Lab 3

Fan Yang (fy2232)

October 2, 2017

Instructions

Before you leave lab today make sure that you upload a .pdf file to the canvas page (this should have a .pdf extension). This should be the PDF output after you have knitted the file, we don't need the .Rmd file (don't upload the one with the .Rmd extension). The file you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. Note, however, in the file you upload you should the above header to have the date, your name, and your UNI. Similarly, when you save the file you should replace **UNI** with your actualy UNI.

Tasks

In this lab, we will use data from homicides in Baltimore City that are collected by the Baltimore Sun newspaper. The data is presented in a map that is publically available at the following website: http://data.baltimoresun.com/news/police/homicides/. I've scraped the data from the website and saved it in a file **BaltimoreHomicides.txt**. Load the data with the following:

```
data <- readLines("BaltimoreHomicides.txt")
```

The data we have corresponds to all the homicides in 2010. There are 224 lines of HTML in the dataset (verify this using **length(data)**) and 224 homicides in 2010.

(1) In the image below, you can see one of the homicides in our dataset. Use a **grep()** call to find which row in the dataset corresponds to the death of Khloe Lewis. (You could search, for example, for her name and the row you should return is 52).

```
grep("Khloe Lewis",data)
```

```
## [1] 52
```

(2) Suppose we wanted to identify the records for all the victims of shootings (as opposed to other causes). Using the following bit of code and some of your own exploration, how many of the homicides in our dataset are the result of shootings?

```
deaths1 <- grep("shooting", data)
deaths2 <- grep("Shooting", data)
length(deaths1)
```

```
## [1] 172
```

```
length(deaths2)
```

```
## [1] 171
```

```
setdiff(deaths1, deaths2)
```

```
## [1] 105
```

```
length(deaths1)+length(deaths2)-setdiff(deaths1, deaths2)
```

```
## [1] 238
```

