Project 1 part 1

This exercise is to be completed and submitted **individually**. Read this document entirely and carefully before starting on the assignment.

Learning goals:

This project will help you learn the following aspects of database management:

- How to work with real world data in CSV or TXT format. (Part 1)
- Understanding data inaccuracies in a large dataset. (Part 1)

Introduction:

The Mecklenburg County Board of Elections maintains a voter data file for registered voters in Mecklenburg County, NC. Go to the link to read the description of the dataset.
DO NOT DOWNLOAD the data from the following link:

https://www.mecknc.gov/BOE/data/Pages/VoterDataFileDetails.aspx

The data set, that you are going to work with, has been taken from this website and mostly cleaned for you. You will still have to do some research on your own on how to clean data for some parts in this dataset. The data is in Comma separated value (CSV) format, and the file size is decently large (160 MB), so consider carefully how you will work with it.

The file is hosted for you at Dr. Singh's Vanderbilt box account. Download the CSV file from the following link:

https://vanderbilt.box.com/s/8swm6pzkbaja0563sfhu95565mhzpk2g

Make sure that your mega_table has only the attributes included in the box file above. Any more or less attributes in your table will incur penalty.

Assignment:

This project has many deliverables. You will first make part 1 submission.

Part 1

You are required to do the following on this part of the project:

1. Download the CSV file from the provided link. Do not open the file in Microsoft Excel. You may open the csv file in a text editor if you want to see the data.

2. Create a database.

3. Create a table structure (mega table) that is appropriately matched to the data file (direct mapping of attributes without any optimization of the structure from the CSV data file). It means that each attribute in the CSV data file should be accounted for in your mega table. I have provided the list of attributes of the table at the end of this document.

4. **(Important)** When you are designing your mega table, be careful about the data types. Deciding on the datatype for attributes by looking at first few rows might not give the correct representation of all the data in each of the row. As a DB designer, think about the strategy that can get the data from CSV file to SQL.

5. **(Important)** Encountering data inaccuracies is common theme in real-world data. When you encounter data loading errors, try to understand the error and find a solution that resolves it.

6. Load the data into the table in MySQL.

    a. You can use LOAD DATA statement to load the file in your database.
        i. See the following link: https://dev.mysql.com/doc/refman/8.0/en/load-data.html

        ii. You can also web search the syntax for LOAD DATA statement.

    b. You can also use MySQL Workbench data import feature, however keep in mind that it might be slow. Students in earlier semesters have reported that this process is not reliable. See the following link:

    https://dev.mysql.com/doc/workbench/en/wb-admin-export-import-table.html

    c. The CSV file has fields enclosed by '$' and separated by ';'
    d. If loading data results in error, identify the errors and do some research on how to resolve them. Get in touch with instructor/TA if you can't resolve it on your own.
    e. **(Important)** If you encounter secure-file-priv issue, see the support file uploaded on Brightspace.

7. If you successfully imported data in your MySQL server, then run the following query:

```
SELECT COUNT(*)
FROM your_table_name (replace with name of your mega table);
```

    Take a screenshot of the MySQL workbench showing the result set and your query.

8. Run some preliminary data analysis to understand the data. For each of the following question,
    a. Write your query in the SQL file. Clearly label each query.
    b. **AND** include the screenshot of query and result set in your PDF file. Please label each question clearly and correctly.

        i. The CSV file doesn't include NC Senate District information (nc_senate_desc) of some voters. Return the number of voters whose NC Senate District information **is absent** in the dataset?

ii. The CSV file doesn't include (or has missing) first_name for many voters. Return the number of voters whose first name is missing/not listed in the database.
iii. Find the number of female voters (sex_code =F) who are affiliated/registered as Democrat (party_cd = DEM).

Submission:

Submit the following three files:

a. p1part1_lastname.sql (The SQL file)
b. p1part1_lastname.pdf (The screenshot)
c. p1part1_lastname.mp4 (A ~1 min long video)
    a. The video should clearly show the evidence of database created in MySQL workbench. You should run a `SELECT * FROM your_mega_table_name`. Then scroll down to result set window to show that you have correctly imported all the required attributes.

These files should contain all of your answers, clearly separated, clearly labeled, and in order. Upload the files on Brightspace. Once uploaded to Brightspace, you should verify your submission. We will grade your work from the files submitted, and if they can't be opened or executed, it will receive no credit.

Grading criteria:

You will be graded on the following criteria:

p1part1_lastname.pdf

- Clear screenshots of MySQL workbench showing each of the queries and result set.

p1part1_lastname.sql

- Clear and adequate comments.
- Correct and appropriate data type usage for data import purpose.
- Correct formatting of SQL clauses in all queries.
- Following style guidelines as discussed in class.

p1part1_lastname.mp4

- ~1 min long video
- Clear evidence of MySQL workbench showing the data imported and result set.

Miscellaneous:

The raw dataset (downloaded csv file) includes the following attributes and has no header information:

precinct_desc, party_cd, ethnic_cd, race_code, sex_code, age, pct_portion, first_name, middle_name, last_name, full_name_mail, mail_addr1, res_city_desc, state_cd, zip_code, registr_dt, voter_reg_num, nc_senate_desc, nc_house_desc, E1, E1_date, E1_VotingMethod, E1_PartyCd, E2, E2_Date, E2_VotingMethod, E2_PartyCd, E3, E3_Date, E3_VotingMethod, E3_PartyCd

Important

If you couldn't successfully import the data and you received some error, get in touch with TA to resolve the issues before the due date. Late penalty will apply on this part of the project as described in late policy in syllabus.