Assignment02. Bigger data and communications

Due by 11:00 am October 26, 2021

Through the first assignment, now you are (should be) familiar with the New York City 311 service request data. You will be using the **Socrata API** to collect MUCH BIGGER data to better understand New York City citizen reports and urban problems in the city.

Assignment Objectives

The purpose of this assignment is to get familiar with the data collection and preparation process using APIs and more advanced Python programming. You will be asked to do 1) a data collection/filtering through APIs, 2) data cleaning and processing, 3) an exploratory analysis based on your research questions, and 4) plotting meaningful visualization.

Data

You will be exploring **full raw data of New York City 311 service requests since 2010**. There were about 2 million data points in the subset of the dataset that you used for the first assignment, and the full dataset for this assignment has **27 million data points** (which means you need approx. or more than 25GB space for storing data). Therefore, you MUST use Socrata APIs to retrieve data from NYC 311 instead of downloading the data.

To facilitate this assignment, here are resources:

1. Understanding what you can do with Socrata APIs specifically for NYC 311 data at https://dev.socrata.com/consumers/getting-started.html
2. Exploring more information about Socrata APIs at https://dev.socrata.com/
3. NYC 311 metadata (311_SR_Data_Dictionary_2018.xlsx) at https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
4. Reading for inspiration at https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0186314&type=printable

*Note: you cannot print out 27 million data points at once. Remember, you have to use specific **queries** to get your 'desired' data.*

Scenario

Here, you will be **exploring the data as an urban data journalist**. Let's assume you are working for the New York Times data analytics team[1]. **You have access to the open data and you are tasked with writing an interesting article (but easy to read) based on data analytics**.

For this assignment, prepare a Medium (https://medium.com/) blog entry of no more than about 1,000 words. You don't actually have to make an actual blog entry, but your tone should be focused generally and use language that is accessible to the general public as if it were shared publicly. It should be pitchy, but engaging, approachable, but complete. You should read some articles with data analytics published in Medium or the NYT**. It should include attractive visuals created by using Matplotlib and Seaborn (try to explore plotting skills and examples yourself)**. Geospatial plots (use your skills learned in GIS class!) are more than welcome, but there is no requirement to use geopandas.

You will choose your own question based on your data exploration process, with some offered below for inspiration and papers with better questions:

- How do New Yorkers use NYC 311 for 10 years?
- Does a certain citizen complaint type correlate with other types of complaints?
- What are the most problematic issues in the city? By neighborhood? How about over the time period?
- Which neighborhoods have the most ice cream truck complaints?
- Is there any seasonality? Differences between Monday vs. weekend?
- Any differences between this year and last year? Before and after the Covid-19 outbreak?
- What are the most reported issues in your neighborhood vs. your classmate's neighborhood?

You are welcome to bring in external data, or just rely on data from NYC 311. *At minimum, your analysis should be pulling in multiple API calls. In other words, you need to pull at least two tables through NYC 311 API queries.* Also, you can focus on one certain neighborhood and/or multiple neighborhoods and/or the entire New York City or boroughs. It applies to the temporal scope as well (certain time period vs. full time period since 2010). **Please state your research scope and questions clearly in your memo and Jupyter notebook using markdown.**

Deliverables

You should work on two deliverables:

- **"Medium" style post** (PDF format)
  Your theme or research question should be expressed in your article. You should state your case study scope, report your findings and interpret them relative to your query

[1] There are tons of great articles with data visualization published in the NYT. One of examples are "Covid World Map" at https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html

(including limitations, if any), and offer a discussion about its impacts. Although short, you should give a sense of the decisions you made and why using a casual and approachable narrative voice.
    a. No more than 1,000 words (excluding visuals)
    b. Your own eye-catching visualization should be included (at least one)
- **Code** (Jupyter notebook AND PDF)
  Please submit your Python code as Jupyter notebook (.ipynb format) and PDF with outputs. Please submit a single notebook, not multiple. Make sure I can run your code and create the same results without any error. Also, think about how to write your code in parts - a collection section, and an analysis section.

Metrics for Success

- Framing a theme and interesting questions (Be an urban data journalist!)
- Clarity of your article
- Attractiveness and effectiveness of your visualization
- Completeness of how you provided descriptive statistics and interpreted the data overall
- Completeness and clarity of your script

An Extra Credit

If you actually post your assignment to your personal blog or create a Medium blog post in a non-anonymized way (in that you are willing to put your name to your work and bravely share it publicly), note the course, and pledge to keep it up, you will receive 1 point added to your grade for this assignment. Add the URL at the end of the assignment memo for evaluation. It should be noted that when you're applying for jobs, people will search for you. This assignment is testament to your analytical chops and therefore, I highly recommend you use this as a way of increasing your digital footprint. (And you're going to be required to post something online later in this course).

Note

- You should use Python based coding.
- Remember, this course is not an Excel class.
- Don't forget to practice good file management.
- If you do something in Python but don't know how to code or you have an error, Google it! Again, this is not a joke.
- If you feel lost, use your colleagues and instructor as resources. But please keep in mind that your code cannot be the exact same as your colleagues'.
- This assignment does not explicitly test your knowledge of statistics or quantitative methods. With the diversity of students and quantitative reasoning abilities in the class, you will not be evaluated on the types of analysis you perform.