

**BOSTON HOUSING ANALYSIS USING R PROGRAMMING**

**SUBMITTED TO: DR. SHILPA BALAN**

PRESENTED BY: KRITHY NANAIAH ATRANGADA

&

ANUPAM SAHAY

**CIS: 5270: BUSINESS INTELLIGENCE**

**OBJECTIVE OF STUDY:** The dataset (Boston Housing Price) was taken from the StatLib library which is maintained at Carnegie Mellon University and is freely available for download from the UCI Machine Learning Repository. The dataset consists of 506 observations of 14 attributes. The Boston housing market is highly competitive and using analysis in R language we want the best real estate agent in the area. Boston Housing provides affordable housing to more than 58,000 residents in and around the city of Boston. Resident are assisted through a combination of public housing and federal and state voucher subsidy programs that provide a wide variety of housing opportunities. Our goal is to use the available data to make the analysis about the home values in Boston area. Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. We also take a look at the Boston crime rate at the various locations, this also helps us in achieving the price rates of various locations in the Boston. Mostly the suburbs area of Boston is more inclined towards the crime. The number of rooms per house is different in different locations, this tells us the various locations in Boston which are thickly populated and are the house are expensive to rent. Another analysis which can be done on this dataset is the distance of the houses at various location to the employment center.

**DATA SETS:** UCI Machine Learning Repository

<https://archive.ics.uci.edu/ml/index.php>

**DATA SET URL:**

<https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>

**BOSTON HOUSING DATASETS:**

This data is taken from UCI. This dataset includes:

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq. Ft
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (1 if tract bounds river; else 0)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
13. LSTAT:  % lower status of the population.
14. MEDV: Median value of owner-occupied homes in $1000’s

**QUESTIONS:**

1. Boston housing value?

2. Crime rate in the suburbs of Boston?

3. Number of rooms in Boston houses?

**REFERENCES:**

1. <https://www.bostonhousing.org/en/Home.aspx>

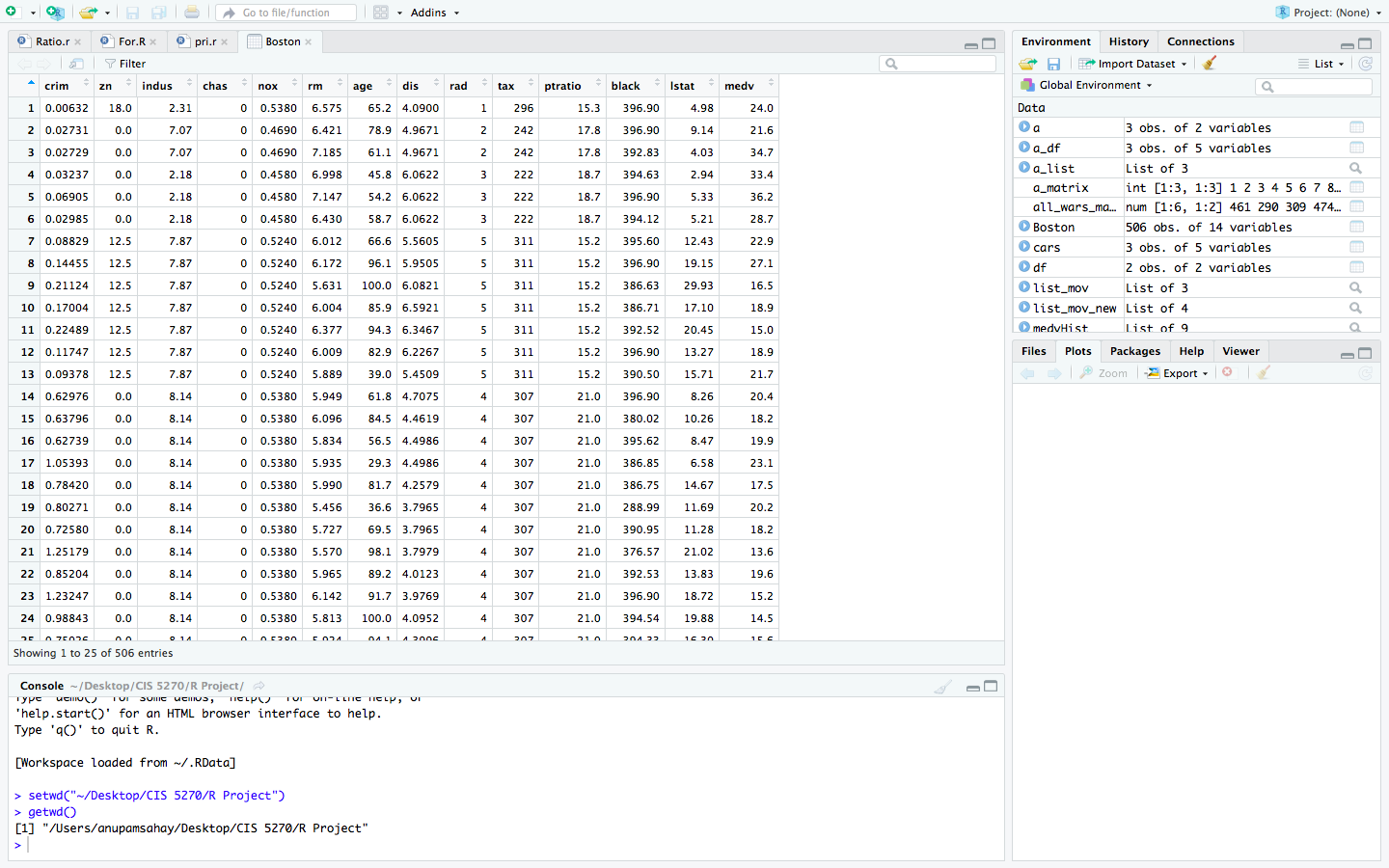
2. <https://www.boston.gov/>

3. <https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names>

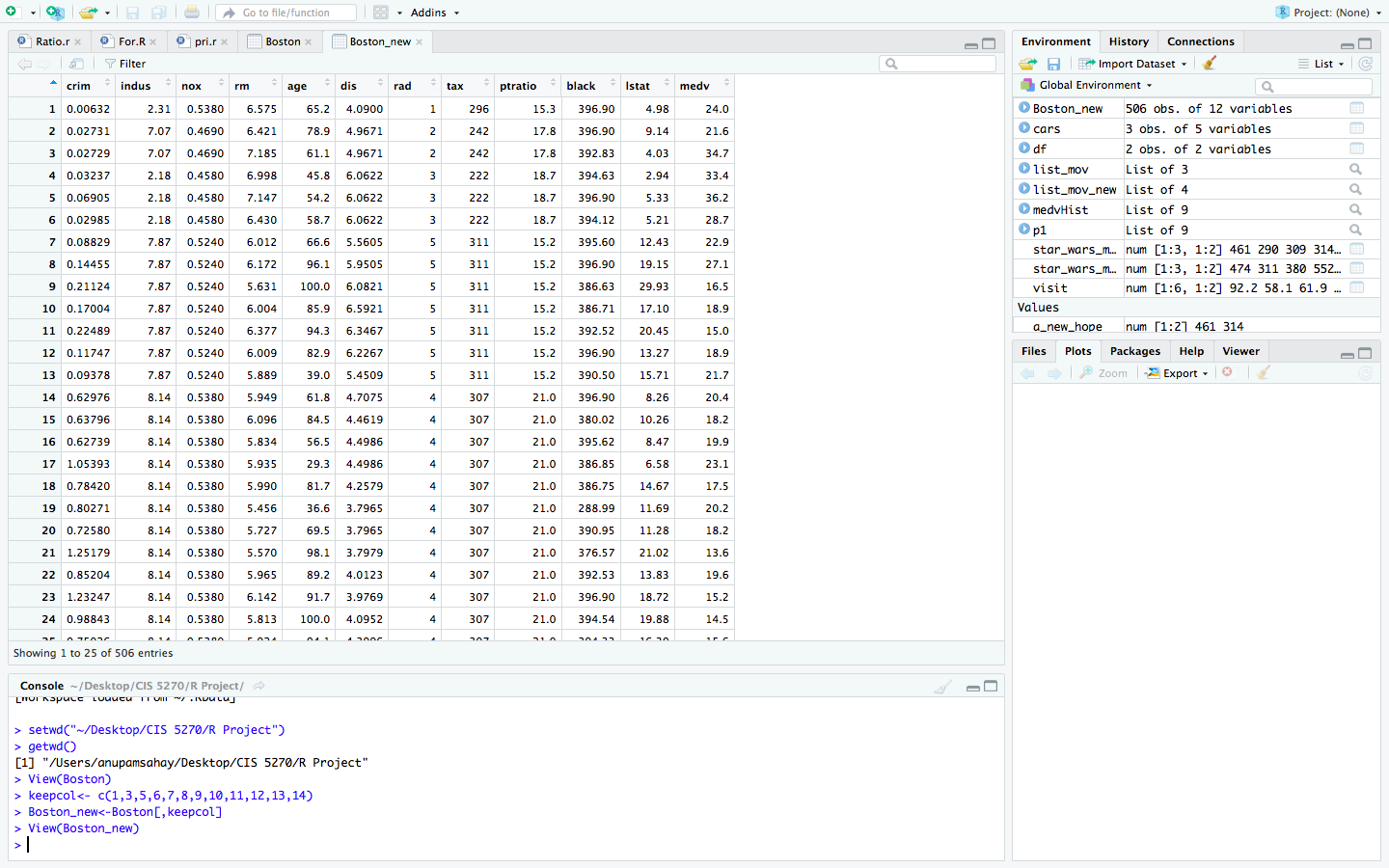
4. [https://www.cityofboston.gov/images\_documents/Guide to the BHA Reports and publications relat\_tcm3-25949.pdf](https://www.cityofboston.gov/images_documents/Guide%20to%20the%20BHA%20Reports%20and%20publications%20relat_tcm3-25949.pdf)

