EXPLORATION OF THE STANFORD OPEN POLICING DATASET ACROSS UNITED STATES: A DATA ANALYSIS PROJECT PROPOSAL

Linni Qin

Student ID: n9632981

Supervisor: Dr. Dimitri Perrin

Unit Code: IFN702

Project Category: A Research Project

QUEENSLAND UNIVERSITY OF TECHNOLOGY

Table of Contents

I. Project Introduction

1.1 Background and Context

This project is going to explore the Stanford open policing dataset across the United States and derive the valuable insights and hidden relationships between police and the public. The whole dataset consists state-level information of daily patrol stops with around 130 million electric records totally ranging from 2004 to 2016 accordingly (Stanford Open Policing Project, 2017). As per the description of the Stanford Open Policing Project, the raw data are collected and gathered by American state and local law enforcement agencies (2017). Meanwhile, the cleaned and standardized data are open freely for the public when the Stanford Open Policing Project launched its website (openpolicing.stanford.edu) on June 19th of 2017. The Stanford Open Policing Project team consists seven researchers respectively from both Stanford Computational Journalism Lab and Stanford School of Engineering since 2014, which is leading by Professor Cheryl Philips (Stanford Libraries, 2017). Their purpose is to supply a platform for any academic researchers, policing policy makers and social journalisms with contributing their efforts together in order to improve police practices in the US (Stanford Open Policing Project, 2017). So far, there is one attribute of the dataset – "Driver Race" has been analysed by Stanford research team and one academic paper has been published with suggesting that police officers are significantly easier to conduct the search and arrest with Hispanic and Black motorist when they pull over the drivers (Emma, P. et al., 2017).

1.2 Problem

Four problems have been detected based on the existing dataset and findings.

- Among 50 states of the United States, there are 31 states have provided the log data and other 19 states claimed that they did not collect the electronic data. If the other 19 states could participate in gathering the data, the researcher might be able to draw larger conclusions about the state of policing in the US.
- Furthermore, over 50,000 motorist are pulled over by police every day and respective interactions are recorded into the various police log formats in different states. With the data cleaning by the Stanford researcher, the raw dataset has been standardised into 11 features for each state and each record as *Table 1*. As mentioned in last section, "Driver Race" has been analysed and there are still 10 features require investigating.

Table 1: Eleven Necessary Features of Open Policing

State	id	Stop Date	Stop Time	Stop Location	Driver Race	Driver Gender	Driver Age	Stop Reason	Search Conducted	Search Type	Contraband Found	Stop Outcome

- It can be learned from a survey conducted in 2016 as Figure 1 that "Texting" is the most common driving behaviour which most likely to be annoying in America (Statista, n.d.). In the meantime, Law Offices of Michael Pines claims that distracted driving such as talking on a phone or sending a text is the top causes of motor accidents in the United State today (n.d.). Therefore, it can be estimated that if "Texting" takes the majority of the "Stop Reason" based on the policing dataset.

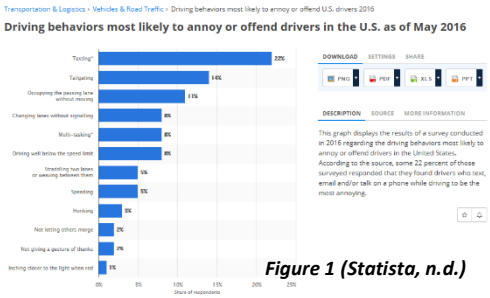


Figure 1 (Statista, n.d.)

Table 2: US Age Structure (IndexMundi, 2018)

Unite States Age Structure	
Age Groups	**Percentage by Total Population**
0-14 years	**18.73%**
15-24 years	**13.27%**
25-54 years	**39.45%**
55-64 years	**12.91%**
65+ years	**15.63%**

- The last issue demands to be figured out is the distribution of common offenses by the drivers. According to the age structure statistics provided by IndexMundi in 2018 as *Table 2*, the majority of the offender might be adult drivers with age from 25 to 54, since their population share the biggest proportion in the country. To draw a comparison with more accurate information, the third party state-level sources about the distribution of drivers by ages in US should be essential for the project.

