Roll No. _____

(Do not write anything on question paper except Roll No.)

[This paper consists of SIX Pages]

Jagan Institute of Management Studies

End-Term Examination, September-October, 2017

Trimester IV – PGDM (IB)/PGDM (RM) 2016-18

Business Analytics with R

ET_IB_RM_BA_0310

Time: 3 Hrs. *M. Marks: 70*

INSTRUCTIONS: Attempt all questions.

Q 1 **a)** A telecom service provider (e.g. Vodafone) wants to predict which customers are going to attrite in near future. Suggest the appropriate modelling technique to solve this business problem. State the dependant variable. List down 5 different customer aspects that you will consider for independent variables. Within each customer aspect list down at least 4 independent variables (for e.g. 'Age' of the customer can be one of the independent variable within customer aspect 'Demographics'.) **7**

b) An auto-insurance service provider wants to predict the amount of claim that can be expected from a future customer. Their data team has already gathered the details of all past customers who have requested for claims. You are hired as a consultant to build a predictive model. What modeling technique will you use. What will be your dependant variable? List down 5 different customer aspects that you will consider for independent variables. Within each customer aspect list down at least 4 independent variables (for e.g. 'Age' of the customer can be one of the independent variable within customer aspect 'Demographics'). **7**

Q 2 **a)** Explain the concept of skewness. Draw the sketch of a frequency distribution and show the positions of the mean, median & mode when the distribution is – (i) Symmetric (ii) Positively Skewed (iii) Negatively Skewed. **4**

b) State your choice of measure of central tendency for a distribution with outliers. What variants of mean are robust to outliers? How do you use both mean & median in conjunction for data exploration? **2**

c) Draw a box plot for the following set of data. Remember to order the data first, if necessary.

4.7, 3.8, 3.9, 3.9, 4.6, 4.5, 5 **2**

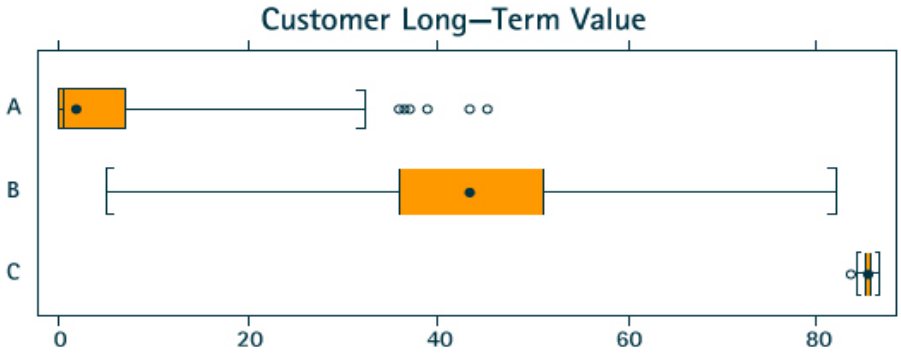
d) The case given below looks at customer long-term value and shows 3 different customer segments (A, B & C). Customer long-term value is a measure of how much the customer is worth to us in the future in thousands of Rs. These box & whisker plots show if the customers are heterogeneous or homogenous and the median that a typical customer in

this segment will spend.

i) Which segment is highly homogenous & why?

ii) Which segment is left skewed & why?

iii) Which segment is right skewed & why?



6

Q 3 The following questions tests your ability to reuse pre-built R codes to build new solutions. Attempt any **FOUR** questions.

a) You have 5 environment objects in your current R sessions – A, B, C, D, E. Write R code to perform following tasks

i) Set a new working directory using function setwd(). You can use any imaginary directory path.

ii) Save the image of your current environment in the working directory by using function save.image(). Give the file a name – my_image.RData.

iii) Save A, B, C objects in a separate file using function save(). Give the file a name – my_selected_objects.RData.

