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 data, a to lower operation was required to be conducted on a few fields.

```
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)
```

The next issue with the intial importing of the CSV file is that the file has not properly converted all the fields to their proper data types. Upon pulling the data frames structure we can see that some fields that should be numbers are characters, as well as fields that should be dates or times are also treated as a character field.

```
head(str(cleanCSV))
```

```
## 'data.frame':    6407 obs. of  18 variables:
##  $ incident_key           : int  138817042 156642467 89216118 138898674 90370145 157022340 166728123
##  $ occur_date             : Factor w/ 1654 levels "01/01/2013 12:00:00 AM",..: 1192 1155 208 1221 51
##  $ occur_time             : Factor w/ 1179 levels "00:01:00","00:02:00",..: 1136 802 202 817 945 488
##  $ boro                   : chr  "brooklyn" "queens" "queens" "brooklyn" ...
##  $ precinct               : int  67 105 102 73 71 45 47 13 43 7 ...
##  $ jurisdiction_code      : int  0 0 0 0 0 0 0 0 2 2 ...
##  $ location_desc          : chr  "" "" "pvt house                    " "" ...
##  $ statistical_murder_flag: Factor w/ 2 levels "false","true": 1 1 1 1 1 2 1 1 1 1 ...
##  $ perp_age_group         : Factor w/ 9 levels "","<18","1020",..: 1 1 1 1 1 1 1 6 1 5 ...
##  $ perp_sex               : chr  "" "" "" "" ...
##  $ perp_race              : chr  "" "" "" "" ...
##  $ vic_age_group          : Factor w/ 6 levels "<18","18-24",..: 3 3 4 1 2 4 3 4 2 3 ...
##  $ vic_sex                : chr  "m" "m" "m" "m" ...
##  $ vic_race               : chr  "black" "black" "asian / pacific islander" "black" ...
##  $ x_coord_cd             : int  996949 1051945 1022214 1007962 1003564 1032140 1025683 986464 10255
##  $ y_coord_cd             : int  176623 180331 188265 183992 181625 242004 261870 208227 236918 2010
##  $ latitude               : num  40.7 40.7 40.7 40.7 40.7 ...
##  $ longitude              : num  -74 -73.8 -73.9 -73.9 -73.9 ...
```

```
## NULL
```

In order to correct these issues, the first column to be targeted was that of the occur_date. From the above structure you can see that this is shown as both a factor and additionally it contained both an unformated date as well as some erronious informationn inserted somewhere along the way during the csv file creation process.

```
removeTime <- function(inputVector){
  inputVector <-gsub(" 12:00:00 AM","",inputVector)
  inputVector <- gsub(" ", "", inputVector)
  inputVector <- as.Date(inputVector, format='%m/%d/%Y')
  return(inputVector)
}
cleanCSV$occur_date<-removeTime(cleanCSV$occur_date)
```

The column titled location_desc is filled with limited information for the purposes of this report. When it is filled with data the information is not consistant or actionable. This column was removed from the data frame and than verified that it was idneed removed.

```
cleanCSV<-cleanCSV[,-6]
colnames(cleanCSV)
```

Finally, after all of this cleaning has been completed only complete rows should remain. This will be rows that do not have any blank values that cannot be used within this report.

```
cleanCSV<-cleanCSV[complete.cases(cleanCSV),]
```

After performing all of these operations to prepare the data for use within the report the data contains 6407 rows and 17 attributes to perform work on for a total of 108,919 data points.

3.4 Data Dictionary

1	incident_key	int	Replace Me
2	occur_date	date	Replace Me
3	occur_time	chr	Replace Me
4	boro	chr	Replace Me
5	precinct	int	Replace Me
6	location_desc	int	Replace Me
7	statistical_murder_flag	chr	Replace Me
8	perp_age_group	chr	Replace Me
9	perp_sex	chr	Replace Me
10	perp_race	chr	Replace Me
11	vic_age_group	chr	Replace Me
12	vic_sex	chr	Replace Me
13	vic_race	chr	Replace Me
14	x_coord_cd	int	Replace Me
15	y_coord_cd	int	Replace Me
16	latitude	num	Replace Me

