TheDataOpen

Presented by Citadel and Citadel Securities
In Partnership with CorrelationOne

Problem Statement

Welcome to the 2020 Correlation One Summer Invitational 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 Olympics games is a quadrennial gathering of the world's best athletes broken into distinct summer and winter games. The modern summer games started in 1896 and their winter counterpart began in 1924. Being the host of an Olympic games is often a source of national pride and is further seen as an investment in the host city due to the increased infrastructure spending and expected boosts to tourism. Because of these perceived benefits, hosting the Olympics became highly competitive, with host cities promising extravagant infrastructure spending to secure their bids. However, these promises resulted in soaring costs. This, coupled with scandals at the 1968 and 1972 Olympics, gave rise to public skepticism about hosting the games. Voters in Dever then went on to reject the city's winning 1976 bid, handing those games to Montreal who took 30 years to pay off the debt associated with hosting its games.

Los Angeles was the only city to place a bid for the 1984 games, allowing them to host without making extravagant spending promises. This also coincided with a rise in Olympic revenues. Since the 1980s, distributing broadcast rights, ticket sales, and awarding sponsorships allowed host cities to generate notable revenues. These factors allowed the Los Angeles Olympics to turn an operating profit of $215 million. This success led to renewed interest in hosting the games, but this was short-lived with the recent Beijing, Sochi, and Rio games each costing over $20 billion. Many of the candidates for the 2024 and 2028 games have withdrawn their bids because of these unjustifiable costs, leaving only bids from Paris and Los Angeles to host this pair of games.

Further background reading:

- The Council on Foreign Relations on the Economics of Hosting the Olympic Games
- Several Research Gate posts provide information on the revenue from the Vancouver and London games and the costs of hosting each olympic games since 1988.

- The International Olympic Committee's Factsheet of the 2012 games, outlining several key facts and figures about the games.

- This Wikipedia page describes the bidding timeline and process of past Olympic games.

- The context surrounding the Rio 2016 Olympics - the outbreak of the Zika virus and the impeachment of the president Dilma Rousseff in August 2016.

Your Task

Your goal is to analyze Olympic host city/country data (described below), potentially in combination with supplementary datasets, in order to increase our understanding of how hosting the Olympics has impacted the respective hosts.

We've provided several pre-cleaned datasets relating to the Vancouver 2010, London 2012, and Rio 2016 Olympics, but you are not limited in your analysis to those Olympics specifically.

You are asked to pose your own question and come up with your own answer 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 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.

Sample Question 1: Is there a significant difference in the economic results of hosting a summer vs. a winter Olympics?

Sample Question 2: Considering factors beyond direct costs and revenues from hosting the Olympics, can a host reasonably expect to be in a better economic position for having hosted?

Sample Question 3: Did the Olympics uniformly affect the boroughs in London? Did they revitalize a specific part of the city?

Sample Question 4: Brazil simultaneously experienced a health and political crisis during the Olympics. How were these circumstances reflected in the socioeconomic impact of the Rio Olympics compared to other games?

Datasets

The provided datasets are stored in the “Datathon Materials” folder on Google Drive and are broken into 3 buckets based on which Olympic games they correspond to. 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 be readily usable in most popular data analysis libraries.

London Datasets

UK_international_visits

Tracking of international visitors to the UK considering duration of stay and estimated spend extrapolated from survey results. Covering from 2002 - 2019.
549,853 rows & 11 columns. Size: ~55MB. Source: Visit Britain

london_earnings_by_borough

Average earnings by week or hour based on the borough of residence, covering from 2002 - 2019. Data from the Annual Survey of Hours and Earnings.
6,768 rows & 7 columns. Size: ~318KB. Source: Office for National Statistics

london_economic_activity

Workforce participation segmented by gender and borough, covering from 2005 - 2019. Created from the Annual Population Survey.
1,470 rows & 8 columns. Size: ~97KB. Source: Office for National Statistics

london_infrastructure_spending

Breakdown of direct expenditures for the London games, including the change in cost estimates over development time.
59 rows & 6 columns. Size: ~4KB. Source: The Guardian

london_sports_participation

Breakdown of the amount of exercise people in London get by number of exercising sessions. Given at the borough level and covers from 2008 - 2016.
1,290 rows & 6 columns. Size: ~60KB. Source: Sport England