iv) Remove all the objects in the environment using rm() function

v) Load the file my_selected_objects.RData in the environment using load() function 5

b) Create a vector containing values 1 - 10 and 21 – 35. Change this vector into a 5 x 5 matrix. Extract a 3 x 3 subset from that matrix. Assign row and column names to the matrix.
(Hint: Use c() function to create vector. Use matrix() function to create a new matrix. Use [] for subsetting. Use the rownames() and colnames() function for assigning row & column names) 5

c) Consider the vectors:
V1 <- c(1, 2, 3, 4, NA)
V2 <- c(1, NA, 4, 3, 9)
Calculate new vectors W1 and W2 such that they contain only the elements that are not NA for both V1 and V2. Multiply each element of W1 by its corresponding element in W2.
(Hint: use is.na() function to identify NA's. Use ! for negation. Use [] for subsetting. Use vector calculation for multiplication) 5

d) Use data frame mtcars. Write a function myfunc() that accepts cylinder

size (cyl) and the name of any other column and then returns the mean of that column for cars having the particular cylinder size. For e.g. if we call myfunc(4, “hp”) than it should return mean of “hp” of all the cars having cylinder size 4.

Use the above function in a loop to generate mean of any particular column for all possible values of cylinder.

Here is the snapshot of mtcars data frame.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4

Sample code for a function that returns sum of 2 numbers –
sample.func <- function(a,b){ c <- a + b return(c) }
Sample code for filtering cars with speed = 40 from data frame “cars” - cars[cars$speed == 40]
Sample code for creating a loop - for(i in 1:10){print(i)}
Sample code for finding unique values in a vector –
sample.vec -> c(1,1,1,2,2,2,3,3,3,4,4,4) unique(sample.vec) 5

e) Create a function that takes as input any data frame. The function then returns the class of each column. (Hint: you can use a loop to pick up each column)

Sample code for a function that returns the class of an input vector –
sample.func <- function(a){answer.class <- class(a) return(answer.class)}

Sample code for creating a loop - for(i in 1:10){ print(i) } **5**

f) Create a table of count of observation (or rows) for data frame mtcars based on am, gear, and carb by using table() function. Generate the same table once again, but this time replace the count of observations (or rows) with sum of horse power (hp) by using xtabs () function. Write a customized function that can take a vector and return 100 times its avg value (100* avg value of vector). Use this function on values of hp broken up by combination of gear and carb by using aggregate() function.

Sample code for table() - table(mtcars[, "gear"])

Sample code for xtabs() - xtabs(hp~gear,mtcars)

Sample code for aggregate() - aggregate(hp~gear, data=mtcars, sum)

Refer to question 6 to see the snapshot of mtcars. **5**

Q 4 Attempt any **TWO** of the following questions.

a) This question focuses on your understanding of linear regression model building. Answer all of the following parts.

i) Write all the steps involved in building a linear regression model. **2**

ii) How do you perform dependent variable exploration? **1**

iii) How do you perform univariate & bivariate analysis for categorical & continuous variables among independent variables (draw illustration of output plots for all cases). **2**

iv) Explain the interpretation of all the aspects of the following output **3**

Regression Statistics	
Multiple R	0.863555563
R Square	0.74572821
Adjusted R Square	0.745697465
Standard Error	1052.537003
Observations	24815

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	80612317390	26870772463	24255.2305	0
Residual	24811	27486472897	1107834.142		
Total	24814	108098790287			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-8.64523376	9.661869424	-0.894778575	0.37091415	-27.5830737	10.29260619	-27.5830737	10.29260619
VARIABLE_A	-0.92119447	0.083693353	-11.00678174	4.1173E-28	-1.085238429	-0.75715051	-1.08523843	-0.75715051
VARIABLE_B	-0.05207561	0.006864746	-7.585947802	3.4164E-14	-0.065530917	-0.03862029	-0.06553092	-0.03862029
VARIABLE_C	0.726601414	0.003044988	238.6221105	0	0.720633056	0.732569771	0.720633056	0.732569771

v) Describe the residual metrics and their interpretation **1**

vi) Describe the checks that you will perform on residual and their interpretation. **2**

b) This question focuses on your understanding of logistic regression model building. Answer all of the following parts.

