Semester Project - Boston Housing Market Data Analysis Project

## Objective: *Analyzing Boston housing market trends for buying and renting houses.*

## Team members: *Deep Vaghani, Sooraj Shetty*

## Course Name: *Collecting, Storing & Retrieving Data*

## Course Number: *DS 6020-01*

## CRN: *18102*

## Semester: *Fall 2016*

### *Purpose of the Project*

Boston ranks among the top 5 most expensive housing markets in the USA. These prices are dependent on various factors like the population of the city, area of the city, type of area etc. Also, Boston is well known for its population of international students as it has more universities per sq. mile than any other city in the US.

Our project caters to the creation of an comprehensive database, which includes detailed information regarding all the aspects of the housing like price, location, crime statistics, demographics, transportation to work etc. Alongwith providing this information, the database will also help in analyzing several trends in the Boston housing markets. These trends include variation of prices with change in zipcode, effect of crime on prices, residents of which zipcode are likely to use pubkic tranportation to work etc.

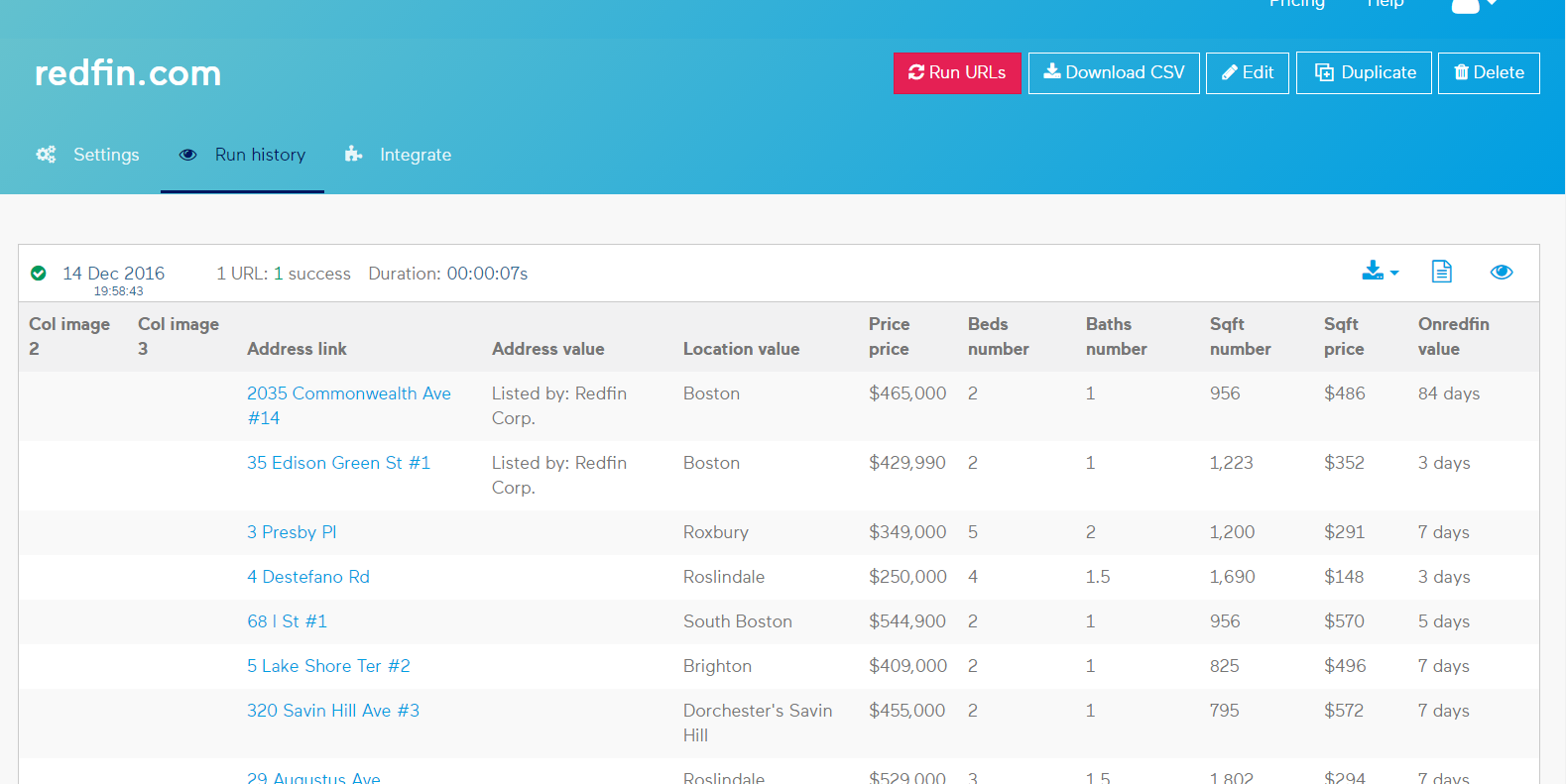
### *Selection of data*

To create the database we are using data obtained from [Redfin](https://www.redfin.com/) and [Moving.com](http://www.moving.com/real-estate/city-profile/zip-selector.asp?State=22&City=Boston) and to validate the obtained data we used [Zillow](http://www.zillow.com/) and [Trulia](https://www.trulia.com/).

### *Process Implemented*

1. Initially we planned to use Zillow API for data collection. However, we faced several issues while attepmting to gather data from Zillow. Hence, for listings we used Redfin.
2. We used import.io to extract data from Redfin. We were able to extract 22,724 listings in Boston. This data is used as the primary data for our database. We scrpaed data using different zipcodes and merged the csv using terminal commands and use that as our main table.

#cleaning the Environment for fresh start  
#rm(list())  
  
#setting the workdirectory to get import the scraped csv into the work environment  
setwd("D:/R")  
  
#importing main table in R environment.  
Housingmain <- read.csv("main Data.csv")  
  
#Using zipcode package to clean zipcodes in mainhousing table so that R doesnt neglect the first "0" in zipcodes  
library(zipcode)  
Housingmain$ZIP <- clean.zipcodes(Housingmain$ZIP)



1. We scraped additional data from Moving.com. This data included crime statistics, demographic information, residential data and income figures.

