LOS ANGELES CRIME DATACLEANING PROJECT

In this project, I embarked on the thorough task of cleaning Los Angeles Crime Data which contained reports of crimes in the city of Los Angeles, CA within the years 2010 to 2023.

This effort aimed to enhance the accuracy and reliability of the dataset. Through rigorous data cleaning techniques, I meticulously sifted through the information, addressing inconsistencies, errors and missing values. The goal was to ensure that the data was primed for analysis and decision making.

For this project, I decided to use Microsoft excel. This decision was backed by the fact that the dataset had about 50,000 rows which excel can perfectly handle and the versatility, ease of use of Microsoft Excel.

For this project, I relied on more than just my data cleaning skills but as well as critical thinking, problem solving abilities and working in a team of other analysts to uncover insights/ideas on areas that were not clear.

Join me as I delve into the intricacies of this project and share insights gained from this transformative data cleaning journey.

ABOUT THE DATA

The Dataset comprises reports of crime committed in Los Angeles between 2010 and 2023. The data was found on the "data.gov" website. The Data had a lot of integrity issues which would have made it difficult to extract meaningful insights from.

The dataset contains 28 columns and 50,939 rows of input. Below are some of the glaring issues I observed before cleaning.

- Wrong Format/Data Types
- Unclear spellings
- Missing Values
- Missing Entries
- Improper spacings
- Inconsistent inputs
- Irrelevant inputs
- Incorrect inputs

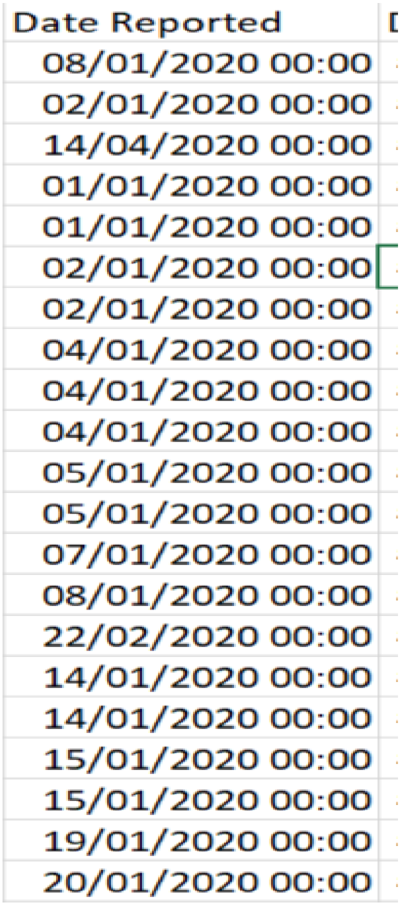
DATA CLEANING PROCESS

I made a new sheet to clean the data, so we could see how it looked before and after. Then, I carefully checked each column for problems like incorrect dates, missing information, or errors in the names of different inputs like crimes or locations.

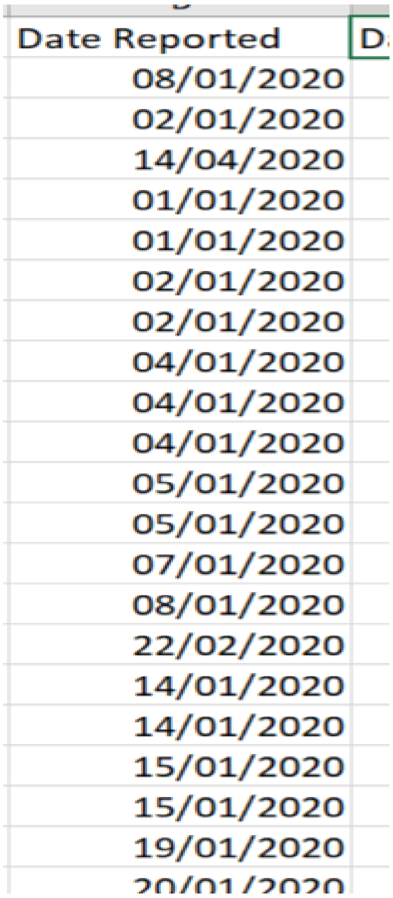
Once I found these issues, I fixed the errors to make them clear. Next, I went through the data in detail to clean it up, and I'll explain the process below based on the columns that had a challenge or more.

Date Reported

A look at this column showed that all inputs contained date with time format 0:00. This is not an ideal format for Date, so I used the "Format cells" function to set it to show the date to display the inputs in this format DD/MM/YYYY only.



