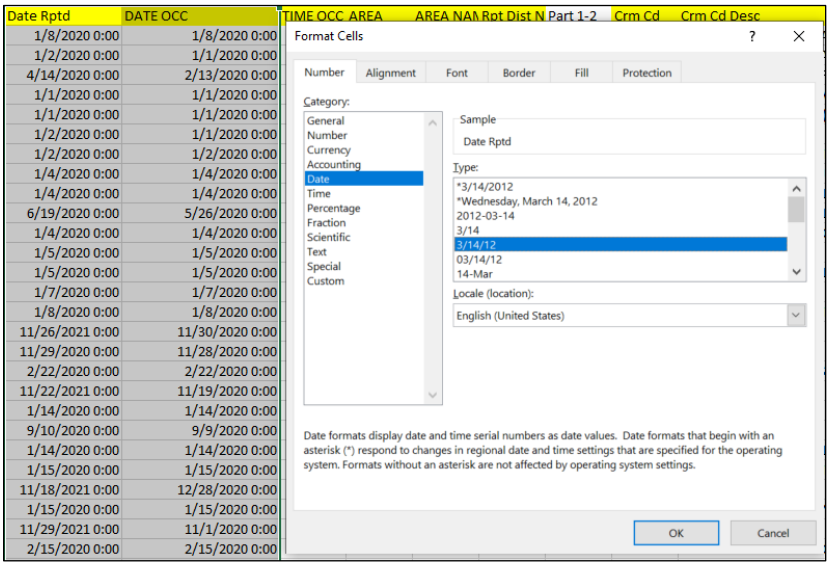
Data Cleaning

The data which we get from the official sites is usually raw data. This data needs to be cleaned to make it easier to analyze data. We need to remove duplicates, combine columns, change the incorrect date format etc. Various data cleaning methods can be used for Data cleaning, these may vary from Dataset to the dataset. Listed below are some of the Data Clearing methods which have been used to clean the LA Crime Dataset. The Data has been cleaned using Microsoft Excel.

1. Embedded values in the field - Date Format:

For the columns “Date Rptd” & “Date OCC”, there are two different values written – date & “0:00”. We have removed the embedded time “0:00” portion so that the entire field has only single type of values. By doing this, we can simplify the data and make it easier for analysis using the date format. Furthermore, leaving the time portion in the date format can cause issues when sorting or grouping data by date.

Pre-Cleaning screenshot:



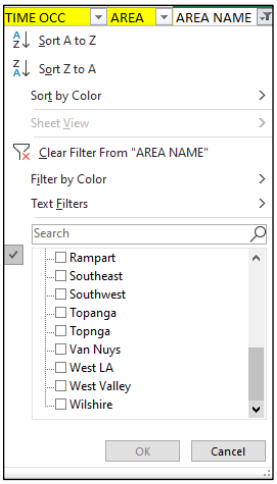
Post-Cleaning Screenshot:

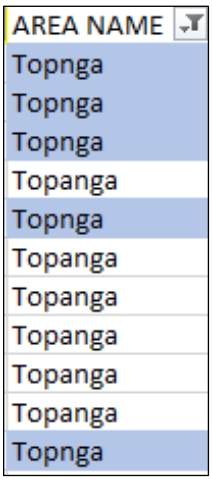
Date Rptd	DATE OCC
1/8/20	1/8/20
1/2/20	1/1/20
4/14/20	2/13/20
1/1/20	1/1/20
1/1/20	1/1/20
1/2/20	1/1/20
1/2/20	1/2/20
1/4/20	1/4/20
1/4/20	1/4/20
6/19/20	5/26/20
1/4/20	1/4/20
1/5/20	1/5/20
1/5/20	1/5/20
1/7/20	1/7/20

