TheData Open

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Problem Statement

Welcome to the 2019 Citadel – Correlation One London 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

The United Kingdom (UK) has been one of the powerhouses in human history and has been a major impact on the rapid growth of the European Union (EU). It was not until June 23, 2016, when the UK held it's a referendum to leave the EU. "British Exit", more commonly known as Brexit, has since left numerous ramifications both within and outside the country. The Brexit Referendum voted "Leave" with just 52% of the ballots; in context, the vote only won by 1.3 million. The vote's result redefined expectations and roiled global markets. The Former Prime Minister, David Cameron, who originally called for the referendum and campaigned for Britain to stay in the EU, resigned the following day which sent the British pound to fall to its lowest level against numerous other currencies to its lowest levels in 30 years.

Britain is scheduled to leave the EU by October 31, 2019. The government has so far extended the negotiation period twice to avoid leaving without formalizing a deal with the EU. The current prime minister of Britain, Boris Johnson was a former Mayor of London, foreign Minister, editor of The Spectator newspaper and hardline Brexit supporter campaigned on a platform to leave the EU by this deadline regardless of any outcome. He has since petitioned to suspend Parliament in hopes of stopping Members of Parliament from a rapid push through legislation blocking his promise to the British citizens. With a few weeks left until this date and doubts of a deal striking, some hope and others fear that the "No-Deal Brexit" might transpire.

Further background reading:

Explanatory news website Vox has a few approachable articles on Brexit, including 9 questions about Brexit you were too embarrassed to ask and the week in Brexit drama, explained.

The Institute for Government published a report, named the *Understanding the economic impact of Brexit*, that aims to help non-economists to interpret the range of available information

Your Task

Your goal is to analyze the UK's economic and demographic data (described below), potentially in combination with supplementary datasets, in order to increase the understanding of how Brexit has impacted various socio-economic, geopolitical or various other factors across localities within the country as well as from a more regional and global perspective.

You are asked to pose your own question and answer it using the available datasets in the available time. 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 will be rewarded overbreadth 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.

Sample Question 1: How can the UK succeed in a "No-Deal Brexit" environment? What are some scenarios that would support Britain over the next few years?

Sample Question 2: How has Brexit shifted the UK and EU unemployment rates and what may this imply for other countries looking to follow suit and leave the EU?

Sample Question 3: How has the Brexit issue affected the British parliaments effectiveness at governing?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive and are spread across 6 tables. Your team should only use the tables 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 "play nice" with each other.

job_listings

Public job listings and associated company information, from 2014 – 2019.
~2.7 million rows & 17 columns. Size: ~740MB. Source: non-public.

employment_by_occupation

UK employment data by occupation, status, and sex. Data is estimated from the Labour Force Survey in Quarter 2 (Apr - Jun) of 2011 – 2018.
29,952 rows & 10 columns. Size: ~2MB. Source: Office for National Statistics

labor_market_statistics

UK labor market statistics from the Labour Force Survey from 1992 – 2019. Refer to *labor_market_stats_legend.csv* for indicator codes and units.
324 rows & 254 columns. Size: ~0.5MB. Source: Office for National Statistics

immigrant_statistics

UK citizenship applications by country of nationality from 2007 – 2018. Refer to *immigrant_legend.csv* for more information.
47 rows & 263 columns. Size: ~0.1MB. Source: Office for National Statistics

lse_historical_data

Historical London Stock Exchange (LSE) company data from 2013 – 2019. Market cap data is reported monthly whereas P/E, EPS, and P/S data is annually since 2014 in most cases.
158,587 rows & 11 columns. Size: 15MB. Source: London Stock Exchange and Financial Times

uk_bill_data

UK Public General Acts, by content and size of bill. All UK legislation passed from 1996 – 2019.
890 rows & 7 columns. Size: ~10MB. Source: UK Legislation

Additional Datasets

You are welcome to scour the Web for custom datasets to supplement your analysis. A couple of places to start are (1) https://ec.europa.eu/eurostat/data/database, (2) https://data.gov.uk, or (3) https://www.ons.gov.uk/, which have many datasets across business, economy, employment, and population data in the EU and UK. All additional data used should be public and should not exceed 2GB unzipped (consult Correlation One's technical product team either in person or 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. 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, which will be judged as follows:

- **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?
- **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 hypotheses 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?

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 raw LaTeX source file.

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 3:30PM. Any submissions received after that time will NOT be evaluated by the judges**.

Tips & Recommendations

You will have ~12 hours total to work on the problem statement. However, you will not have access to the actual data until the morning of the competition. As such, we recommend you split your time as follows:

- Friday evening, ~7:00PM – 12:00AM: You will receive a copy of the problem statement, data table schema, and data table heads. This gives you the opportunity to study the available data fields, think about suitable questions to tackle, and plan out your exploration process. Additionally, the data table heads should be sufficient for you to begin putting together some data wrangling & cleaning scripts.
- Saturday, 8:30AM – 3:30PM: You will receive the actual data. If you set up your data munging scripts already, you should be able to quickly apply them and immediately begin working with the data. You should spend most of your day investigating the data, performing qualitative & quantitative analysis, and writing up your process & results.

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 appertain 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 3 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.