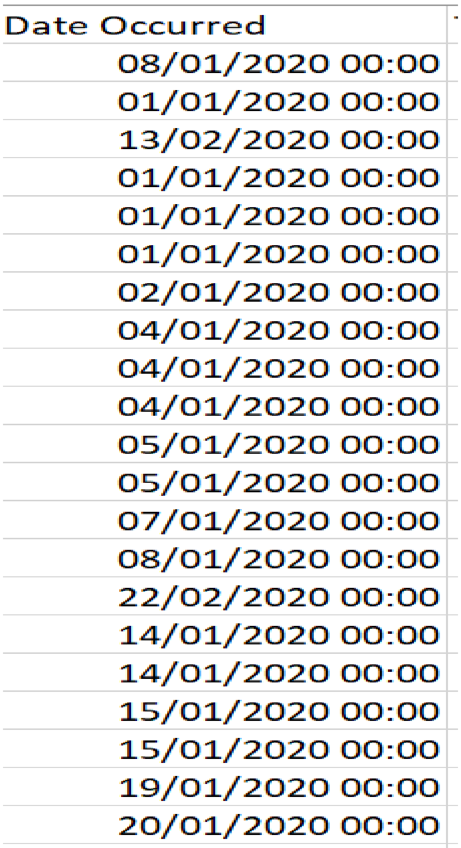
Date Reported
08/01/2020 00:00
02/01/2020 00:00
14/04/2020 00:00
01/01/2020 00:00
01/01/2020 00:00
02/01/2020 00:00
02/01/2020 00:00
04/01/2020 00:00
04/01/2020 00:00
04/01/2020 00:00
05/01/2020 00:00
05/01/2020 00:00
07/01/2020 00:00
08/01/2020 00:00
22/02/2020 00:00
14/01/2020 00:00
14/01/2020 00:00
15/01/2020 00:00
15/01/2020 00:00
19/01/2020 00:00
20/01/2020 00:00



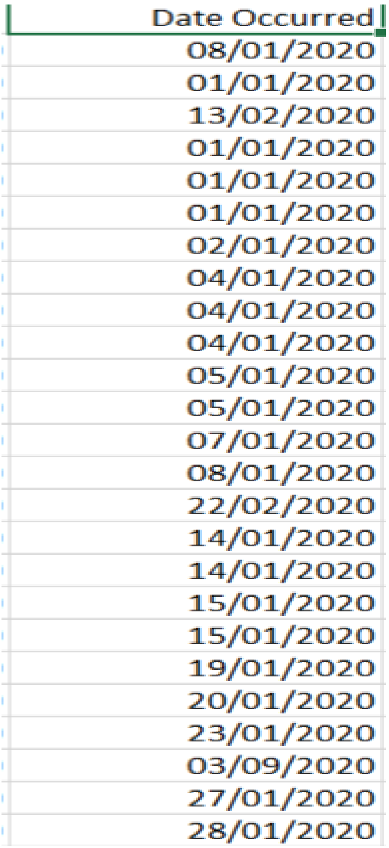
Date Reported
08/01/2020
02/01/2020
14/04/2020
01/01/2020
01/01/2020
02/01/2020
02/01/2020
04/01/2020
04/01/2020
04/01/2020
05/01/2020
05/01/2020
07/01/2020
08/01/2020
22/02/2020
14/01/2020
14/01/2020
15/01/2020
15/01/2020
19/01/2020
20/01/2020

Date Occurred

I used the same method as before and followed the same steps. I formatted the column to display the date in the format “DD/MM/YY” only.



Date Occurred
08/01/2020 00:00
01/01/2020 00:00
13/02/2020 00:00
01/01/2020 00:00
01/01/2020 00:00
01/01/2020 00:00
02/01/2020 00:00
04/01/2020 00:00
04/01/2020 00:00
04/01/2020 00:00
05/01/2020 00:00
05/01/2020 00:00
07/01/2020 00:00
08/01/2020 00:00
22/02/2020 00:00
14/01/2020 00:00
14/01/2020 00:00
15/01/2020 00:00
15/01/2020 00:00
19/01/2020 00:00
20/01/2020 00:00



Date Occurred
08/01/2020
01/01/2020
13/02/2020
01/01/2020
01/01/2020
01/01/2020
02/01/2020
04/01/2020
04/01/2020
04/01/2020
05/01/2020
05/01/2020
07/01/2020
08/01/2020
22/02/2020
14/01/2020
14/01/2020
15/01/2020
15/01/2020
19/01/2020
20/01/2020
23/01/2020
03/09/2020
27/01/2020
28/01/2020

Time Occurred

The following were the challenges with this column:

The time was wrongly formatted as HHMM, while time is supposed to be HH:MM to be understood by Excel.

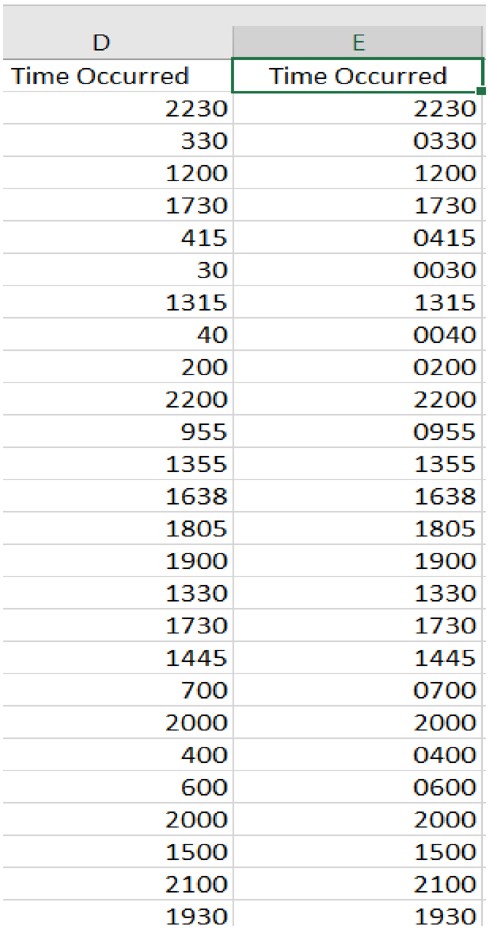
The inputs in some of the rows within the column were not 4 digits as standard 24 hours is recorded, some were in 3, 2 and 1 digit(s) which made it hard to read or understand.