2. Misspellings:

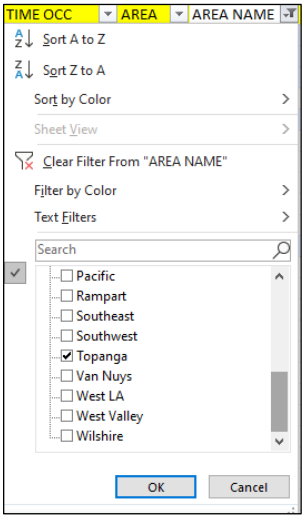
It is very important to fix the Misspellings in a data set, as this may cause ambiguity. As shown in the screenshot below, the column "Area Name" has a value – "**Topnga"** is spelled Incorrectly. Hence, we replaced the incorrect value "**Topnga"** with the Correct Value **"Topanga",** for the whole Column- 'Area Name'. After correcting the Misspelled value, we can see only one filter with the 'Area Name'- **Topanga** in the post-cleaning screenshot.

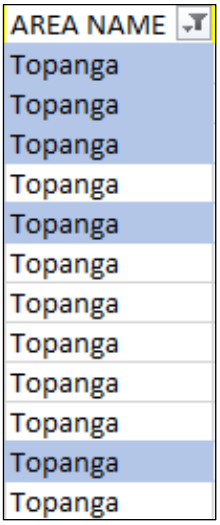
Pre-Cleaning Screenshots:





Post-Cleaning Screenshots:





3. Duplicated Records:

Duplicate records cause confusion while analyzing the data as well as visualizing it leading to unnecessary wastage of memory. Therefore, removing duplicate records is essential before conducting any data analysis to ensure accurate and reliable results.

Here, we have 2 Fields “Crm Cd” & “Crm Cd1”, both contain the same data. Hence, we have removed the column “Crm Cd1” and kept the other column to avoid the duplicate records.

Pre-cleaning Screenshot:

Crm Cd	Crm Cd 1	Crm Cd 2
624	624	
624	624	
845	845	
745	745	998
740	740	
121	121	998
442	442	998
946	946	998
341	341	998
341	341	
330	330	
930	930	
341	341	
648	648	998
442	442	
626	626	
626	626	
440	440	624

Post-cleaning Screenshot:

Crm Cd 1	Crm Cd 2
624	
624	
845	
745	998
740	
121	998
442	998
946	998
341	998
341	
330	
930	
341	
648	998
442	
626	
626	
440	624

4. Combining Columns:

Combining columns helps to standardize the data with similar data types or formats and make it more meaningful.

Here, we have two fields “LOCATION” & “Cross Street”, which can be merged and give us the precise location. So have merged these two columns by using the CONCAT formulas in excel. We have also used “TRIM” formulas to remove unwanted blank spaces from the “LOCATION” column.

Pre-cleaning Screenshots:

LOCATION	Cross Street
1100 W 39TH	PL
700 S HILL	ST
200 E 6TH	ST
5400 CORTEEN	PL
14400 TITUS	ST
700 S BROADWAY	
700 S FIGUEROA	ST
200 E 6TH	ST
700 BERNARD	ST
11900 BALBOA	BL
15TH	OLIVE
800 N ALAMEDA	ST
800 S OLIVE	ST
700 W 7TH	ST
100 S LOS ANGELES	ST
14200 BERG	ST

LOCATION	Cross Street	=U1&" "&V1
1100 W 39TH	PL	
700 S HILL	ST	
200 E 6TH	ST	
5400 CORTEEN	PL	
14400 TITUS	ST	
700 S BROADWAY		
700 S FIGUEROA	ST	
200 E 6TH	ST	
700 BERNARD	ST	
11900 BALBOA	BL	
15TH	OLIVE	
800 N ALAMEDA	ST	
800 S OLIVE	ST	
700 W 7TH	ST	
100 S LOS ANGELES	ST	
14200 BERG	ST	
3200 W AVENUE 32		
PACIFIC COAST	VERMONT	

