Data Cleaning

Anju Menon

July 12, 2016

A. "Real Property Taxes" Dataset

Download the "real property taxes" dataset from the website (via OpenBaltimore), he data is located here (note you don't need to unzip it to read it into R): http://sisbid.github.io/Module1/data/real_property_tax.csv.gz

1. Read the Property Tax data into R and call it the variable `tax`

```
tax<-read.csv("real_property_tax.csv.gz")
```

2. How many addresses pay property taxes?

```
tax$cityTax <- as.numeric(gsub(pattern = "$", replacement="",
                               tax$cityTax, fixed=TRUE))
tax$stateTax<- as.numeric(gsub(pattern = "$", replacement="",
                               tax$stateTax, fixed=TRUE))
head(tax)
```

```
##                     propertyAddress           lotSize cityTax stateTax
## 1 1205 CLARKSON ST                   0.018 ACRES             0        0
## 2 1207 CLARKSON ST                   0.018 ACRES             0        0
## 3 1209 CLARKSON ST                   0.018 ACRES             0        0
## 4 1203 CLARKSON ST                   0.018 ACRES             0        0
## 5 4111 OLD YORK ROAD                 0.291 ACRES             0        0
## 6 101 W PRESTON ST                   0.909 ACRES             0        0
##                     resCode
## 1 NOT A PRINCIPAL RESIDENCE
## 2 NOT A PRINCIPAL RESIDENCE
## 3 NOT A PRINCIPAL RESIDENCE
## 4 NOT A PRINCIPAL RESIDENCE
## 5 NOT A PRINCIPAL RESIDENCE
## 6 NOT A PRINCIPAL RESIDENCE
```

```
#tax$proptax<-tax$cityTax + tax$stateTax
#head(tax$proptax)
#nrow(filter(tax,proptax > 0))
nrow(tax) - length(which(tax$cityTax ==0 & tax$stateTax==0 ))
```

```
## [1] 238299
```

There are 238299 addresses that pay property taxes.

3. What is the total city and state tax paid?

The total city tax paid is $ 782,158,186

The total state tax paid is $ 39,584,834

4. How many observations/properties are in each residence code?

```
table(tax$resCode)
```

```
##
## NOT A PRINCIPAL RESIDENCE PRINCIPAL RESIDENCE
##                    122675                    115696
```

5. What is the 75th percentile of city and state tax paid by residence code?
6. Subset the data to only retain those houses that are principal residences and describe the distribution of property taxes on these residences.
7. Convert the 'lotSize' variable to a numeric square feet variable. Tips:

- Look at the data
- Assume hyphens represent inches within square foot meassuremnts
- Assume decimals within acreage measurements
- 1 acre = 43560 square feet
- Look at the data

B. "Baltimore Salary 2015" Dataset

Download the "Salary 2015" Dataset from the website (via OpenBaltimore), which is located here: http://sisbid.github.io/Module1/data//Baltimore_City_Employee_Salaries_FY2015.csv

8. Make an object called health.sal using the salaries data set, with only agencies of those with "fire" (or any forms), if any, in the name
9. Make a data set called trans which contains only agencies that contain "TRANS".
10. What is/are the profession(s) of people who have "abra" in their name for Baltimore's Salaries?
11. What is the distribution of annual salaries look like? What is the IQR?
12. Convert HireDate to the `Date` class - plot Annual Salary vs Hire Date
13. Plot annual salary versus hire date. Hint: first convert to numeric and date respectively
14. Create a smaller dataset that only includes the Police Department, Fire Department and Sheriff's Office. How many employees are in this new dataset?