housekeeping

Tianyue Wang

2017/11/30

Airbnb boston dataset

```
#Read Airbnb datasets collected in year 2016
boston5<-fread("tomslee_airbnb_boston_0282_2016-01-16.csv", header=TRUE)
boston6<-fread("tomslee_airbnb_boston_0314_2016-02-16.csv", header=TRUE)
boston7<-fread("tomslee_airbnb_boston_0344_2016-03-18.csv", header=TRUE)
boston8<-fread("tomslee_airbnb_boston_0386_2016-04-14.csv", header=TRUE)
boston9<-fread("tomslee_airbnb_boston_0420_2016-05-18.csv", header=TRUE)
boston10<-fread("tomslee_airbnb_boston_0461_2016-06-18.csv", header=TRUE)
boston11<-fread("tomslee_airbnb_boston_0489_2016-07-16.csv", header=TRUE)
boston12<-fread("tomslee_airbnb_boston_0524_2016-08-19.csv", header=TRUE)
boston13<-fread("tomslee_airbnb_boston_0566_2016-09-16.csv", header=TRUE)
boston14<-fread("tomslee_airbnb_boston_0610_2016-10-18.csv", header=TRUE)
boston15<-fread("tomslee_airbnb_boston_0649_2016-11-21.csv", header=TRUE)
boston15<-data.frame(boston15[,c(1:3,5:8,10:13)])
#Merge all datasets, only keep repeated observations
boston2016<-Reduce(function(x, y) merge(x, y, all=TRUE),list(boston5,boston6,boston7,boston8))
boston2016<-Reduce(function(x, y) merge(x, y, all=TRUE),list(boston2016,boston9,boston10,boston11,boston
boston2016<-Reduce(function(x, y) merge(x, y, all=TRUE),list(boston2016,boston13,boston14))
boston2016<-data.frame(boston2016[,c(1:3,5:8,10:13)])
boston2016<-Reduce(function(x, y) merge(x, y, all=TRUE),list(boston15,boston2016))
#Remove observations that overall statisfaction is NA
boston2016<-subset(boston2016,!is.na(overall_satisfaction))
#Keep unique observations
boston2016df<-unique(boston2016)
```

Airbnb data cleaning

```
#Create airbnb housing districts by boston police department districts
boston2016df$District=boston2016df$neighborhood
boston2016df$District<-suppressWarnings(recode(boston2016df$District, "c('Downtown','Charlestown','China
boston2016df$District<-suppressWarnings(recode(boston2016df$District,"c('Allston','Brighton')='Allston/
boston2016df$District<-suppressWarnings(recode(boston2016df$District,"c('Roxbury','Mission Hill')='Roxbu
boston2016df$District<-suppressWarnings(recode(boston2016df$District,"c('South End','Back Bay','Fenway'
boston2016df$District<-suppressWarnings(recode(boston2016df$District,"c('West Roxbury','Roslindale')='We
boston2016df$District[boston2016df$District=="South Boston Waterfront"]<-"South Boston"
boston2016df$District[boston2016df$District=="Mattapan"]<-"Mattapan/North Dorchester"
boston2016df$District<-as.factor(boston2016df$District)
boston2016df$room_type<-as.factor(boston2016df$room_type)
boston2016df$room_id<-as.numeric(boston2016df$room_id)
boston2016df$host_id<-as.numeric(boston2016df$host_id)
air<-data.frame(boston2016df[,c(1:3,5,6,8:12)])
#Set 'Na's in minimum stay as 1 day.
```

```r
air$minstay[is.na(air$minstay)]<-1
#Take averages of price, minimum stay, reviews and overall statisfication.
airdat<-air%>%group_by(room_id,host_id,room_type,District)%>%summarise_all(funs(mean))
airbnb<-subset(airdat,!is.na(host_id))
airbnb[,c(5,7,8)]<-round(airbnb[,c(5,7,8)])
airbnb[,c(6)]<-round(airbnb[,c(6)],digits = 1)
#cleaned airbnb dataset
head(airbnb)
```

```
## # A tibble: 6 x 10
## # Groups:   room_id, host_id, room_type [6]
##   room_id host_id       room_type                District reviews
##     <dbl>   <dbl>          <fctr>                  <fctr>   <dbl>
## 1    3353    4240    Private room        Allston/Brighton      31
## 2    3781    4804 Entire home/apt             East Boston       8
## 3    5453    8021    Private room           Jamaica Plain      53
## 4    5506    8229    Private room    Roxbury/Mission Hill      35
## 5    6695    8229 Entire home/apt    Roxbury/Mission Hill      46
## 6    6976   16701    Private room West Roxbury/Roslindale      38
## # ... with 5 more variables: overall_satisfaction <dbl>, price <dbl>,
## #   minstay <dbl>, latitude <dbl>, longitude <dbl>
```

