

## School of Computing Science and Digital Media

### Coursework Assignment

Surname	UGWU
First name	ARINZE
Matriculation Number	1813624
Contact phone number	07741944240
Course + Year	2019/2020
Module Co-ordinator	Kit Ying Hui
Module Number + Name	CMM524 Advanced Data Management
Coursework Title	<b>Analysing Datasets of Your Choice</b>
Coursework Part	Part 3 of 3
Due Date	22 November 2019 (16:00)
Feedback Due Date	By 30 December 2019


The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so. Further details are available at:  
[www.rgu.ac.uk/academicregulationsstudentforms](http://www.rgu.ac.uk/academicregulationsstudentforms)

**Declaration** \*\* This **must** be affirmed by adding your name below with the date of submission

**I acknowledge that by submitting the work, accompanied by this front cover, I take responsibility for the ownership of the submitted work.**

**I confirm:**

- that the work undertaken for this assignment is entirely my own and that I have not made use of any unauthorised assistance
- that the sources of all reference material has been properly acknowledged.

Student Signature	
Date Submitted	22 <sup>nd</sup> November, 2019

#### Marker's Comments

Marker	Grade
--------	-------

\*\* An extract from the University Regulations

6. **Academic Misconduct**

Refer also to Schedule 3.3 of this Regulation for guidance on this procedure.

6.1 **Academic Misconduct** is defined as any attempt by students to gain an unfair advantage in assessments and examinations. Examples of academic misconduct include plagiarism, cheating, falsifying data, collusion, bribery or attempted bribery, personation or any other activity intended to provide an unfair advantage.

(i) **Plagiarism** is the practice of presenting the thoughts or writings of another or others as original, without acknowledgement of their source(s). All material used to support a piece of work should be carefully referenced and should not normally be copied directly unless as an acknowledged quote. Text translated into the words of the individual student should in all cases acknowledge the source.

(ii) **Cheating** includes:

- the taking of any unauthorised material into an examination;
- obtaining copy of “unseen” papers in advance of an examination;
- communicating or attempting to communicate in any way with another student during an examination;
- copying or attempting to copy from another student during an examination or in the production of coursework;
- willful deception in any element of an examination or assessment.

(iii) **Falsification of data** consists of the misrepresentation of the results of experimental work or the presentation of results from fictitious work.

(iv) **Collusion** is the representation of unauthorized group work as that of an individual student.

(v) **Bribery** is the paying, offering or attempted exchange of an inducement for information or material intended to advantage the recipient in an examination or assessment.

(vi) **Personation** consists of a substitute taking the place of a student in an examination.

**A student who aids and abets a fellow student to commit academic misconduct shall be deemed to have committed academic misconduct and will be dealt with accordingly.**

## Introduction

The part 3 of the coursework entails choosing a dataset and formulating a set of analysis that could give some insight into the chosen dataset.

The chosen dataset recorded the police stops under *Terry v. Ohio*, 392 U.S. 1 (1968) which is a decision by the Supreme Court of the United States. This is to guide police officers when making stops to search a suspect who has committed, is committing or is about to commit a crime. As shown in figure 1, the dataset contains the demographics of the suspect and officer involved. This is done to aid employment purposes when background check is done on an applicant. Over a period of years, the analysis will give an insight into how police stops and searches are conducted across suspects races, and the time periods those stops took place.

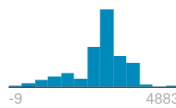
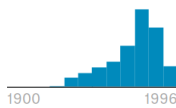
terry-stops.csv (5.3 MB)							
11 of 23 columns							
Views							
	Subject Age Group	Subject ID	Stop Resolution	Officer ID	Officer YOB	Officer Gender	Officer Race
	Subject Age Group (10 year increments) as reported by the officer.	Key, generated daily, identifying unique subjects in the dataset using a character to character match of first	Resolution of the stop as reported by the officer.	Key identifying unique officers in the dataset.	Year of birth, as reported by the officer.	Gender of the officer, as reported by the officer.	Race of the officer, as reported by the officer.
	26 - 35 33%	NULL 3%	GO Report 37%			M 89%	White
	18 - 25 22%	4747 0%	Street Check 37%			F 11%	Hispanic or Latino
	Other (5) 45%	Other (19821) 97%	Other (4) 26%			Other (1) 0%	Other (7)
1	36 - 45	NULL	-	1735	1977	M	White
2	26 - 35	NULL	GO Report	1561	1984	M	White
3	18 - 25	NULL	Street Check	1539	1973	M	White
4	18 - 25	NULL	Street Check	1539	1973	M	White

Figure 1: Screenshot of few rows and columns of dataset

## **Dataset Description**

Number of Rows: 26,043

Number of columns: 23

Size: 5MB

Source: [Seattle Police Stops, Open Data](#)

-----

Loaded Rows: 26,043

Loaded Columns: 11

Size: 2.17MB

## **Column Description**

1. Subject Age Group: Range of age groups for the suspects.
2. Stop Resolution: The resolution made by the officer after the stop.
3. Officer ID: Key identifying unique officers in the dataset.
4. Officer YOB: Year of Birth of the reporting Officer.
5. Officer Gender: Gender of the reporting Officer.
6. Officer Race: Race of the reporting Officer.
7. Subject Perceived Race: The age of the Suspect as reported by the Officer.
8. Subject Perceived Gender: The gender of the Suspect as perceived by the Officer.
9. Reported Date: The date the report was entered into the Record Management System (RMS).
10. Reported Time: This is the time the record was entered into the RMS.
11. Arrest Flag: Takes account of whether an arrest was made or not for the intended suspect.