And I took these steps to rectify it:

a. Changing the column to text and populating the column using TEXT function to return inputs in the format HHMM.

I used this to achieve this:

=TEXT(D2, “0000”)

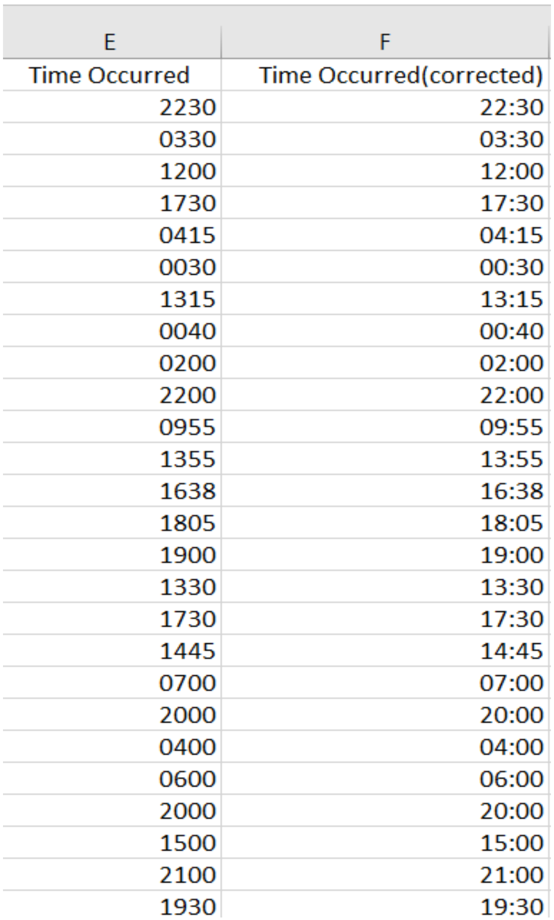


D	E
Time Occurred	Time Occurred
2230	2230
330	0330
1200	1200
1730	1730
415	0415
30	0030
1315	1315
40	0040
200	0200
2200	2200
955	0955
1355	1355
1638	1638
1805	1805
1900	1900
1330	1330
1730	1730
1445	1445
700	0700
2000	2000
400	0400
600	0600
2000	2000
1500	1500
2100	2100
1930	1930

b. I also wrote another TEXT function to populate the column with inputs for the correct 24 hrs time format HH:MM.

To achieve this, I used:

=TEXT(D2, “00\:00”)



E	F
Time Occurred	Time Occurred(corrected)
2230	22:30
0330	03:30
1200	12:00
1730	17:30
0415	04:15
0030	00:30
1315	13:15
0040	00:40
0200	02:00
2200	22:00
0955	09:55
1355	13:55
1638	16:38
1805	18:05
1900	19:00
1330	13:30
1730	17:30
1445	14:45
0700	07:00
2000	20:00
0400	04:00
0600	06:00
2000	20:00
1500	15:00
2100	21:00
1930	19:30

After confirming that the inputs are now in the HH:MM format, I changed its format to time.

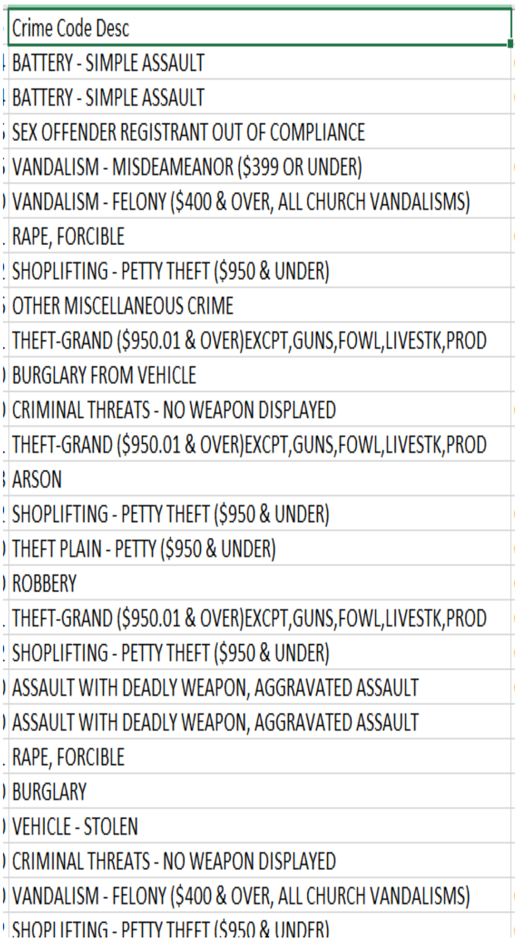
Crime Code Desc

For this column, I corrected the Column Title from **Crime Code Desc -> Crime Code Description.**

