquality)
head(airquality,5)
```

Output:

```
> head(airquality,5)
  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

iii.Extract all columns from airquality except Temp and Wind

Program:

```
data(airquality)
airquality[,c("Ozone", "Solar.R", "Month", "Day")]
```

Output:

```
> airquality[,c("Ozone", "Solar.R", "Month", "Day")]
```

	Ozone	Solar.R	Month	Day
1	41	190	5	1
2	36	118	5	2
3	12	149	5	3
4	18	313	5	4
5	NA	NA	5	5
6	28	NA	5	6
7	23	299	5	7
8	19	99	5	8
9	8	19	5	9
10	NA	194	5	10
11	7	NA	5	11
12	16	256	5	12
13	11	290	5	13
14	14	274	5	14
15	18	65	5	15
16	14	334	5	16
17	34	307	5	17
18	6	78	5	18
19	30	322	5	19
20	11	44	5	20
21	1	8	5	21
22	11	320	5	22
23	4	25	5	23
24	32	92	5	24
25	NA	66	5	25
26	NA	266	5	26
27	NA	NA	5	27
28	23	13	5	28
29	45	252	5	29
30	115	223	5	30
31	37	279	5	31
32	NA	286	6	1
33	NA	287	6	2
34	NA	242	6	3
35	NA	186	6	4

36	NA	220	6	5
37	NA	264	6	6
38	29	127	6	7
39	NA	273	6	8
40	71	291	6	9
41	39	323	6	10
42	NA	259	6	11
43	NA	250	6	12
44	23	148	6	13
45	NA	332	6	14
46	NA	322	6	15
47	21	191	6	16
48	37	284	6	17
49	20	37	6	18
50	12	120	6	19
51	13	137	6	20
52	NA	150	6	21
53	NA	59	6	22
54	NA	91	6	23
55	NA	250	6	24
56	NA	135	6	25
57	NA	127	6	26
58	NA	47	6	27
59	NA	98	6	28
60	NA	31	6	29
61	NA	138	6	30
62	135	269	7	1
63	49	248	7	2
64	32	236	7	3
65	NA	101	7	4
66	64	175	7	5
67	40	314	7	6
68	77	276	7	7
69	97	267	7	8
70	97	272	7	9
71	85	175	7	10
72	NA	139	7	11
73	10	264	7	12
74	27	175	7	13
75	NA	291	7	14
76	7	48	7	15
77	48	260	7	16
78	35	274	7	17
79	61	285	7	18
80	79	187	7	19
81	63	220	7	20
82	16	7	7	21
83	NA	258	7	22

84	NA	295	7	23
85	80	294	7	24
86	108	223	7	25
87	20	81	7	26
88	52	82	7	27
89	82	213	7	28
90	50	275	7	29
91	64	253	7	30
92	59	254	7	31
93	39	83	8	1
94	9	24	8	2
95	16	77	8	3
96	78	NA	8	4
97	35	NA	8	5
98	66	NA	8	6
99	122	255	8	7
100	89	229	8	8
101	110	207	8	9
102	NA	222	8	10
103	NA	137	8	11
104	44	192	8	12
105	28	273	8	13
106	65	157	8	14
107	NA	64	8	15
108	22	71	8	16
109	59	51	8	17
110	23	115	8	18
111	31	244	8	19
112	44	190	8	20
113	21	259	8	21
114	9	36	8	22
115	NA	255	8	23
116	45	212	8	24
117	168	238	8	25
118	73	215	8	26
119	NA	153	8	27
120	76	203	8	28
121	118	225	8	29
122	84	237	8	30
123	85	188	8	31
124	96	167	9	1
125	78	197	9	2
126	73	183	9	3
127	91	189	9	4
128	47	95	9	5
129	32	92	9	6
130	20	252	9	7
131	23	220	9	8

132	21	230	9	9
133	24	259	9	10
134	44	236	9	11
135	21	259	9	12
136	28	238	9	13
137	9	24	9	14
138	13	112	9	15
139	46	237	9	16
140	18	224	9	17
141	13	27	9	18
142	24	238	9	19
143	16	201	9	20
144	13	238	9	21
145	23	14	9	22
146	36	139	9	23
147	7	49	9	24
148	14	20	9	25
149	30	193	9	26
150	NA	145	9	27
151	14	191	9	28
152	18	131	9	29
153	20	223	9	30

iv.Which was the coldest day during the period?