**DATA CLEANING:**

**A) Removing Columns:**

Before: 

After:

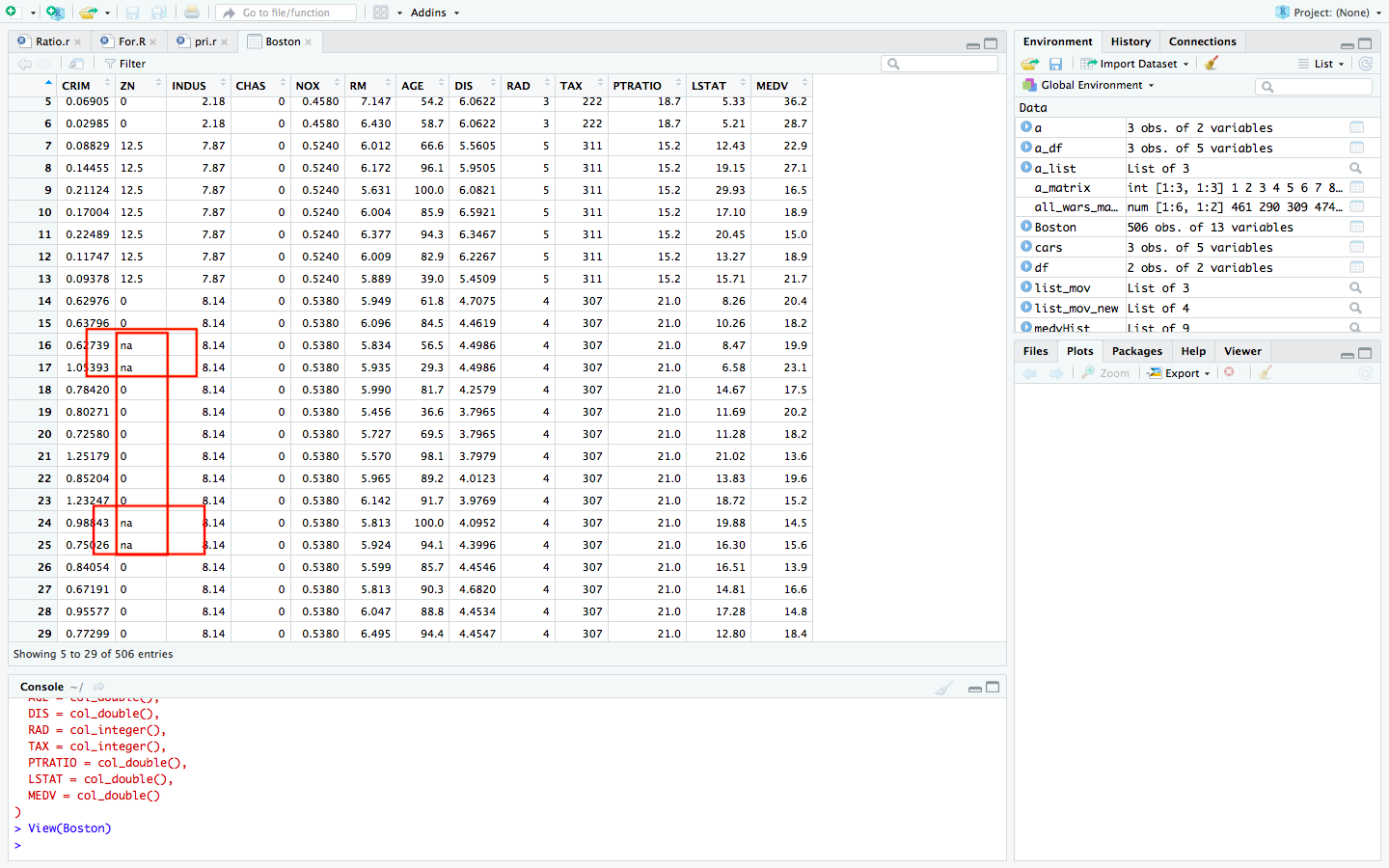


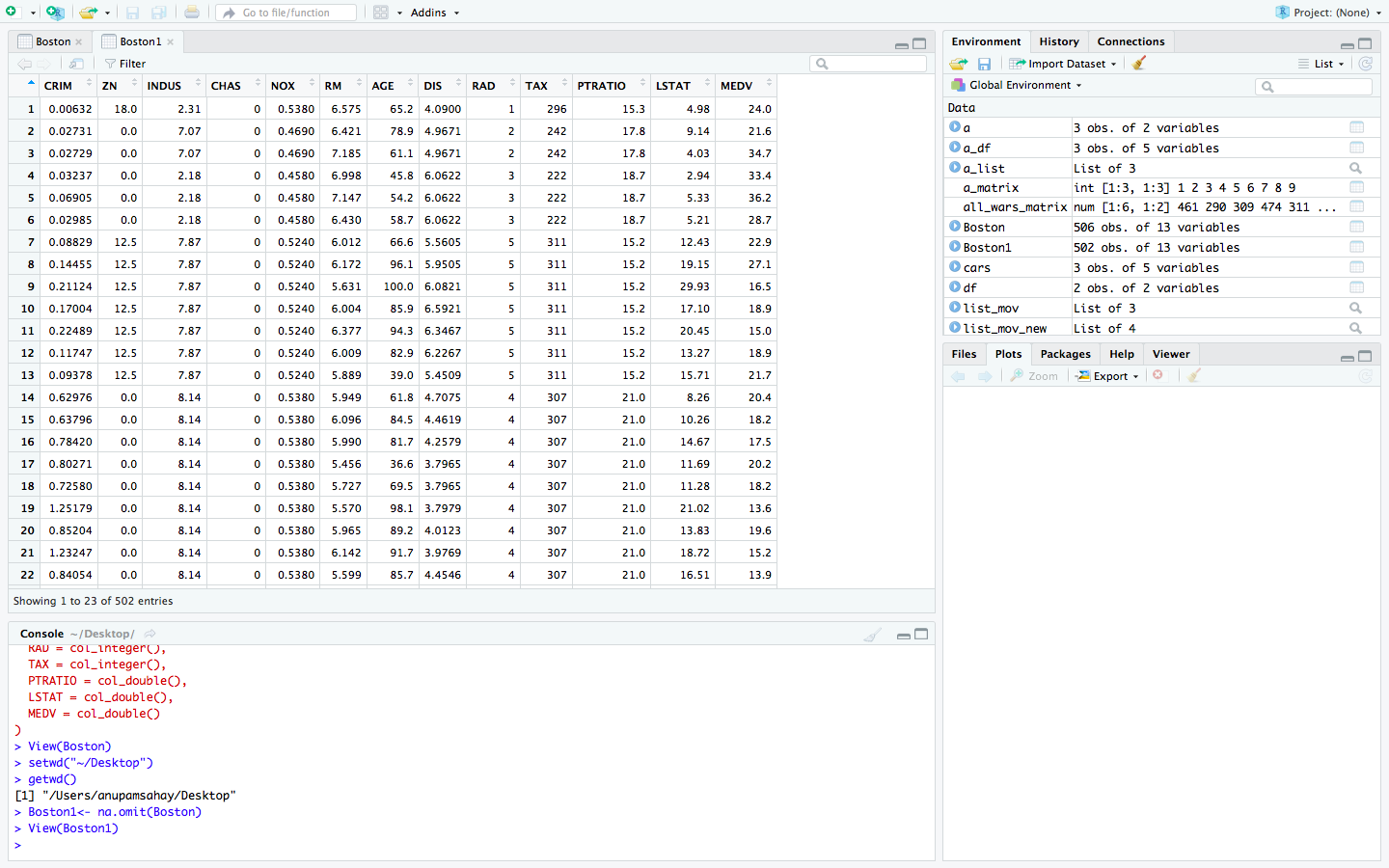
In data cleaning part (A) I have removed the column ZN i.e. proportion of residential land zoned for lots over 25,000 sq. Ft and CHAS: Charles River dummy variable (1 if tract bounds river; else its 0).

Note: All the columns are necessary for the analysis and visualization purposes. The data cleaning is for demo purpose only.