1.3 Purpose

The purposes of this project is to examine the possible contribution to the platform of Stanford Open Policing project by investigating the offending behaviours among different aged drivers in the United Sates. In general, the main objectives of this project are:

- To claim an effective project management approach, a data analysis workflow and a collection of tools and skills for data mining;
- To examine the correlation between the driver ages and the different traffic citation;
- To determine the distribution rate of traffic offenders with comparing with the state-level and year-level demography based on different age groups;
- To identify a well-fit prediction model of patrol stop trend and determine the proper data size for model training and testing with applying advanced evaluation algorithms;

1.4 Scopes

Given the four research problems, *Table 3* represents the significance of the requirements through the whole project with applying prioritisation tool MoSCoW. "Must", "Should" and "Could" requirements would be determined in the scope and "Would" items should be out of the scope as the limited time factor for this data analysis task (Business Analyst Learnings, 2013). Those items will be scored as well to be allocated into weekly sprint tasks. Sprint tasks can be referred to section three *Project Management Approach*.

Table 3: MoSCow Prioritised Requirement List

ID	*Requirement List*	*Deliverables Priority*	*Effort Points*	*Reason*
1	As the project generator, the Stanford Open Policing team would like to apply the dataset on the website in the research.	Must (Guaranteed)	2	The aim of this project is to investigate those datasets.
2	As a data analysis, selecting the suitable size of the dataset and the targeted features, gathering the proper third party data such as demography and the number of driver license holder are essential.	Must (Guaranteed)	23	These data are solid foundation to implement the analysis.
3	Analyse the collected data to find out the common distribution pattern so as to draw larger conclusions.	Should (Expected)	18	The research assumption is aimed to discover the depth insight of those datasets.
4	To compare the rate of different driving behaviours offender such as speeding, DUI, seat belts based on various age groups.	Should (Expected)	20	Data analysis report should be persuasive and detailed.
5	A forecast model to estimate the offenses rate each year based on age groups might be helpful for other researchers, policy maker such as United States Department of Transportation	Could (Possible)	17	The ideal prediction model could definitely predict the required offenses rate and distribution trends. However, its effectiveness should be evaluated accordingly.
		In-Scope Points:	***80***	
6	As the project generator and promoter, the Stanford Open Policing team wants the participants to do contribution with analysing as many features as possible to drive varied value insights.	Would (Maybe)	20	As the time limitation in one semester, it has less chance to go over all the features, except focus mostly on "Driver Age" and "Search Reason".
		Out-Scope Points:	***20***	
		Project Total Points:	***100***	

1.5 Approach Overviews

Table 4 indicates the commonly four phases with totally eight steps involved in the workflow for facilitating data analysis project. The four phases are data preparation, data analysis, results reflection and results dissemination. Data preparation is the time-consuming part in the workflow. However, it plays as the fundamental base for further analysis. Data analysis is the core activity to execute the parameter and analyse the data for obtaining the insightful information. The phases of results reflection and dissemination will determine the quality for the final outcomes by continually adjusting the experiments with collecting the helpful feedbacks from the supervisor and comparing various outputs value. As to the details of the project execution approach, it can be referred to section two *Project Methodology*.

Table 4: Data Analysis Workflow

Phase One ***Data Preparation***	***Phase Two*** ***Data Analysis***	***Phase Three*** ***Results Reflection***	***Phase Four*** ***Results Dissemination***
Step 1: Defining problem Step 2: Identifying ideal datasets to answer analysis problems Step 3: Acquiring data Step 4: Cleaning data	Step 5: Exploring data Step 6: Statistical prediction and modelling data	Step 7: Interpreting results	Step8: Communicating and distributing results

1.6 Outcome and the Expected Significance

Two main tangible outcomes will be returned in the end of this data exploration task. One of them is a data analysis report that represents the depth insight of the Stanford open policing datasets building on a considerably large amount of patrol stops data across 31 states in the United States, along with the third party demography data. It is going to demonstrate the visualized relationship patterns and detailed statistic reports between the drivers in different age groups and the various driving offenses they commit. The other one is a rational prediction model of patrol stop trend relying on the driver structures and their common driving behaviours.