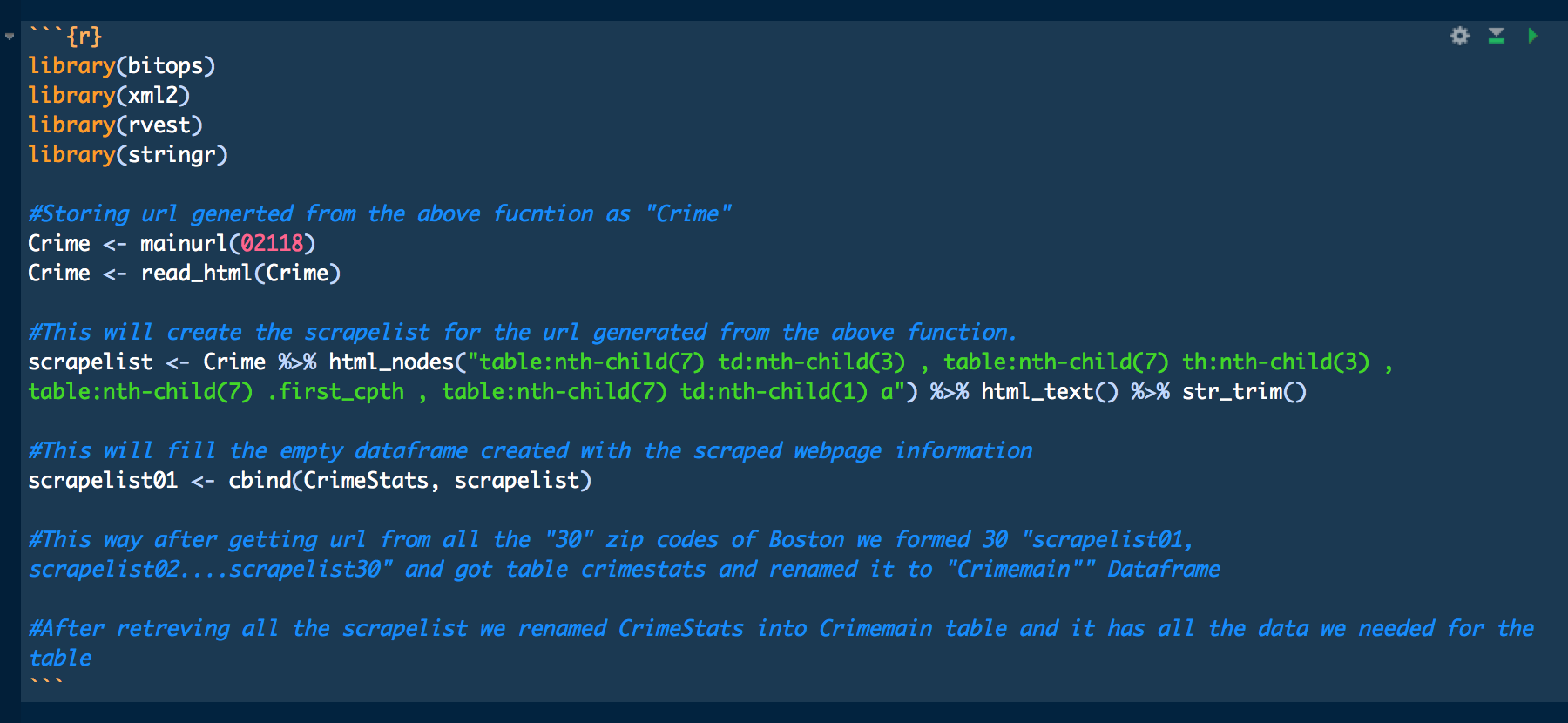
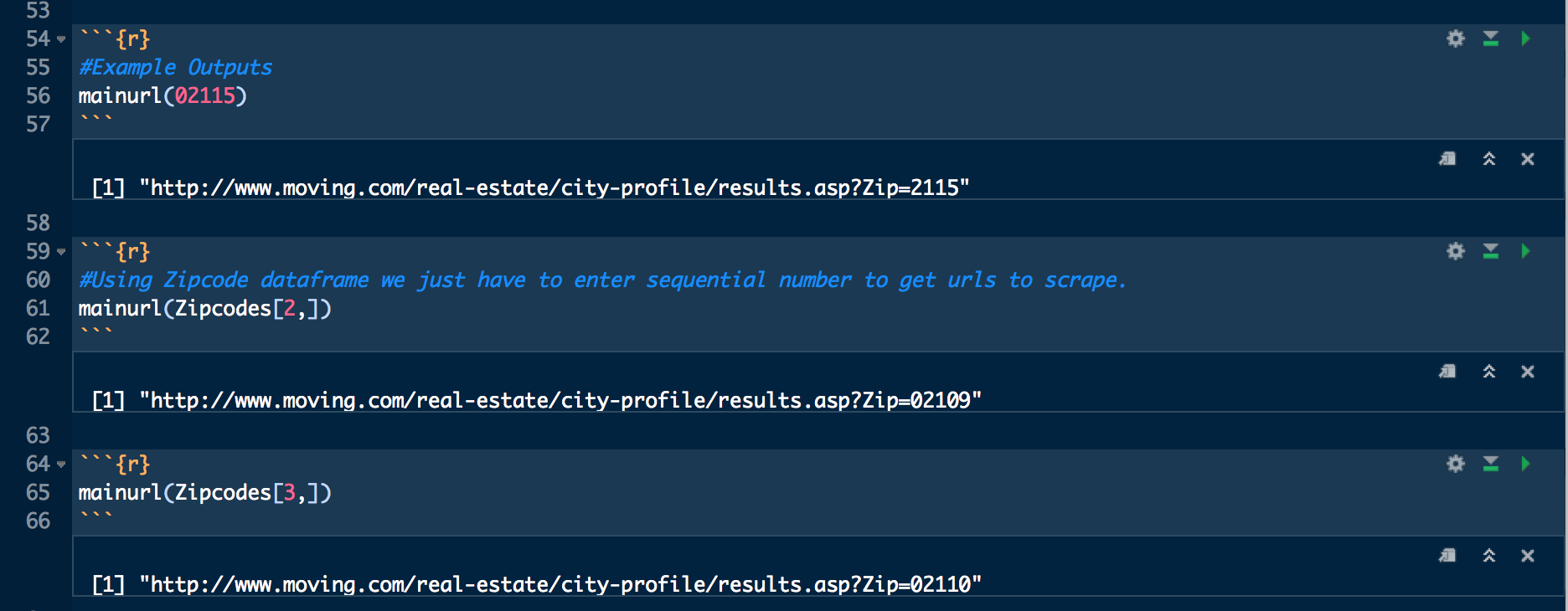
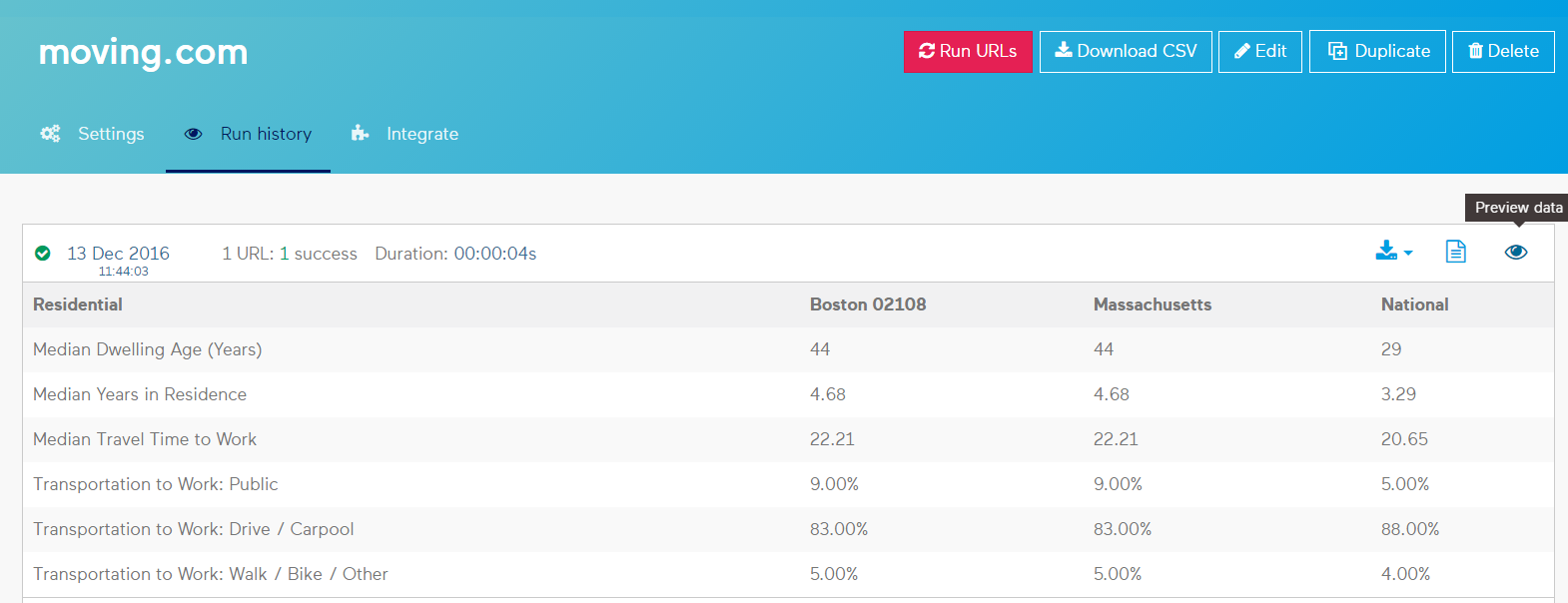
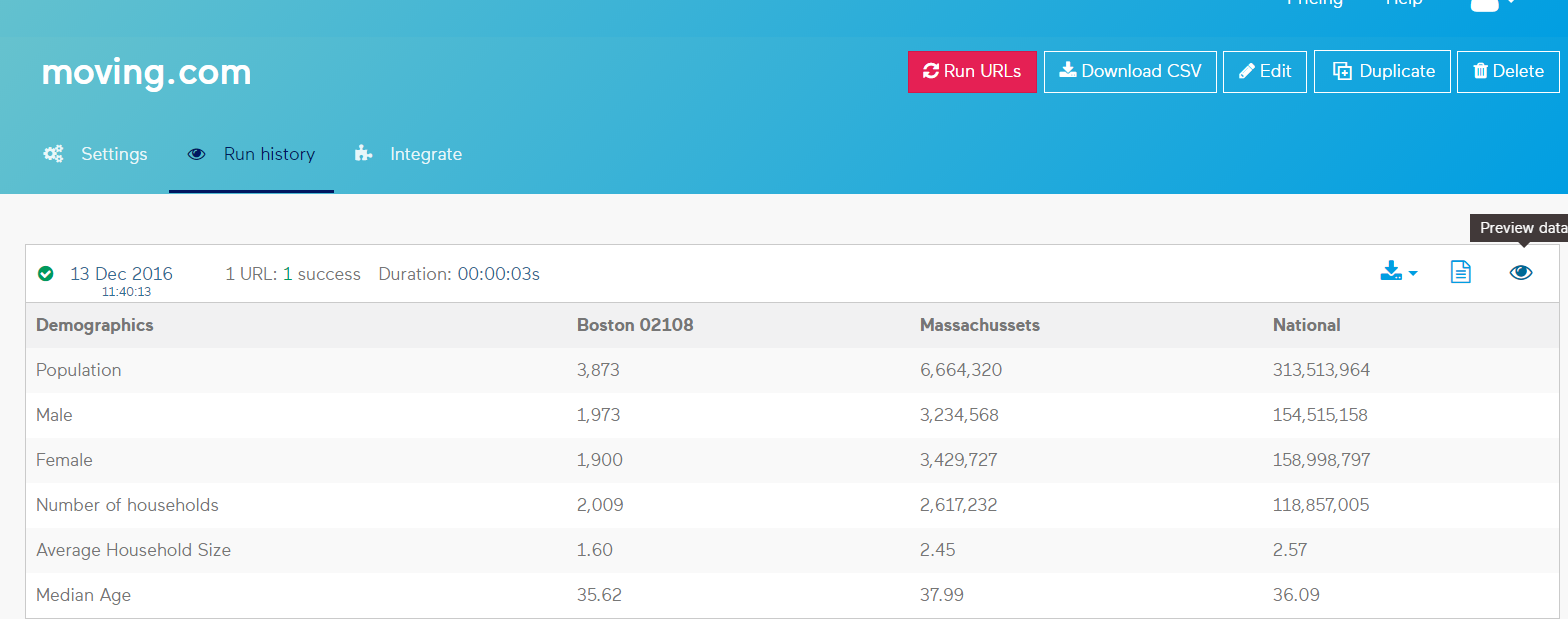
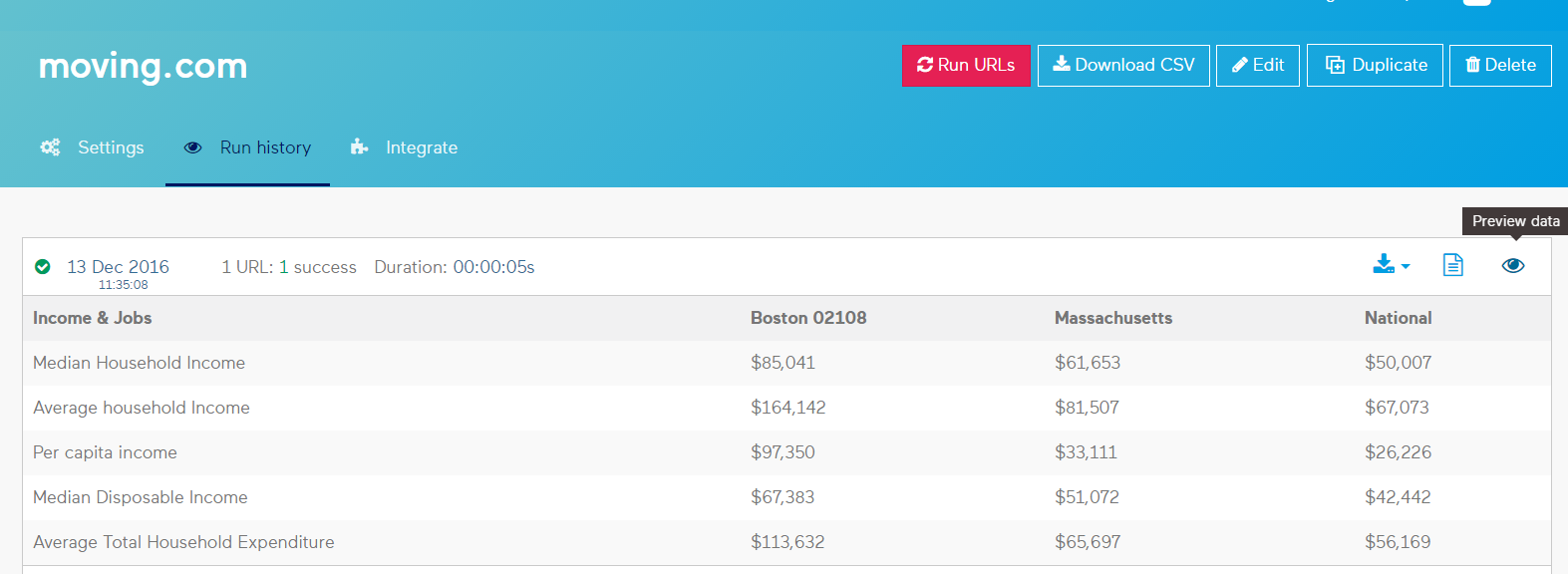
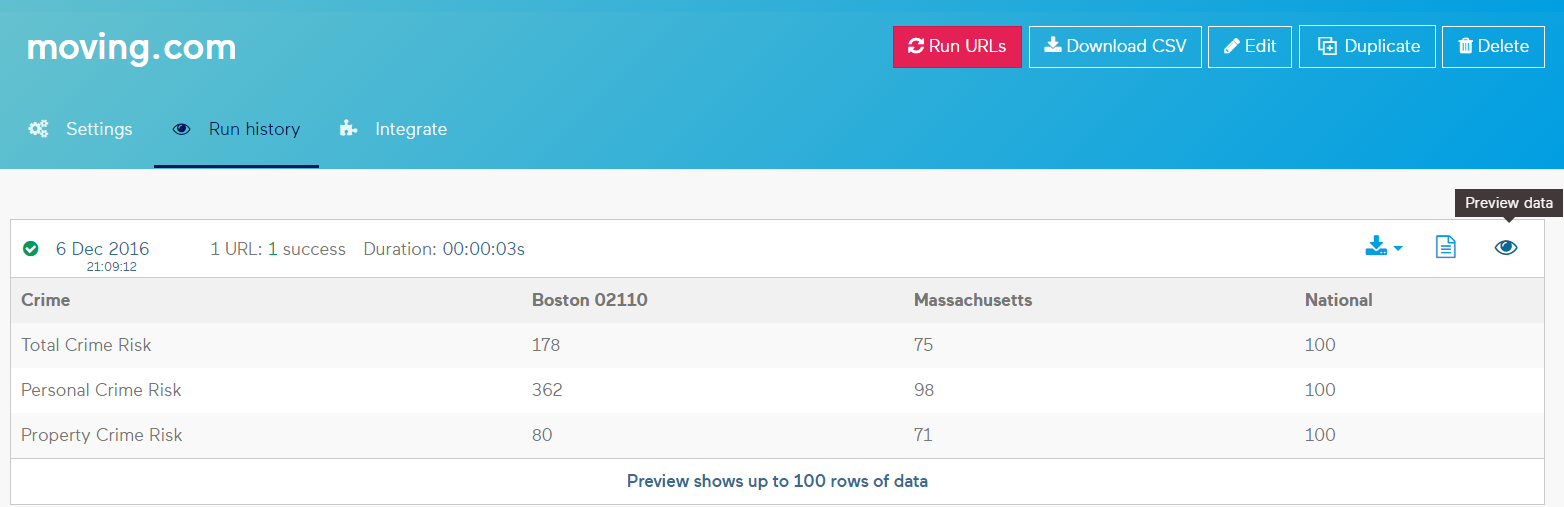
#Empty DataFrame to collect scraped data for the crime stats.  
CrimeStats <- data.frame(Zipcodes=character(),   
 TotalCrimeRisk=numeric(),   
 PersonalCrimeRisk=numeric(),   
 PropertyCrimeRisk=numeric(),   
 stringsAsFactors=FALSE)