**B) Replacing Null Values:**

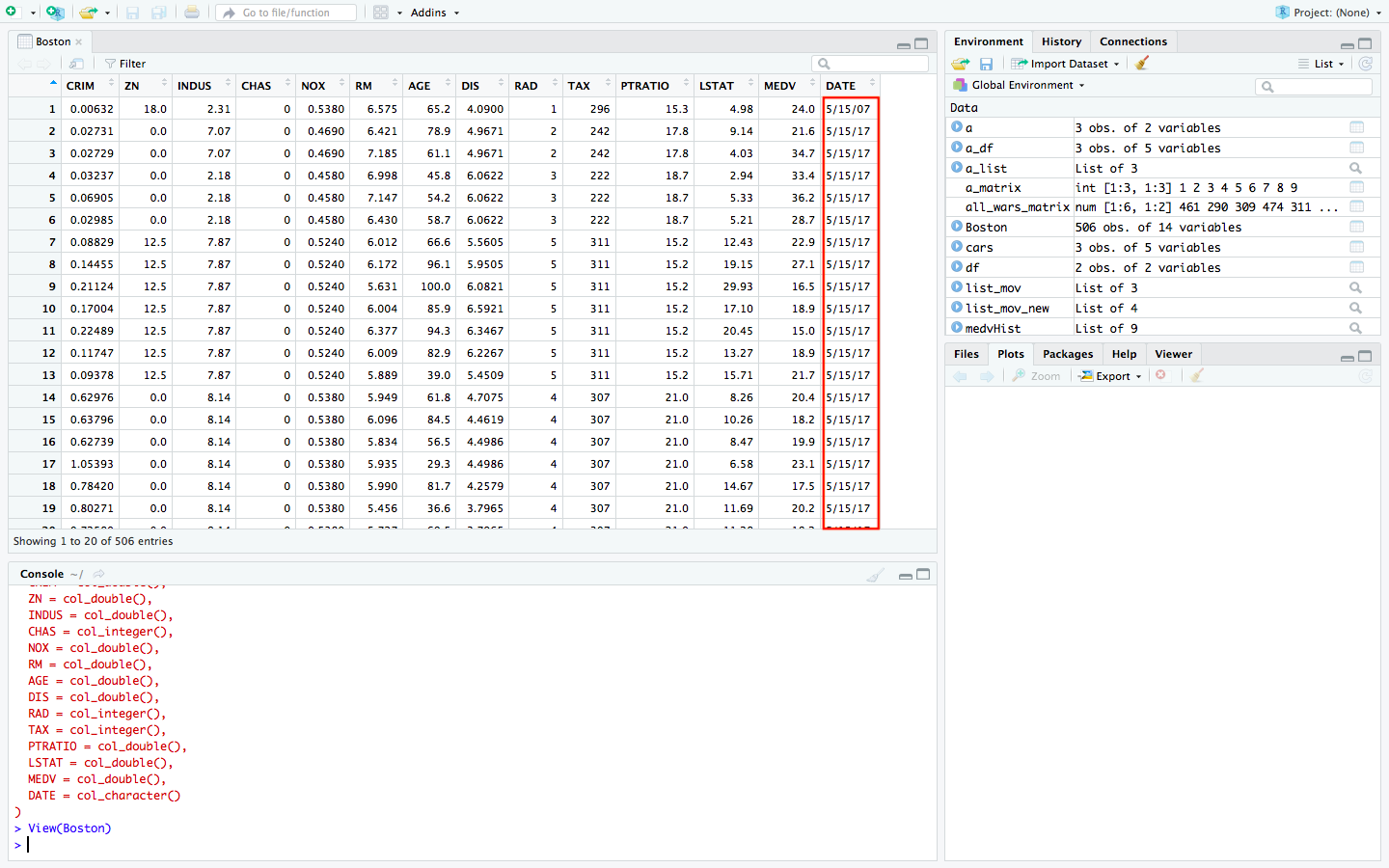
Before:



After:

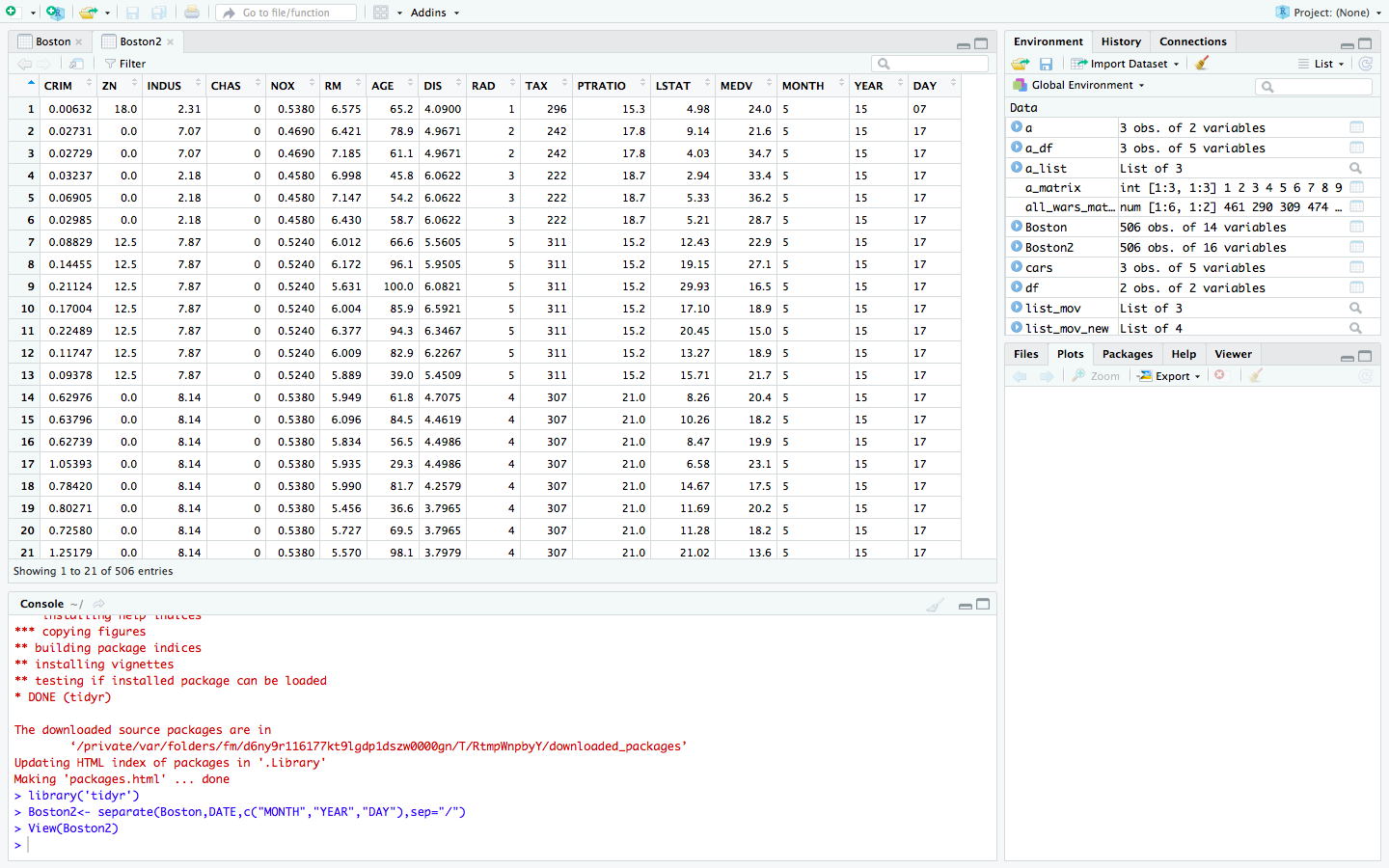
In this data cleaning process, we have cleaned all the NA values from the dataset.

**C) Splitting Values:**

Before:

I have added date column in the dataset. Now, I will be splitting the column in Month, Year and Date.

After:



**Exploring the Boston-Dataset:**

Let’s Install and load the required packages:

> install.packages('readr')

> install.packages('ggplot2')

> install.packages('mlbench')

> install.packages('corrplot')

> install.packages('Amelia')

> install.packages('caret')

> install.packages('plotly')

> install.packages('caTools')

> install.packages('reshape2')

> install.packages('dplyr'**)**

> install.packages("pheatmap")

> library(readr)

> library(ggplot2)

> library(corrplot)

> library(mlbench)

> library(Amelia)

> library(plotly)

> library(reshape2)

> library(caret)

> library(caTools)

> library(dplyr)

> library(pheatmap)

> View(Boston).

> summary(Boston). #To understand the basics statistics of each individual variable.

> class(Boston). #Display the class of R object.

> names(Boston). # The function gives the variable names in the Boston dataset; the

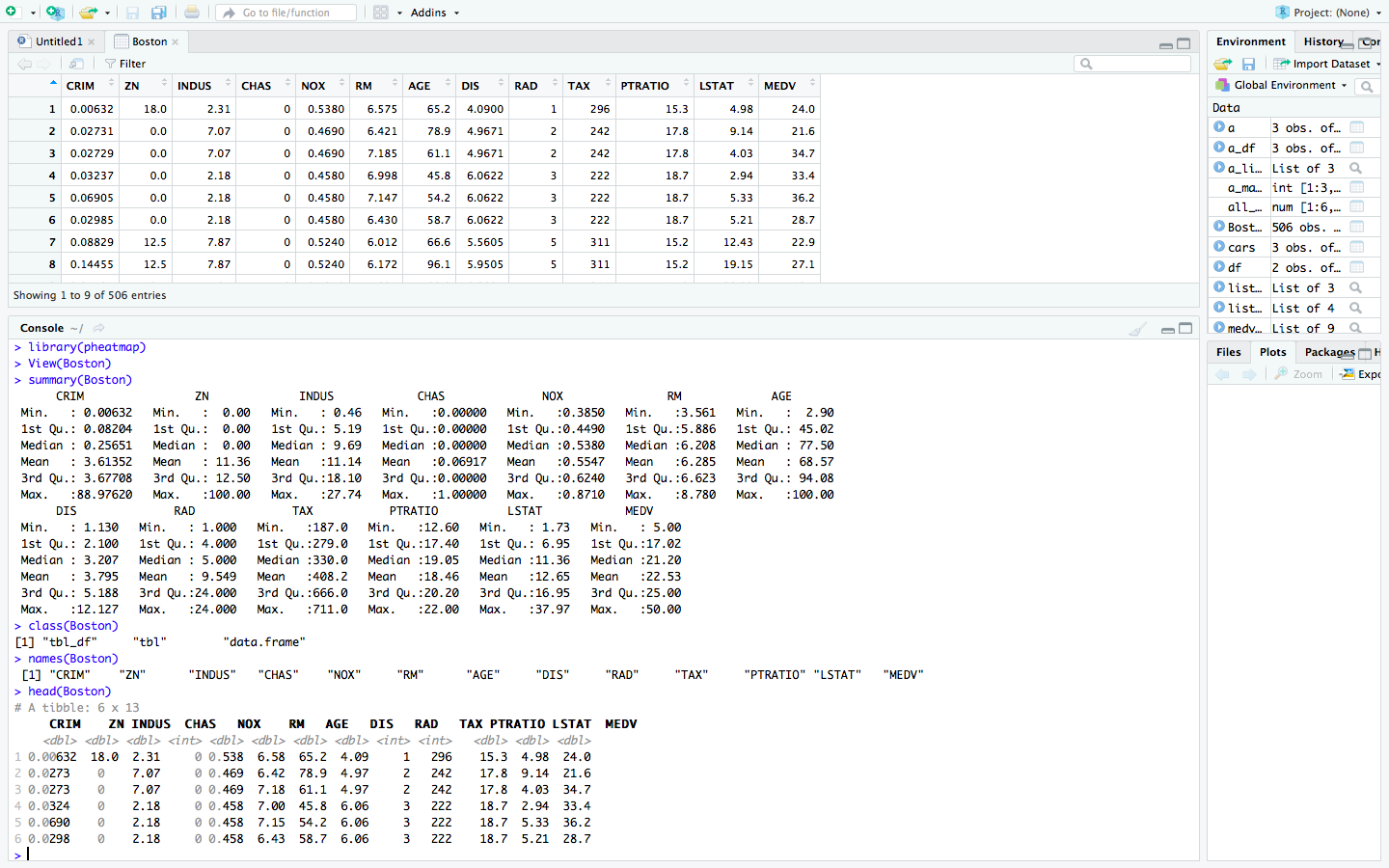
definition has been provided at the beginning for the reference.

