# Identification cell type marker genes of the brain and their use in estimation of cell type proportions

**Thesis Proposal for Doctor of Philosophy(PhD) Degree**
UBC bioinformatics Graduate Program

Ogan Mancarci, B.Sc

**Thesis Supervisor**
Dr. Paul Pavlidis

**Committee Members**
Dr. Clare Beasley
Dr. Shernaz Bamji
Dr. Sara Mostafavi

**Chair**
Dr. Ryan Brinkman

**Examination Date**
June 19, 2015

# Contents

# 1 Introduction

As a once long-term resident of the great state of New York. The author aims to learn more about its most populous city through this project. The City of New York, New York (NYC) has generated an Open Data imitative in order to be fair and open with its constituents. As the author may be taking jobs in the city as early as 2020, there must be a way to determine the dangerous areas of the city from the crime data. In order to do this the student has referenced the New York Police Department (NYPD) public facing records for shooting crime incidents for the period of 2013 through 2018. Throughout this project chunks of **R** code will be displayed inline as a reference to better understand the process.

---

# 2 Business Questions

This project will attempt to answer the following three questions:

- Does time of day and location have a direct correlation to a shooting event?
- Has any part of the city been consistently high in shooting incidents?
- Is any one particular demographic group at high risk ?

As the purpose of this project is to determine the dangerous parts of NYC. Understanding the time of day that most shootings occur as well as which areas of the city these incidents occur in will be key to a new resident choosing their domicile. Additionally, as a bi-product of the information we can generate information on at risk demographics. While project will not predict the housing market prices or the cost of living, it will give a reader a reasonable glimpse at the public order of the city.

---

# 3 Data Acquisition, Cleansing Transformation, Munging

## 3.1 Problem Definition

The purpose of this project is to generate actionable data for a potential future resident of NYC. Actionable data is defined by this project as tables, charts, and modeling to provide graphical information for the reader to understand where the shooting hotspots in the last five years of shooting incidents in NYC. The data from New York Police Department is assigned several flags that can better help define the demographics of the project. There are some key attributes within this dataset that can be used to generate actionable data.

Exmaple Attributes:

- Ages of Accused & Victims
- Sex of Accused & Victims
- Location of the Crime
- Time and Date of the Crime

### 3.2 Data Acquisition

The **NYPD Shooting Incident Data**, which is available freely from the NYC Open Data website contains one set of data. This project will utalize the powerful data science oriented programing language of **R** to analyze this data. Upon intial review of the information provided to the author via the NYPD. Reading the source documentation for this file it was determined that the best method to import would be a Comma Seperated Values (CSV). Addtionally the document retrieved should contain a total of 6,407 rows with 18 attributes for a total of 115,326 data points. In order to start processing the data it was first imported into the system and stored in the variable rawCSV.

```
urlToImport <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
rawCSV <- data.frame((read.csv(urlToImport)))
```

Generating the vector of rawCSV, we created a data frame with the contents of the download csv button from our selected source. Wtih this data now stored within our project, a data scieentist must confirm that the information has arrived correct and that it is able to be used within the project.

### 3.3 Data Clensing Process

Although the data retrived is in relativly good shape there are some peculiarities that need to be addressed during the porocess in order to properly utalize this data within the project. The first piece that could be easily recognized just in a look at the data is that alot of the information is in capital letters. Due to several proceses within this project requiring the lowercase columns the data was brought down in case.

```
cleanCSV <- rawCSV
colnames(cleanCSV) <- tolower(colnames(cleanCSV))
cleanCSV$boro<-tolower(cleanCSV$boro)
cleanCSV$location_desc<-tolower(cleanCSV$location_desc)
cleanCSV$perp_sex<-tolower(cleanCSV$perp_sex)
cleanCSV$perp_race<-tolower(cleanCSV$perp_race)
```

```
cleanCSV$vic_sex<-tolower(cleanCSV$vic_sex)

cleanCSV$vic_race<-tolower(cleanCSV$vic_race)
```

### 3.4 Data Dictionary

---

## 4 Descriptive Statistics

### 4.1 Summary Statistics

### 4.2 Data Structure

### 4.3 Re-shaping the Data

### 4.4 Graphs, Charts, Tables

---

## 5 Modeling Techniques

---

## 6 Data Interpretation

---

## 7 Summary

Actionable ideas or insights

---

# 8 Appendix

## A.1 R Code

```r
#### Generate a Function to Ensure a Package is Installed ###
EnsurePackage<-function(x){
  x<-as.character(x)
  if (!require(x,character.only=TRUE)){
    install.packages(pkgs=x, repos="http://cran.r-project.org")
    require(x, character.only=TRUE)
  }
}
###########Import All Required Packages######################
EnsurePackage("ggplot2")
EnsurePackage("ggmap")
EnsurePackage("gridExtra")
EnsurePackage("maptools")
EnsurePackage("RJSONIO")

##############Gather and Load the Data########################
urlToImport <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
rawCSV <- data.frame((read.csv(urlToImport)))

########Create the Data Dictionary###########################
varName<-c("list of names")
varType<-c("list of types")
varDesc<-c("list of descriptions")
dataDict<-data.frame(varName,varType,varDesc)
colnames(dataDict)<-c("Variable Name", "Variable Type", "Variable Description")
grid.table(dataDict)

########Create the map to plot on############################
```

### A.2 Notes