TheDataOpen

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Problem Statement

Welcome to the Fall 2020 Europe Regional Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

Background

Gentrification is the process of changing the character of a neighbourhood through the influx of more affluent residents and businesses. Although at first glance it seems very positive, it is actually a controversial topic in politics and urban planning. It is unquestionable that a neighbourhood can benefit from physical improvements of its buildings or the creation of new business; however, gentrification normally comes at the price of the displacement (or replacement) of the working-class population by the new middle class as the cost of living in the area tends to increase.

A BuzzFeed News article published in 2018 entitled, "They Played Dominoes Outside Their Apartment For Decades. Then The White People Moved In And Police Started Showing Up, expands on that fact. Residents of a historically Latino neighbourhood in NYC noticed that the police presence in the area increased significantly "over the past few years". This fact was corroborated by the dramatic increase in 311 quality-of-life complaints on their block starting in 2015 (the majority of the complaints were about noise).

The gentrification process is not new. Historians say the first time it was observed was in ancient Rome when large villas replaced small shops by the 3rd century, AD. The actual term, however, was first used by a British sociologist in 1964 to describe the "influx of middle-class people displacing lower-class worker residents in urban neighbourhoods of London". After 1964, scholars have applied several definitions of gentrification. Some of those focused more on the good side of the process, such as improvement of the space for more affluent users. Others focused on the less desirable consequences, such as the displacement of people. The overall process has had both positive and negative outcomes.

Similar to its definition, there are also different methodologies to determine if a tract (which corresponds to a neighbourhood boundaries) became gentrified over a given time period. One of the most accepted methodologies was proposed by "Governing Magazine" who tackles the problem in two steps. First, they developed a test to determine if a tract is eligible to gentrify by checking 3 different criteria based on its population, household income, and median property value. Those which were considered eligible then proceeded to step 2, which focused not only on median home value but also on the resident's educational attainment. More details on both tests proposed by this methodology can be found here.

Your Task

Your goal is to use census data for every tract in the New York-Newark-Jersey City, NY-NJ-PA Metropolitan Statistical Area (MSA) in order to discover and analyze patterns related to the gentrification process. More broadly, you should consider if and how quickly gentrification tends to happen and where does (or doesn't) it happen.

We have curated census data for all the tracts mentioned above for the years 2009 to 2018 using the Census API from the United States Census Bureau. The API offers thousands of variables but we selected only a handful of them as providing all would not be feasible for this event. Feel free to enhance or recreate these datasets as you see fit - just be careful that not all variables are available for all years. A list of available fields for the year 2009 can be found here.

If you would like to enhance your analysis, we are also providing a sample of "311 Service Requests" data from the NYC open data portal. The original dataset contains over 24 million rows from the years of 2010 to 2018. The files you are being provided were built from the original dataset via splitting it by year based on the "Created Date" field, and then taking a 25% random sample from each year.

You are asked to pose your own question and answer it using the available datasets as well as any supplementary datasets you may find. What is important is both the creativity of your question and the quality of your data analysis. **You need not be comprehensive; depth of insight is more important over breadth of the question posed.**

Submissions may be predictive, using machine learning and/or time series analysis to investigate your research topic. Submissions may also be illuminating, through the use of data visualizations or through sound statistical tests.

Consider exploring one of the sample questions below, or creating your own variation. Creativity in formulating your own question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigour, which are far more important.**

Sample Question 1: Which tracts have gentrified and how do they compare with the ones which didn't?

Sample Question 2: Does the change in tracts demographics over time correlate with its increase in median home values and\or 311 complaints?

Sample Question 3: Do the incidents mentioned in the BuzzFeed article tend to happen more in tracts that gentrified over time?

Sample Question 4: Of all the tracts that gentrified, which proceeded fastest and slowest? Why?

Datasets

The provided datasets are stored in the “Datathon Materials” folder on Google Drive. Your team should only use the datasets that are relevant to your chosen question/topic. The raw data sources are noted; however, we encourage you to use our tables since they have been organized and cleaned to be readily usable in most popular data analysis libraries.

