# Description of Data:

The data chosen for this exercise is Houston crime data, which I have used earlier for Project-2. The reason I chose this data set is because there were a few things I wanted to explore in Project-2, which I wasn't able to, mainly due to the way the data was presented and not so adequate sql querying skills.

The original data includes the following columns:

1. Date (M/DD/YY)
2. Hour
3. Offense_Type
4. Beat
5. Premise
6. BlockRange
7. StreetName
8. Type
9. Suffix
10. Number_Offenses

# Transformation:

- Split Date column –
  - The original file had date in the format (M/DD/YY), which made it very difficult to sort out which months contributed to most crimes or the type of crime distribution.
  - Splitting the date into Month | Date | Year, now there are 3 separate columns. Once loaded into the database we can easily do a query to sour out which month had the most crime, etc.

- Rename new Date columns –
  - This is to rename the newly split date columns – Month, Date, Year.

- Replace Offense_Type data –
  - The original Offense_Type column included many two word entries, such as Auto-Theft or Aggravated-Assault. Now, in the project I did a query to sort and count the theft crimes but because Auto-Theft was entered in a different format, I wasn't able to include them.
  - This allowed me to replace all two worded Offense_Type entries into one word (eg: Auto-Theft = Theft and Aggravated-Assault = Assault)
  - Now the same query can include all types of theft crimes.

- Boolean filter on Theft –
  - This is Boolean filter (if Offense_Type = Theft, then 1 else 0). The purpose for this is to efficiently sort theft crimes, upon which we can do further operations on the results.

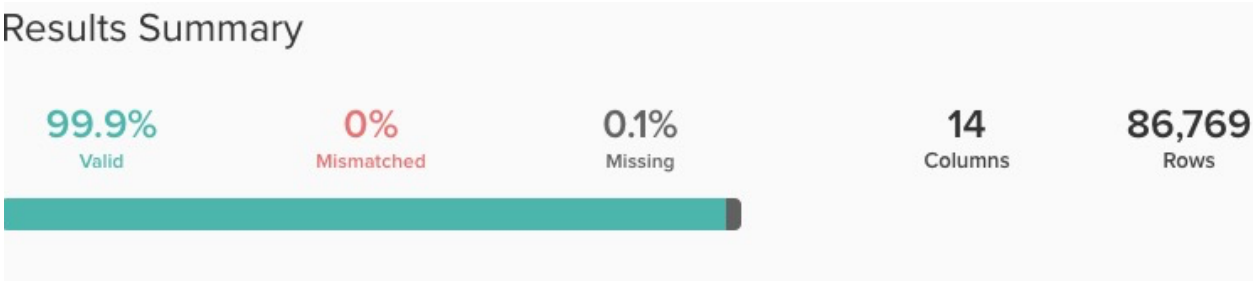
- Rename boolean filter –
  - Renamed the newly created Boolean filter column to "Theft"

- Boolean filter on specific hour –
  - We can even create a filter on a specific hour, say 8:00 P.M. and have it available for efficient sorting. What I really wanted to do was create a filter for Day hours and Night hours. This would allow me to explore the data in terms of day and night. But I wasn't able to figure out how to add multiple hours in the same filter.

## Trifecta Script:

```
splitrows col: column1 on: '\r\n' quote: '\"'
split col: column1 on: ',' limit: 9 quote: '\"'
header
split col: Date on: '\/' limit: 2
rename col: Date1 to: 'Month'
rename col: Date2 to: 'Date'
rename col: Date3 to: 'Year'
replace col: Offense_Type with: '' on: `{start}{alpha}+ `
countpattern col: Offense_Type on: `{start}{alpha}{5}{end}`
countpattern col: Hour on: `20`
rename col: countpattern_Offense_Type to: 'Theft'
```

## Evaluation of Results:




I was very surprised to see that almost all the date was updated accurately. But then again the data was clean to begin with and the operations would have followed through. If, however, the data had discrepancies (eg: a string in the date column or an wrong entry in the Offense_Type column), then we would find misclassifications.

I evaluated the newly created csv file as well and it has been sorted as scripted. I am very impressed with this tool. I don't know why we didn't use this before. This tool can be very useful for our exploration projects.