APAC 2023 Datathon: Roadway Vigilantism

Presented by Correlation One

Problem Statement

Welcome to the 2023 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

In the world of safety, the City of Philadelphia is not regarded. In fact, Philadelphia has one of the highest crime rates among cities with >1 million people within the USA. On top of this, Philadelphia has one of the highest auto-fatalities rates within the USA even though it has one of the lowest car ownership demographics for a city of its size.

For these reasons, it makes a perfect case study to practice your vigilantism!

Today you are all tasked with becoming data driven vigilantes! The Correlation One staff has presented you all with 5+ years of crime, traffic stop, car crash data and more, from the city of Philadelphia. You are tasked with the requirement of providing novel insights or practical recommendations on how to improve road safety andor crime mitigation within the city!

Your Task

You are asked to pose your own question and answer it using the available datasets, as well as any supplementary datasets that you find to aid your analysis. Both the creativity of your question, and the quality of your analysis are of paramount importance. **You need not be comprehensive; depth of insight is more important than breadth of question posed**.

Submissions may be predictive, using machine learning to classify or predict patterns. Submissions may also be illuminating by way of data visualization or sound statistical inference.

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 rigor, which are far more important.*

Sample Question 1: How can we better allocate the locations of presence of Philadelphia traffic police to help mitigate crime?

Sample Question 2: What types of car crashes can most easily be forecasted? How can the Philadelphia traffic police be used to mitigate these?

Sample Question 3: Are there any relationships between crimes, traffic stops and the number of car crashes that occur in a given area? What implications can be derived if a relationship exists?

Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive. Your team need only use the data / datasets that are relevant to your chosen question / topic.

Traffic & Investigations

traffic_stops_philadelphia.csv – A dataset of traffic stops in the city of Philadelphia from December 2013- April 2018.

investigations.csv – A dataset of investigations taken by Philadelphia police between 2014-2022 on pedestrians and automobiles.

crimes.csv – A dataset of crimes committed in Philadelphia from 2013-2022.

police_stations.csv – A dataset of the location details of police stations within Philadelphia

police_districts.csv - A dataset containing the location of details of police districts within Philadelphia

Crashes:

crash_info_general.csv – Information related to car crashes in Philadelphia from 2010-2021.

crash_info_commericial_vehicles.csv – Information related to car crashes with commercial vehicles in Philadelphia from 2010-2021.

crash_info_motorocycle.csv – Information related to car crashes with motorcycles in Philadelphia from 2010-2021.

crash_info_people.csv – Information related to the people in car crashes in Philadelphia from 2010-2021.

crash_info_roadway.csv – Roadway information related to car crashes in Philadelphia from 2010-2021.

crash_info_trailed_vehicles.csv – Information related to the trailers in car crashes in Philadelphia from 2010-2021.

crash_info_flag_variables.csv – General information related to car crashes in Philadelphia from 2010-2021 but encoded with flag variables.

crash_info_vehicles.csv – Information related to specific cars involved in car crashes in Philadelphia from 2010-2021.

Other:

philadelphia_population_metrics.csv – Population demographic information of the city of Philadelphia.

hourly_weather_philadelphia.csv – A dataset of hourly weather in Philadelphia from 2010-2022. Data is the weather from the centroid of the city.

Additional Datasets

Participants are welcome and **encouraged** to scour the internet for their own custom datasets to supplement their analysis. All additional data used should be public and reputable. Additionally, any supplementary datasets should not exceed <2 GB unzipped (consult Correlation One's R&D team if you believe your idea is worthy of an exception).

Other Materials

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

Submissions: Content

Submissions should have two 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 are their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged to help explain your thought process
   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/or 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, otherwise 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: Evaluations

The competition will have multiple rounds of evaluation. Your Report will judged as follows:

- **Technical Executive Summary**
  - o *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose that question? Are your conclusions precise and nuanced, as opposed to over-generalizations?
- **Technical Expositions**
  - o *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.
  - o *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 those tests and analyses? What patterns did you notice, and how did you use these to make subsequent decisions?
  - o *Analytics & Modeling Rigor.* What assumptions and choices did you make, and how did you justify them? How did you perform feature selection? If you build 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 models you build, and what did you tell you?

Submissions: Formats

Make sure to send your submissions in an email to **datathonsubmissions@correlation-one.com**.

The email subject line should be in the following format: "Team [n] Submission - APAC Datathon 2023" where [n] is your team number.

Only one of your team members should send us the email (feel free to CC your teammates). You can resend the email before 11:59PM if you would like to make any changes - we will look at your latest submission before the deadline.

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.

Please 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.

Code should be submitted in a single zipped collection of files separate from your report.

Please ensure to give yourselves extra time to send your submissions, especially if the size of your files is large.

Tips and Recommendations

Since this is a single day Datathon, time is of the essence. It will be important to settle on a tractable question early on to ensure that you have time to both thoroughly explore the problem that you pose and write up your results. The outcome of this Datathon, and your overall success, will largely be a product of the quality of the question that you choose to answer and the depth within which you explore that question.

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 stick to tools and techniques that you have previously used. Learning new skills is certainly valuable, but it can consume a large portion of your available time, leaving less time for completing the task at hand.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls that successful teams will actively avoid. 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 communications through both visualizations and writing	**2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so it is best to make sure those conditions are sufficiently met
3. Engage in proper causal analysis. Just because your model passes standard cross-validation checks does not demonstrate (or even necessarily 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 true or worthwhile

Ask for Help

Correlation One’s R&D 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 forward.