> head(Boston). # To view the first few rows in the data in order to check the data as a

Sample.

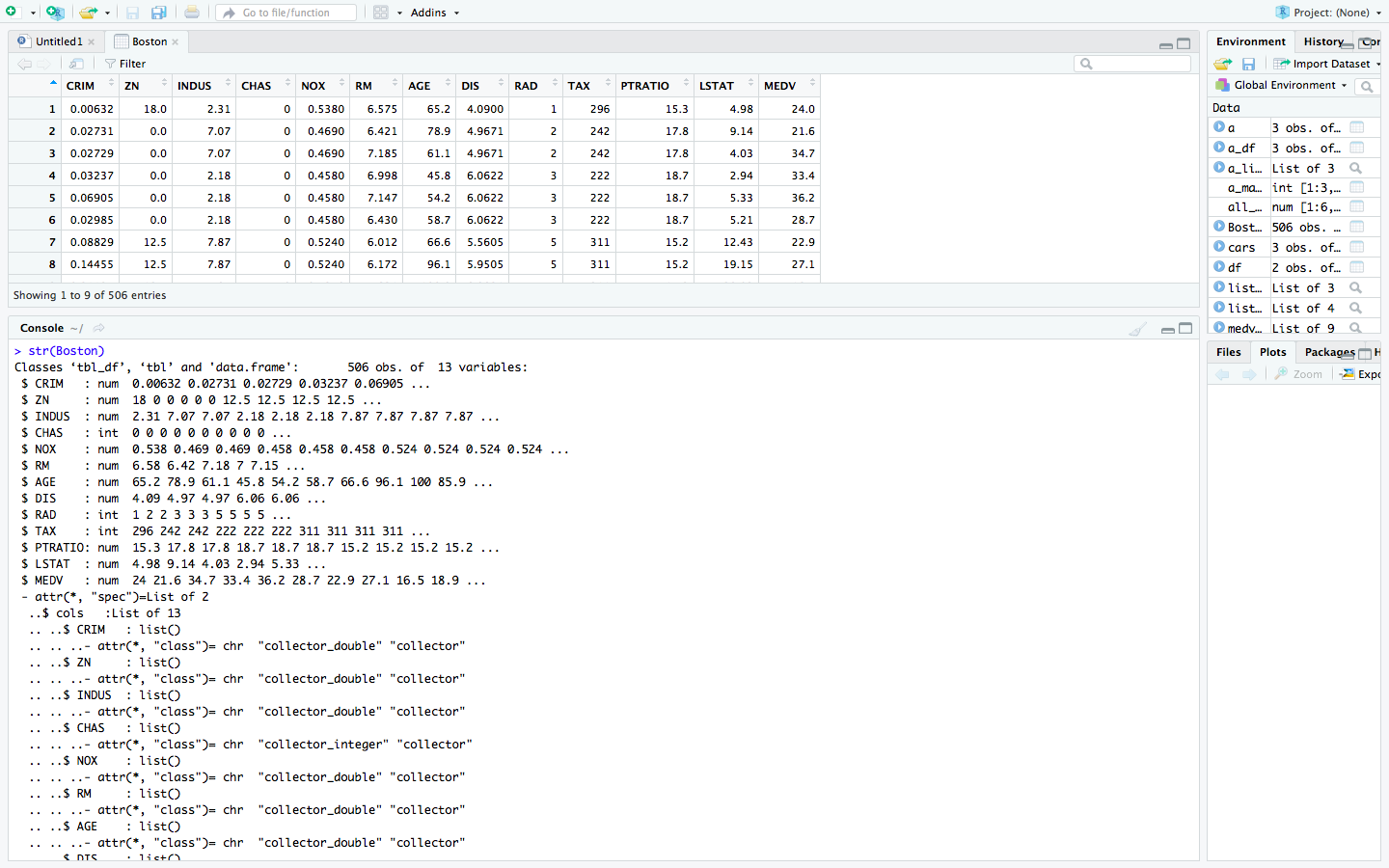
> data(Boston,package="MASS") #This function reads the Boston dataset from the MASS

package in R



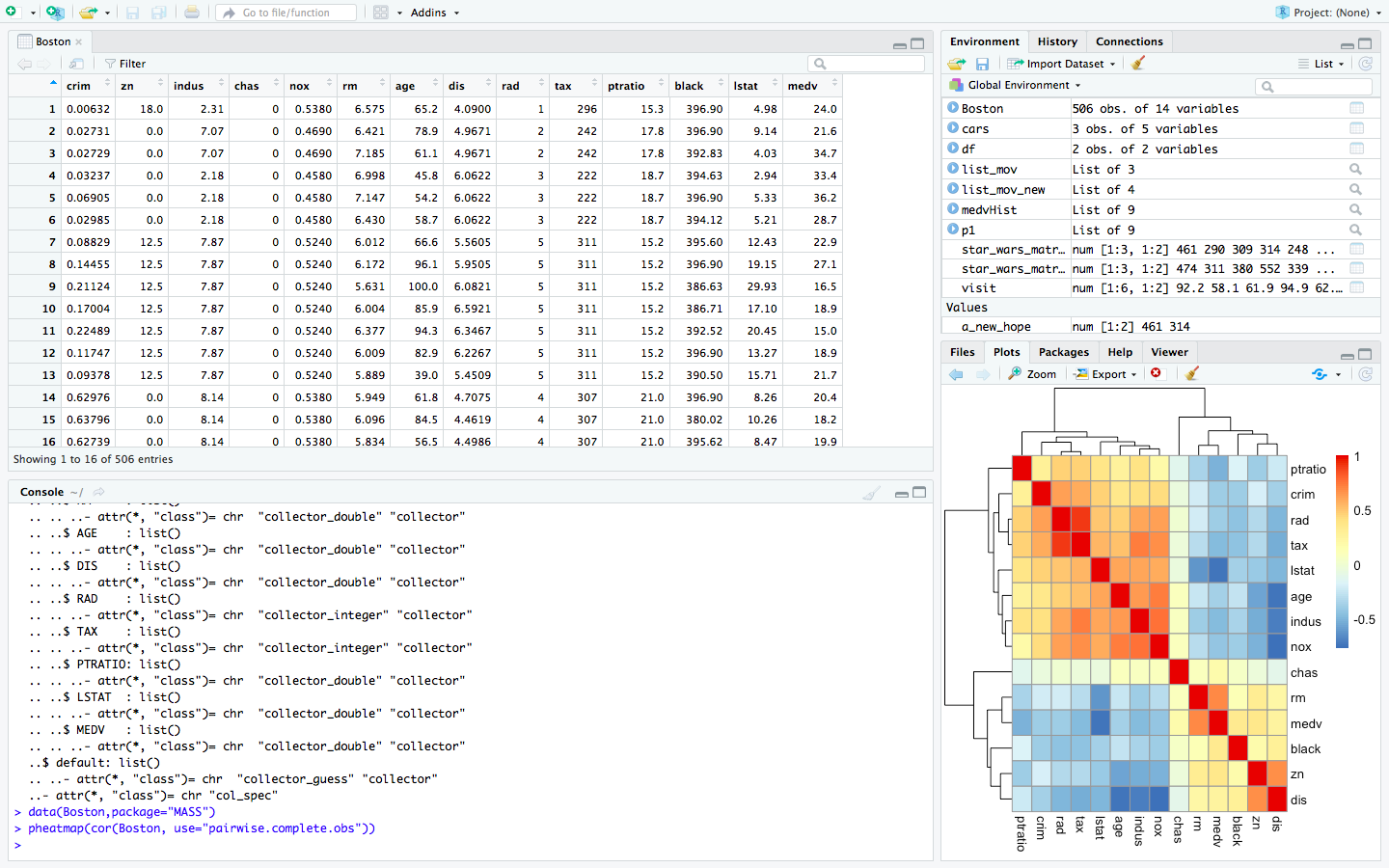
> str(Boston). # This function gives the structure of the data, the total number of variables