#Scraping data for various zip codes from moving.com website.  
  
View(Zipcodes) #Dataframe of thirty zipcodes used in project

#function to get url from moving.com for various zipcode to scrape data.   
mainurl <- function(zipcode) {   
 url <- paste("http://www.moving.com/real-estate/city-profile/results.asp?Zip=", zipcode, sep="")  
 print(url)  
}

#Example Outputs  
mainurl(02115)  
  
#Using Zipcode dataframe we just have to enter sequential number to get urls to scrape.  
mainurl(Zipcodes[2,])  
  
mainurl(Zipcodes[3,])

library(bitops)  
library(xml2)  
library(rvest)  
library(stringr)  
  
#Storing url generted from the above fucntion as "Crime"  
Crime <- mainurl(02118)  
Crime <- read\_html(Crime)  
  
#This will create the scrapelist for the url generated from the above function.   
scrapelist <- Crime %>% html\_nodes("table:nth-child(7) td:nth-child(3) , table:nth-child(7) th:nth-child(3) , table:nth-child(7) .first\_cpth , table:nth-child(7) td:nth-child(1) a") %>% html\_text() %>% str\_trim()  
  
#This will fill the empty dataframe created with the scraped webpage information  
scrapelist01 <- cbind(CrimeStats, scrapelist)  
  
#This way after getting url from all the "30" zip codes of Boston we formed 30 "scrapelist01, scrapelist02....scrapelist30" and got table crimestats and renamed it to "Crimemain"" Dataframe  
  
#After retreving all the scrapelist we renamed CrimeStats into Crimemain table and it has all the data we needed for the table



1. As we are considering Listing ID as the primary key, it cannot be NA. Hence, we are replacing all NA values with unique values

Housingmain$LISTING.ID[is.na(Housingmain$LISTING.ID)] <- sample(1:4579, size=sum(is.na(Housingmain$LISTING.ID)), replace=F)

1. Reshaping the data we obtained as per our needs

#Removing "." from the column names of different data frames so that it will not trouble during quering in SQLite Database while retriving data.  
  
names(Housingmain) <- gsub(x = names(Housingmain),  
 pattern = "\\.",  
 replacement = " ")

names(Income) <- gsub(x = names(Income),  
 pattern = "\\.",  
 replacement = " ")

names(Residential) <- gsub(x = names(Residential),  
 pattern = "\\.",  
 replacement = " ")

names(Demographics) <- gsub(x = names(Demographics),  
 pattern = "\\.", replacement = " ")

#Removing Whitespaces from the Dataframes  
names(Housingmain) <- gsub(" ", "", names(Housingmain))

names(Residential) <- gsub(" ", "", names(Residential))

names(Income) <- gsub(" ", "", names(Income))

names(Demographics) <- gsub(" ", "", names(Demographics))

#Replacing NA values in main table to random unique values so that we can set listing.ID as Primary key during making database design.  
Housingmain$LISTINGID[is.na(Housingmain$LISTINGID)] <- sample(1:4579, size=sum(is.na(Housingmain$LISTINGID)), replace=F)

#Replacing NA in Days in Market column with the avg. number of days of the column so that intergrity of data is maintained and we can ask some of the questions in the project we had.  
  
Housingmain$DAYSONMARKET[is.na(Housingmain$DAYSONMARKET)] <- round(mean(Housingmain$DAYSONMARKET, na.rm = TRUE))

Calcuating if there are any other NA in the Days on Market table

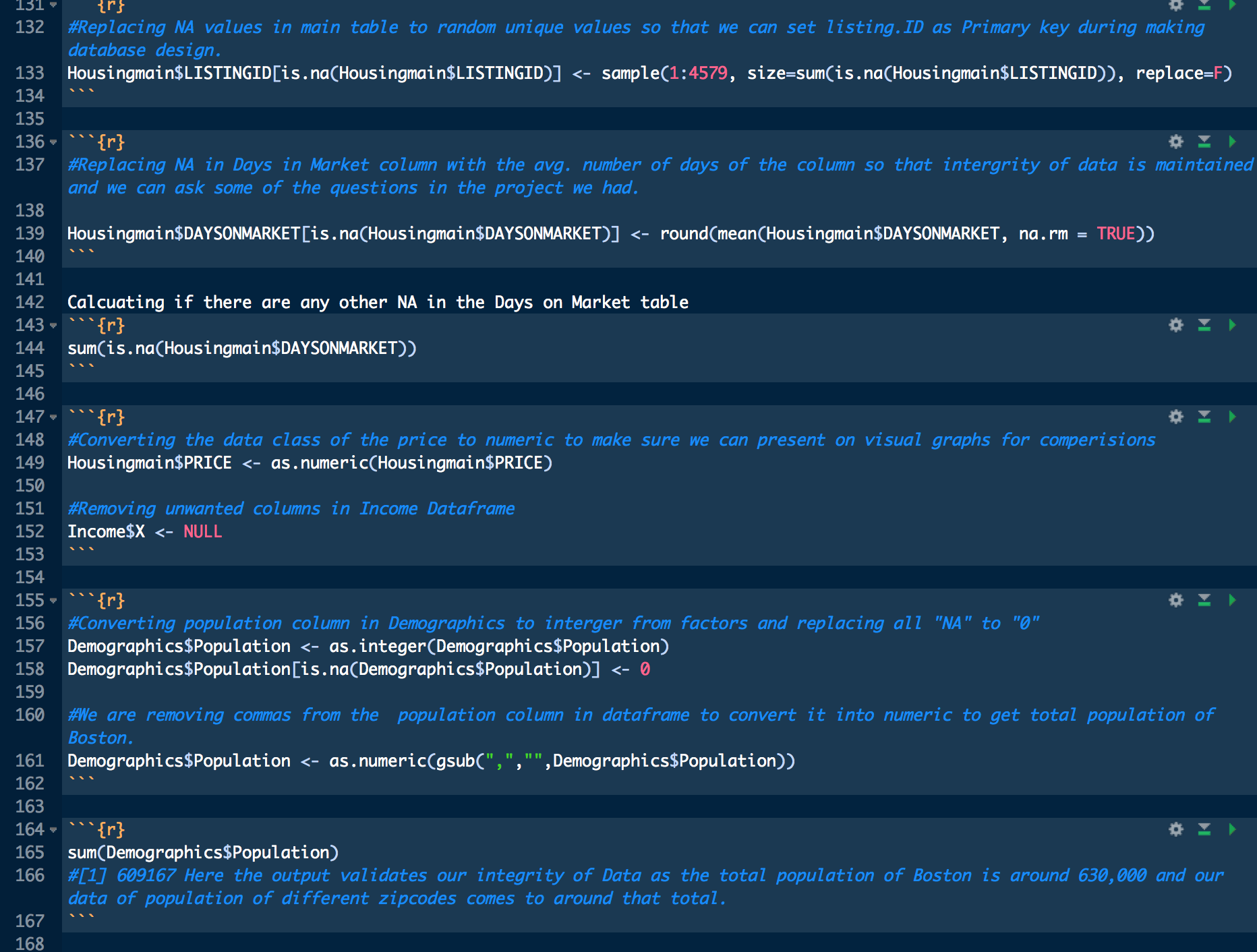
sum(is.na(Housingmain$DAYSONMARKET))