LOCATION	Cross Street	LOCATION	Cross Street
1100 W 39TH	PL	1100 W 39TH	PL
700 S HILL	ST	700 S HILL	ST
200 E 6TH	ST	200 E 6TH	ST
5400 CORTEEN	PL	5400 CORTEEN	PL
14400 TITUS	ST	14400 TITUS	ST
700 S BROADWAY		700 S BROADWAY	
700 S FIGUEROA	ST	700 S FIGUEROA	ST
200 E 6TH	ST	200 E 6TH	ST
700 BERNARD	ST	700 BERNARD	ST
11900 BALBOA	BL	11900 BALBOA	BL
15TH	OLIVE	15TH OLIVE	
800 N ALAMEDA	ST	800 N ALAMEDA	ST
800 S OLIVE	ST	800 S OLIVE	ST
700 W 7TH	ST	700 W 7TH	ST
100 S LOS ANGELES	ST	100 S LOS ANGELES	ST
14200 BERG	ST	14200 BERG	ST
3200 W AVENUE 32		3200 W AVENUE 32	
PACIFIC COAST	VERMONT	PACIFIC COAST VERMONT	
14700 FRIAR	ST	14700 FRIAR	ST
7TH	HILL	7TH HILL	
13600 LEADWELL	ST	13600 LEADWELL	ST

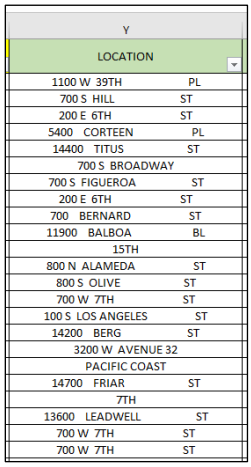
Post-Cleaning Screenshot:

Location	
1100 W 39TH	PL
700 S HILL	ST
200 E 6TH	ST
5400 CORTEEN	PL
14400 TITUS	ST
700 S BROADWAY	
700 S FIGUEROA	ST
200 E 6TH	ST
700 BERNARD	ST
11900 BALBOA	BL
15TH OLIVE	
800 N ALAMEDA	ST
800 S OLIVE	ST
700 W 7TH	ST
100 S LOS ANGELES	ST
14200 BERG	ST
3200 W AVENUE 32	
PACIFIC COAST VERMONT	
14700 FRIAR	ST
7TH HILL	
13600 LEADWELL	ST

5. Trimming extra “blanks” from the string:

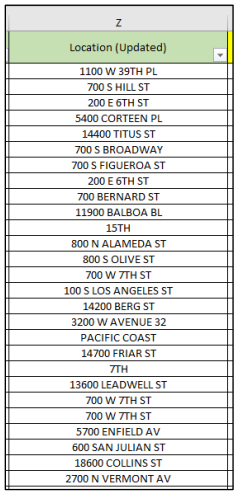
Trimming extra blanks from a string is important because it can help to improve the readability and usability of the data. Extra blanks can be added accidentally during data entry, data manipulation or file formatting, and can make it difficult to analyze and process the data accurately. Trimming extra blanks from strings can help to ensure that the data is accurate, consistent, and easily processed.

Pre-cleaning Screenshot:



Y
LOCATION
1100 W 39TH PL
700 S HILL ST
200 E 6TH ST
5400 CORTEEN PL
14400 TITUS ST
700 S BROADWAY
700 S FIGUEROA ST
200 E 6TH ST
700 BERNARD ST
11900 BALBOA BL
15TH
800 N ALAMEDA ST
800 S OLIVE ST
700 W 7TH ST
100 S LOS ANGELES ST
14200 BERG ST
3200 W AVENUE 32
PACIFIC COAST
14700 FRIAR ST
7TH
13600 LEADWELL ST
700 W 7TH ST
700 W 7TH ST

Post-cleaning Screenshot:



