**JAIPURIA INSTITUTE OF MANAGEMENT, INDORE**

**PGDM**

**FOURTH TRIMESTER (Batch 2019-21)**

**END TERM EXAMINATION, OCTOBER-2020**

|  |  |  |  |
| --- | --- | --- | --- |
| Course Name | **Fundamentals of Business Analytics** | Course Code | **IT402** |
| Max. Time | **2 hours** | Max. Marks | **40** |

**INSTRUCTIONS:**

1. Attempt all questions of Part A, B, and C.
2. Use “R” software to complete all the steps, write the code wherever it has been asked.
3. There are two ways to submit the answers,
   1. After attempting the questions in R or word document, submit both the files on Moodle- a separate link will be provided 30 minutes before the end of the examination
   2. After attempting each question in R, directly copy it in the space provided after each question. There is a place to submit the R code file also.
4. Data sets for all the parts (A, B, and C) along with the descriptions will be provided before the start of the examination.

**Part-A (Max Marks-10)**

**Research Objective:**

To explore the given dataset (***StudentSurvey.csv***) to answer the relevant questions.

**Preparation steps**

1. Download dataset first (e.g. xyz.csv) into your working directory

2. Set working directory: setwd("<enter the fullpath>")

3. Read the file: data\_file <- read.csv("xyz.csv")

**Answer the following questions by writing appropriate code in R:**

1. What is the data type of the following relevant variables? (***Max 2 marks)***
   1. Gender :
   2. Age :
   3. Birth\_month :
   4. Like\_austin :
2. From the first 10 students, what is the ID of a student who commute by bus and attended at least one concert. (***Max. 1 Mark***)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Change the appropriate data types of the following variables - ID and birth\_month (***Max. 2 Marks)***

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Write appropriate R code and answer the following: (***Max. 5 marks – 1+1+2+1***)
2. List the ID, gender, age, birth month, like\_austin of all the students.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Create a dataframe of male students to include ID, gender, age, birth month and like austin.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. From this new dataframe, list the male students as per the birth month (1 to 12).

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Draw a graph to show the distribution of male student’s ages. What is the shape of the distribution?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Part B: Descriptive analytics (Max marks -20)**

Primary Research Question

**Is Identity status of the comic character is related to alignment (Good, Bad or neutral).**

**Check the Data**

Download the comics.csv file and place it in working directory. Then read it using the following functions:

setwd("write complete path of your working directory")

comics <- read.csv("comics.csv")

1. Look at the spreadsheet view of the data to answer the following questions. (***Max. 2 Marks***)

1. How many comic characters are in this dataset? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. How many of the first 10 characters in the dataset have neutral alignment? \_\_\_\_\_\_\_\_\_
3. How many of the last 15 characters in the dataset are having Secret identity? \_\_\_\_\_\_\_
4. What is the identity status of the first female character in the dataset? \_\_\_\_\_\_\_\_\_\_\_\_\_

**Check the Variables of Interest**

**2**. Let’s find the variables we need to answer the question. ( ***2 Marks***)

1. Which variable tells us the identity status of each character? What type of variable is this? How many possible values are there for this variable? **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
2. Which variable tells us whether the character is Good, Bad or Neutral? What type of variable is this? How many possible values are there for this variable?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Reflect on the Method**

*3. Which method should we be using for this analysis and why? (****4 Marks – 2+1+1****)*

(a) We will generate a contingency table of \_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_ to help us with this analysis. Why?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(b) To draw the bar chart with both a legend and side-by-side bars for each id, the following code is used: barplot(twoway, legend=TRUE, beside=TRUE)

What would the code look like if we wanted to keep the legend but stack the bars (instead of set them side-by-side)?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(c) We will compare marginal and conditional probabilities to determine if characters of different identity tend to have different kinds of alignment. Why?

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Conduct the Analysis in R**

**4**. Answer the following questions by writing appropriate code in R. (***10 Marks)***

* + 1. How many characters are having secret identity in the comics dataset?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. How many characters are having public identity in the comics dataset? ?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. Write the code for the table you should look at to determine how many characters are Good, Bad or Neutral?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. How many characters are aligned as “Good”?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. Write a line of code that provides the probability that a randomly selected character from the dataset is aligned as “Good”?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. What is the probability that a randomly selected character from the dataset is “Good”?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. Write a line of code that provides the probability that a randomly selected Public character is aligned Good?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. What is the probability that a randomly selected Public character is aligned Good?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* + 1. For align and id to be independent, which of the following statements must be true?

1. P(Good)\*P(Public) = P(Good) + P(Public)
2. P(Public) = P(Good)/P(Public)
3. P(Good) = P(Good/Public)
4. P(Good)=P(Public/Good).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Conclusions**

Answer the question and support your answer with statistics: (2 Marks)

**5**. The alignments of various characters in comics was grouped into four aligns, including Good, bad, neutral and Reformed. **\_\_\_\_\_\_\_\_\_\_\_\_\_** alignmentwas depicted **\_\_\_\_\_\_\_\_\_\_\_\_\_\_** % of the comics characters. Among public characters, however, only **\_\_\_\_\_\_\_\_\_\_\_\_\_** % of the characters are aligned as “Good”. This difference between the marginal and **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** probabilities suggests that id and align **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** independent.

**Part C: EDA (Max marks - 10)**

Primary Research Question

**Develop a linear model to predict End\_scores using the other variables in the given dataset.**

Download the StudentsPerformance.csv file and place it in working directory. Then read it using the following functions:

setwd("write complete path of your working directory")

comics <- read.csv("StudentsPerformance.csv")

Do the following steps and write the appropriate R script: (10 Marks- 1+2+2+2+2+1)

* + - 1. Identify the one dependent variable(target) and other independent variables (predictor)
      2. Graphically, do univariate analysis of the dependent and some of the relevant independent variables.
      3. Identify the variable that has missing values and treat the missing value appropriately.
      4. Identify the variable that has outliers and treat them appropriately.
      5. Create a linear model where as End\_scores is a function of Mid\_scores.
      6. Predict the End\_score when Mid\_score is 33.