#Converting the data class of the price to numeric to make sure we can present on visual graphs for comperisions  
Housingmain$PRICE <- as.numeric(Housingmain$PRICE)  
  
#Removing unwanted columns in Income Dataframe  
Income$X <- NULL

#Converting population column in Demographics to interger from factors and replacing all "NA" to "0"   
Demographics$Population <- as.integer(Demographics$Population)  
Demographics$Population[is.na(Demographics$Population)] <- 0

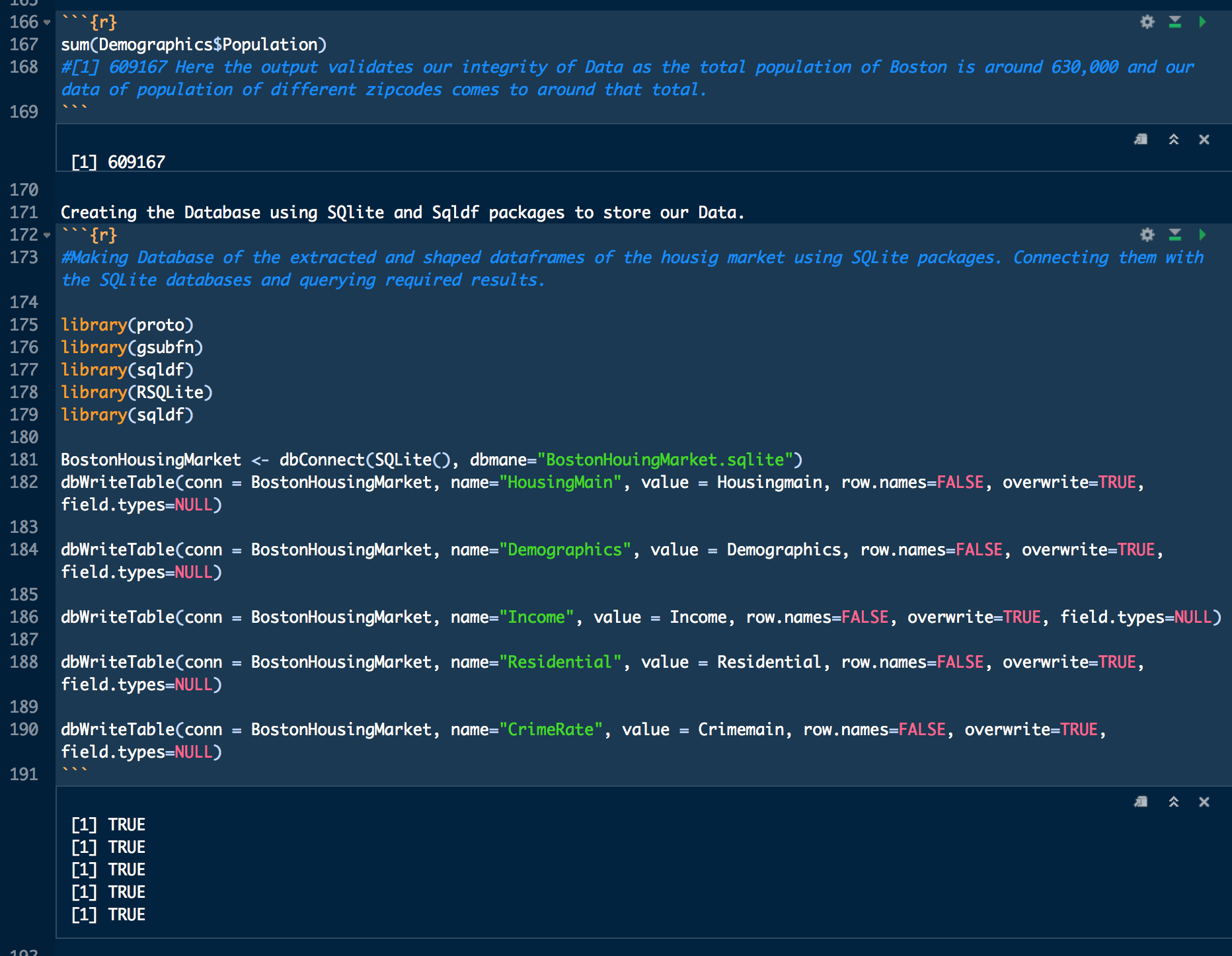
#We are removing commas from the population column in dataframe to convert it into numeric to get total population of Boston.   
Demographics$Population <- as.numeric(gsub(",","",Demographics$Population))

sum(Demographics$Population)  
#[1] 609167 Here the output validates our integrity of Data as the total population of Boston is around 630,000 and our data of population of different zipcodes comes to around that total.



1. We created the SQLite database and inserted all the designed data frames into the Boston Housing Market database. Connecting them with the SQLite databases and querying required results.

library(proto)  
library(gsubfn)  
library(sqldf)  
library(RSQLite)  
  
BostonHousingMarket <- dbConnect(SQLite(), dbmane="BostonHouingMarket.sqlite")  
dbWriteTable(conn = BostonHousingMarket, name="HousingMain", value = Housingmain, row.names=FALSE, overwrite=TRUE, field.types=NULL)  
  
dbWriteTable(conn = BostonHousingMarket, name="Demographics", value = Demographics, row.names=FALSE, overwrite=TRUE, field.types=NULL)  
  
dbWriteTable(conn = BostonHousingMarket, name="Income", value = Income, row.names=FALSE, overwrite=TRUE, field.types=NULL)  
  
dbWriteTable(conn = BostonHousingMarket, name="Residential", value = Residential, row.names=FALSE, overwrite=TRUE, field.types=NULL)  
  
dbWriteTable(conn = BostonHousingMarket, name="CrimeRate", value = Crimemain, row.names=FALSE, overwrite=TRUE, field.types=NULL)



Listing the tables we created in BostonHousingMarket Database

dbListTables(BostonHousingMarket)