---

4 Descriptive Statistics

4.1 Summary Statistics

```
##   incident_key         occur_date            occur_time
##  Min.   : 88354616   Min.   :2013-01-01   23:30:00:  45
##  1st Qu.:109340977   1st Qu.:2014-03-30   01:30:00:  42
##  Median :142679416   Median :2015-05-05   21:00:00:  42
##  Mean   :135265119   Mean   :2015-05-13   00:30:00:  39
##  3rd Qu.:154851942   3rd Qu.:2016-07-13   04:00:00:  39
##  Max.   :173129246   Max.   :2017-12-31   02:00:00:  36
##                                           (Other) :6164
```

```
##      boro              precinct      location_desc
##  Length:6407        Min.   :  1.00   Length:6407
##  Class :character   1st Qu.: 44.00   Class :character
##  Mode  :character   Median : 69.00   Mode  :character
##                     Mean   : 66.69
##                     3rd Qu.: 81.00
##                     Max.   :123.00
##
##  statistical_murder_flag perp_age_group   perp_sex
##  false:5260                     :3028   Length:6407
##  true :1147              18-24  :1490   Class :character
##                          25-44  :1374   Mode  :character
##                          <18    : 319
##                          45-64  : 130
##                          UNKNOWN:  44
##                          (Other):  22
##   perp_race         vic_age_group    vic_sex            vic_race
##  Length:6407        <18    : 556   Length:6407        Length:6407
##  Class :character   18-24  :2463   Class :character   Class :character
##  Mode  :character   25-44  :2847   Mode  :character   Mode  :character
##                     45-64  : 470
##                     65+    :  43
##                     UNKNOWN:  28
##
##    x_coord_cd        y_coord_cd        latitude       longitude
##  Min.   : 922884   Min.   :132099   Min.   :40.53   Min.   :-74.22
##  1st Qu.: 999970   1st Qu.:181481   1st Qu.:40.66   1st Qu.:-73.94
##  Median :1007605   Median :193359   Median :40.70   Median :-73.92
##  Mean   :1008757   Mean   :206795   Mean   :40.73   Mean   :-73.91
##  3rd Qu.:1016320   3rd Qu.:239453   3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :1063056   Max.   :269205   Max.   :40.91   Max.   :-73.72
##
```

4.2 Data Structure

```
## 'data.frame':    6407 obs. of  17 variables:
##  $ incident_key          : int  138817042 156642467 89216118 138898674 90370145 157022340 166728123
##  $ occur_date            : Date, format: "2014-09-21" "2016-09-12" ...
##  $ occur_time            : Factor w/ 1179 levels "00:01:00","00:02:00",..: 1136 802 202 817 945 488
##  $ boro                  : chr  "brooklyn" "queens" "queens" "brooklyn" ...
##  $ precinct              : int  67 105 102 73 71 45 47 13 43 7 ...
##  $ location_desc         : chr  "" "" "pvt house                     " "" ...
##  $ statistical_murder_flag: Factor w/ 2 levels "false","true": 1 1 1 1 1 2 1 1 1 1 ...
##  $ perp_age_group        : Factor w/ 9 levels "","<18","1020",..: 1 1 1 1 1 1 1 6 1 5 ...
##  $ perp_sex              : chr  "" "" "" "" ...
##  $ perp_race             : chr  "" "" "" "" ...
##  $ vic_age_group         : Factor w/ 6 levels "<18","18-24",..: 3 3 4 1 2 4 3 4 2 3 ...
##  $ vic_sex               : chr  "m" "m" "m" "m" ...
##  $ vic_race              : chr  "black" "black" "asian / pacific islander" "black" ...
##  $ x_coord_cd            : int  996949 1051945 1022214 1007962 1003564 1032140 1025683 986464 10255
##  $ y_coord_cd            : int  176623 180331 188265 183992 181625 242004 261870 208227 236918 2010
##  $ latitude              : num  40.7 40.7 40.7 40.7 40.7 ...
##  $ longitude             : num  -74 -73.8 -73.9 -73.9 -73.9 ...
```

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

```
#### 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