BUAN 6320

Database Foundations for Business Analytics

Project 1

Due: 11/04 – 11:59pm

Project 1 includes 5 steps as follows. For each step follow the description and requirements of the step and provide your answers/results in the report.

- The report should not be longer than 10 pages
- Make sure to document every step you take and explain what you have done
- Document all your SQL queries in the report (wherever applicable)

Step 1: Choose a Dataset

Find a data set of your choice (publicly available) that meets the requirements for this project. The requirements are as follow:

Dataset Requirements

- Dataset should be publicly available and free to use
  - You can use websites such as Kaggle.com to find an appropriate dataset for this project
- The size of the dataset should be at least **1 GB**
- The dataset should contain **structured** data or could be easily converted into structured data
  - Make sure to check the column types and values they contain before choosing the dataset
- You should be able to create at least **3-4 tables** out of the dataset (your final schema should contain at least 3-4 tables that are related to each other via some foreign keys)
- Data **quality**: the chosen dataset should not have a significant amount of missing values
- Dataset **domain**: the chosen dataset should be related to a business (please refer to step 2 (business understanding) to make sure your dataset meets this requirement

Step 2: Business Understanding [5]

Assume you are the business who owns the data and has provided it. On that perspective, try to ask yourself questions such as what follows and provide answers to them.

1. Why this data has been gathered?
2. What can be done with data? What can we achieve?
3. What are some of the goals/targets we have regarding the business that we can achieve by investigating this data?
4. What insightful information can this data provide us that can be used to improve the business?
5. Why are we studying this data?
6. Are there any problems in our business (based on the given data)?
7. Can we find any solutions to these problems by studying this data?
8. What are some of the things we can optimize/improve in our business by studying this data?
9. …

You should ask yourself questions of this type from the business owner point of view and provide answers to them as well.

At the end of this step, you should be able to come up with at least a couple of goals you try to achieve (or a couple of questions you try to answer) by studying this data.

Step 3: Data Understanding [5]

Now assume you are a group of data engineers/data analysts/data scientists/ business analysts hired by the business to investigate/study the data they have provided you and answer the questions they have (questions you asked in the previous step). In this step you need to familiarize yourself with the data they have provided you. This data understanding should include the followings:

- What information each column of the data contains
- The data types of each column
- What are some of the values each column contains
    - Describe the values, scales, the range of the data, …
- Verify the data quality
    - Verify the quality of the name of the columns
        - Do you need to change any of the column names? Propose proper column names if the names does not look good to you
    - Are there any missing values? If yes, then what columns and what percentage?
    - Are there any duplicate data?
    - Do you believe there are outliers in the data (optional)
- Provide simple statistics of the data for each column (do this either here or in step 5)
    - For example: range, mode, mean, median, variance, counts (frequency)
    - Describe what these values mean especially if you found something interesting
- Try to understand the relationships between the columns of the data
    - What relationships can you find between the columns?
    - Are there any functional dependencies in the data?
-

Step 4: Design a Database [10]

In this step you should design a database given the business requirements (discussed in step 2) and your findings (from step 3). Follow the same steps discussed in the class for the database design process.

1. Requirement Analysis (this is already done in step 2)
2. Data Understanding (this is already done in step 3)
3. Schema Design
    a. Find entities, their attributes, their primary keys, and relationships between them
    b. Model all the constraints you believe should be there in your schema
    c. Draw and ER diagram of your dataset
    d. Translate your ER diagram into relations
4. Schema Normalization
    a. Find all the functional dependencies you can from your schema
    b. Check if the keys you have chosen for your relations are minimal
    c. Check if your schema is in BCNF (Boyce-Codd Normal Form)
    d. If your schema violates BCNF, bring it to BCNF by decomposing it
    e. Update your ER diagram with the latest schema
5. Create your database in MySQL using the latest version of your schema
6. Import the data into your database
    a. If there are errors while importing, document these errors in your report and mention how you dealt with them

Step 5: Data Cleaning and Database Testing [10]

Now that you have your database ready and the data in it, you can start working with the data and querying it.

In this step you will investigate the data quality and deal with what you find. Your data might need some data cleaning which you will take care of in this step.

- For each table in your database, check all the columns and the values they contain
- For numeric columns, check for the statistics, and see what you find
    - You can use the same information you found in step 3 (you don't have to copy it here again if you have already done it step 3, just point to it)
    - You should be looking for missing values, values that seem to be outliers (typically far away from the mean), or data errors or any values that does not seem to be valid (like a typo)
    - Make sure all the values of these columns are from the same type (all numeric)
    - Document the problems you find; fix them and explain how you dealt with them
- For character columns, check for all the values they contain
    - You should be looking for missing values or data errors or values that does not seem to be valid (e.g., sometimes there are white spaces in some of the cells either before or after the value)
    - Make sure all the values are from the same type and domain
    - Document the problems you find, fix them, and explain how you dealt with them
- Try to query your database especially from more than one table (by joining them) to see if the results make sense or not
    - Check if the results of these queries match what you expect
    - Check if the constraints are working properly