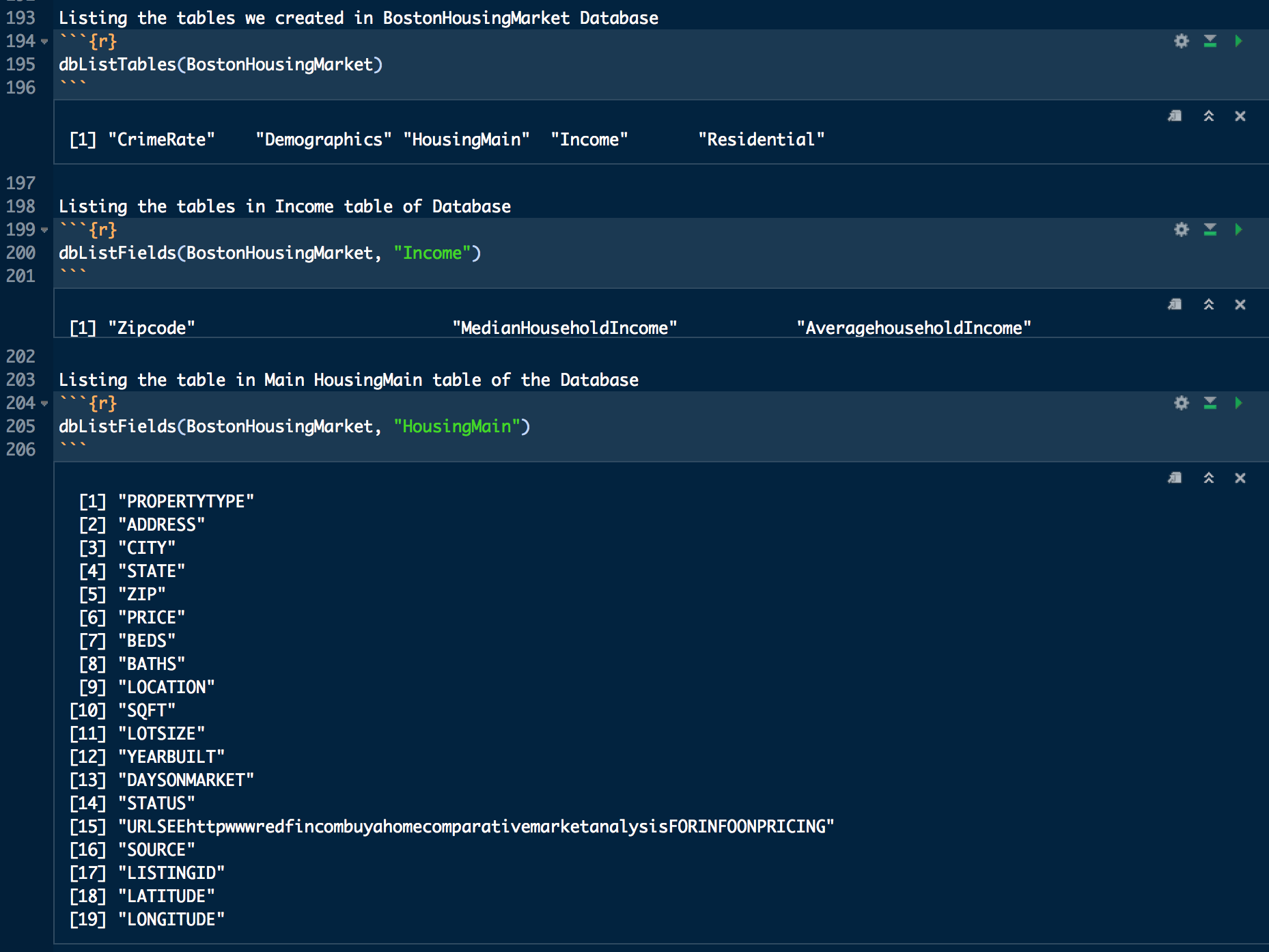
Listing the tables in Income table of Database

dbListFields(BostonHousingMarket, "Income")

Listing the table in Main HousingMain table of the Database

dbListFields(BostonHousingMarket, "HousingMain")

dbGetQuery(conn = BostonHousingMarket, 'Select count(\*) from Demographics')

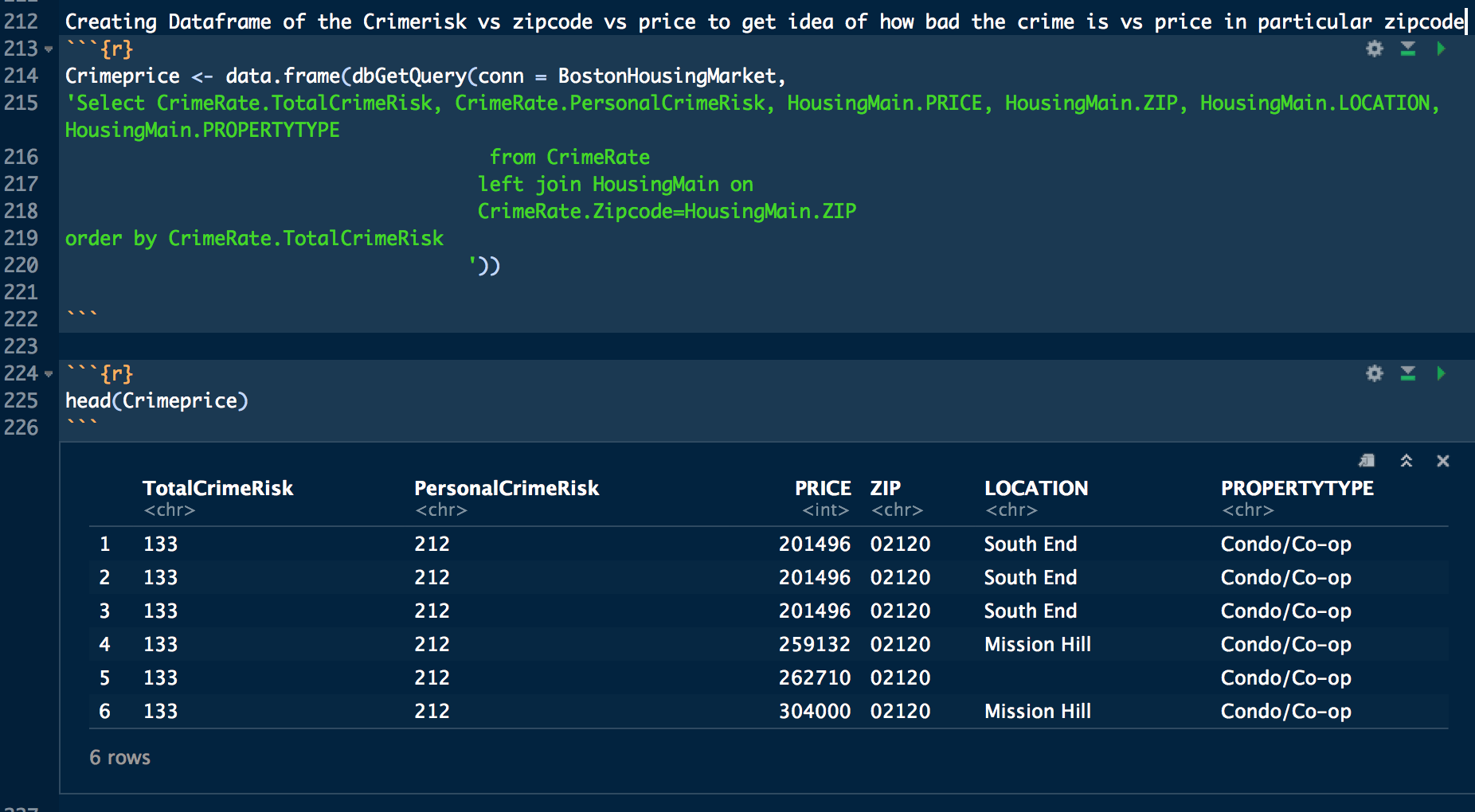


1. The queries that we will be answering with their solutions

Creating Dataframe of the Crimerisk vs zipcode vs price to get idea of how bad the crime is vs price in particular zipcode