The main expected significances of the relating outcome from this project will be able to:

- Improve the performance of data collection, data report and data analysis among the state and local law enforcement agencies;

- Contribute to the academic research, social commentary related to the policing data in the near future;
- Facilitate the policy maker and the police practices in diver communities across the country;
- Develop the driver guideline including safety tips for diverse driving groups and improve the traffic warning signs;

II. Project Methodology

2.1 Programing Tools: R and RStudio

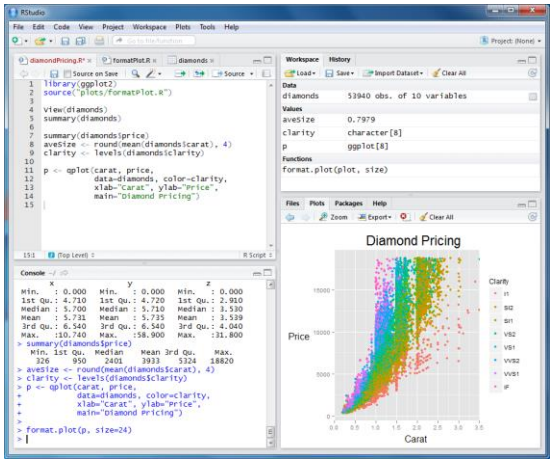


Figure 2: RStudio Interface

R language and RStudio will be the data analysing and code compiling tool for this project as R is an open source development environment for statistical analysis and visualization (R, n.d.). Also, as shown as *Figure 2*, RStudio is integrated with four windows with console, editor and a wide range of visual plotting tool (RStudio, n.d.). These environments are effective for the data analysis.

Given the experience of unit IFN701 with successfully implementing the dataset with millions of real estate dataset and building a multiple regression forecasting model, the computing power of desktop computer at QUT and personal laptop are workable to handle Open Policing dataset. Meanwhile, the data analysis methodologies and techniques taught by Dr. Guido Zuccon in his lecture of unit IFN509 – Data Manipulation in 2017 at Queensland University of Technology will be applied in facilitating this project. The following subsections will explain the methodologies mentioned in *Table 4: Data Analysis Workflow* in detail and present the relating justification.

2.2 Phase one: Data Preparation

Step 1: Defining the problem

Before proceeding, it needs to figure out the categories of data analysis questions that are to be solved. As per the search problems and the purposes, the project is to define the correlation and distribution between the driver ages and the different traffic citation, then to identify a well-fit prediction model of patrol stop trend. Thus, this project is going to solve a hybrid analysis problem - an inferential and predictive analysis.

Step 2: Identifying the ideal dataset to answer the analysis problem

Although there are 130 million of records, the ideal dataset will be the states that provide value for feature “Driver Age” and “Stope Reason”. Then, the third party statistic sources about the driving aging distribution is highly demanded. This step is essential to implement the data analysis.

Step 3: Acquiring the data

Valid data will be acquired from the Stanford Open Policing Project website and other official channels. Those data will be stored in working computer hard disk and other stable backup working stations. A couple of risks related to data storage will be explained in detail in section five *Risk Management Plan*.

Step 4: Cleaning the data

Most datasets contain value for 11 features at least. To reduce the size of operating dataset, other irrelevant data should be cleaned. The executing functions for data munging can be R packages such as plyr and dplyr (Hadley, W., n.d.).

2.3 Phase two: Data Analysis

Step 5: Exploring the data

Starting from this step, it is going to explore the hypotheses for the relationships between the driver age and their stope reasons. The functions can be used with R packages, like ggplot to plot or cluster the datasets (Hadley, W., n.d.).

Step 6: Statistical prediction and modelling the data

Implement the research to find out prediction models, which are appropriate to forecast the development trend. Deploy the models into the experiment and monitor their performance. Repeat the cycle of model building, model deploying and model monitor until finding the most effective one.

2.4 Phase three: Results Reflection

Step 7: Interpreting the results

Analysis and reflection phases should be frequently alternative, especially when making comparisons among the various outputs. Adjust the code and parameters to rerun the processes with testing data. Also, discuss the outputs with the supervisor to get his helpful feedbacks.