Program:

```
data(airquality)
coldest_day<-airquality[which.min(airquality$Temp),]
coldest_day
```

Output:

	NA	NA	14.3	56	5	5

v.How many days was the wind speed greater than 17 mph?

Program:

```
data(airquality)
sum(airquality$Wind>17)
```

Output:

```
> sum(airquality$Wind>17)
[1] 3
```

4. (i)Get the Summary Statistics of air quality dataset

Program:

```
data("airquality")
summary(airquality)
```

Output:

```
Ozone           Solar.R
```

```
Min.   :  1.00   Min.   :  7.0
1st Qu.: 18.00   1st Qu.:115.8
Median : 31.50   Median :205.0
Mean   : 42.13   Mean   :185.9
3rd Qu.: 63.25   3rd Qu.:258.8
Max.   :168.00   Max.   :334.0
NA's   :37       NA's   :7
     Wind          Temp
Min.   : 1.700   Min.   :56.00
1st Qu.: 7.400   1st Qu.:72.00
Median : 9.700   Median :79.00
Mean   : 9.958   Mean   :77.88
3rd Qu.:11.500   3rd Qu.:85.00
Max.   :20.700   Max.   :97.00

    Month          Day
Min.   :5.000   Min.   : 1.0
1st Qu.:6.000   1st Qu.: 8.0
Median :7.000   Median :16.0
Mean   :6.993   Mean   :15.8
3rd Qu.:8.000   3rd Qu.:23.0
Max.   :9.000   Max.   :31.0
```

(ii)Melt airquality data set and display as a long – format data?

Program:

```
data("airquality")
library(reshape2)
airquality_melted <- melt(airquality, id = c("Month", "Day"))
airquality_melted
```

Output:

```
Month Day variable value
1         5   1    Ozone 41
2         5   2    Ozone 36
3         5   3    Ozone 12
4         5   4    Ozone 18
5         5   5    Ozone NA
6         5   6    Ozone 28
7         5   7    Ozone 23
8         5   8    Ozone 19
9         5   9    Ozone 8
10        5  10    Ozone NA
11        5  11    Ozone 7
12        5  12    Ozone 16
13        5  13    Ozone 11
14        5  14    Ozone 14
15        5  15    Ozone 18
```

16	5	16	Ozone	14
17	5	17	Ozone	34
18	5	18	Ozone	6
19	5	19	Ozone	30
20	5	20	Ozone	11
21	5	21	Ozone	1
22	5	22	Ozone	11
23	5	23	Ozone	4
24	5	24	Ozone	32
25	5	25	Ozone	NA
26	5	26	Ozone	NA
27	5	27	Ozone	NA
28	5	28	Ozone	23
29	5	29	Ozone	45
30	5	30	Ozone	115
31	5	31	Ozone	37
32	6	1	Ozone	NA
33	6	2	Ozone	NA
34	6	3	Ozone	NA
35	6	4	Ozone	NA
36	6	5	Ozone	NA
37	6	6	Ozone	NA
38	6	7	Ozone	29
39	6	8	Ozone	NA
40	6	9	Ozone	71
41	6	10	Ozone	39
42	6	11	Ozone	NA
43	6	12	Ozone	NA
44	6	13	Ozone	23
45	6	14	Ozone	NA
46	6	15	Ozone	NA
47	6	16	Ozone	21
48	6	17	Ozone	37
49	6	18	Ozone	20
50	6	19	Ozone	12
51	6	20	Ozone	13
52	6	21	Ozone	NA
53	6	22	Ozone	NA
54	6	23	Ozone	NA
55	6	24	Ozone	NA
56	6	25	Ozone	NA
57	6	26	Ozone	NA
58	6	27	Ozone	NA
59	6	28	Ozone	NA
60	6	29	Ozone	NA
61	6	30	Ozone	NA
62	7	1	Ozone	135
63	7	2	Ozone	49

