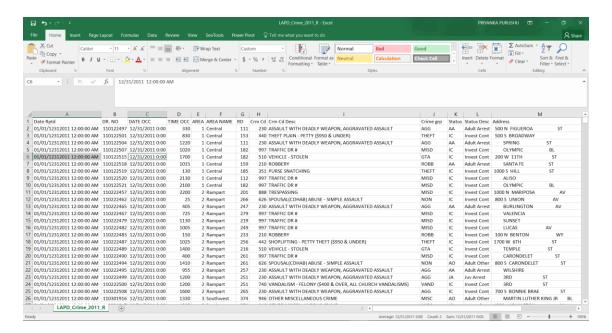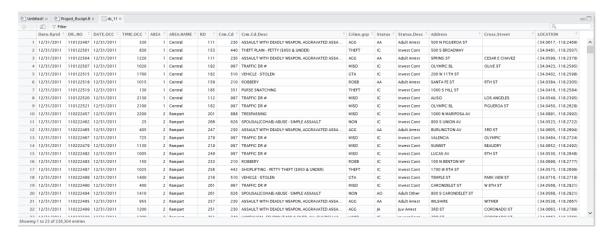# DATA CLEANING

The dataset required very little cleaning to be done. The corrections made on the existing dataset are as follows:

1. Cleaning LA Crime 2011 dataset – Date field cleaning (Removing the time data and the invalid data "01/01/" the date.rptd column & Date.occ column)
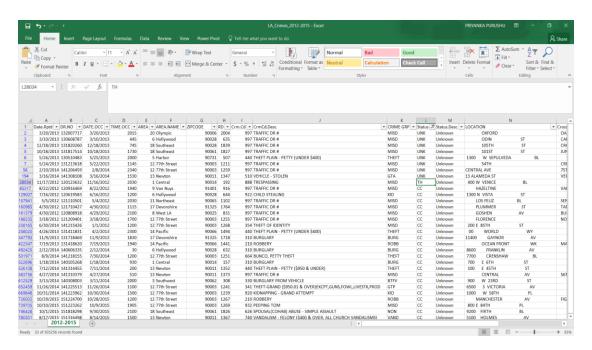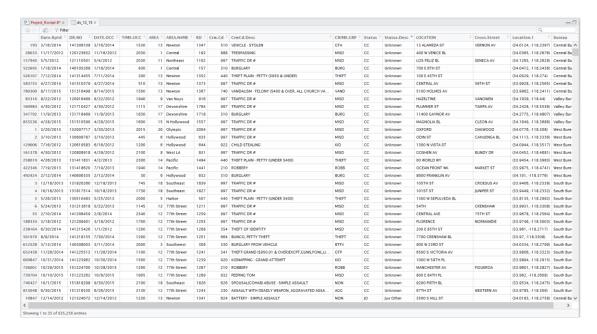
Actual Dataset:

After cleaning in R :

2. Mismatch in the status value.

- UNK is the description for status type CC. TH is unknown. Hence status UNK and TH has been replaced with CC because of the same description with respect to LA crime dataset 2012-2015
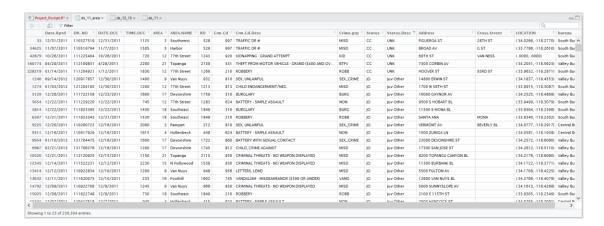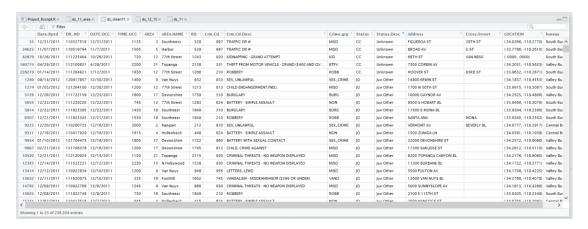
Actual Dataset:

After cleaning in R:

