RICE | D2K
Data to Knowledge Lab

Rice Data Science Consulting Clinic

Client Report

March 20, 2021

Client Information

Clinic Date:	8 Feb 2021
Client Name:	Arko Barman
Client Email:	arko.barman@rice.edu
Client Department or Affiliation:	Rice EE and D2K Lab
Returning Client:	No

Consultant Team

Duarte "Eddie" Costeira	Masters	COMP	duarte.costeira@gmail.com
Em Gamboa	BA	STAT	evg1@rice.edu
Armin Khamoshi	PhD	PHYS	ak88@rice.edu
Pranav Khemka	BS	BIOE	pbk2@rice.edu
Howard Lan	Masters	COMP	dw52@rice.edu

Problem & Objectives

Client Problem

On a typical day, police make over 50,000 traffic stops in the United States. The Stanford Open Policing project has gathered records from millions of traffic stops over multiple years from over 40 states. The data is available for download from:

https://openpolicing.stanford.edu/data/.

The client wants to use Stanford's datasets to address the question of whether there is systemic bias present in the policing system in the US. Initially, the client intends to analyze stop data for the state of Texas, since it is easier to reach out to legislators and request information. Long-term, the client intends to analyze the data for all the states to come up with a more significant conclusion for their study.

Primary Objectives

Our client had two objectives:

1. How to open and process the data given that the dataset is large (millions of rows)? Specially, what are some of the libraries and software that could make the data processing easier (preferably in Python)?

2. Brainstorm ideas on how to use the dataset to test for evidence of systemic bias in policing in novel ways.

Data Science Problem

The problem has two parts: (1) The client needs some good tools to create a pipeline through which the dataset can be fetched and preprocessed for further analysis. The main challenges are that the dataset could be large and that the pipeline should be sufficiently robust to handle different datasets arising from different locations. To this end, we suggest some software packages in python that the client can consider to accomplish this goal.

(2) The client's aim is to use the data provided by openpolicing.stanford.edu/data to determine whether systemic bias is present in the policing system in the US. To that aim, they require novel hypotheses to support/disprove the theory. Some hypotheses regarding systemic bias in policing that have been answered in previous studies include:

- Determining the effect of race on the rates at which individuals are stopped
- Checking if the time of day influences the rates at which individuals of different races are stopped
- Analysing whether the proportion of steps that result in arrest or search vary by driver race
- If stop rates by driver race have changed over time or in response to new policy implementation

In this report, we suggest other hypotheses (and relevant tests) that the client could consider to test for systemic bias in policing.

Recommendations

1. Data Processing

   (a) DB Browser for SQLite: SQL is a language used for managing data in relational database systems (usually in a dedicated server). SQLite is a language library that implements a local, self-contained SQL database. While we recommend the use of R and Python for data analysis, SQLite offers a solution to the problem of storing and opening very large datasets in a stable manner. There exist command-line interfaces for SQLite as well as GUI applications (we recommend DB Browser

for SQLite: https://sqlitebrowser.org/). Furthermore, both Python and R can interface directly with SQLite via the `sqlite3` module in Python, and the `RSQLite` package in R, allowing either language to send SQL commands to the database.

Recommended Reading:
https://www.sqlitetutorial.net/sqlite-python/
https://docs.python.org/3/library/sqlite3.html

(b) PySpark: The client can consider using Apache Spark to open and process the dataset. Spark is a data analytics engine that is made specifically to handle large datasets. It has a highly parallelized backend with high-level APIs available in multiple programming languages including Java, R, and Python. In particular, PySpark is the API package for python which features multiple subpackages for data processing and building machine learning models. For example `pyspark.sql` allows the passing of SQL like queries as texts to python functions which might be advantageous if there is an exiting program that preprocesses the data. The documentation and programming guides can be found at: https://spark.apache.org/docs/3.0.1/api/python/

Note that it is not necessary to create the entire pipeline in PySpark. The client might prefer to use PySpark for the initial data cleaning and exploratory data analysis, then pass a smaller and more manageable dataset to other packages (e.g. Pandas or scikit-learn) for which more user-friendly APIs and extensive documentation is available.

2. Hypotheses Choices and Related Tests

We would recommend performing some initial exploratory data analysis to determine if the data is normally distributed. If the data is normally distributed, we recommend using parametric tests such as one-way ANOVA, coupled with paired t-tests to determine if there is a significant difference in the test groups.

For the case that the data is not normally distributed, for each of the hypotheses, we recommend using the Kruskal-Wallis one-way ANOVA test, since it is a non-parametric test (it does not assume your data comes from a particular distribution). For pairwise testing, we recommend using the Mann-Whitney test, since it is also a non-parametric test. In general, parametric tests are more robust than non-parametric tests, so if the data is normally distributed, we recommend using parametric tests.