london_taxpayer_income

Breakdown of the mean and median annual income of taxpayers based on what borough of London they reside in. Covering from 1999 - 2018. Data is from the Survey of Personal Incomes.
846 rows & 6 columns. Size: ~39KB. Source: HM Revenue & Customs

london_ticket_sales

Sales figures for select Olympic events, outlining how many tickets were sold and at what price for each event. These sales do not include group tickets (such as those purchased by the families of athletes).
938 rows & 5 columns. Size: ~29KB. Source: London Organising Committee of the Olympic & Paralympic Games

london_tickets_for_sale

Outlines all tickets that were available for sale at the Olympic games and what their pricing was. This dataset does not indicate sales figures, only the sales offerings.
1,128 rows & 15 columns. Size: ~129KB. Source: The Guardian

UK_inflation

UK inflation data using 2018 as a base year, covering back to 1751.
268 rows & 3 columns. Size: ~4KB. Source: Kate Rose Morley

london_underground_activity

Tracking of entry and exit figures from London's various underground subway stations based on the type of day. Covering from 2007 - 2017.
2,953 *rows & 11 columns*. Size: ~207KB. Source: Transport for London

london_underground_station_info

Information about London's underground stations. It outlines what local authority the station belongs to, in what zone(s) it operates and when the station first opened.
269 *rows & 5 columns*. Size: ~14KB. Source: Transport for London

Vancouver Datasets

vancouver_tourism_indicators

High-level indicators relating to tourism in Vancouver, including revenue, GDP, employment, salarties, and regional estimates. Covering from 2000 - 2015.
16 rows & 14 columns. Size: ~2KB. Source: British Columbia

vancouver_employment_by_industry

British Columbia employment by North American Industry Classification System (NAICS) industry. Please refer to *vancouver_employment_by_industry_legend.csv* for the corresponding industry names. Covering from 1987 - 2019.
4654 rows & 3 columns. Size: ~116KB. Source: British Columbia

vancouver_business_size

Business locations by employee size, by region and industry. Covering from 1999 - 2019.
15,909 rows & 10 columns. Size: ~882KB. Source: British Columbia

vancouver_small_business_by_region

Small business counts segmented by region. Covering from 2007 - 2018.
85 rows & 3 columns. Size: ~1KB. Source: British Columbia

vancouver_visits

Monthly overnight visitors to Vancouver. Covering from 1994 - 2018.
301 rows & 3 columns. Size: ~1KB. Source: Tourism Vancouver

vancouver_room_revenues_2000_2010

Room revenues, property counts, and room counts by municipal jurisdictions. Covering from 2000 - 2010.
58,212 rows & 7 columns. Size: ~1MB. Source: British Columbia

vancouver_room_revenues_2010_2019

Room revenues by municipal jurisdictions. Covering from 2010 - 2019. **Note**: the region codes were manually mapped to best match up with the 2000 - 2010 data, but discrepancies may arise.
5,280 rows & 5 columns. Size: ~147KB. Source: British Columbia

region_codes

Mappings of region code and region names, as well as their relationships and descriptions.
75 rows & 4 columns. Size: ~3KB. Source: British Columbia

Rio Datasets

brazil_gdp

Gross domestic product by region. Covering from 2002 - 2017.
2,193 rows & 4 columns. Size: ~100KB. Source: Link

brazil_monthly_income

Nominal average income, usually received per month from the main job, segmented by occupational groups. Covering from 2012 - 2020.
9,802 rows & 5 columns. Size: ~659KB. Source: Link

brazil_unemployment

Unemployment dataset, by employment type and state. Covering from 2012 - 2019.
2,674 rows & 5 columns. Size: ~118KB. Source: Link

brazil_tourism_jobs

Government dataset from IPEA (Institute of Applied Economic Research) with data about the population working in tourism roles. Covering from 2006 - 2018.
33,661 rows & 6 columns. Size: ~2MB. Source: Link

brazil_international_arrivals

Government dataset from Brazil's Ministry of Tourism detailing the number of international tourists that arrived in Brazil. Covering from 1989 - 2018.
32,761 rows & 6 columns. Size: ~689KB. Source: Link

Additional Datasets

You are welcome to scour the Web for custom datasets to supplement your analysis. Here are a couple of places to find datasets across business, economy, employment, and population:

- **London**: starting point 1, starting point 2, starting point 3
- **Vancouver**: starting point 1, starting point 2, starting point 3
- **Rio:** starting point 1, starting point 2, starting point 3, starting point 4

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 on Sunday, July 19th. 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 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.