2.5 Phase four: Results Dissemination

Step8: Communicating and distributing the results

The final step is to disseminate the results. The data analysis report generated in R Markdown format will visualize the relationships of the targeted features and provide the statistic information while comparing with third party sources. The trend forecast model will be able to estimate the future development of driver behaviours properly.

III. Project Management Approach

3.1 Scrum

Scrum will be adopted as project management approach for this project. The essential reasons are highlighted as followings:

- With applying Cynefin framework to analyse the project context, it is a complex project producing products that require “Probe-Sense-Respond” (MindTools, n.d.).

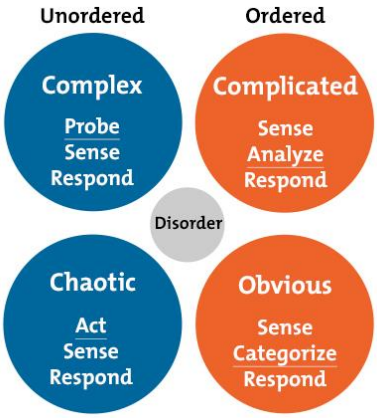
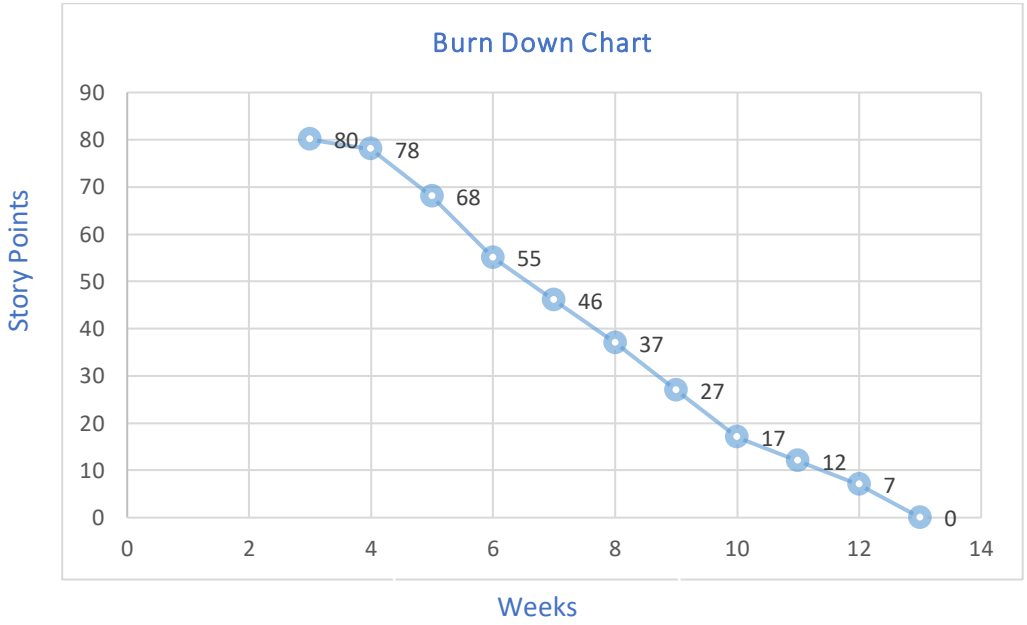


Figure 3: MindTools

- Transparent plan and visible progress made the communication effective between supervisor and me.
- “To do”, “done” “undo” backlogs enabled me to control the progress on the right track.
- Sprint retrospect allowed me to adjust progress as an increment.
- It enabled the project to be completed on time.
- It helped to reduce the risk of submitting failure outcomes.

3.2 Burn Down Chart and Sprint Backlog

With applying prioritised *Table 1: MoSCow Prioritised Requirement List*, the requirement list will be transferred into the product backlog and the estimated effort points will be the measurement scale for Scrum management. 80 story points of the project backlog will be allocated into each two-week sprint backlog until they are completed on time as indicated in the *Burn Down Chart*.



