Project2

Alvaro Bueno

10/8/2017

Project 2

1st data set - energy usage

Although the data seems tidy at first, the first thing to do is to set correct values to the columns the other thing we can do is to group by regions and analyze their consumption rates.

```
NE_edata <- dplyr::filter(energydata, state %in% c("Maine", "Vermont", "New Hampshire", "Massachusetts"

MW_edata <- dplyr::filter(energydata, state %in% c("South Dakota", "North Dakota","Nebraska", "Kansas",

SO_edata <- dplyr::filter(energydata, state %in% c("New Mexico", "Texas", "Arkansas", "Louisiana", "Alal

MO_edata  <- dplyr::filter(energydata, state %in% c("Montana", "Idaho", "Wyoming", "Utah", "Colorado"))

NW_edata  <- dplyr::filter(energydata, state %in% c("Washington", "Oregon", "Nevada", "Arizona", "Califo

NE_edata_new <- factorsNumeric(NE_edata[2:ncol(NE_edata)])
NE_edata_new$state <- NE_edata$state

MW_edata_new <- factorsNumeric(MW_edata[2:ncol(MW_edata)])
MW_edata_new$state <- MW_edata$state

SO_edata_new <- factorsNumeric(SO_edata[2:ncol(SO_edata)])
SO_edata_new$state <- SO_edata$state

MO_edata_new <- factorsNumeric(MO_edata[2:ncol(MO_edata)])
MO_edata_new$state <- MO_edata$state

NW_edata_new <- factorsNumeric(NW_edata[2:ncol(NW_edata)])
NW_edata_new$state <- NW_edata$state
```

```
NE_total <- dplyr::summarise(NE_edata_new, suma = sum(NE_edata_new$`energy total`))
MW_total <- dplyr::summarise(MW_edata_new, suma = sum(MW_edata_new$`energy total`))
SO_total <- dplyr::summarise(SO_edata_new, suma = sum(SO_edata_new$`energy total`))
MO_total <- dplyr::summarise(MO_edata_new, suma = sum(MO_edata_new$`energy total`))
NW_total <- dplyr::summarise(NW_edata_new, suma = sum(NW_edata_new$`energy total`))

totals <- c(NE_total,MW_total,SO_total,MO_total,NW_total)
names(totals) <-c("NE","MW","SO","MO","NW")
kable(as.data.frame(totals))
```

NE	MW	SO	MO	NW
29220.8	28615.6	21929.9	3710.8	12710.8

Not surprisingly, the Northeast consumes more energy than other regions, what's a bit shocking is that NW (that includes california) do not consume as much energy, but this is also an area with only 5 states, instead, the MidWest area has 15 states and those have some heavy industry there.

2nd data set - NYC collision data

For this data set i would like to know the borough that's more dangerous for cyclists

```
collisiondata <- read.csv("https://raw.githubusercontent.com/delagroove/dataScience/master/city_colissi

## remove first 3 lines
collisiondata <- dplyr::slice(collisiondata, 4:nrow(collisiondata))

#use the first column as the column name
names(collisiondata) <- lapply(collisiondata[1, ], as.character)
collisiondata <- collisiondata[-1,]



#convert all numeric columns
coldata <- factorsNumeric(collisiondata[3:14])
coldata$GeoCode <-collisiondata$GeoCode
coldata$GeoCodeLabel <-collisiondata$GeoCodeLabel

coldata_boroughs <- dplyr::filter(coldata, GeoCode %in% c("M", "B","Q", "K", "S"))

coldata_boroughs <- dplyr::mutate(coldata_boroughs, cyclist_kill_ratio = CyclistsKilled/Bicycle)
coldata_boroughs <- dplyr::mutate(coldata_boroughs, cyclist_injury_ratio = CyclistsInjured/Bicycle)
coldata_boroughs <- dplyr::mutate(coldata_boroughs, motorist_kill_ratio = MotoristsKilled/Number_of_Mot
coldata_boroughs <- dplyr::mutate(coldata_boroughs, motorist_injury_ratio = MotoristsInjured/Number_of_

kable(coldata_boroughs[,14:ncol(coldata_boroughs)])
```

