**Case Study: We want you to investigate whether there is relationship between playground availability and household income in New York City. We are most interested in how you think and articulate your process and not a specific answer.**

**DATA:** You can find relevant data sets here (and we recommend that you start with them), but you may use any additional publicly available data you wish so long as you document it clearly and obey the terms and conditions set forth by the provider.

<https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2016-zip-code-data-soi>

<https://data.cityofnewyork.us/Environment/Directory-of-Playgrounds/59gn-q4ai>

<https://data.cityofnewyork.us/Recreation/Directory-of-Parks/79me-a7rs>

**Project:** Please respond to the following concisely in Word; well-chosen charts and tables are most welcome. Please include all supporting code (preferably in Python or R), data files, and output. We should be able to reproduce your work exactly. So, if there is a manual step, you must document it.

1. Load the data (or a relevant subset), organize it into a useable form. You are going to have to connect playgrounds to income somehow; one way involves using zip codes and the directory of parks (playgrounds that are not part of a school are also parks). Provide some summary statistics.
2. Explore the data. Do you see anything interesting? Are there obvious outliers, errors, or gaps? What, if anything, should you do to exclude or fill problem rows?
3. What are some natural checks you can do for data completeness? E.g. What is the population of New York City? ###
4. Does income in an area have a relationship to the availability of playgrounds? Do you have a hypothesis to explain any of the patterns you observe? What are coincident or normalizing factors you should consider (e.g. population in an area)? Is average income the right thing to consider, or some other measure?

**Communication:** In PowerPoint, create one or two slides which communicate your most powerful finding to a non-technical audience with no prior knowledge of the questions asked.

**Conceptual Questions (no coding, in Word):**

1. What were some avenues you explored which did yield insights you used in the main write-up? What would you do differently if you could start from the beginning?
   1. Add hypothesis testing … if one thing is not the same as the other.
2. What biases might you have introduced into your results as a part of your methodology? How might you reduce them with more time / information? === illegal immigrants not present in data, playground availability for schools
3. What are areas for further development of our understanding of this and related relationships?
   1. What other kinds of data would you like to consider?
   2. Are there other kinds of visualizations or models that might be enlightening? Clustering
4. How would you go about automating your process to update when new parks are added to the data? How would you detect these updates? What changes might you have to make to your code to do this?