PGDM-RM, 2020-22
Retail Analytics, RM-406

Trimester IV, End Term Examination, September 2021

Time allowed: 2 hrs. 30 mins
Max Marks: 50

Section	Questions to attempt	Marks	Total Marks
A	either 1a **or** 1b either 2a **or** 2b either 3a **or** 3b	3*10	30
B	Compulsory	20	20
			50

Download the file from the github repository using the commands given below.

```
library(readr) # you need this library package; install it first
urlf <- "https://raw.githubusercontent.com/amarnathbose/Datafiles/master/auto-mpg.csv"
auto <- read_csv(url(urlf))
```

The data columns are as follows.

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model.year: multi-valued discrete
8. origin: multi-valued discrete
9. car.name: string (unique for each instance)

Answer the following questions using R wherever required.
You are required to submit your answers on Moodle, preferably as a pdf file.

Section A

1a. How do you find whether a regression model is good? How can you compare two regression models? What does r.squared represent in a regression model?
1b. Create a regression model to predict *mpg* using columns *horsepower*. Interpret the result.

2a. Can you use categorical variables as predictors in a regression model? If so, describe how. How would you interpret the regression model thus created?
2b. Treating number of cylinders as a categorical variable (factor), create a regression model to predict *mpg* using columns *horsepower* and the newly created categorical variable. Comment.

3a. Describe how interaction terms can be used to improve a regression model.
3b. Treating number of cylinders as a categorical variable (factor), create a regression model to predict *mpg* using the interaction of columns *horsepower* and the number of cylinders (treated as a categorical variable). Comment.

Section B

a. Create a regression model to predict *mpg* using columns 2-8,
i.e. excluding the column for *car.name*. Comment on the model.

b. Create a factor variable called *mcat* from *model.year*, with the following levels:

1. "early 70s" values 70-73
2. "late 70s" values 74-79
3. "80s" values 80-82

Create a new regression model by replacing *model.year* with *mcat*. Comment on this model.
In what ways, if at all, is this better than the model from part (a)?

c. Augment the model in part (b) with the addition of the interaction term of *acceleration* and *mcat*. Compare and comment on this regression model with the model in part (b).
How many regression equations does this model include? Write down these different regression models.