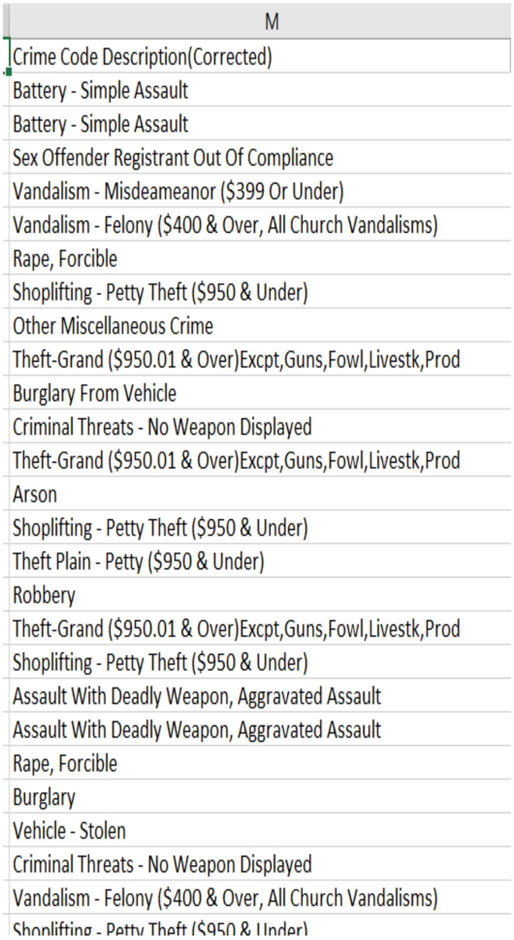
It was also noted that the column contained inputs that were all-in upper-case letters. To enhance readability, I used the PROPER function to correct this.

To achieve this, I used

=PROPER(L2)



Crime Code Desc
BATTERY - SIMPLE ASSAULT
BATTERY - SIMPLE ASSAULT
SEX OFFENDER REGISTRANT OUT OF COMPLIANCE
VANDALISM - MISDEAMEANOR ($399 OR UNDER)
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
RAPE, FORCIBLE
SHOPLIFTING - PETTY THEFT ($950 & UNDER)
OTHER MISCELLANEOUS CRIME
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
BURGLARY FROM VEHICLE
CRIMINAL THREATS - NO WEAPON DISPLAYED
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
ARSON
SHOPLIFTING - PETTY THEFT ($950 & UNDER)
THEFT PLAIN - PETTY ($950 & UNDER)
ROBBERY
THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD
SHOPLIFTING - PETTY THEFT ($950 & UNDER)
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
RAPE, FORCIBLE
BURGLARY
VEHICLE - STOLEN
CRIMINAL THREATS - NO WEAPON DISPLAYED
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
SHOPLIFTING - PETTY THEFT ($950 & UNDER)



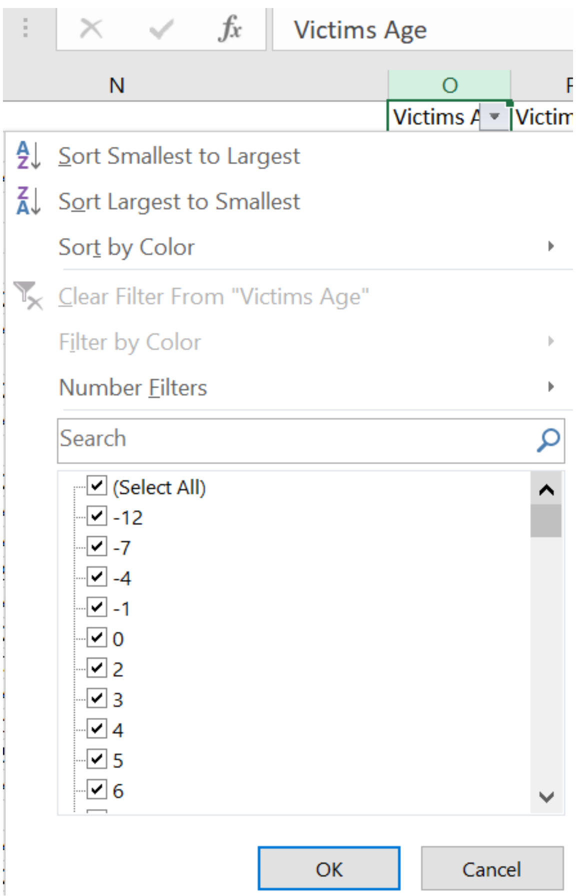
M
Crime Code Description(Corrected)
Battery - Simple Assault
Battery - Simple Assault
Sex Offender Registrant Out Of Compliance
Vandalism - Misdeameanor ($399 Or Under)
Vandalism - Felony ($400 & Over, All Church Vandalisms)
Rape, Forcible
Shoplifting - Petty Theft ($950 & Under)
Other Miscellaneous Crime
Theft-Grand ($950.01 & Over)Excpt,Guns,Fowl,Livestk,Prod
Burglary From Vehicle
Criminal Threats - No Weapon Displayed
Theft-Grand ($950.01 & Over)Excpt,Guns,Fowl,Livestk,Prod
Arson
Shoplifting - Petty Theft ($950 & Under)
Theft Plain - Petty ($950 & Under)
Robbery
Theft-Grand ($950.01 & Over)Excpt,Guns,Fowl,Livestk,Prod
Shoplifting - Petty Theft ($950 & Under)
Assault With Deadly Weapon, Aggravated Assault
Assault With Deadly Weapon, Aggravated Assault
Rape, Forcible
Burglary
Vehicle - Stolen
Criminal Threats - No Weapon Displayed
Vandalism - Felony ($400 & Over, All Church Vandalisms)
Shoplifting - Petty Theft ($950 & Under)