64	7	3	Ozone	32
65	7	4	Ozone	NA
66	7	5	Ozone	64
67	7	6	Ozone	40
68	7	7	Ozone	77
69	7	8	Ozone	97
70	7	9	Ozone	97
71	7	10	Ozone	85
72	7	11	Ozone	NA
73	7	12	Ozone	10
74	7	13	Ozone	27
75	7	14	Ozone	NA
76	7	15	Ozone	7
77	7	16	Ozone	48
78	7	17	Ozone	35
79	7	18	Ozone	61
80	7	19	Ozone	79
81	7	20	Ozone	63
82	7	21	Ozone	16
83	7	22	Ozone	NA
84	7	23	Ozone	NA
85	7	24	Ozone	80
86	7	25	Ozone	108
87	7	26	Ozone	20
88	7	27	Ozone	52
89	7	28	Ozone	82
90	7	29	Ozone	50
91	7	30	Ozone	64
92	7	31	Ozone	59
93	8	1	Ozone	39
94	8	2	Ozone	9
95	8	3	Ozone	16
96	8	4	Ozone	78
97	8	5	Ozone	35
98	8	6	Ozone	66
99	8	7	Ozone	122
100	8	8	Ozone	89
101	8	9	Ozone	110
102	8	10	Ozone	NA
103	8	11	Ozone	NA
104	8	12	Ozone	44
105	8	13	Ozone	28
106	8	14	Ozone	65
107	8	15	Ozone	NA
108	8	16	Ozone	22
109	8	17	Ozone	59
110	8	18	Ozone	23
111	8	19	Ozone	31

112	8	20	Ozone	44
113	8	21	Ozone	21
114	8	22	Ozone	9
115	8	23	Ozone	NA
116	8	24	Ozone	45
117	8	25	Ozone	168
118	8	26	Ozone	73
119	8	27	Ozone	NA
120	8	28	Ozone	76
121	8	29	Ozone	118
122	8	30	Ozone	84
123	8	31	Ozone	85
124	9	1	Ozone	96
125	9	2	Ozone	78
126	9	3	Ozone	73
127	9	4	Ozone	91
128	9	5	Ozone	47
129	9	6	Ozone	32
130	9	7	Ozone	20
131	9	8	Ozone	23
132	9	9	Ozone	21
133	9	10	Ozone	24
134	9	11	Ozone	44
135	9	12	Ozone	21
136	9	13	Ozone	28
137	9	14	Ozone	9
138	9	15	Ozone	13
139	9	16	Ozone	46
140	9	17	Ozone	18
141	9	18	Ozone	13
142	9	19	Ozone	24
143	9	20	Ozone	16
144	9	21	Ozone	13
145	9	22	Ozone	23
146	9	23	Ozone	36
147	9	24	Ozone	7
148	9	25	Ozone	14
149	9	26	Ozone	30
150	9	27	Ozone	NA
151	9	28	Ozone	14
152	9	29	Ozone	18
153	9	30	Ozone	20
154	5	1	Solar.R	190
155	5	2	Solar.R	118
156	5	3	Solar.R	149
157	5	4	Solar.R	313
158	5	5	Solar.R	NA
159	5	6	Solar.R	NA

