Department of Computer Science (New Campus)
University of Engineering & Technology, Lahore

Subject: **Introduction to Data Science** (6th Semester, 2018 Session) **Total Marks: 15**

Semester Project Due Date: 18-06-21 (11:59PM)

CLOs	Questions	Marks
2	**Instructions:** 1. If your registration number is ODD, then you will work on Dataset_1 otherwise on Dataset_2. 2. You need to perform Exploratory Data Analysis (EDA) on the given dataset. Your aim should be to explore and analyze the data using all the methods / techniques you have learned for EDA. 3. You should draw **at least 5 graphs** and explain those graphs in detail. (Ref: Lecture 11 delivered on May 07, 2021). This is the minimum number of graphs needed, in principle, you should draw as many as graphs as required to fully explore / explain the data. 4. While explaining the graphs, follow the instructions / examples given in Lecture 11, May 07, 2021. 5. You need to write a report for this purpose. In the report, you will add all the graphs as Figures (e.g, Figure 1, Figure 2 etc.) and then explain the Figure / Graph. (For reference, you can see the attached PDF file, EDA.pdf) 6. Your report should not exceed 15 pages including the title page. On the title page, you must mention your Name and Registration number. 7. You can use MS Word for writing the report and then convert it to a PDF file. The name of the PDF file should be your registration number in the format 2018CS000. **(If you fail to follow this format, you will lose marks)**. 8. All the graphs should have your registration number as a watermark. Similarly, all the pages of the report should be watermarked with your registration number. 9. Before submitting, make sure that the PDF file can be opened. You can save graphs as PNG or JPEG files and then insert in the Word Document. 10. Both the datasets i.e., Dataset_1 and Dataset_2 are described below. 11. The students who did not appear in Quiz#1 should draw at least 7 graphs. They should also mention in the report that graph #6, 7 are in lieu of Quiz # 1. (This is only for those students who were on leave)	15

Dataset_1

A dataset which contains some customers who are withdrawing their account from the bank due to some loss and other issues.

Total Columns = 12
Total Rows = 10,000

	Column	Description
1	CustomerId	ID of a customer
2	Surname	Family name of a customer
3	CreditScore	Total Credit points a customer has.
4	Geography	Country / Nationality of customers
5	Gender	Gender
6	Age	Age
7	Tenure	How many months he / she remained the customer
8	Balance	Customer balance in the account (in USD)
9	Number of products	How many different products of the bank a customer was using?
10	HasCrCard	Whether a customer had credit card or not (1 = YES, 0=NO)
11	IsActiveMember	If a customer is still using the bank. (1=YES, 0=NO)
12	EstimatedSalary	Approx. salary of the customer (in USD)

Dataset_2

Data of ATM transaction of XYZ bank

Total Columns = 10
Total Rows = 11,589

	Column	Description
1	ATM Name	Name of the ATM where transaction occurred
2	Transaction Date	Date of Transaction
3	No Of Withdrawals	Total number of withdraws from the ATM
4	No Of XYZ Card Withdrawals	Number of withdraws when XYZ bank card was used
5	No Of Other Card Withdrawals	Number of withdraws when another bank (i.e., other than XYZ) card was used
6	Total amount Withdrawn	Total amount withdrawn (in USD)
7	Amount withdrawn XYZ Card	Amount withdrawn using XYZ bank card (in USD)
8	Amount withdrawn Other Card	Amount withdrawn with another bank's card
9	Weekday	Day of the week when transaction occurred
10	Working Day	Whether it was a Holiday (H) or a working day (W)