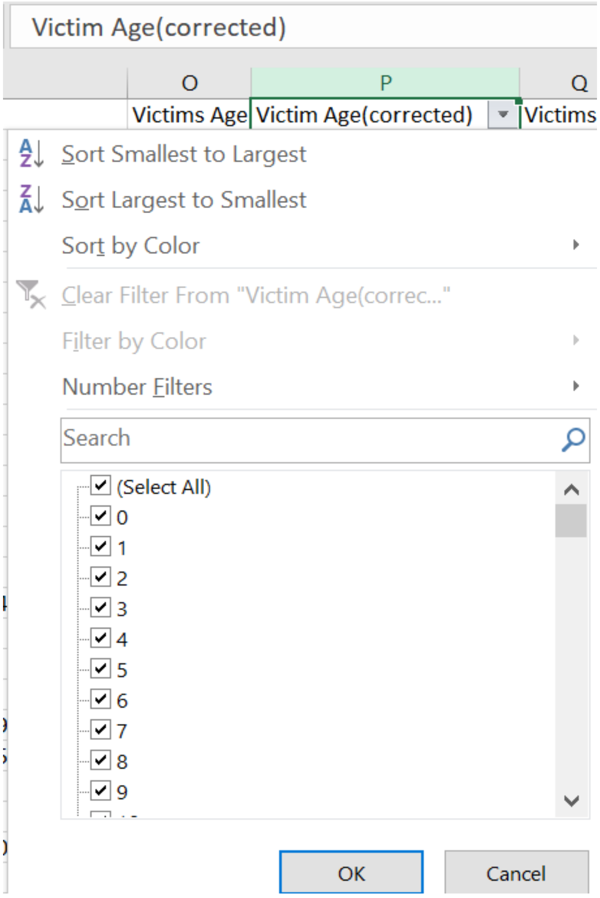
Victim’s Age

The victim ages column had values that ranged from -12 to 99, which didn’t make sense because ages can’t be negative. So, I used the “ABSOLUTE” function to make all the ages positive. Now, the ages range from 0 to 99. I decided to keep 0 for babies or anyone under the age of 1.

To achieve this, I used

=ABS(O2)





NB: After using the ABS formula to effect change in one Cell, I used Autofill to populate the rest of the column.

Victims Sex

The inputs in this column were F, H, M, X and blanks.

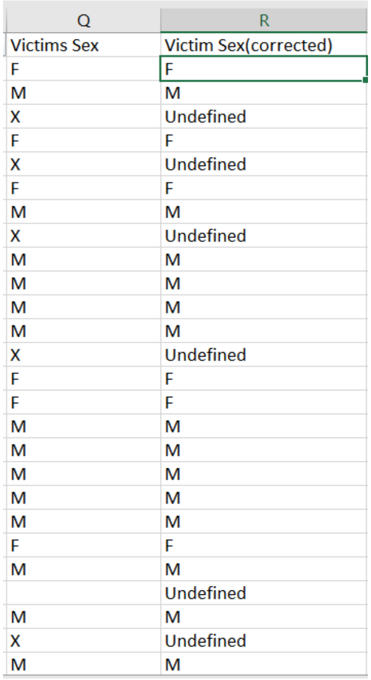
Firstly, we only recognize two genders. M and F represent Male and Female.

I replaced the other inputs (H, X, Blanks) with “Undefined”.

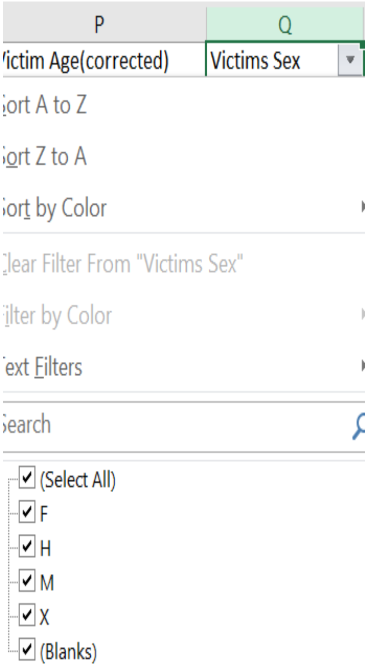
I used the IF(OR) statement to retain the F, M then specified the others as “Undefined”.

