National University of Computing and Emerging Sciences, Islamabad

Assignment 1: Exploratory Data Analysis using Visualizations

Total Marks: 40

Data description:

You are required to use the attached NUCES admission stats of 2019. File contains data applicants of all over Pakistan for all the disciplines. There are many empty fields you should devise a mechanism to handle them. Meaning of each column is as follows:

L: Unique ID

CITY: Applicant city of residence

STATUS_ID: Current completion status of the application, there are 10 values 1-10. 1 is only personal information is added while 10 is the applicant is admitted to university after completing all the requirements.

STATUS: Description of the STATUS_ID.

Campus Pref (1^{st}, 2^{nd}, 3^{rd}, 4^{th}): Applicants campus preferences. Note that NUCES has five campuses all over Pakistan.

Discipline Pref. (1^{st}, 2^{nd}, 3^{rd}, 4^{th}): Applicants discipline preferences.

Matriculation: Self explanatory

Intermediate: Self explanatory

NTS_MARKS: Marks secured by the student in the NTS exam. NU allows admission on the basis of NTS exam as well.

NU_TEST_CENTER: Self explanatory

NU Test Marks: NU conducts two different entry tests for BBA and other BS programs.

SAT1: Self explanatory

SAT2: Self explanatory

NTS Merit Marks: Overall merit of the student who is seeking admission based on NTS. This field is calculated by using the matric, intermediate and NTS test score by using a formula.

NU Merit Marks: Same as "NTS Merit Marks". In this field instead of NTS score, NU test score (BBA/BS)
was used.

NTS Admission Offer: Offer made by the NU to the student. Campus and program are mentioned in the

offer.

NU Admission Offer: Same as "NTS Admission Offer". **Admission:** The program in which the applicant is admitted.**DEGREE_STATUS:** Current degree status.

Tasks to perform:

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used. 'Understanding the dataset' can refer to a number of things including but not limited to.

- Extracting important variables and leaving behind useless variables.
- Identifying outliers, missing values, or human error.
- Understanding the relationship(s), or lack of, between variables.
- Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process

In short, EDA refers to performing visualizations & identifying significant patterns, such as correlated features, missing data, and outliers. EDA's are also essential for providing hypotheses for why these patterns occur.

As described earlier, the provided dataset is NUCES admission dataset. University administration is interested in knowing answers to the questions mentioned below to define better future policies. You are required to answer the following questions by using the tools taught in the class. None of the questions can be answered as **Yes** or **No**. You are required perform EDA to reach a conclusion. You may use 1D, 2D, and/or multivariate (visual) analysis to answer the below given questions. You have to draw conclusions from your analysis and write/explain your answers as well.

Questions (30):

Question 1. Do the students performing good in Matric/O-level exam have a better chance of getting admission in NUCES?

Question 2. Do the students performing good in FSC/A-level exam have a better chance of getting admission in NUCES?

Question 3. Do the students who could not get the first preference of their choice of discipline failedto join NUCES?

Question 4. Students who have passed the NTS exam instead of the NU test have a better chance of joining NUCES?

Question 5. Which BS program is a popular choice for admission (campus wise analysis required)?

Question 6. Is there any correlation between the Matric/O-level/F.Sc./A-Level and the NU test score? (Students with the NTS test score can be ignored here)

Question 7. Which Intermediate Boards were successful in getting admission in NU?

Question 8. Which region's Intermediate board performed best in the NU test? You can categorize the boards into five regions (4 provinces + Federal). Is there any correlation between the Board andthe performance in the NU test?

Question 9. It was commonly believed that NU gives admission to students with at least 70% marks in F.Sc. Do you concur with this hypothesis based on the given data?

Question 10. Administration was told that all applicants given admission in NU had scored more than 50% in NU test. Administration is skeptical about this claim. Can you accept or reject it based on the given data?

Questions (10):

These marks will be awarded to the student who has identified two more important/interesting questions (other than above 10 questions) and answers them. We shall decide the importance of the question(s) raised.

Many concepts asked in the asked questions are subjective and are left to you to define. For example, "performing good", "popular" and so on. Note that in some cases, you may need to look at multiple columns to get the required information.

Submission:

You are required to submit a report with response to each question with supporting graphs, tables and stats. You have to write analysis and conclusion you draw from each graph and visualization you use to answer your queries. Python code for the experiment must be submitted separately for each question. *Note: Furthermore, you have to mention your definition of the subjective terms used while answering.*

Use variable name "admission" for importing excel file. Please note, we shall run your code ourselves toverify your graphs and test results. Your submission must include all the steps you have performed.