COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

# In-Class Assignment 20

## Examining Baby Names from the Social Security Administration

Follow the instructions below and answer the questions that follow. Add your solutions to the **In Class Assignment 20.sql** script and submit to Canvas by the deadline listed above. Save your file frequently to avoid losing work!

Two data files have been pulled from the Social Security Administration website (https://www.ssa.gov/oact/babynames/limits.html) representing baby names from 2019 and 2020. The datasets contain the following fields:

**yob** – year of birth (4 digits)
**first_name** – first name given to baby (up to 20 characters)
**sex** – biological sex of the baby (values of 'M' or 'F' only)
**frequency** – number of babies given that first name in that year (values up to 10s of thousands)

**Note:** any names reported in less than 5 babies in a given year are excluded from the data due to privacy concerns.

## Part 1 – Creating the Schema in MySQL:

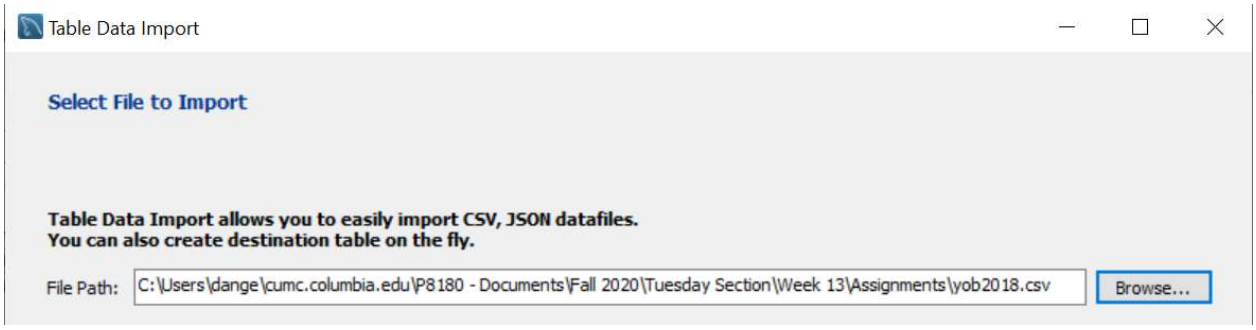
In order to import all data into one table, do the following steps:

1. Create a new schema called **baby_names** and apply all subsequent code to that schema.

2. Using SQL code, create an empty table called **yob2019** in which to import the name data from 2019. No primary key is necessary, but be sure to choose correct data types.

3. Import the 2019 csv file by right-clicking on the new table, then selecting **Table Data Import Wizard.**

   Browse to the file location and then click **Next** through the remaining prompts. **The file may take a couple of minutes to import due to size.**



# In-Class Assignment 20

4. Using SQL code, create an empty table called **yob2020** in which to import the name data from 2020. No primary key is necessary, but be sure to choose correct data types.

5. Import the 2020 csv file by following the steps in #3.

6. Using SQL code, create a new table called **ssa_names** by combining the data from both tables above. Use the following code sample as a guide. (Note, square brackets [ ] refer to optional terms)

**13.1.20.4 CREATE TABLE ... SELECT Statement**

You can create one table from another by adding a `SELECT` statement at the end of the `CREATE TABLE` statement:

```
1  CREATE TABLE new_tbl [AS] SELECT * FROM orig_tbl;
```

From: https://dev.mysql.com/doc/refman/8.0/en/create-table-select.html

## Questions

Use the new **ssa_names** table that you created to answer the questions that follow. Include an answer to each question in a comment under its respective query in the script.

1. How many unique names were reported in 2019 and 2020 combined?

2. How many babies were included in the data for each year?

3. List the most popular boy and girl names of each year and include their frequencies. (Use only one query to achieve these results.)

4. List the top 10 ranked boy names for both years and give the total number of babies they represent per year. (Hint: final result set should have 20 rows and 5 columns.)

5. What is the most popular boy name from either year that begins with R? Give the year, name, and frequency. Only return the row showing the information for the most popular name (but account for possible ties in frequency).

6. How many unique boy names that end in 'I' or 'O' were reported in 2019 and 2020 combined, and how many babies were given those names in total? Report both numbers in one result set.
(Hint: use the RIGHT() function https://www.w3schools.com/sql/func_mysql_right.asp)

# In-Class Assignment 20

7. Combine all name frequencies from 2019 and 2020 and then rank each name by sex based on combined frequency. Then, query those results to see where your name ranks.

8. Produce a list of distinct names that were given to both boys and girls. Restrict results to names that were given to at least 1000 girls and 1000 boys, and order the resulting names alphabetically. (For the answer to this question, list only the first 5 names in the list.)