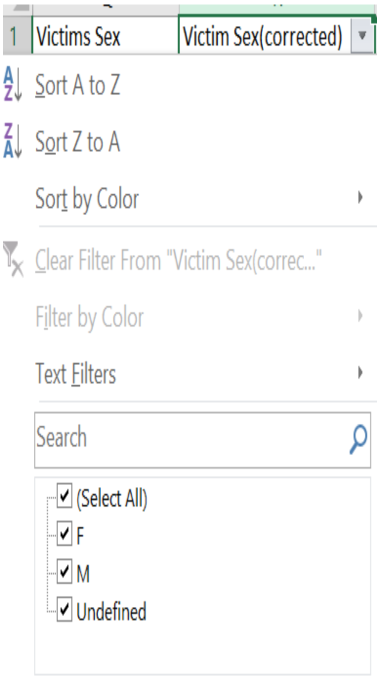
To achieve this, I used

=IF(OR(Q2=”M”, Q2=”F”), Q2, “Undefined”)



Q	R
Victims Sex	Victim Sex(corrected)
F	F
M	M
X	Undefined
F	F
X	Undefined
F	F
M	M
X	Undefined
M	M
M	M
M	M
M	M
X	Undefined
F	F
F	F
M	M
M	M
M	M
M	M
M	M
F	F
M	M
	Undefined
M	M
X	Undefined
M	M





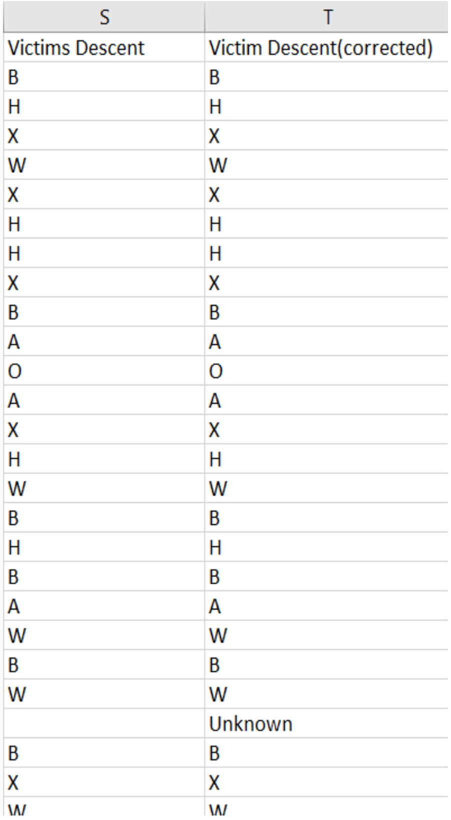
NB: After writing an IF statement to effect change in one Cell, I used Autofill to populate the rest of the column.

Victims Descent

This column had 19 distinct inputs as well as blank cells, I replaced the blank cells with “Unknown” to show that there was no record of their descent.

To achieve this, I used;

=IF(S2= “ ”, “Unknown”, S2)



S	T
Victims Descent	Victim Descent(corrected)
B	B
H	H
X	X
W	W
X	X
H	H
H	H
X	X
B	B
A	A
O	O
A	A
X	X
H	H
W	W
B	B
H	H
B	B
A	A
W	W
B	B
W	W
	Unknown
B	B
X	X
W	W

NB: After writing an IF statement to effect change in one Cell, I used Autofill to populate the rest of the column.

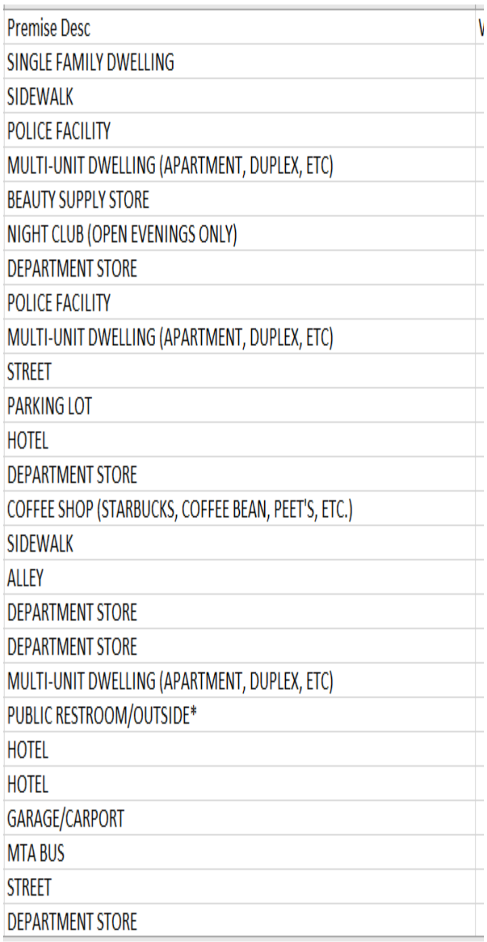
Premise Desc

For this column, I corrected the Column Title from **Premise Desc -> Premise Description.**

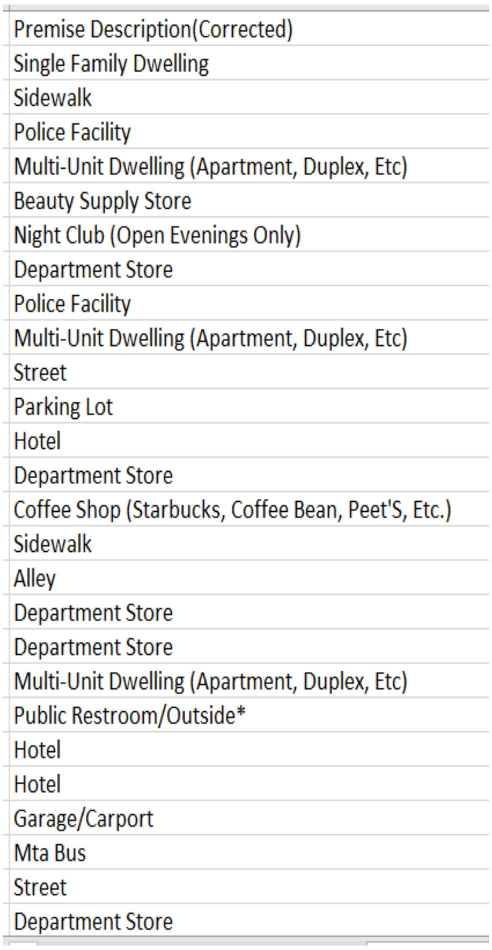
It was also noted that the column contained inputs that were all-in upper-case letters. To enhance readability, I used the PROPER function to correct this.