Z
Location (Updated)
1100 W 39TH PL
700 S HILL ST
200 E 6TH ST
5400 CORTEEN PL
14400 TITUS ST
700 S BROADWAY
700 S FIGUEROA ST
200 E 6TH ST
700 BERNARD ST
11900 BALBOA BL
15TH
800 N ALAMEDA ST
800 S OLIVE ST
700 W 7TH ST
100 S LOS ANGELES ST
14200 BERG ST
3200 W AVENUE 32
PACIFIC COAST
14700 FRIAR ST
7TH
13600 LEADWELL ST
700 W 7TH ST
700 W 7TH ST
5700 ENFIELD AV
600 SAN JULIAN ST
18600 COLLINS ST
2700 N VERMONT AV

6. Standardizing the Time Format:

Time data becomes difficult to analyses if it is not in standard format. Time-related data can be complex and small errors or inconsistencies will affect the accuracy in analysis. Cleaning the data ensures it is in a more reliable and usable format. We have used Formulas to convert the numeric format to Time format.

We have a column “TIME OCC” which contains the value in military 24 Hr. time format. It is OK for the time like 22:30 (having all 4 digits). However, with values like “330” (03:30) and “30” (12:30), it is difficult to perform visualization. Hence, we have changed the format of the values to the 24 hrs. format.

Pre-Cleaning Screenshot:

TIME OCC
2230
330
1200
1730
415
30
1315
40
200
1925
2200
955

Post-Cleaning Screenshots:

We have used two Formulas, first (Time (left (c2, LEN(c2)-2), RIGHT(c2,2),0)) to convert the 4- or 3-digit integers to Time format. Second (A7/1440) to convert the 1 or 2-digit integers into time format.

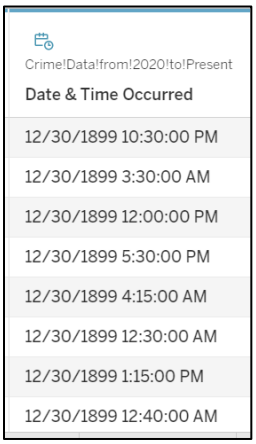
Old Time	Time Occurred
2230	10:30 PM
330	3:30 AM
1200	12:00 PM
1730	5:30 PM
415	4:15 AM
30	#VALUE!
1315	1:15 PM

30	0.020833333
40	0.027777778
1	0.000694444
55	0.038194444
30	0.020833333
40	0.027777778
15	0.010416667
1	0.000694444
1	0.000694444

30	12:30 AM
40	12:40 AM
1	12:01 AM
55	12:55 AM
30	12:30 AM
40	12:40 AM
15	12:15 AM
1	12:01 AM

1638	16:38
1805	18:05
730	7:30
2018	20:18
1900	19:00
1200	12:00
1330	13:30
1735	17:35
1730	17:30
1445	14:45
1	0:01

For visualization in the Tableau, these values are taken in “datetime” datatype as below:



To extract the default date “12/30/1899”, we have added some calculated fields in Tableau to split “date” & “time” and we have used “time” part in our visualization.

7. Rephrasing the Column Names:

The given Column Names in the dataset were more meaningful & understandable for LAPD employees. However, for us, the names were unclear & ambiguous, it was difficult to determine - what data represents. Hence, we have renamed the column names for better interpretation of data, we have made them more descriptive and informative, which can help to clarify the meaning and purpose of the data.

For example, we have renamed the existing columns:

“Crm Cd” to “Primary Crime Code”, “Status” to “Crime Status”, “Vict Age” to “Victim Age”, “LAT” to “Latitude”, “LON” to “Longitude”

Pre-Cleaning Screenshot:

H	O	R	AB	AC
Crm Cd	Status	Vict Age	LAT	LON
624	AO	36	34.0141	-118.2978
624	IC	25	34.0459	-118.2545
845	AA	0	34.0448	-118.2474

Post-Cleaning Screenshots:

I	P	R	AC	AD
Primary Crime Code	Crime Status Code	Victim Age	Latitude	Longitude
624	AO	36	34.0141	-118.2978
624	IC	25	34.0459	-118.2545
845	AA	0	34.0448	-118.2474