160	5	7	Solar.R	299
161	5	8	Solar.R	99
162	5	9	Solar.R	19
163	5	10	Solar.R	194
164	5	11	Solar.R	NA
165	5	12	Solar.R	256
166	5	13	Solar.R	290
167	5	14	Solar.R	274
168	5	15	Solar.R	65
169	5	16	Solar.R	334
170	5	17	Solar.R	307
171	5	18	Solar.R	78
172	5	19	Solar.R	322
173	5	20	Solar.R	44
174	5	21	Solar.R	8
175	5	22	Solar.R	320
176	5	23	Solar.R	25
177	5	24	Solar.R	92
178	5	25	Solar.R	66
179	5	26	Solar.R	266
180	5	27	Solar.R	NA
181	5	28	Solar.R	13
182	5	29	Solar.R	252
183	5	30	Solar.R	223
184	5	31	Solar.R	279
185	6	1	Solar.R	286
186	6	2	Solar.R	287
187	6	3	Solar.R	242
188	6	4	Solar.R	186
189	6	5	Solar.R	220
190	6	6	Solar.R	264
191	6	7	Solar.R	127
192	6	8	Solar.R	273
193	6	9	Solar.R	291
194	6	10	Solar.R	323
195	6	11	Solar.R	259
196	6	12	Solar.R	250
197	6	13	Solar.R	148
198	6	14	Solar.R	332
199	6	15	Solar.R	322
200	6	16	Solar.R	191
201	6	17	Solar.R	284
202	6	18	Solar.R	37
203	6	19	Solar.R	120
204	6	20	Solar.R	137
205	6	21	Solar.R	150
206	6	22	Solar.R	59
207	6	23	Solar.R	91

```
208     6  24  Solar.R   250
209     6  25  Solar.R   135
210     6  26  Solar.R   127
211     6  27  Solar.R  47
212     6  28  Solar.R  98
213     6  29  Solar.R  31
214     6  30  Solar.R   138
215     7   1  Solar.R   269
216     7   2  Solar.R   248
217     7   3  Solar.R   236
218     7   4  Solar.R   101
219     7   5  Solar.R   175
220     7   6  Solar.R   314
221     7   7  Solar.R   276
222     7   8  Solar.R   267
223     7   9  Solar.R   272
224     7  10  Solar.R   175
225     7  11  Solar.R   139
226     7  12  Solar.R   264
227     7  13  Solar.R   175
228     7  14  Solar.R   291
229     7  15  Solar.R  48
230     7  16  Solar.R   260
231     7  17  Solar.R   274
232     7  18  Solar.R   285
233     7  19  Solar.R   187
234     7  20  Solar.R   220
235     7  21  Solar.R  7
236     7  22  Solar.R   258
237     7  23  Solar.R   295
238     7  24  Solar.R   294
239     7  25  Solar.R   223
240     7  26  Solar.R  81
241     7  27  Solar.R  82
242     7  28  Solar.R   213
243     7  29  Solar.R   275
244     7  30  Solar.R   253
245     7  31  Solar.R   254
246     8   1  Solar.R   83
247     8   2  Solar.R   24
248     8   3  Solar.R   77
249     8   4  Solar.R   NA
250     8   5  Solar.R   NA
```

(iii)Melt airquality data and specify month and day to be “ID variables”?

Program:

```
data("airquality")
```

```
library(reshape2)
airquality_melted1 <- melt(airquality, id = c("Month", "Day"),
                    variable.name = "Measurement", value.name = "Value")
head(airquality_melted1)
```

Output:

```
Month Day Measurement Value
1         5   1    Ozone 41
2         5   2    Ozone 36
3         5   3    Ozone 12
4         5   4    Ozone 18
5         5   5    Ozone NA
6         5   6    Ozone 28
```

(iv)Cast the molten airquality data set with respect to month and date features

Program:

```
data("airquality")
library(reshape2)
airquality_cast <- dcast(airquality_melted2, Month + Day ~ Measurement,
                    mean)
head(airquality_cast)
```

(v) Use cast function appropriately and compute the average of Ozone, Solar.R , Wind and temperature per month?

Program:

```
data("airquality")
library(reshape2)
airquality_average <- dcast(airquality_melted2, Month ~ Measurement,
                           mean, fun.aggregate = mean)
head(airquality_average)
```

5.(i) Find any missing values(na) in features and drop the missing values if its less than 10% else replace that with mean of that feature.

Program:

```
data("airquality")
df[is.na(df)] <- ifelse(sum(is.na(df))/nrow(df) < 0.1, df[is.na(df)],
                    apply(df, 2, mean, na.rm = TRUE)[is.na(df)])
```

(ii) Apply a linear regression algorithm using Least Squares Method on “Ozone” and “Solar.R”

(iii)Plot Scatter plot between Ozone and Solar and add regression line created by above model

6. Load dataset named ChickWeight,

(i).Order the data frame, in ascending order by feature name “weight” grouped by feature

“diet” and Extract the last 6 records from order data frame.

