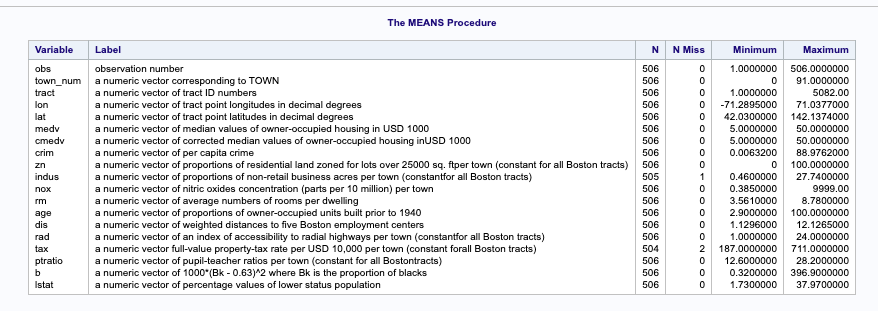
STAT 440 Midterm Boston Housing Data

Wanyu Bi

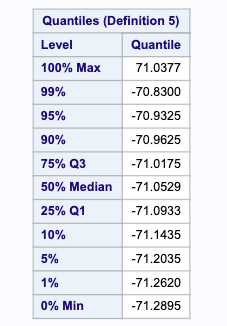
The original data file uses the modified list input style. There are 21 variables. They are: obs town town\_num tract lon lat medv cmedv crim zn indus chas nox rm age dis rad tax ptratio b lstat, in which chas and town are chars, the rest of the variables are numeric.

The objective of this project is to use format and label to make the data more readable, find the potential problem existed in the data like missing value and wrong value.

I use **proc mean** find the missing values in **tax** and **indus**. Then I need to fill in those missing values. First, I use where statement to find the **town\_num** of these missing values. Then I print out their **tax** and **indus** value. I approximate the missing values with the existing the tax and indus value in the same town\_num.

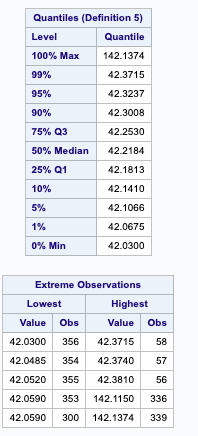


I use **univariate** to find the lowest 5 extreme observations and the highest 5 extreme observations. There are weird values in Lon and Lat and nox.



Nox (next page left)

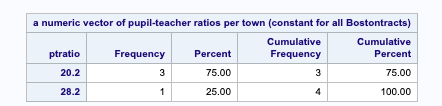
Lat(next page right)



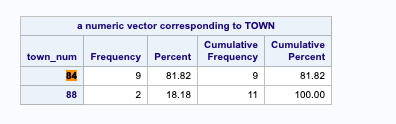
Then I all check the value of Lon and Lat and nox, they are approximately in a range or the same within the same town, then I changed the value within reasonable range.

I printed the **frequency** of the variables, and pay attention to the entry where the frequency is 1, find that there are wrong spelling in the Town name.

Given the description in the midterm, within a town, zn, indus, tax, rad, ptratio, town\_num are the same. So I run the frequency table by town to check whether there are two values of those variables within a town. Under the Boston Hyde Park, there are two kinds of values in the ptratio. So I then update the value 28.2 with frequency 1 to the other value 28.2 with frequency 3.



Under the town name **Boston Dorchester,** there are two town\_num



I updated all the town number of town **Boston Dorchester** to 88.

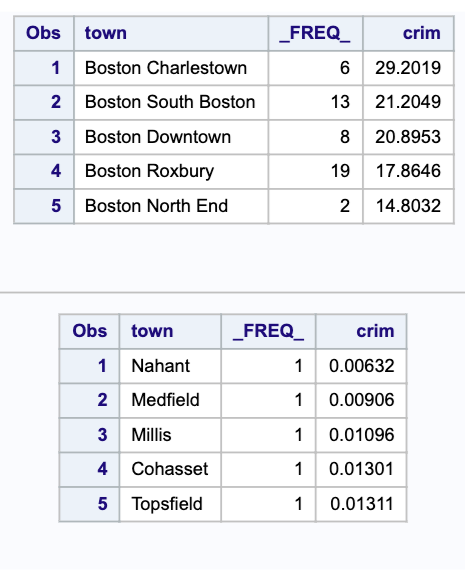
**Exercises Section:**

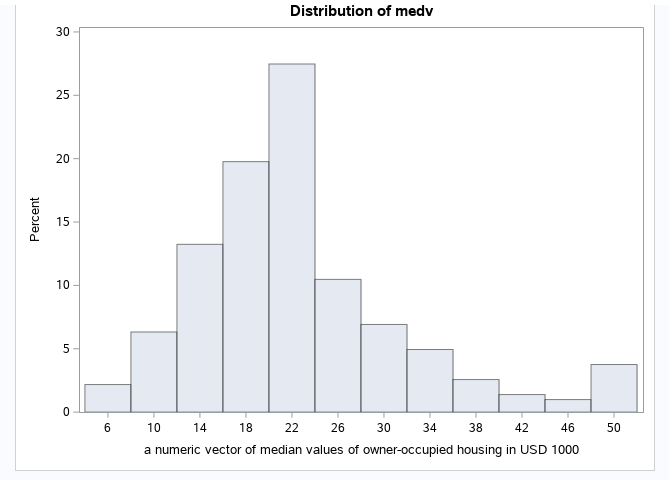
**The two represented by the most tracts is Cambridge**

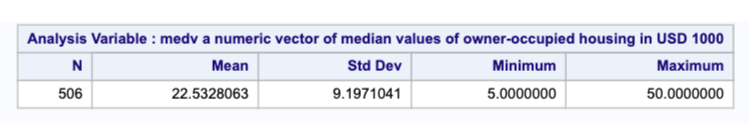
The number of tracts is 30.

The following table shows that Boston Charlestown, Boston South Boston, Boston Downtown, Boston Roxbury, and Boston North End have the top 5 per capita crime rates.

Nahant, Medfield, Millis, Cohasset and Topsfield have have the lowest per capita crime rates.







The mean of MEDV is 22.53, the standard deviation is 9.20.

Town number 20-24 has the highest proportion of the distribution of MEDV.