Recommended Reading:
https://www.statisticshowto.com/kruskal-wallis/

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html

(a) *Hypotheses*: The rates at which citations are issued for traffic stops is different for different driver races.

Rationale: A difference in the rates at which citations are issued by race could indicate systemic bias in policing.

Explanation: Divide the data by driver race, and calculate the proportion of citations issued for each driver race for every stop location and time period. The proportion can be expressed as number of citations issued per 100 drivers stopped of a particular race. Use the recommended tests depending on the distribution of the data to check if the rates of citations being issued is significantly different for any one driver race than the others.

In the case that at least one of the groups is significantly different from the others, we recommend conducting paired t-tests or pairwise Mann-Whitney tests to pinpoint which driver race pairings account for the significant difference.

(b) *Hypotheses*: The rates at which certain reasons are issued for traffic stops are different for different driver races.

Rationale: The intuition is, if systemic bias exists, we would expect a large variation in reasons seen for one race, but less for the other races (indicating absurd reasons).

Explanation: Categorise the data by 'reason for stop', and count each reason by driver race. Depending on the intention, the client could either aggregate the counts across locations and times (and risk losing information for a bigger picture view), or they could keep the counts separate by location and time, preserving information but requiring a larger number of statistical tests to be conducted to disprove the null hypothesis.

We recommend using either one-way ANOVA or the Kruskal-Wallis one-way ANOVA to test for any significant difference in reasons cited for stopping drivers of certain races. We also recommend using paired t-tests or pairwise Mann-Whitney tests to further pinpoint which driver race pairings for a given stop reason are significantly different.

(c) *Hypotheses*: The rates at which searches are conducted, and the proportion of searches that turn up contraband are different for different driver races.

Rationale: Let's take 2 races, A and B. If A is searched 90% of the time, and turns up contraband 10% of the time, while B is searched 10% of the time and turns up contraband 10% of the time, that could indicate systemic bias in policing.

Explanation: Divide/Categorise the dataset by driver race. Determine the proportion of stops that result in a search being conducted and express this as searches conducted per 100 drivers stopped. From this, further determine the proportion of stops that turn up contraband. We recommend separating the results for different stop locations and stop times. To determine if a difference in proportions is significant, we recommend using either one-way ANOVA or the Kruskal-Wallis one-way ANOVA test.

If the Kruskal-Wallis test indicates at least one race pairing is significantly different, we recommend conducting paired t-tests or pairwise Mann-Whitney tests to indicate which pair comparisons are yielding significantly different results.

(d) *Hypotheses*: The rates at which citations are issued for a given stop reason are different for different driver races.

Rationale: If the rates at which citations are issued is different for a given stop reason for drivers of different races, then it could indicate systemic bias in handing out penalties.

Explanation: First divide the dataset by reason for stop. Then divide it by driver race. For a given combination of stop reason and driver race, determine the proportion of stops that result in a citation being issued, and present it as citations issued per 100 stops. The data can be presented for different stop reasons, comparing the proportion of citations issued for each driver race. The data can be further separated by location and time. We recommend performing either one-way ANOVA or a Kruskal-Wallis one-way ANOVA on each stop reason dataset for a given time and location to determine if there is a significant difference in citation issuance rate for driver races. If there is a significant difference detected, we recommend proceeding with paired t-tests or pairwise Mann-Whitney tests to identify particular driver race pairings that are significantly different.

Summarised recommendations:

We suggest DB browser for SQLite and PySpark to open and process the large CSV files containing traffic stop data. We recommend 4 novel hypothesis that the client could test with the data:

1. The rates at which citations are issued for traffic stops is different for different driver races
2. The rates at which certain reasons are issued for traffic stops are different for different driver races
3. The rates at which searches are conducted, and the proportion of searches that turn up contraband are different for different driver races
4. The rates at which citations are issued for a given stop reason are different for different driver races

For each of the tests, we recommend using either one-way ANOVA or the Kruskal-Wallis one-way ANOVA to determine if any of the driver race pairings are significantly different from each other. To determine which of the pairings are significantly different, we recommend using paired t-tests or the Mann-Whitney test. For all tests, we recommend using a significance level of 0.05.

A word of caution, the Stanford Open Policing dataset is not exhaustive. It does not contain information about location demographics, demographics of the police officers performing the stops, etc. We recommend looking into supplementary datasets to complement the information obtained from Stanford Open Policing if possible, to counter some of the limitations that may arise from interpreting the results of the current study.