To achieve this, I used

=PROPER(V2)



Premise Desc	V
SINGLE FAMILY DWELLING	
SIDEWALK	
POLICE FACILITY	
MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	
BEAUTY SUPPLY STORE	
NIGHT CLUB (OPEN EVENINGS ONLY)	
DEPARTMENT STORE	
POLICE FACILITY	
MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	
STREET	
PARKING LOT	
HOTEL	
DEPARTMENT STORE	
COFFEE SHOP (STARBUCKS, COFFEE BEAN, PEET'S, ETC.)	
SIDEWALK	
ALLEY	
DEPARTMENT STORE	
DEPARTMENT STORE	
MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	
PUBLIC RESTROOM/OUTSIDE*	
HOTEL	
HOTEL	
GARAGE/CARPORT	
MTA BUS	
STREET	
DEPARTMENT STORE	



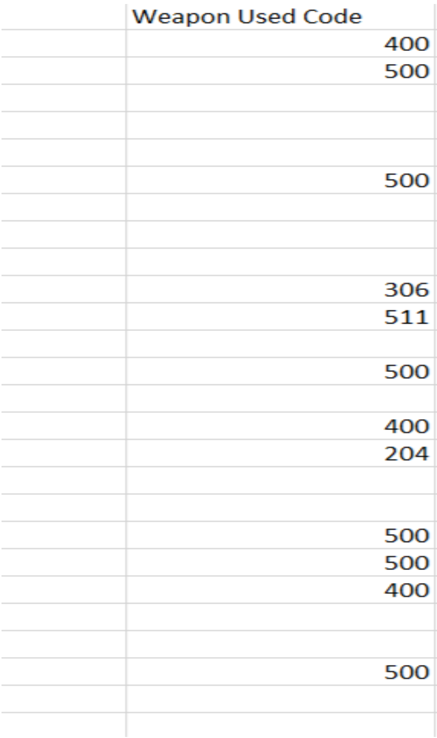
Premise Description(Corrected)
Single Family Dwelling
Sidewalk
Police Facility
Multi-Unit Dwelling (Apartment, Duplex, Etc)
Beauty Supply Store
Night Club (Open Evenings Only)
Department Store
Police Facility
Multi-Unit Dwelling (Apartment, Duplex, Etc)
Street
Parking Lot
Hotel
Department Store
Coffee Shop (Starbucks, Coffee Bean, Peet'S, Etc.)
Sidewalk
Alley
Department Store
Department Store
Multi-Unit Dwelling (Apartment, Duplex, Etc)
Public Restroom/Outside*
Hotel
Hotel
Garage/Carport
Mta Bus
Street
Department Store

Weapon Used Code

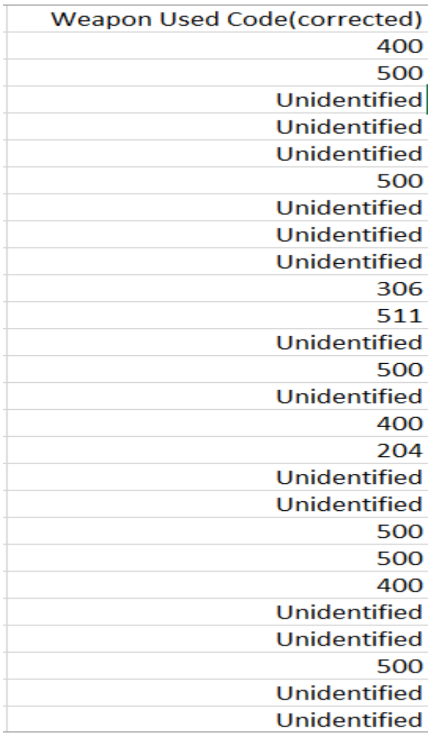
This column contained a lot of blank cells which I eventually replaced with "unidentified" which showed that the weapon code wasn't known.

To achieve this, I used the formula below to populate the column via auto fill.