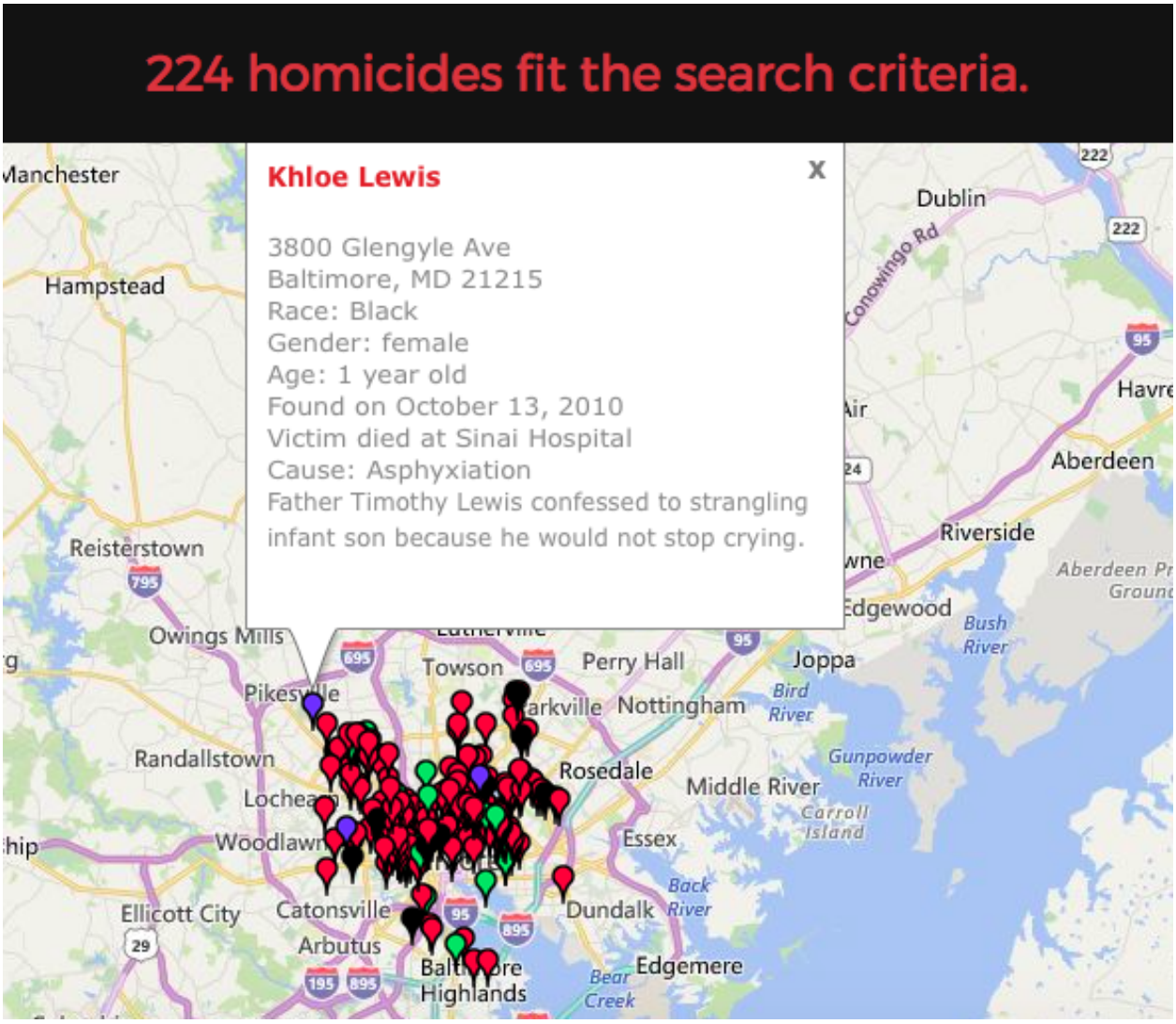

Figure 1: image

(3) The following code creates a vector of the ages of the homicide victims in the dataset:

```
r       <- regexpr(">Age:\\s.*years old<", data)
age_vec <- regmatches(data, r)
age_vec <- substring(age_vec, 2, nchar(age_vec) - 1)
head(age_vec)
```

```
## [1] "Age: 50 years old" "Age: 25 years old" "Age: 20 years old"
## [4] "Age: 23 years old" "Age: 14 years old" "Age: 43 years old"
```

(a) Explain what the regular expression ">Age:\s.*years old<" searches for.

This expression searches for phrases that start with ">Age:" and followed by a blank space. And this phrase should also ends with "years old<".

(b) Explain in words what the output of the first line of code provides.

The first line finds the first location of phrases that matched with rule in part(a)for each line; and also provides the length of each match.

(c) The second line of code.

The second line gives the exact matched phrase in each line.

(d) The third line of code.

Since the phrases begin with ">Age:" and end with "years old<", the third line extracts words between ">" and "<".

(4) Use the same strategy as we used in question (3) to create a vector holding each victims' name. Hint: I had to use the fact that the first letter of the name is capitalized and the specific structure of the HTML code after the name in writing my regular expression.

```
r       <- regexpr(">[A-Z].*</a>", data)
age_vec <- regmatches(data, r)
age_vec <- substring(age_vec, 2, nchar(age_vec) - 4)
head(age_vec,20)
```

```
##  [1] "Bernard Clowney"              "David King"
##  [3] "Raymond Woodland"             "Keith L. Robinson"
##  [5] "Issac Joyner"                 "Mustafa Malik"
##  [7] "Juan Carlos Santos-Hernandez" "Brian Taylor"
##  [9] "Karen Ferrell"                "Ramon Uceda"
## [11] "Ellison McCall"               "Alethea Hawkins"
## [13] "Micha Crane"                  "Cherrie Gammon"
## [15] "Travis Baltimore"             "David Carter"
## [17] "Dante Sweeney"                "Troy Thomas"
## [19] "Tammy Madison"                "Raquan Campbell"
```