GeoCodeLabel	cyclist_kill_ratio	cyclist_injury_ratio	motorist_kill_ratio	motorist_injury_ratio
MANHATTAN	0.0000000	0.7103825	0.0002554	0.0635853
BRONX	0.0000000	0.8382353	0.0010070	0.1503860
BROOKLYN	0.0042553	0.8000000	0.0003726	0.1265139
QUEENS	0.0000000	0.8256881	0.0005534	0.1496034
STATEN ISLAND	0.0000000	1.0909091	0.0000000	0.1321839

We can see that Brookyln is the most dangerous borough to drive a bike, the only one with casualties. the one with most injuries is the Bronx, which can make some sense because it does not have as many Bike Lanes as other boroughs (Source: http://www.nycbikemaps.com/wp-content/uploads/2016/04/bikeroutedetailscy06-cy15.pdf)

Also the Bronx is the borough Download NYC borough mapwith more percentage of motorists injured in accidents.

3rd data set - OATH Hearings

I'm going to plot around that map the balnce due of the violations to see which zipcode (or which area within a Borough) gets the most cumulative tickets.

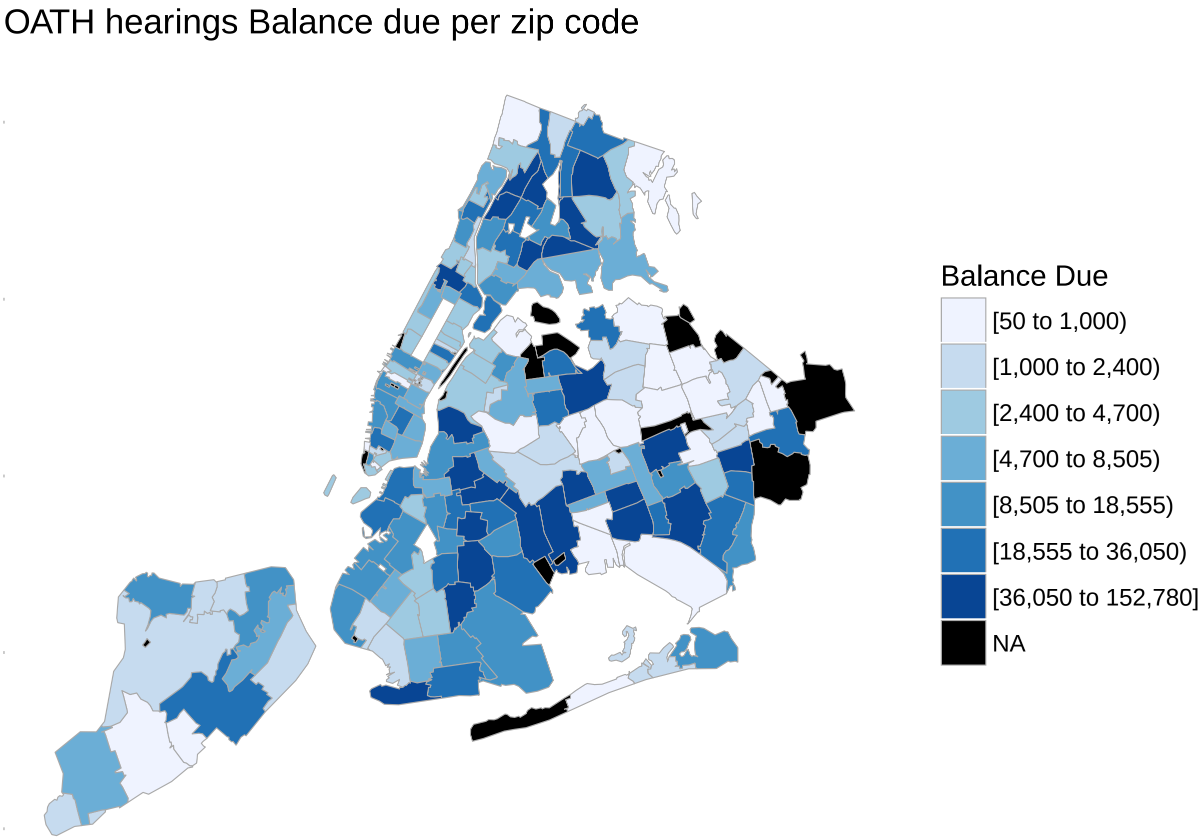
```r
#loading ny county data from cloropleth package
ec_states <- c("new york")
data(county.regions)

hearingdata <- read.csv("https://raw.githubusercontent.com/delagroove/dataScience/master/OATH_Hearings_
hearingdata$Balance.Due <- as.numeric(gsub("\\$|,| ","",hearingdata$Balance.Due))
hearingdata$Violation.Location..Zip.Code. <- as.numeric(gsub("\\$|,| ","",hearingdata$Violation.Locatio


res =""
res$region <-hearingdata$Violation.Location..Zip.Code.
```

```