(ii).a Perform melting function based on “Chick", "Time", "Diet" features as ID variables

b. Perform cast function to display the mean value of weight grouped by Diet

c. Perform cast function to display the mode of weight grouped by Diet

7. a. Create Box plot for “weight” grouped by “Diet”

Program:

```
library(ggplot2)

ggplot(data, aes(x = Diet, y = weight)) +

  geom_boxplot() +

  xlab("Diet") +

  ylab("Weight") +

  ggtitle("Box plot of Weight by Diet")
```

b. Create a Histogram for “weight” features belong to Diet- 1 category

Program:

```
data<-read.csv("ChickWeight.csv")

chick<-subset(data,Diet==1)

hist(chick$Diet)
```

Output:

c. Create Scatter plot for “ weight” vs “Time” grouped by Diet

Program:

8. a. Create multi regression model to find a weight of the chicken , by “Time” and “Diet” as as

predictor variables

b. Predict weight for Time=10 and Diet=1

c. Find the error in model for same

9 .For this exercise, use the (built-in) dataset Titanic.

a. Draw a Bar chart to show details of “Survived” on the Titanic based on passenger Class

b. Modify the above plot based on gender of people who survived

c. Draw histogram plot to show distribution of feature “Age”

10. Explore the USArrests dataset, contains the number of arrests for murder, assault, and rape for each of the 50 states in 1973. It also contains the percentage of people in the state who live in an urban area.

(i) a. Explore the summary of Data set, like number of Features and its type. Find the number of records for each feature. Print the statistical feature of data

b. Print the state which saw the largest total number of rape

c. Print the states with the max & min crime rates for murder

(ii).a. Find the correlation among the features

b. Print the states which have assault arrests more than median of the country

c. Print the states are in the bottom 25% of murder

(iii). a. Create a histogram and density plot of murder arrests by US stat

b. Create the plot that shows the relationship between murder arrest rate and proportion

of the population that is urbanised by state. Then enrich the chart by adding assault

arrest rates (by colouring the points from blue (low) to red (high)).

c. Draw a bar graph to show the murder rate for each of the 50 states .

11. a. Create a data frame based on below table.

b. Create a regression model for that data frame table to show the amount of sales(Sales) based on the how much the company spends (Spends) in advertising

c. Predict the Sales if Spend=13500

Set 2

1.(i) Write a R program to extract the five of the levels of factor created from a random sample from the LETTERS (Part of the base R distribution.)

(ii)Write R function to find the range of given vector. Range=Max-Min

Sample input, C<-(9,8,7,6,5,4,3,2,1),

output=8

(iii)Wirte the R function to find the number of vowels in given string

Sample input c<- “matrix”, output<-2

2.Load inbuild dataset “ChickWeight” in R

(i) Explore the summary of Data set, like number of Features and its type. Fins the number of records for each features

(ii)Extract last 6 records of dataset

(iii) order the data frame, in ascending order by feature name "weight" grouped by feature "diet"

(iv)Perform melting function based on "Chick","Time","Diet" features as ID variables

(v)Perform cast function to display the mean value of weight grouped by Diet

3.(i)Get the Statistical Summary of "ChickWeight" dataset

(ii)Create Box plot for "weight" grouped by "Diet"

(iii)Create a Histogram for "Weight" features belong to Diet- 1 category

(iv) Create a Histogram for "Weight" features belong to Diet- 4 category

(v) Create Scatter plot for weight vs Time grouped by Diet

4.(i) Create multi regression model to find a weight of the chicken , by "Time" and "Diet" as as predictor variables

(ii) Predict weight for Time=10 and Diet=1

(iii)Find the error in model for smae

Program:

```
chick<-read.csv("ChickWeight.csv")

View(chick)

input<-chick[,c("weight", "Time","Diet")]

model <- lm(weight~Time+Diet,data=input)

print(model)
```

```
A<- coef(model)[1]

print(A)

xtime<- coef(model)[2]

xdiet<- coef(model)[3]

y = A+xtime+xdiet

print(y)

z= A+xtime*10+xdiet*1

z
```

Output:

```
> print(A)
(Intercept)
   1.542804
> xtime<- coef(model)[2]
> xdiet<- coef(model)[3]
> y = A+xtime+xdiet
> print(y)
(Intercept)
   22.08676
> z= A+xtime*10+xdiet*1
> z
(Intercept)
   100.9748
```