- The CC status description is UNK in the crime dataset 2011 – This has been changed

  to Unknown to make sure there is no mismatch of value while combining dataset
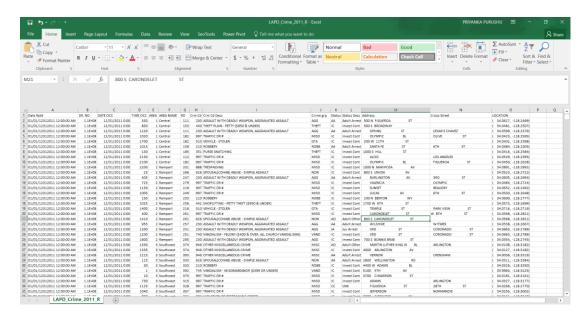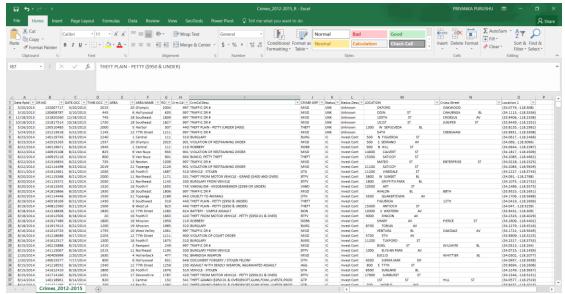
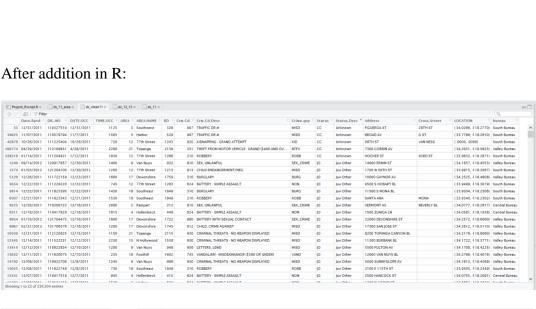Actual Dataset:

After cleaning in R:

3. The area stations have been grouped into 4 bureaus. This is done to perform analysis based on the bureau.
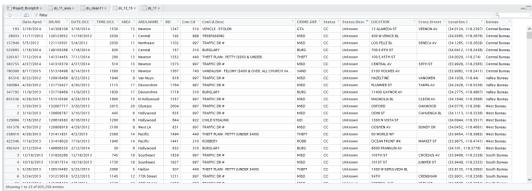
Actual Dataset:
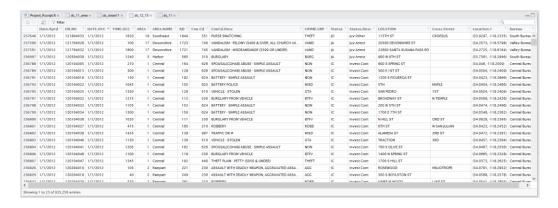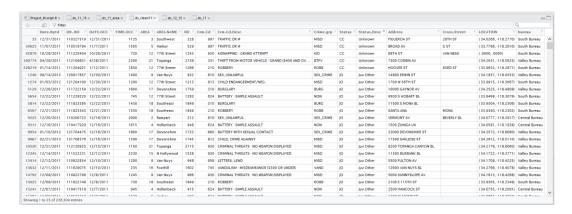
After addition in R:

4. Column Name change & Dataset Join:

In order to combine both the dataset the column names have to be same. So column name

of dataset 2012-2015 was changed with respect to 2011 dataset column names.
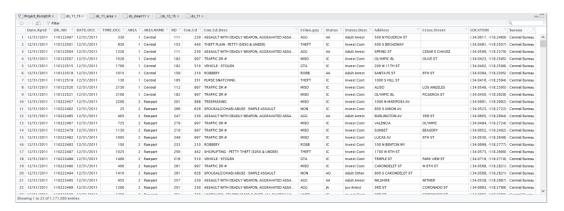
Actual Column Names: 2012-2015 dataset

2011 Dataset:

After changing in R: Common column name and combined dataset

5. Addition of a new columns named Crime group and zip code in Excel. This was done to

group the various crime codes which would help in portraying a better data visualization.