and observation in the data and the datatype of each variable

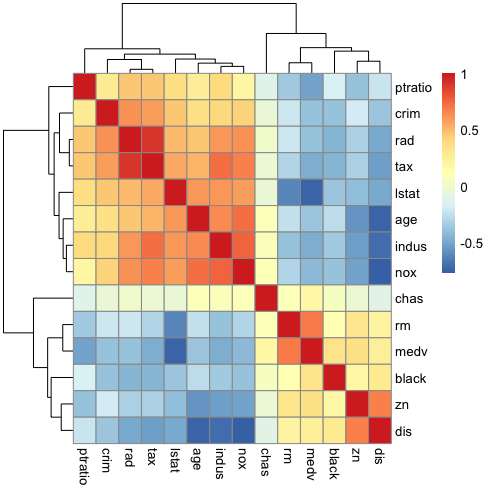


**Understanding the dataset using heatmap:**

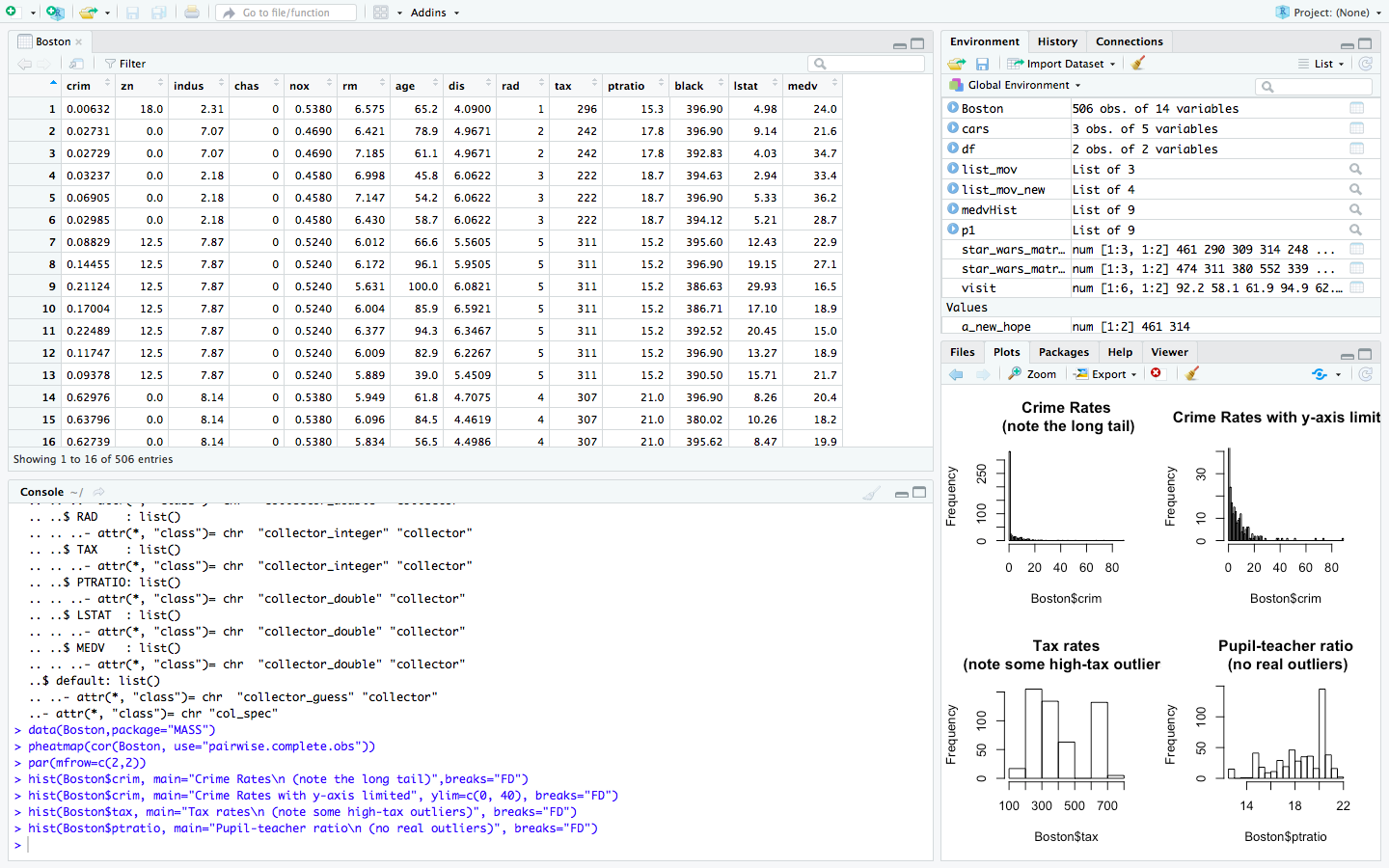
> pheatmap(cor(Boston, use="pairwise.complete.obs"))

****

We, use the heatmap of correlations, calculating correlations using pairwise complete observations (for a given pair of variables, neither has a missing value). It looks like there are a number of variables associated with “crim”: ptratio, rad, tax, lstat, age, indus and nox.

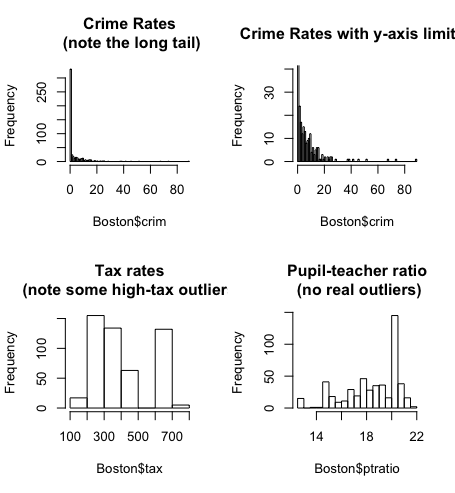
****

**VISUALIZATION -1:**

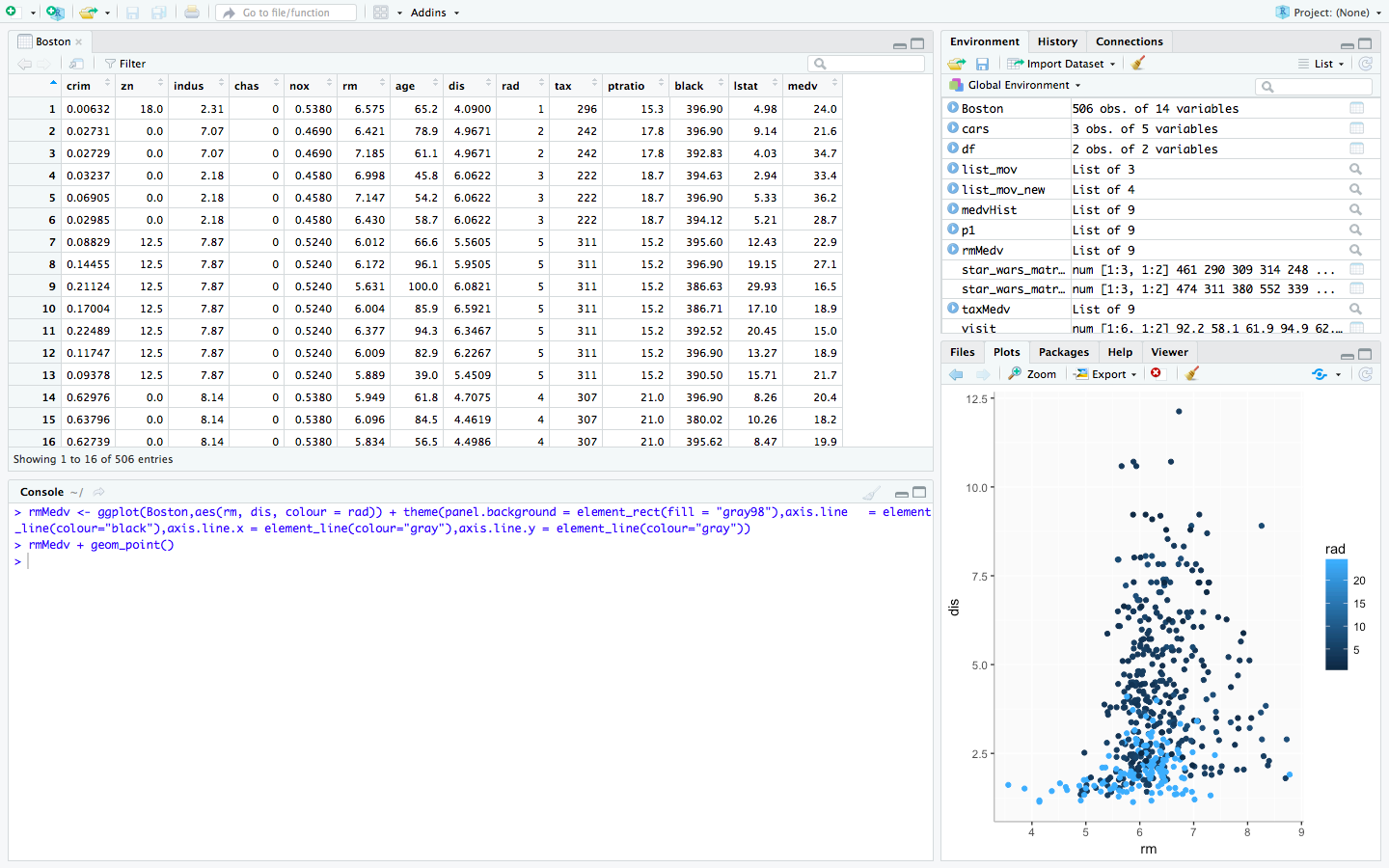
****

In Visualization-1 we find out by analysis that the crime rate in Boston’s suburbs is very high when we put a limit to y-axis. The pupil-teacher ratio is below the average which indicates the lack of good education quality in these suburbs. Some places the pupil-teacher ratio is very less. By increasing this ratio, we can educate the children and it brings down the crime rate.