## Warning in res$region <- hearingdata$Violation.Location..Zip.Code.:
## Coercing LHS to a list
```

```r
res$value <-hearingdata$Balance.Due

#converting result to data frame and cleaning it
res <- as.data.frame(res)
res <- na.omit(res)
res <- ddply(res,"region",numcolwise(sum))

res$region <- as.character(res$region)
nyc_county_names = c("kings", "bronx", "new york", "queens", "richmond")
nyc_county_fips = county.regions %>% filter(state.name == "new york" & county.name %in% nyc_county_name

zip_choropleth(res,
               state_zoom = ec_states,
               county_zoom = nyc_county_fips$region,
               title      = "OATH hearings Balance due per zip code",
               legend     = "Balance Due") + coord_map()
```

```
## Warning in super$initialize(zip.map, user.df): Your data.frame contains the
## following regions which are not mappable: 10000, 10105, 11249
```

```
## Warning in self$bind(): The following regions were missing and are being
## set to NA: 11040, 10280, 10174, 10119, 11371, 10110, 10271, 11003, 11370,
## 10171, 10069, 10162, 10177, 10152, 10279, 11363, 10115, 11697, 10111,
## 10112, 10167, 11351, 11359, 11366, 11424, 11425, 11451, 11360, 10169,
## 10103, 10311, 10154, 10199, 10165, 11239, 10168, 10278, 10044, 10173,
## 11109, 10170, 10172, 11005
```



A lot of balance due tickets are in Poor areas of brookyln (Canarsie and East NEw York, Flatbush) and in inner parts of bornx, let's see if these balances are paid per zip code with a similar analysis.

