Data Driven Decisionmaking **Project 2**

Retrieving Data from Non-traditional Sources

There is a well known correlation between the location of certain classes of commercial establishments and the socio-economic status of individuals living nearby (i.e., fast food, pawn shops). Across the US, it has been hypothesized that this relationship will hold irrespective of the geographic location being studied. This is particularly relevant for individuals trying to reverse-engineer the demographic strategies companies are employing to select the locations of new establishments. In this project:
(A) You will scrape information off of the web providing the geographic location of a commercial entity you believe to be correlated with socio-economic status for one US State[1].
(B) You will use this in conjunction with US Census information, making an argument for what criteria a given establishment used to select locations. You may have to explore multiple criteria.

1 Project Deliverables

You will need to turn in three deliverables as a part of this project:
(1) A 2-page report summarizing your findings, including the following elements:

- Summary of Findings (1 paragraph summarizing everything you did and the key take-away)
- Figure(s) detailing your findings
- Data and Methods, with enough information for another practitioner to reproduce your approach (2-3 paragraphs). Make sure you detail the two predictive approaches you selected.
- Results, with a written description of any tables or figures you produce (1-2 paragraphs)
- Table(s) detailing your findings
- A discussion and conclusion, covering limitations of your approach, take-aways, and next steps. (1-2 paragraphs)
- A bibliography with any literature you cite.

(2) A brief one-paragraph description of the code you used to produce 1 and 2.
(3) You do not need to turn in your python code, as it will already be on the shared Jupyter hub!

[1] Excluding starbucks, as this has been used in previous studies on the topic

1.1 Getting the Data

1.1.1 Web Scraping

Web scraping is a common approach to retrieving data, but can be uniquely challenging for some sources. As an example case, we will seek to retrieve information on the location of Starbucks within Virginia. There are two general approaches we can use to this: first, we could use the Starbucks website itself and write a custom scraping routine - however, this would be very time consuming. Second, we can leverage other databases that have already done this for us - in this case, our example leverages data already collated by Google as a part of Google Maps.

The Jupyter notebook that comes with this assignment provides details on how to approach this. Note that you will need to register for a free Google Places API key, which is limtied to 1000 requests per day. Thus, be careful not to write code that will lock you out of the google system!

1.1.2 Census Data

While you can choose to download and re-create census data on your own using the US Census website and the ”Tiger” spatial zone files, an easier alternative is to leverage an online platform built by the University of Minnesota - NHGIS https://www.nhgis.org/ . Similar to GeoQuery, this system can take up to a day to process data requests, so it is recommended you get started early!

1.1.3 Simulation

Just like in the first project, you will be creating a simulation by adding uncertainty into your census data to verify that there are true trends in your data. In this case, the data you download from the census will have estimates of the uncertainty (at the 90th percentile) for each variable you use. Make sure you leverage these estimates in your selection of distributions for simulation.

2 Stretch Goals

These goals are optional, and worth a very small amount (up to 5% total of your assignment grade for all goals in total) of extra credit. Completing any one stretch goal gives you the opportunity to receive all 5 points of extra credit.
(1) Scrape directly from a companies website, instead of using the Google API.
(2)Conduct your analysis at multiple geographic scales using different census units and contrast your results.
(3) Conduct all spatial analysis steps of this assignment in python, rather than in Q.