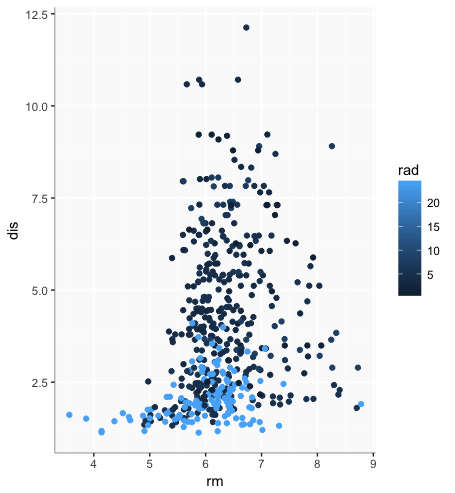
The tax rates of certain places are very high and the highest paying tax outliers are also included the graph.

****

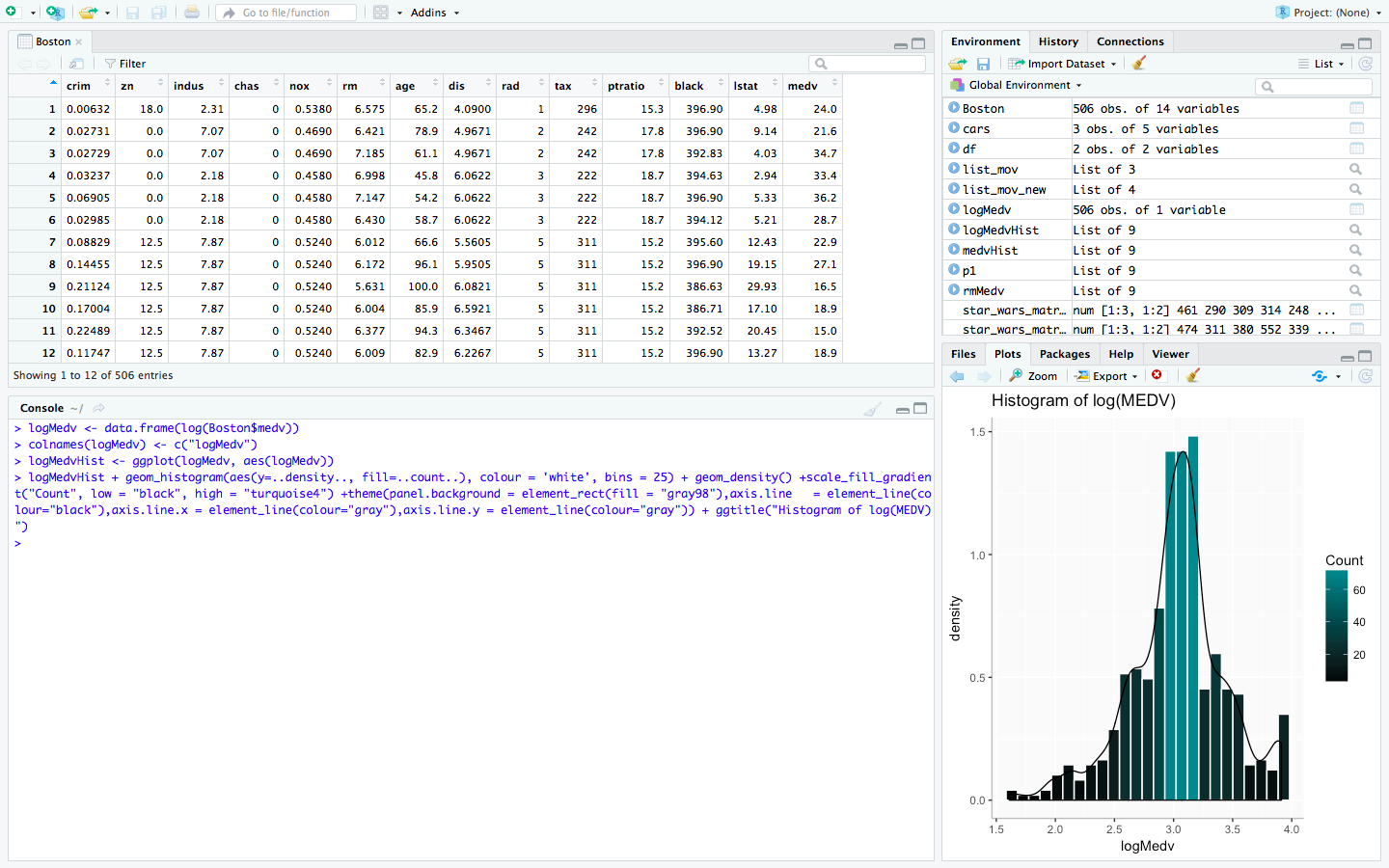
**VISUALIZATION -2:**

****

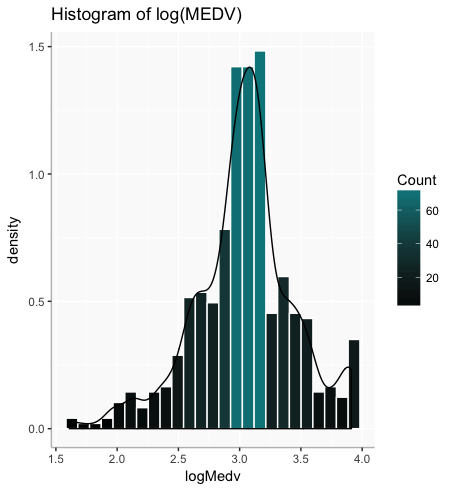
In this analysis we see the that the average number of rooms per dwelling is more near to five Boston employment centers. As the distance increases the number of room per dwelling increases. Within the range of 0-2.5 the room per dwelling is thickly populated in the graph. This shows the emerging business in Boston. The rad shows us the accessibility of the highways from the rooms.

****

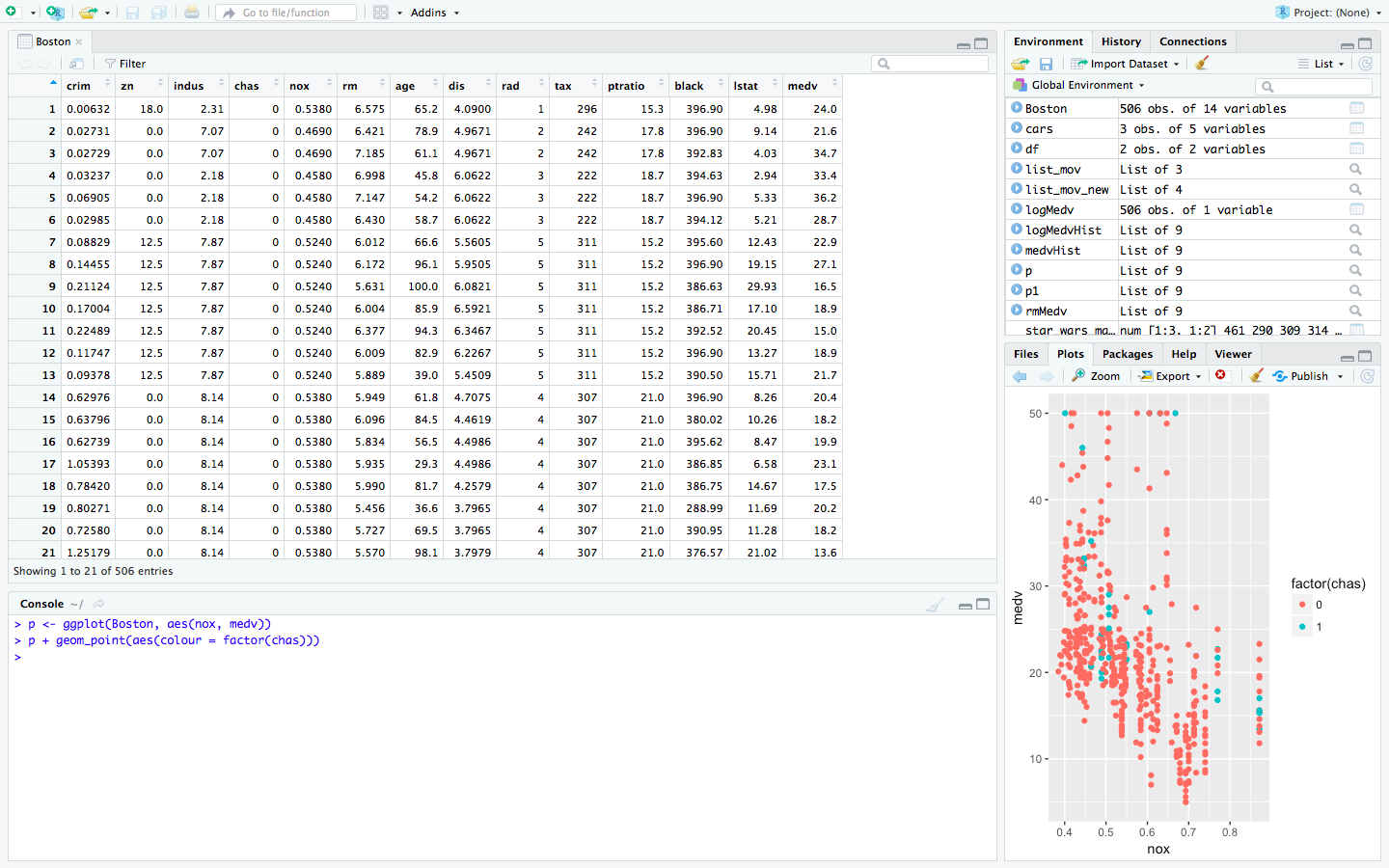
**VISUALIZATION -3:**

****

In this visualization we are emphasize on the distribution of medv in the Boston. We are looking in the log of the median value of owners which has occupied homes at the Boston. This distribution gives us the clarity of number of owner-occupied houses in Boston City in $1000 dollars. This also shows that the count of the log medv increases between 2.7-3.4. It tells us that there are large number of owners who prefer to stay in their homes and these houses are having rent less than $1000.