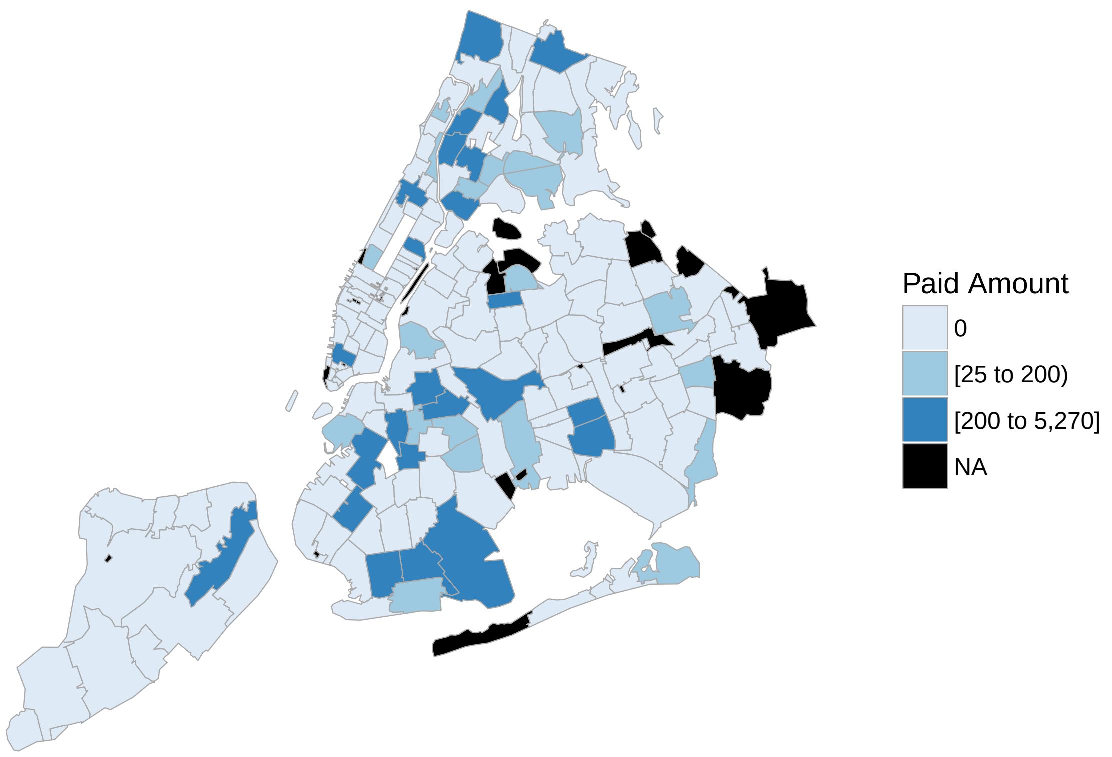
```
#loading ny county data from cloropleth package

hearingdata$Paid.Amount <- as.numeric(gsub("\\$|,| ","",hearingdata$Paid.Amount))
#hearingdata$Violation.Location..Zip.Code. <- as.numeric(gsub("\\$|,| ","",hearingdata$Violation.Locati



res2 =""
res2$region <-hearingdata$Violation.Location..Zip.Code.
```

```
## Warning in res2$region <- hearingdata$Violation.Location..Zip.Code.:
## Coercing LHS to a list
```

```
res2$value <-hearingdata$Paid.Amount

#converting result to data frame and cleaning it
res2 <- as.data.frame(res2)
res2 <- na.omit(res2)
res2 <- ddply(res2,"region",numcolwise(sum))

res2$region <- as.character(res2$region)

zip_choropleth(res2,
               state_zoom = ec_states,
               county_zoom = nyc_county_fips$region,
               title      = "OATH hearings Balance due per zip code",
                legend      = "Paid Amount") + coord_map()
```

```
## Warning in super$initialize(zip.map, user.df): Your data.frame contains the
## following regions which are not mappable: 10000, 10105, 11249

## Warning in self$bind(): The following regions were missing and are being
## set to NA: 11040, 10280, 10174, 10119, 11371, 10110, 10271, 11003, 11370,
## 10171, 10069, 10162, 10177, 10152, 10279, 11363, 10115, 11697, 10111,
## 10112, 10167, 11351, 11359, 11366, 11424, 11425, 11451, 11360, 10169,
## 10103, 10311, 10154, 10199, 10165, 11239, 10168, 10278, 10044, 10173,
## 11109, 10170, 10172, 11005
```

OATH hearings Balance due per zip code



A lot of the owed balance remains unpaid, The poor areas we mentioned above might pay some of the amount, but what they owe to the city is way more than the amounts paid to those balances. Staten island is one of the areas that carry a bigger balance and don't pay anything to the city, surprisingly, the area with the least income is the one that pays more of the debt, while the southern part of the island have don't pay at all, and that's the area with most income (source: http://www.silive.com/news/2014/05/how_much_income_does_your_zip.html)