Note that the provided datasets are split by year. In some cases, the sources of the datasets and information on how they were built is noted in case you want to enhance or recreate the datasets.

Counties
List of counties in the New York-Newark-Jersey City, NY-NJ-PA Metropolitan Statistical Area.
26 rows & 3 columns. Size: < 1MB per year. Source: Public.

Census_YYYY
Census data for year YYYY (2009 to 2018) for each tract in the New York-Newark-Jersey City, NY-NJ-PA Metropolitan Statistical Area.
4700 rows & 17 columns. Size: < 1MB per year. Source: Census API.

311Calls_YYYY
Sample (25% per year) of “311 complaints” from New York City's open data portal from 2010 - 2018.
Size: ~60MB zipped, ~300MB unzipped per year. Source.

Additional Datasets

Participants are welcome to scour the Web for their own custom datasets to supplement their analysis. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One’s technical product team via Slack if you believe your idea is worthy of an exception).

Other Materials

We will provide you the schema for each of the data tables in another packet.

Submissions: Content

Submissions should have three components:

1. Report – this should have two main sections:

   a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what is their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged if they help explain your thoughts.

   b. Technical Exposition – What was your methodology/approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and modeling steps. Again, the use of visualizations is highly encouraged when appropriate.

2. Datafolio – a story-driven visual snapshot of your analysis (please see our guidelines). To start, make your own copy of the template slides provided. **Note that this must be accessible to someone with only a rudimentary understanding of economics, statistics, and machine learning.**

3. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it or your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself"**. Please ensure that your main findings are clear and that any visualizations are functionally labeled.

Submissions: Evaluation

The competition will have multiple rounds of evaluation. The most important component of this evaluation will be your Report. Of secondary importance is your Datafolio. These components are judged as follows:

- **Report Non-Technical Executive Summary**

   ○ *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose it? Are your conclusions precise and nuanced, as opposed to blanket (over)generalizations?

- **Report Technical Exposition**

   ○ *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.

  - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of these? What patterns did you notice, and how did you use these to make subsequent decisions?

  - *Analytical & Modeling Rigor.* What assumptions and choices did you make, and what was your justification for them? How did you perform feature selection? If you built models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular ones you built, and what do they tell you?

- **Datafolio**

  - *Data Storytelling*. Did you find a compelling narrative that effectively communicates and highlights your analysis? How clear is the phrasing of the question and main insights? Does the layout of the Datafolio complement the flow of the analysis? Please don't just copy and paste text from your report.

  - *Visualizations*. How heavily does the datafolio rely on text when graphics, flowcharts, or other visual representations would have been more effective? How well are the charts and graphs interpreted by accompanying text?

Submissions: Format

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

However, please also include the source file used to generate your report. For example, if you submit a PDF with math-type, equations, or symbols, please include your LaTeX source file.

Datafolios should also be submitted in **a universally accessible format (PDF, PPT, etc.).** Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to upload and send in your submitted content. **Submissions MUST be received by 5:00PM ET on Sunday, October 25th, 2020. Any submissions received after that time will NOT be evaluated by the judges**.

Tips & Recommendations

This will be a weeklong event, however, you should try to complete as much of your work as possible before the weekend. The extra time may lull you into a false sense of security. Additionally, with your extra time, you should really think about what problem you want to solve. The outcome of this Datathon for you will likely be decided by how well you planned your work.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time development environment that eliminates many pain points of the standard "terminal + text editor" environment, and is compatible with both Python and R.

We also recommend that your team not try to learn new tools if possible; instead, leverage your existing skills to extract as much insight from the data as you can.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls of unsuccessful teams. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

Tips for Success	Try to Avoid
1. Focus on hypothesis testing when brainstorming your research question	**1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy
2. Spend at least 4 hours on your report to ensure strong communication through visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so make sure those conditions are sufficient
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality	**3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not even true or worthwhile

Ask for Help

Correlation One's technical product team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move your analysis forward.