i) Write all the steps involved in building a logistic regression model. **2**

ii) How do you perform dependent variable exploration? **1**

iii) How do you perform univariate & bivariate analysis for categorical & continuous variables among independent variables (draw illustration of output plots for all cases). **2**

iv) Summarize any SIX of the following model statistics and their interpretation

- Model Chi
- Pseudo R Square
- Concordance
- Hosmer Lemshow Test
- Confusion matrix
- Model accuracy
- ROC curve
- Lift Chart **6**

c) i) Create a function that given a vector and an integer will return how many times the integer appears inside the vector. (Hint: use length() to find the length of a vector)

Sample code of a function –

```
sample.func <- function(a){answer.class <-
class(a)return(answer.class)}
```

Sample code for creating a loop -

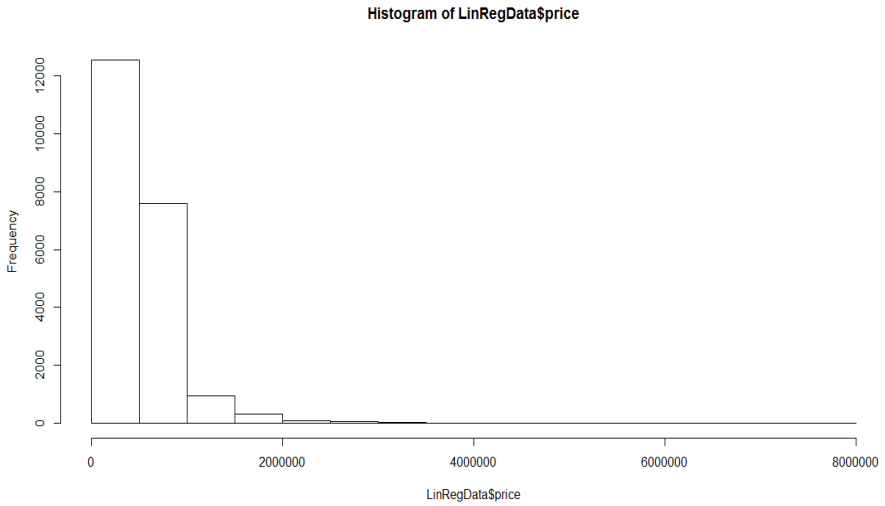
```
for(i in 1:10){print(i)}
```

Sample code for conditional statement –

```
If(a>b){print("A is greatest")}else{print("B is greatest")}
```

7

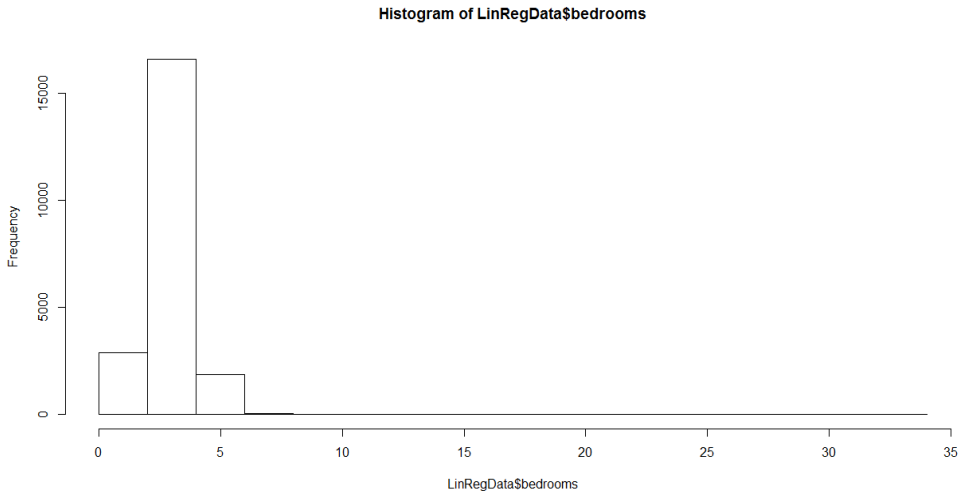
ii) Consider the following histogram of a dependent variable for linear regression. What is your interpretation? List the actions that you will take on this dependent variable.



2

iii) Consider the following histogram of an independent variable. What

is your interpretation? List the actions that you will take on this independent variable.



2

****