****

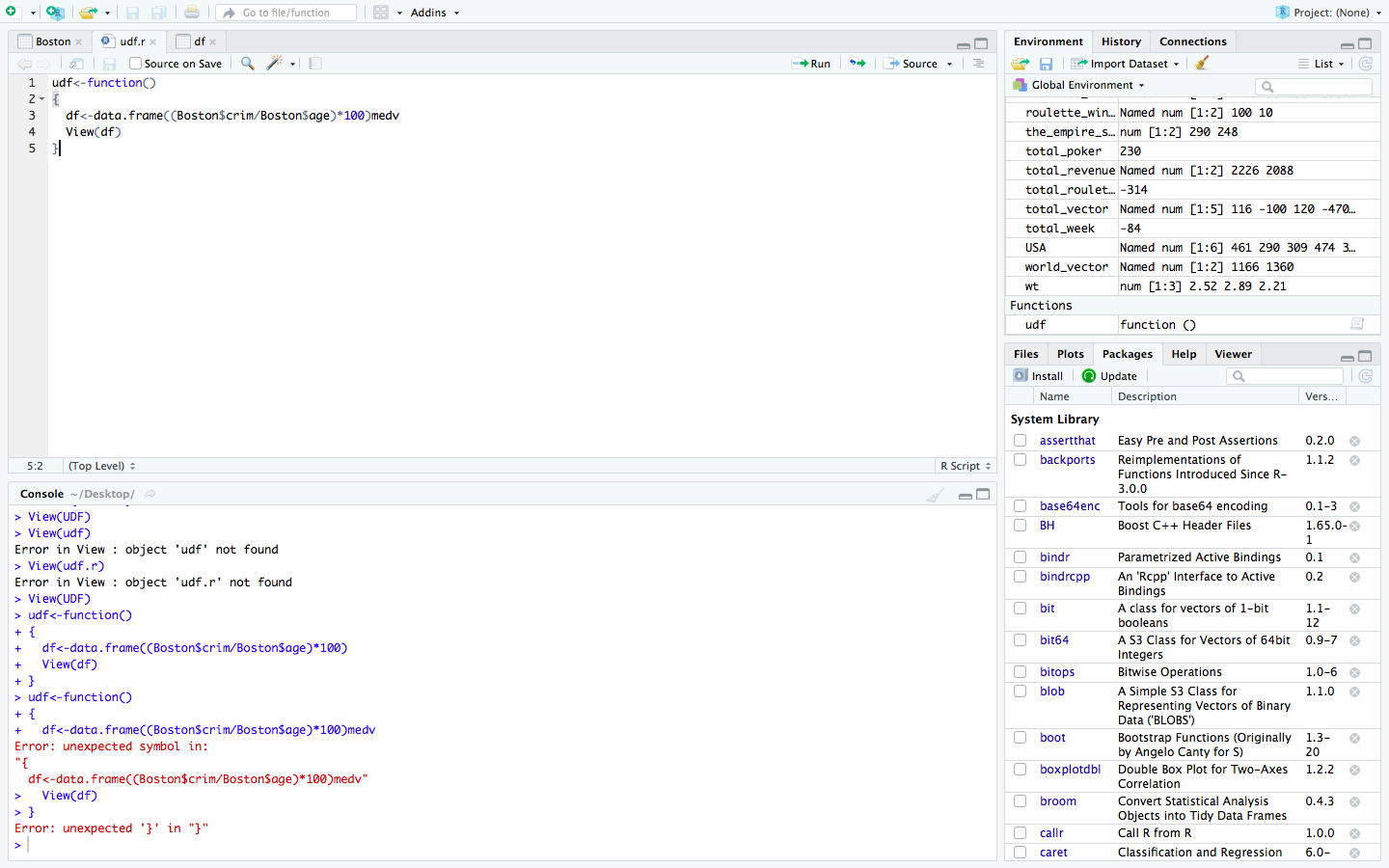
**VISUALIZATION -4:**

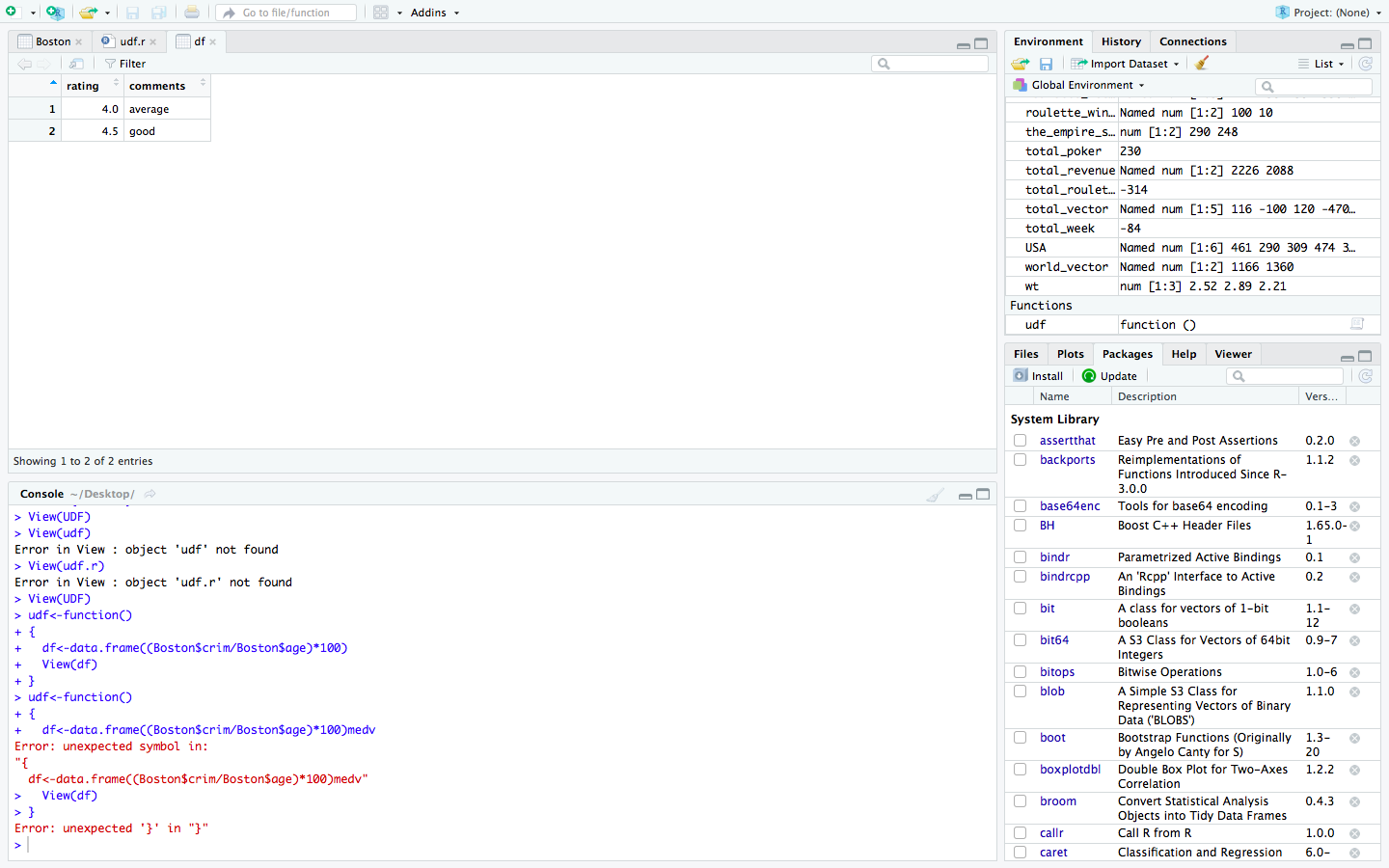
****

In this visualization we see the distribution of nitric oxide concentration in the v/s the median value of owners-occupied homes. In this analysis we know that in most of the places the concentration of nitric oxide is less. The Charles River is one of the main factors. The pink color represents that there is no nitric oxide concentration whereas the blue represents the concentration of nitric oxide in the area. This represents the air quality in Boston. This indirectly affects the health of the owners.

