**JAIPURIA INSTITUTE OF MANAGEMENT, INDORE**

**PGDM**

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

**END TERM Re-EXAMINATION, NOVEMBER-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. 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
4. Data sets for all the parts (A and B) will be provided before the start of the examination.

**Part-A (Max Marks-15)**

**Note: Answer the questions using the given dataset in R**

**Data set: World Record standings.**

Description

This data contains time, event, and information of World Record athletic events across several years.

Usage: WorldRecords

Format: This dataframe contains the following information:

Event Name of event

Type Is it a timed event or a distance event?

Record Record time (in seconds) or distance (in meters)

Athlete Name of the athlete

Nationality Nationality of the athlete

Location Location of the event

Year Year record was broken

**Research Objective:**

To explore the given dataset (***WorldRecords.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. How many variables and observations are there in the dataset? (***Max 1 mark)***

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

1. What is the data type of the following relevant variables? (***Max 1.5 marks)***
	1. Event :
	2. Year :
	3. Record :
2. How many different types of events (e.g. "Mens 100m," "Womens shotput," etc.) are represented in the dataset? (***Max. 1 Mark***)

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

1. In what year did Usain Bolt first break the world record for the men's 100m dash? Display the relevant fields (Name of the athlete, Event, Record and Year) only. (***Max. 2 Marks***)

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

1. Who was the first woman to break the women’s 1-mile world record with a time of less than 260 seconds? Display the relevant fields (Name of the athlete, Event, Record and Year) only. (***Max. 2.5 Marks)***

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

1. Do the following operations: (***Max. 8 Mark: 1+2+2+3)***
2. Display all the athlete of “Rome, Italy” Olympics.
3. Display all the athlete from “Rome, Italy” Olympics of 1984.
4. Display the list of athlete who are winner of distance type events.
5. Sort on year and event and display the athlete details as (Year, Event and Athlete Name).

**Part-B (Max Marks-15)**

**Case : Cycling in Austin**

In 2011, researchers at the Texas Transportation Institute and the Center for Transportation Research at UT Austin ran an advertising campaign aimed at recruiting Austin cyclists to join the South Congress Bike Mapping Project. As members of the project, cyclists downloaded and used *Cycle Tracks*, a smartphone app developed by the San Francisco County Transportation Authority to track where people are riding their bikes based on their GPS points. The goal was to gain new information about bike commuting patterns and this data set is based on the results of the study: 3600 trips tracked from 315 users over a 6 month period. Data includes distances travelled, speed of travel, and reasons for travel among other variables.

Data on bicycle commuters in Austin Texas

**Description**

Data on bicycle commuters in Austin Texas - Data has been cleaned, identified, and simulated.

**Data frame:** BikeData

This data frame contains the following:

user\_id ID of subject.

Age Age of Subject

Gender Gender of subject (M=Male, F=Female)

Student Indicator if subject is a student (1=Yes, 0=No)

Employed Indicator if subject is employed (1=Yes, 0=No)

cyc\_freq Frequency of cycling - (Daily, Less than a month, Several times per month, Several times per week)

distance How far user rides based on average data collecion

time Average length in minutes based on collected data

speed Average speed in mph based on collected data

**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")

**Research Question – (*Max. 15 Marks)***

***How many of the cyclists were students, how often did they ride, and what was the average distance they rode?***

To do: When we first load a dataset, it is a good idea to immediately rename it something simple and familiar to you.

**bike <- BikeData**

Look at the spreadsheet view of the data to answer the following questions

1. Answer the following by looking at the data frame ***– Max. 3 Marks***

1. What is the age of the 7th rider in the dataset? \_\_\_\_\_\_\_\_\_\_\_
2. How many of the first 10 riders in the dataset ride daily? \_\_\_\_\_\_\_\_\_\_\_
3. What is the speed of the first female who cycles less than one time per month (in miles/hour)? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. Examine the variables of interest and correctly categorize them. What type of variable (categorical (nominal/ordinal), numeric/Integer is? ***– Max. 2 Mark***

1. gender \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. cyc\_freq \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
3. student \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
4. distance \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Analyzing the Data

Follow the directions below.

3. We want to know about the students in the dataset. Write the proper code and answer the questions:

How many students are in the dataset? ***– Max. 1 Mark \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

1. Create a new data frame that only includes students. Run this code:

#Pull out student data into a new data frame

students <-bike[bike$student==1,]

When you look in your workspace, how many variables do you see in this new data frame, students? ***– Max. 1 Mark*** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. We want to know how often the students ride. Notice we are referencing the data frame students now.

#Find how often the students ride.

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

What is the most frequently observed answer? Tick any one***– Max. 2 Marks***

* Daily
* Several times per week
* Several times per month
* Less than once a month
1. We also want to know how far the students travel on average. Let's create a vector of just the distances. Write the code:

#Create vector for the variable distance

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

How is this vector "distances" described in the workspace? Tick any one ***– Max. 2 Mark***

* num[1:14]
* 9 variables
* 14 obs. of 9 variables
1. Now let's find the average distance ridden by the students, using the mean function. Run this code:

#Find average distance ridden

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

How far do the students ride on average? Tick any one***– Max. 2 Mark***

* 3.74 miles
* 6.26 miles
* 12.9 miles

**Write Your Conclusion- *– Max. 2 Mark***

Write an answer using the results of your analysis as support. A template is provided to guide you. Just need to fill in the blanks.

We examined data on \_\_\_\_\_\_\_\_ student riders. Most of the student riders ( a total of \_\_\_\_\_\_\_ out of \_\_\_\_\_\_\_\_\_\_\_\_ ) rode their bikes \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ . On average, the students rode about \_\_\_\_\_\_\_\_\_\_\_ miles on each trip.

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

Write the answers of the following questions (4 + 6)

1. What are various steps followed in exploratory data analysis. (Word limit: 100 words)
2. Briefly explain how do an analyst identify an outlier for a variable in a data set and what are the ways to handle it to build an robust model for predictions.