Crimeprice <- data.frame(dbGetQuery(conn = BostonHousingMarket,   
'Select CrimeRate.TotalCrimeRisk, CrimeRate.PersonalCrimeRisk, HousingMain.PRICE, HousingMain.ZIP, HousingMain.LOCATION, HousingMain.PROPERTYTYPE  
 from CrimeRate  
 left join HousingMain on   
 CrimeRate.Zipcode=HousingMain.ZIP   
order by CrimeRate.TotalCrimeRisk  
 '))

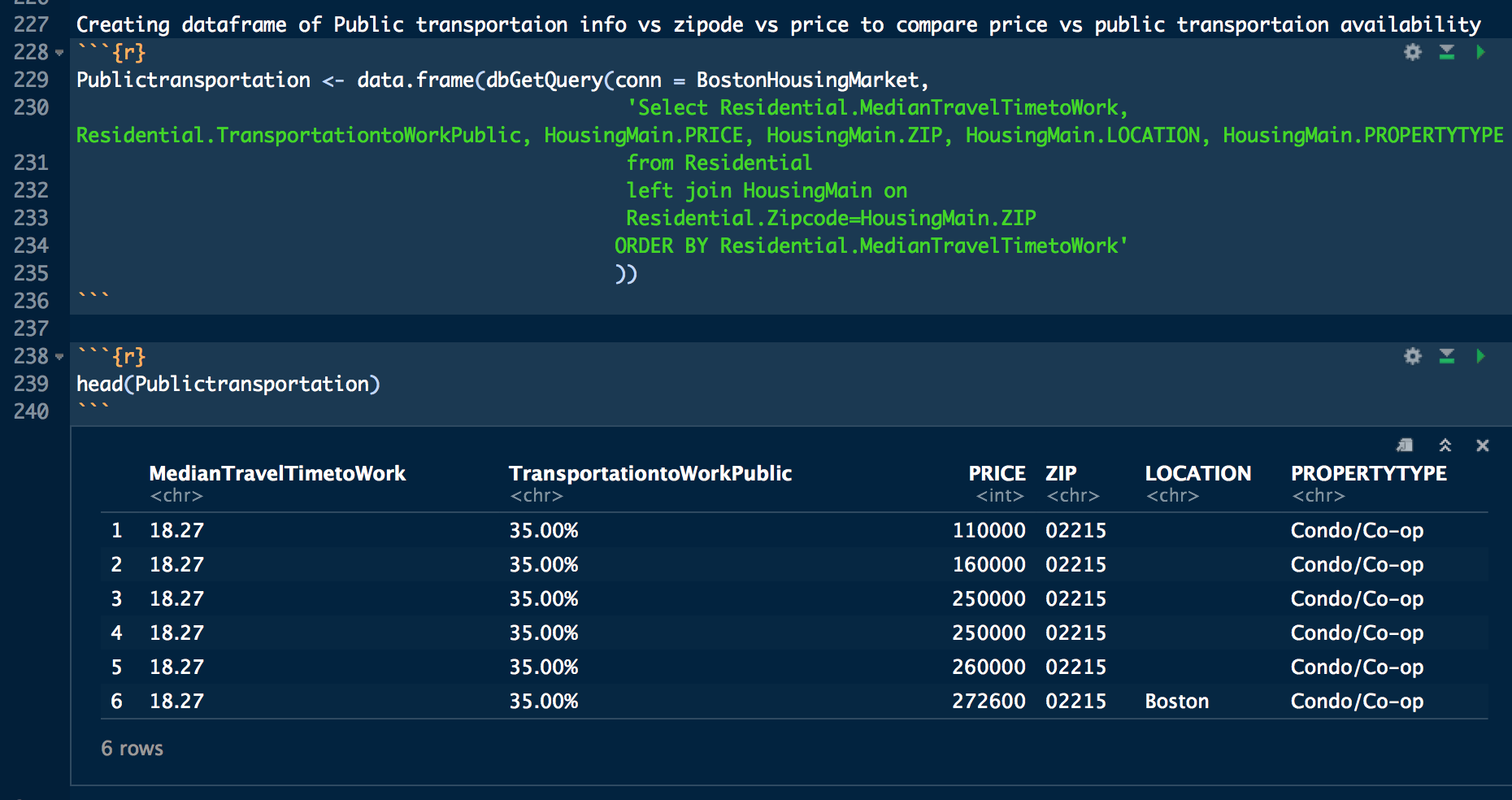
head(Crimeprice)



Creating dataframe of Public transportaion info vs zipode vs price to compare price vs public transportaion availability

Publictransportation <- data.frame(dbGetQuery(conn = BostonHousingMarket,   
 'Select Residential.MedianTravelTimetoWork, Residential.TransportationtoWorkPublic, HousingMain.PRICE, HousingMain.ZIP, HousingMain.LOCATION, HousingMain.PROPERTYTYPE  
 from Residential   
 left join HousingMain on   
 Residential.Zipcode=HousingMain.ZIP  
 ORDER BY Residential.MedianTravelTimetoWork'  
 ))

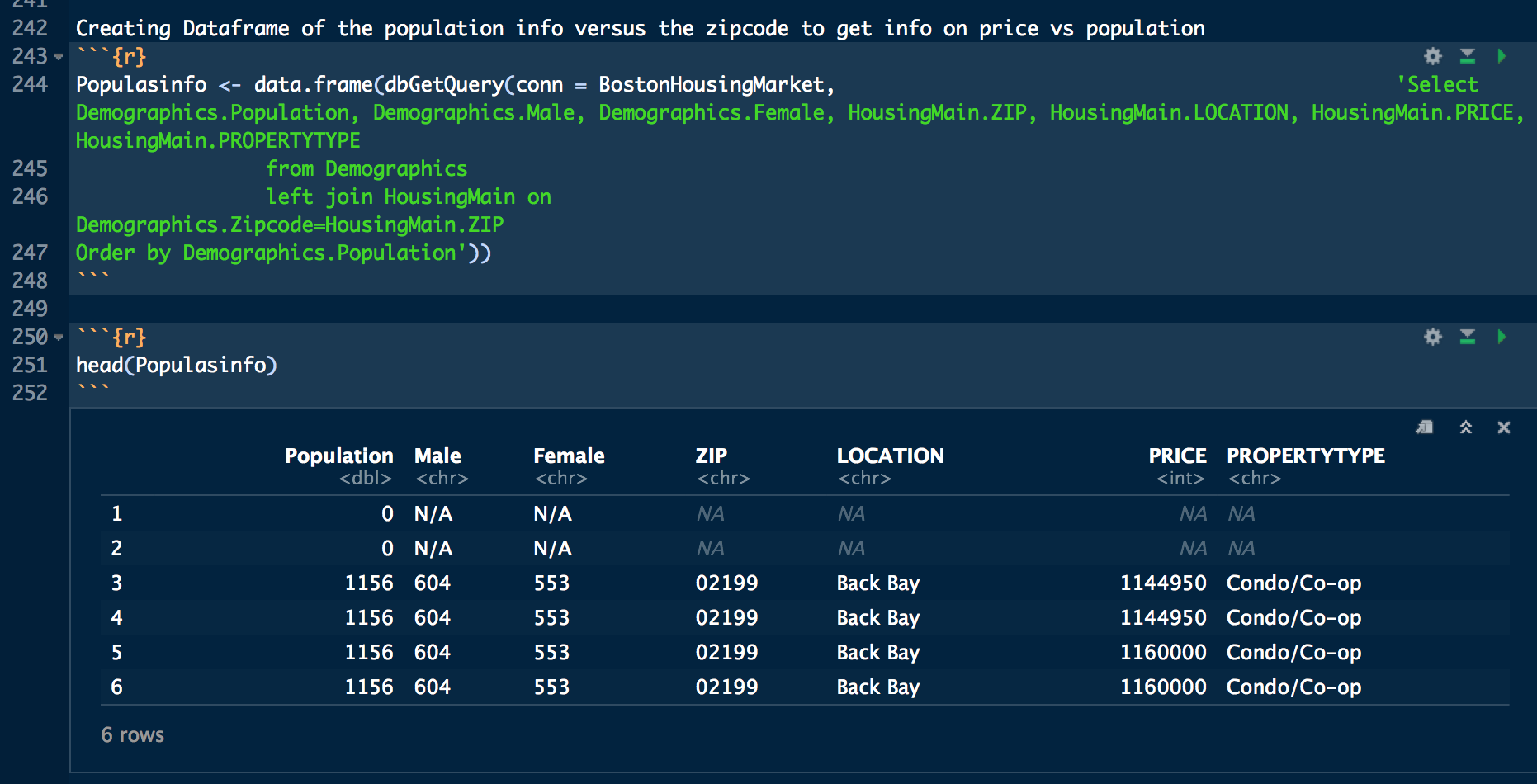
head(Publictransportation)