## **Formulated Analysis**

1. To find out the number of stops made to various races according to their age range.
  - For instance, how many **Whites** within the age of **26 – 35** have been stopped across the whole record and so on for other **races** and **age ranges**.
2. Using the time stamps, divide them into four periods of the day (Early Morning, Morning, Afternoon and Evening). Obtain the number of stop-search occurrences within these time periods.
  - For instance, in how many stops were made in the mornings, evenings or afternoons of 2015. Compute for other years respectively.
3. Compute the percentage of each Suspects race, stopped by an Officer race of a particular race or another.
  - For instance, for all stops by an Asian Officer, what percentage of the Suspect is White?

### **Task 1 Implementation and result**

This task is to find out the proportion of races that are stopped and searched, according to their age ranges.

The pig script for *implementation* is contained in a well annotated *part3\_question1.pig* file.

The output in table 1 below, shows the top 10 order of the Suspects Races and the number of stops according to their age ranges. Whites between the ages of 26 – 35 had the highest stop count of 4652, followed by those within the ages of 36 – 45 and subsequently 18 – 25 years of age. Blacks and other Races subsequently followed suit. The rest of the result is in a *part3\_question1\_answer.txt* file. The results show that whites were more likely to be stopped and searched compared to other races. Across the whole output, Suspects between the age of 26 – 35, across all races are most likely to be stopped and searched.

Subject_Perceived_Race	Subject_Age_Group	Count
White	26 – 35	4652
White	36 – 45	2736
White	18 - 25	2576
Black	26 – 35	2283
Black	18 - 25	2018
White	46 - 55	1671
Black	36 - 45	1420
Black	46 - 55	1045
Black	1 - 17	710
White	56 and Above	643

*Table 1: Top 10 list of Races and stop counts by Age group*

## **Task 2 Implementation and Result**

This task is geared towards getting an insight into the period of the day when most of the searches likely occur. The time stamps were divided into four parts from which the count of the various stop-searches is recorded.

The pig script for *implementation* is contained in a well annotated *part3\_question2.pig* file.

A *macros* in *yearTimeMacro.pig* file is defined to compute the *counts* for the various years. This helps to avoid repeating the computation 4 times, for the various time periods.

Knowing the times that most searches took place could give an insight into when suspects are more likely engaging in a crime. From the result in table 2 below, it shows that most of the searches were at the time of the day when it is still dark. However, in 2018, the number of overall stop-search reduced significantly.

Year	earlyMorningCount	morningCount	afternoonCount	eveningCount
2015	1789	1336	1957	1967
2016	1863	1512	2077	2035
2017	2045	1292	1880	2038
2018	1146	751	1158	1196

*Table 2: years and count of the various search periods in the day*

### **Task 3 Implementation and Result**

For this task, there is need to have an insight into how often a police officer is likely to stop a suspect based on race.

A *part3\_question3.pig* contains the code used to implement the result formulated analysis. The rest of the result is contained in a *part3\_question3\_answer.txt* file.

<b>officerRace</b>	<b>Subject_Perceived_Race</b>	<b>Suspects race(%)</b>
American Indian/Alaska Native	Black	43.6
American Indian/Alaska Native	White	37.0
American Indian/Alaska Native	Multi-Racial	7.0
American Indian/Alaska Native	American Indian / Alaskan Native	5.3
American Indian/Alaska Native	Asian	4.5
American Indian/Alaska Native	Hispanic	2.1
American Indian/Alaska Native	Other	0.4
Asian	White	57.8
Asian	Black	27.6
Asian	Hispanic	5.4
Asian	American Indian / Alaskan Native	3.6
Asian	Asian	3.0
Asian	Multi-Racial	2.3
Asian	Other	0.2
Black or African American	White	47.8
Black or African American	Black	35.8
Black or African American	Hispanic	7.5
Black or African American	American Indian / Alaskan Native	4.1
Black or African American	Asian	2.6
Black or African American	Multi-Racial	1.5

*Table 3: First 20 list of Officer races and the percentage of stops for various Suspects race.*



The result in table 3 shows that an (American Indian/Alaska Native) is more likely to stop a Black suspect, from the result the officer stopped a Black suspect 43.6% of the time. A (Black or African American) officer is more likely to stop a White suspect 47.8% of the time and a White Officer is more likely to stop a white suspect 52.9% of the time and so on for other Officer races. It can be inferred that an officer may be biased in deciding on which race of suspect to stop in order to make a search.

## **Conclusion**

The analysis performed on this dataset using pig is quite insightful. Suspects within the ages of 26 – 35 are more likely to be stopped. An Officer is more likely to stop a suspect that belongs to a race. For the time periods, suspects were stopped more at dark hours. More analysis can still be done to gain insight into the dataset. Like looking at the age of the officers and type of reports they gave after a stop-search.

## **Bibliography**

White, Byron Raymond, and Supreme Court Of The United States. *U.S. Reports: Terry v. Ohio*, 392 U.S. 1. 1967. Periodical. Accessed 11<sup>th</sup> November 2019, [www.loc.gov/item/usrep392001](http://www.loc.gov/item/usrep392001)

David S. Kemp *U.S. Reports: Terry v. Ohio*, 392 U.S. 1. 1967. Periodical. Accessed 11<sup>th</sup> November, 2019, <https://supreme.justia.com/cases/federal/us/392/1/>.