The detailed weekly plan to complete the target is demonstrated as *Table 5: Two-Week Sprint Backlogs*. One GitHub repository for this specific project will be used as the platform to record and manage the project processes, together with serving as a backup repository to save the works.

Table 5: Two-Week Sprint Backlogs

Sprint 1	***Sprint 2***	***Sprint 3***	***Sprint 4***	***Sprint 5***	***Sprint 6***
Week 3 – Week 4	***Week 5 – Week 6***	***Week 7 – Week 8***	***Week 9 – Week 10***	***Week 11 – Week 12***	***Week 13***
Points:2	***Points:23***	***Points:18***	***Points:20***	***Points:10***	***Points:7***
Meet with supervisor weekly; Define the problem and project methodology; Identify the ideal dataset;	Meet with supervisor weekly; Acquire the data; Clean the data; Explore the data;	Meet with supervisor weekly; Present the draw of data analysis report; Build forecast model;	Meet with supervisor weekly; Deploy and evaluate the prediction model;	Meet with supervisor weekly; Analysis results refection, prediction model refection;	Final Outcomes: Data analysis report, forecast model, final report.

IV. Communication Plan

As there are totally four students under the supervision of Dr Dimitri Perrin with handling the same dataset, even though every student will propose different research problems, it has great chance to share the mutual data manipulating techniques and problem-solving experience. Therefore, the stakeholders involved in this project will be supervisor, other three researchers, and me.

Table 6: Communication Plan

Phases		Aim	Attendee	Type of Communication	Location	Frequency	Duration
Preparation		Ensure the research assumption is workable, scopes and data are proper.	Supervisor, Other 3 researchers, Me	Face-to-Face	QUT GP S1201	Weekly	1 hour
		Ensure understanding of the dataset and their features is on the right track.	Supervisor, Other 3 researchers, Me	Online chat	Slack	Depends	Depends
					Email	Depends	Depends
				Face-to-Face	QUT GP S1201	Weekly	1 hour
Data Analysis & Result Reflection	Sprint Planning	Discuss the sprint goals	Supervisor, Other 3 researchers, Me	Online chat	Slack	Depends	Depends
					Email	Depends	Depends
				Face-to-Face	QUT GP S1201	Weekly	1 hour
	Sprint Review	Review the achievement of each sprint and adjust the sprint goals for next sprint if necessary	Supervisor, Other 3 researchers, Me	Online chat	Slack	Depends	Depends
					Email	Depends	Depends
				Face-to-Face	QUT GP S1201	Weekly	1 hour
Data Analysis & Result Reflection		Discuss the outcomes and update it if necessary	Supervisor, Other 3 researchers, Me	Face-to-Face	QUT GP S1201	Weekly	1 hour
Result Dissemination		Ensure the outcomes to be presented in the right way	Supervisor, Other 3 researchers, Me	Online chat	Slack	Depends	Depends
					Email	Depends	Depends
				Face-to-Face	QUT GP S1201	Weekly	1 hour

V. Risks Management Plan

With considering the risk occurrence probability and the consequence severity level, *Table 7: Risk Assessment Matrix* will be used to access the serious level of the potential risks and define the relevant mitigations. The green tables with low 1 represent the risk and consequence are acceptable without doing any response. The yellow tables with medium 2 show the risk and consequence are also acceptable but need certain kind of monitoring to keep it from being worse. For example, when there is missing demography data for in-scope states, it would be ideal to select the states with available driver data. The orange tables with high 3 mean the risk and the results are unacceptable then it needs to stop the project execution until finding the workable solution. *Table 8* will indicate mainly the orange unacceptable risks together with their mitigations. The last red table with extreme 5 show the immediate project close is a compulsory to prevent the catastrophe happening. For instance, the data available on Stanford Open Policing website is hacked and replaced with malware programs. This will cause catastrophe for the private workstation and QUT public network environment. When encountering this kind of extreme situation, the project should be stop immediately.