Creating Dataframe of the population info versus the zipcode to get info on price vs population

Populasinfo <- data.frame(dbGetQuery(conn = BostonHousingMarket, 'Select Demographics.Population, Demographics.Male, Demographics.Female, HousingMain.ZIP, HousingMain.LOCATION, HousingMain.PRICE, HousingMain.PROPERTYTYPE  
 from Demographics   
 left join HousingMain on Demographics.Zipcode=HousingMain.ZIP   
Order by Demographics.Population'))

head(Populasinfo)



* Please note: In our data we are comparing the index number with the Masschusetts and National average for all the parameters we have mentioned in our data that we scraped from moving.com

*Demographics*

* Population MA - 6,664,320
* Population National - 313,513,964
* Male MA - 3,234,568
* Male National - 154,515,158
* Female MA - 3,429,727
* Female National - 158,998,797
* Number of households MA - 2,617,232
* Number of households National - 118,857,005
* Average Household Size MA - 2.45
* Average Household Size National - 2.57
* Median Age MA - 37.99
* Median Age National - 36.09

*Residential*

* Median Dwelling Age (Years) MA - 44
* Median Dwelling Age (Years) National - 29
* Median Years in Residence MA - 4.68
* Median Years in Residence National - 3.29
* Median Travel Time to Work MA - 22.21
* Median Travel Time to Work National - 20.65
* Transportation to Work: Public MA - 9.00%
* Transportation to Work: Public National - 5.00%
* Transportation to Work: Drive / Carpool MA - 83.00%
* Transportation to Work: Drive / Carpool National - 88.00%
* Transportation to Work: Walk / Bike / Other MA - 5.00%
* Transportation to Work: Walk / Bike / Other National - 4.00%

*Income*

* Median Household Income MA - $61,653
* Median Household Income National - $50,007
* Average household Income MA - $81,507
* Average household Income National - $67,073
* Per capita income MA - $33,111
* Per capita income National - $26,226
* Median Disposable Income MA - $51,072
* Median Disposable Income National - $42,442
* Average Total Household Expenditure MA - $65,697
* Average Total Household Expenditure National - $56,169

*CrimeRate*

* Total Crime Risk MA - 75
* Total Crime Risk National - 100
* Personal Crime Risk MA - 98
* Personal Crime Risk National - 100
* Property Crime Risk MA - 71
* Property Crime Risk National - 100

### *Issues faced*

* For every query to obtain data from Zillow, we have to input complete adress with zipcode as input, which was a tedious task. We also attempted to use chron job for this purpose, however it was not feasible.
* For obtaining data we also tried city-data, however to scrape data from this website we needed a paid subscription.
* To extract data from openstreet.map was very complex, hence we did not attempt it any further.
* We tried to scrape information from moving.com, however we were not able to extract the data properly. Hence, we extracted the data using import.io.

### *Future Scope*

* We can add the distance between nearby amenities to the address of the residences to our database by tracking the addresses with Google Maps API.
* Using machine learning as an added feature in our project we can forecast future prices of houses in a particular zipcode and based on that prices we can create rent expectations for the houses.
* We can use ggplot2 packaqe in R to map the population density in Boston.

### *Appendix*

This appendix is for the term we used in our project as parameters in our data.

*Demographic*

* Average Household Size - Average size of area households in number of people.
* Median Age (in Years) - Half of the total population is above the listed age, half is below it.

*Income*

* Median Household Income (in Dollars) - Half of the households have an income above the amount listed, half have an income below it.
* Average Household Income (in Dollars) - Average combined income of households in the area.
* Per Capita Income (in Dollars) - Average individual income of people in the area.
* Median Disposable Income (in Dollars) - Half the households have disposable income above the amount listed; half have disposable income below it.
* Average Total Household Expenditure (in Dollars) - Dollars. Includes total retail and nonThe number of people working in this area employed in this role.-retail expenditures per household. Examples of non-retail expenditures are mortgages, rent, insurance, repairs and maintenance--the type of things not typically obtained from retail establishments. They do not include income taxes, savings, and deferred income plans, such as an IRA or 401(k).

*CrimeRate*

* Total Crime Risk - A score that represents the combined risks of rape, murder, assault, robbery, burglary, larceny and vehicle theft compared to the national average of 100. A score of 200 indicates twice the national average total crime risk, while 50 indicates half the national risk. The different types of crime are given equal weight in this score, so murder, for example, does not count more than vehicle theft. Scores are based on demographic and geographic analyses of crime over seven years.
* Personal Crime Risk - Index score (100=National Average) that represents the combined risks of rape, murder, assault and robbery.
* Property Crime Risk - Index score (100=National Average) that represents the combined risks of burglary, larceny and motor vehicle theft.

*Residential*

* Median Dwelling Age (Years) (in Years) - Half the dwellings in the area are older than this, half are newer.
* Median Years in Residence (in Years) - Half the households in the area have had the same primary householder for longer than this; half have had the same primary household for less time.
* Median Travel Time to Work (in Minutes) - Half the people living in this area travel fewer minutes than this to get to work, half travel more minutes.
* Transportation to Work: Public (in Percent) - The percentage of workers living in the area who commute by train, subway, ferry, bus or taxi.
* Transportation to Work: Drive / Carpool (in Percent) - The percentage of workers living in the area who commute as solo drivers, by carpool, or by motorcycle.
* Transportation to Work: Walk / Bike / Other (in Percent) - The percentage of workers living in the area who commute by walking, bicycle or other non-motorized means.

### References

* <http://www.zillow.com/howto/api/APIOverview.html>
* <https://developers.google.com/maps/>
* [http://dj.github.io/boston/#](http://dj.github.io/boston/)
* <https://www.r-bloggers.com/>
* <https://www.redfin.com/?utm_source=google&utm_medium=ppc&utm_campaign=1003170&utm_term=kwd-844252101&utm_content=153163340069>
* <https://www.openstreetmap.org/>
* <http://www.city-data.com/>
* <https://www.walkscore.com/>
* <http://www.moving.com/real-estate/city-profile/>
* [https://www.trulia.com/?cid=sem|google|tbw\_br\_nat\_x\_x\_nat!53f9be4f|Trulia-Broad](https://www.trulia.com/?cid=sem%7Cgoogle%7Ctbw_br_nat_x_x_nat!53f9be4f%7CTrulia-Broad)
* <https://www.jumpshell.com/posts/average-rent-in-boston>
* <http://www.areavibes.com/>
* <https://www.neighborhoodscout.com/?gclid=COye7vOY99ACFcVWDQodJMQBdQ>