Boston Crime dataset

```r
#read City of Boston crime dataset
crime<-read.csv("crime.csv")
#Translate boston police districts
town = c(A1 = 'Downtown/Charlestown',
         A15= 'Downtown/Charlestown',
         A7= 'East Boston',
         B2= 'Roxbury/Mission Hill',
         B3= 'Mattapan/North Dorchester',
         C6= 'South Boston',
         C11= 'Dorchester',
         D4= 'South End/Back Baay/Fenway',
         D14= 'Allston/Brighton',
         E5= 'West Roxbury/Roslindale',
         E13= 'Jamaica Plain',
         E18= 'Hyde Park')
crime$DISTRICT = as.factor(town[as.character(crime$DISTRICT)])
#Only select observations from 2016.
crime1<-subset(crime,crime$YEAR=="2016")
crime1<-subset(crime1,!is.na(DISTRICT)) #delete NAs in districts.
#Define violent crimes according to UCR violent crime definition:
crime1$violent<-crime1$OFFENSE_CODE_GROUP
crime1$violent<-recode(crime1$violent,"c(
        'Aggravated Assault',
        'Homicide',
        'Manslaughter',
        'Robbery',
        'Larceny From Motor Vehicle',
        'Auto Theft',
```

```
        'Commercial Burglary',
        'Residential Burglary',
        'Other Burglary',
        'Arson')='violent'")
crime1$violent=ifelse(crime1$violent=="violent","violent","non-violent")
crime2<-crime1[,c(1,5,7,15,16,18)]
crime2<-unique(crime2)
#For same incident appears with different offense code, I define this incident as a violent crime.
crime3<-crime2%>%group_by(DISTRICT,INCIDENT_NUMBER)%>%filter("violent"%in%violent)
crime3$violent<-"violent"
crime3<-unique(crime3) #violent crimes
#total number of crimes by region
number_crime<-as.data.frame<-crime3%>%group_by(DISTRICT)%>%count_()%>%arrange(desc(n))
colnames(number_crime)<-c("District","Num_crime")
#population added for each boston police department district
population<-as.data.frame(as.matrix(c(76917,77773,91982,55971,36480,37468,74997,35200,30631,40508,50983
colnames(population)<-c("population")
crimerate<-data.frame(population,number_crime)
#Personal crime rate = (number of crimes / (population))*1000
crimerate$rate<-1000*(crimerate$Num_crime/(crimerate$population))
crimeindex<-data.frame(crimerate[,c(2,4)])
```

Combine Airbnb and Crime datasets by boston police districts

```
#cleaned data
mydat<-merge(airbnb,crimeindex,by="District")
head(mydat)
```

```
##           District  room_id  host_id       room_type reviews
## 1 Allston/Brighton     3353     4240    Private room      31
## 2 Allston/Brighton  9572941 36388924    Private room       4
## 3 Allston/Brighton 14253149 26956083    Private room       6
## 4 Allston/Brighton  8388481 26956083    Private room       5
## 5 Allston/Brighton  7346767  4198683 Entire home/apt       7
## 6 Allston/Brighton  7346767  4198683    Private room       6
##   overall_satisfaction price minstay latitude longitude     rate
## 1                  4.5    37       7 42.35502 -71.12759 10.65376
## 2                  5.0    62       1 42.35081 -71.14172 10.65376
## 3                  5.0    79       2 42.36121 -71.12649 10.65376
## 4                  4.8    81       2 42.35857 -71.14060 10.65376
## 5                  4.5    54       1 42.34715 -71.14043 10.65376
## 6                  4.5    54       1 42.34715 -71.14043 10.65376
```