**MongoDB Mini-project #4**

**REVISED VERSION**

You will utilize 3 real-world data sets for this project.

**Data Set 1:**

Use the restaurants data set as your starting collection for this project.

<https://raw.githubusercontent.com/mongodb/docs-assets/primer-dataset/primer-dataset.json>

**Data Set 2:**

I looked at <https://www.data.gov/> (there are over 200,000 data sets) to find NYC restaurant inspections:

<https://catalog.data.gov/dataset/dohmh-new-york-city-restaurant-inspection-results>

I recommend that you use the CSV and Compass to upload the data. Note that the data set will need to be cleaned and prepared for usage by:

1. changing the commas to semicolons before uploading,
2. creating a new column for the name to convert it to title case.

**Data Set 3:**

Here is the information page for NYC OpenData:

<https://opendata.cityofnewyork.us/data/>

I looked at the city government data sets (there are more than 600 of them) and picked this one that gives demographics by zipcode:

<https://data.cityofnewyork.us/City-Government/Demographic-Statistics-By-Zip-Code/kku6-nxdu>

Note: I am assuming the data is uploaded as strings since that is the default. You will not need to convert any fields to numeric types for the queries below.

**Project Outline**

1. Your job is to download the data sets and store them in MongoDB. What format you choose and how you organize the data is up to you (but it must all ultimately be in MongoDB). Describe your approach (did you combine everything into one collection? did you create separate collections?) How many documents are in your collection(s)?
2. Answer the data demands given below by writing MongoDB queries. Give the query and an example of the result (along with a total number of documents retrieved.)
3. Create a new data demand and solution (query and example result with total number of documents) that utilizes all three data sets. Your data demand should be a descriptive English sentence and your query should be in MongoDB. Your new data demand should be different than the given ones.

**Data Demands** (entire section changed)

Use the first data set in all of these queries, plus the other two as needed. Give the MongoDB queries and results for each one.

* Show the first three resulting documents for the final answer, but run the query for the entire data set and give the resulting number of documents.
* If there are nested fields and you are using Compass, expand the nested fields to show the values.
* If you are using the mongo shell to execute your queries, use **pretty()** for the final answer.

Hint: create smaller collections before doing a join.

1. Give the restaurant ID (not the document ID), restaurant name, and street name (*address.street*) where the restaurant is in the borough (also BORO) of Manhattan and the street is “Madison Avenue” where the restaurant is not listed in the inspection data set. Sort by restaurant name ascending.
2. Give borough, restaurant name, and restaurant ID (sorted by borough and name) for restaurants with evidence of mice or live mice (use violation code = “04L”) in zipcodes that have a fifty-fifty split by gender (use percent female = “0.5”).

**Submission**

Due date: Wednesday 4/24/19 at noon.

Submit both an electronic copy and a hardcopy. Only one submission per team (online and paper) is needed.