Table 7: Risk Assessment Matrix

Likelihood to happen ↑	Very likely (90% to occur)	Acceptable Risk Medium 2	Unacceptable Risk High 3	Unacceptable Risk Extreme 5
	Likely (50% to occur)	Acceptable Risk Low 1	Acceptable Risk Medium 2	Unacceptable Risk High 3
	Unlikely (20% to occur)	Acceptable Risk Low 1	Acceptable Risk Low 1	Acceptable Risk Medium 2
		Minor	Moderate	Major
		The severity of the risk →		

Table 8: Risks of Level 3 and Mitigations

Risks	Solutions
Loss of the work The hard disk of the private workstation or the public workstation at QUT might break down unexpectedly.	To prevent losing the efforts and the analysis results, it needs to store the work in a stable repository such as GitHub. GitHub does not only offer the free repository for the public account user, but also provide the users the tools and bridges to manage and share the code. Supervisor can also check regularly my progress of data analysis there.
Limited computing power The main workstation for the task is on private laptop and public computer at QUT. Their hardware specifications, such as the graphic processor and flash memory, are not capable to deal with GB level of data calculation and complex 3D plots.	Discuss with supervisor or together with the faculty to get more powerful resources and rational working environments if necessary.
Unavailable of Supervisor Supervisor might be on the sick leave or participating conference. When encounter this hardship especially on the key data analysis phase and refection phase, it is hardly to finish the task with high quality without the supports from the supervisor.	Alternative communication channels, such as Slack and email, are able to increase the chances of getting supports from supervisor. The other way is to get the supports from the project coordinator who holding the rich resources to solve the problem.

Reference

Business Analyst Learnings (2013, March 5). MoSCoW: Requirements Prioritization Technique. Retrieved from https://businessanalystlearnings.com/ba-techniques/2013/3/5/moscow-technique-requirements-prioritization

Emma, P., Camelia, S., Jan, O., Sam, C, Vignesh, R, Cheryl, P. & Shard, G (2017). A large-scale analysis of racial disparities in police stops across the United States. Retrieved from https://5harad.com/papers/traffic-stops.pdf

Guido, Z. (2017). IFN509: Data Manipulation – Week 1 Lecture. Retrieved from https://blackboard.qut.edu.au/bbcswebdav/pid-6726760-dt-content-rid-7967236_1/courses/IFN509_17se1/lecture_w1.pdf

Hadley, W. (n.d.). ggplot2 v2.2.1. Create Elegant Data Visualisations Using the Grammar of Graphics. Retrieved from https://www.rdocumentation.org/packages/ggplot2/versions/2.2.1

Hadley, W. (n.d.). plyr v1.8.4. Tools for Splitting, Applying and Combining Data. Retrieved from https://www.rdocumentation.org/packages/plyr/versions/1.8.4

Law Offices of Michael Pines (n.d.). Top 25 Causes of Car Accidents. Retrieved from https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/

MindTools (n.d.). The Cynefin Framework. Retrieved from https://www.mindtools.com/pages/article/cynefin-framework.htm

RStudio (n.d.). Why RStudio? The technology to amass data exceeds our abilities to make use of it. Retrieved from https://www.rstudio.com/about/

Stanford Libraries (2017). The Stanford Open Policing Project. Retrieved from https://exhibits.stanford.edu/data/feature/the-stanford-open-policing-project

Stanford Open Policing Project (2017). The Stanford Open Policing Project. Retrieved from https://openpolicing.stanford.edu/

Statista (n.d.). Driving behaviors most likely to annoy or offend drivers in the U.S. as of May 2016. Retrieved from https://www.statista.com/statistics/301157/most-offensive-driving-behaviors-in-the-us/

Appendix: Project Proposal Statement – Supervisor Sign-Off

I, Dimitri Perrin <*name of supervisor*>, confirm that I have gone through the project plan made by Linni Qin <*student name*> holding student ID number: n9632981 on the project titled: "EXPLORATION OF THE STANFORD OPEN POLICING DATASET ACROSS THE UNITED STATES" for IFN702 <unit code>.

I confirm that I have been consulted in deriving this project proposal and that I approve of the suggested scope and tasks described in this project plan and that I am satisfied with the identified risk mitigation and communication plans articulated here.

_____ _____

Supervisor signature Date