****

**USER DEFINED FUNCTION: Rating and Comments**

****

****

**Codes:**

udf<-function()

{

df<-data.frame((Boston$crim/Boston$age)\*100)medv

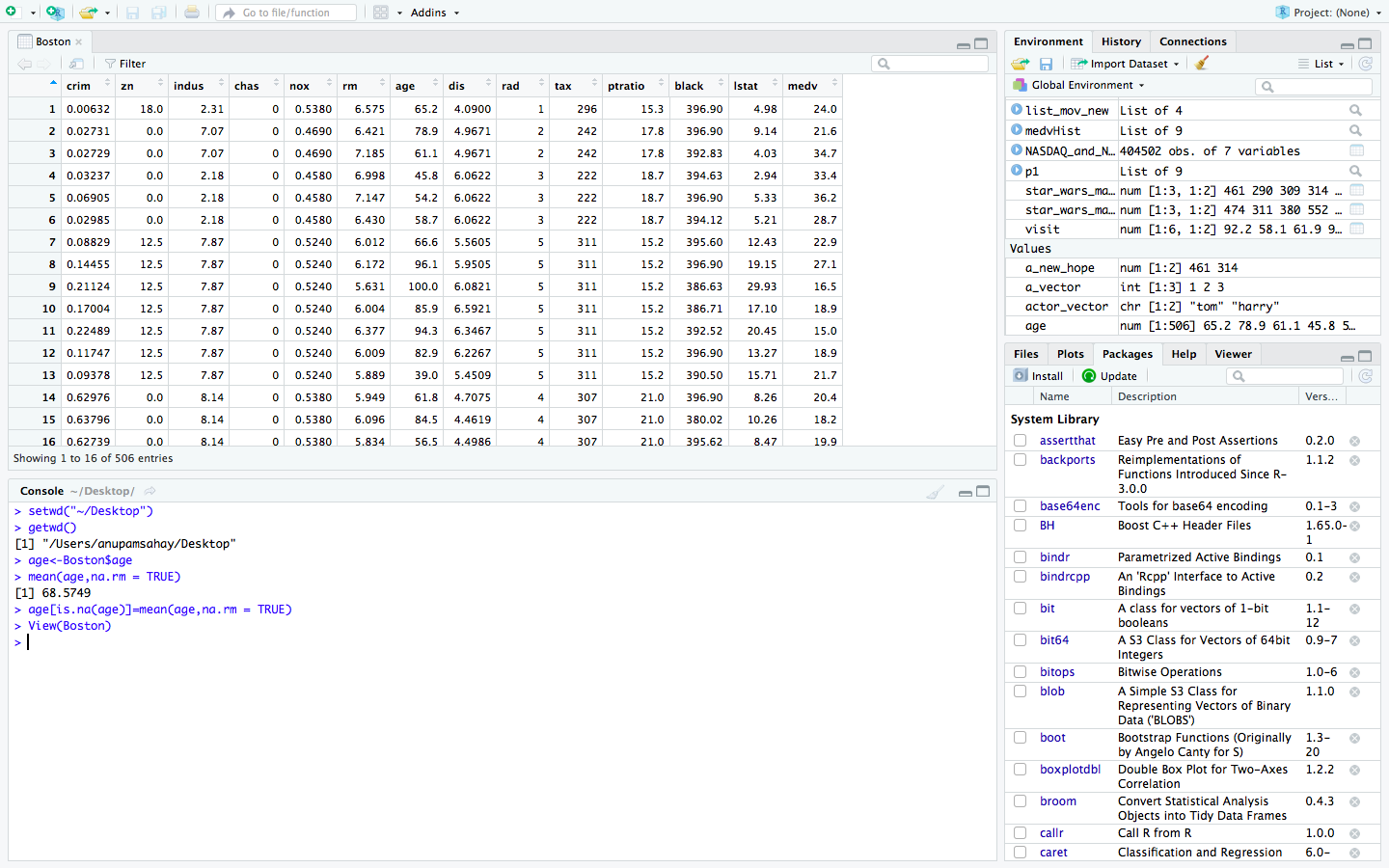
View(df)

}

By this user-defined function we can come to know the rating and the comments of the various places in Boston.

**STATISTICAL FUNCTION:**

1. Calculating Mean for the Age Col:



**Codes:**

> setwd("~/Desktop")

> getwd()

[1] "/Users/anupamsahay/Desktop"

> age<-Boston$age

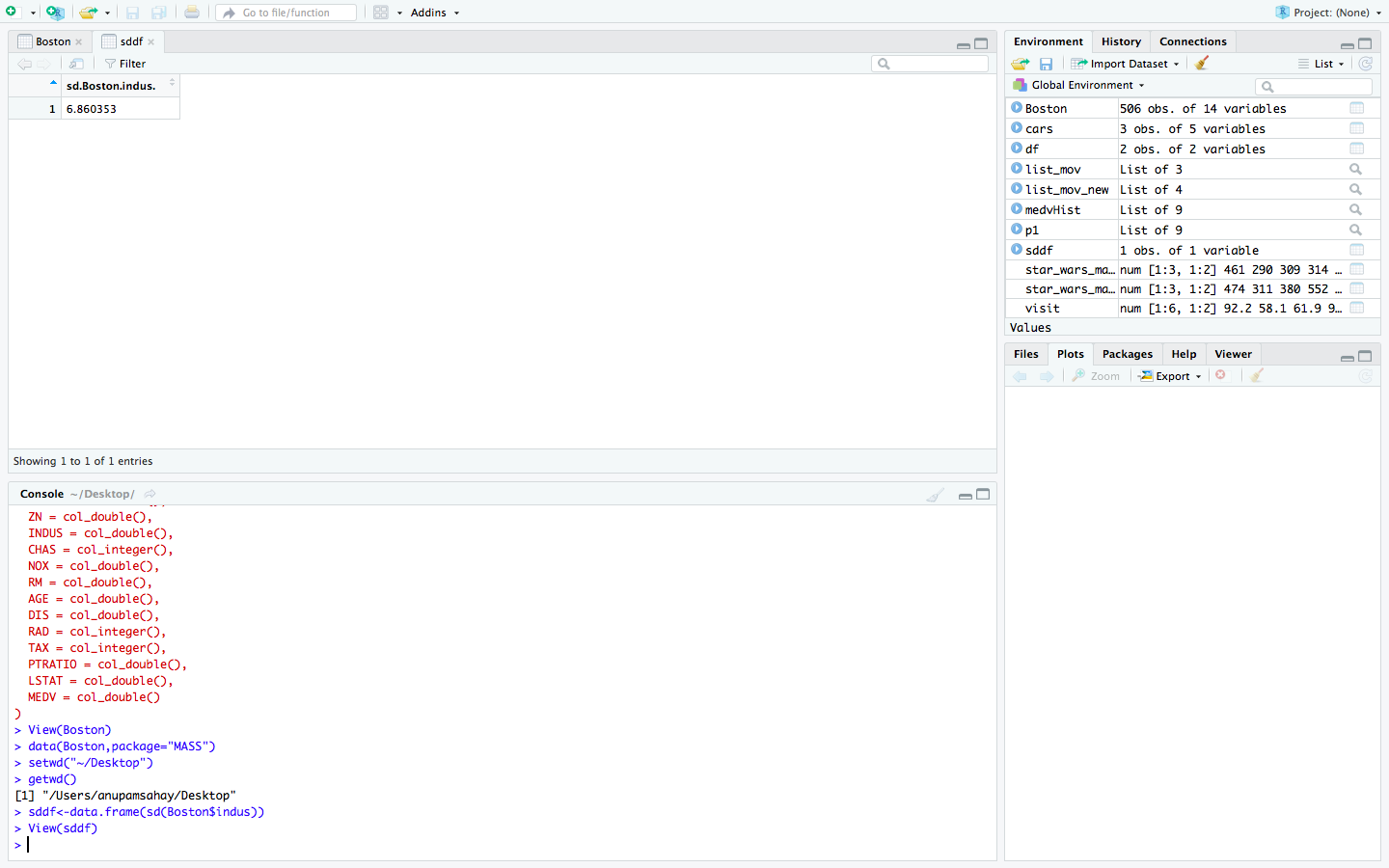
> mean(age,na.rm = TRUE)

[1] 68.5749

> age[is.na(age)]=mean(age,na.rm = TRUE)

> View(Boston)

2.Calculating the standard deviation for Indus col:



**CODES:**

> View(Boston)

> data(Boston,package="MASS")

> setwd("~/Desktop")

> getwd()

[1] "/Users/anupamsahay/Desktop"

> sddf<-data.frame(sd(Boston$indus))

> View(sddf)

**DATA CLEANING - (A) CODES:**

> setwd("~/Desktop/CIS 5270/R Project")

> getwd()

[1] "/Users/anupamsahay/Desktop/CIS 5270/R Project"

> View(Boston)

> keepcol<- c(1,3,5,6,7,8,9,10,11,12,13,14)

> Boston\_new<-Boston[,keepcol]

> View(Boston\_new)

**DATA CLEANING - (B) CODES:**

> View(Boston)

> setwd("~/Desktop")

> getwd()

[1] "/Users/anupamsahay/Desktop"

> Boston1<- na.omit(Boston)

> View(Boston1)

**DATA CLEANING - (C) CODES:**

> View(Boston)

> install.packages("tidyr")

> library('tidyr')

> Boston2<- separate(Boston,DATE,c("MONTH","YEAR","DAY"),sep="/")

> View(Boston2)

**VISUALIZATION -1:**

> par(mfrow=c(2,2))

> hist(Boston$crim, main="Crime Rates\n (note the long tail) ", break="FD")

> hist(Boston$crim, main="Crime Rates with y-axis limited", ylim=c(0, 40), breaks="FD")

> hist(Boston$tax, main="Tax rates\n (note some high-tax outliers)", breaks="FD")

> hist(Boston$ptratio, main="Pupil-teacher ratio\n (no real outliers)", breaks="FD")

**VISUALIZATION -2:**

**>** rmMedv <- ggplot(Boston,aes(rm, dis, colour = rad)) + theme(panel.background = element\_rect(fill = "gray98"),axis.line = element\_line(colour="black"),axis.line.x = element\_line(colour="gray"),axis.line.y = element\_line(colour="gray"))

> rmMedv + geom\_point()

**VISUALIZATION -3:**

>logMedv <- data.frame(log(Boston$medv))

> colnames(logMedv) <- c("logMedv")

> logMedvHist <- ggplot(logMedv, aes(logMedv))

> logMedvHist + geom\_histogram(aes(y=..density.., fill=..count..), colour = 'white', bins = 25) + geom\_density() +scale\_fill\_gradient("Count", low = "black", high = "turquoise4") +theme(panel.background = element\_rect(fill = "gray98"),axis.line = element\_line(colour="black"),axis.line.x = element\_line(colour="gray"),axis.line.y = element\_line(colour="gray")) + ggtitle("Histogram of log(MEDV)")

**VISUALIZATION -4:**

>p<- ggplot(Boston, aes(nox, medv))

>p + geom\_point(aes(colour = factor(chas)))