=IF(W4="", "Unidentified", W4)



Weapon Used Code
400
500
500
306
511
500
400
204
500
500
400
500



Weapon Used Code(corrected)
400
500
Unidentified
Unidentified
Unidentified
500
Unidentified
Unidentified
Unidentified
306
511
Unidentified
500
Unidentified
400
204
Unidentified
Unidentified
500
500
400
Unidentified
Unidentified
500
Unidentified
Unidentified

Weapon Description

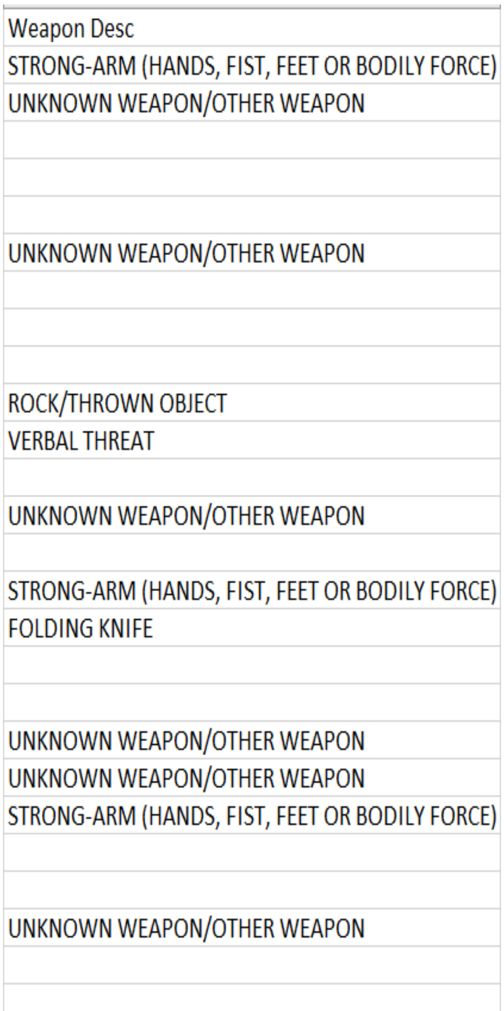
For this column, I corrected the Column Title from **Weapon Desc -> Weapon Description.**

This column contained a lot of blank cells which I eventually replaced with **"unidentified Weapon"** which showed that the weapon used in committing the crime wasn't known.

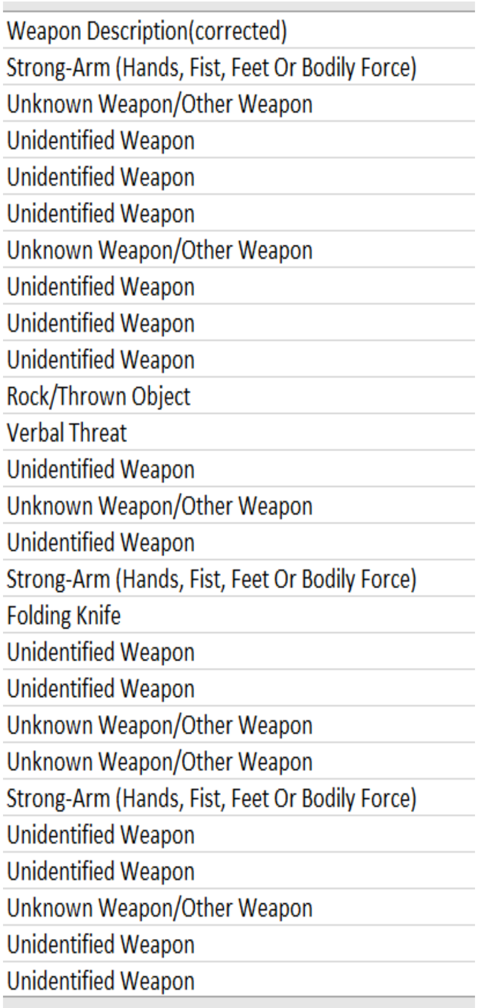
The column also had inputs that were majorly upper case and this affects the readability of the inputs. I wrapped the IF statement used in filling the blanks with a **PROPER** formula.

To achieve this, I used the formula below to populate the column via auto fill.

=PROPER(IF(Y2=“ ”, “Unidentified Weapon”, Y2))



Weapon Desc
STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)
UNKNOWN WEAPON/OTHER WEAPON
UNKNOWN WEAPON/OTHER WEAPON
ROCK/THROWN OBJECT
VERBAL THREAT
UNKNOWN WEAPON/OTHER WEAPON
STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)
FOLDING KNIFE
UNKNOWN WEAPON/OTHER WEAPON
UNKNOWN WEAPON/OTHER WEAPON
STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)
UNKNOWN WEAPON/OTHER WEAPON

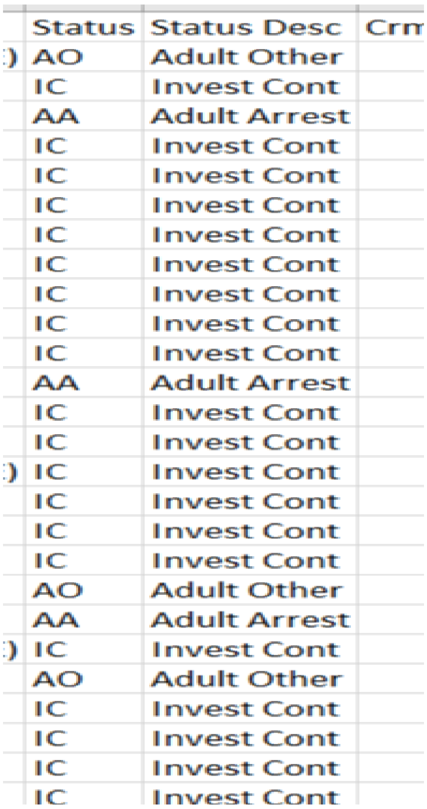


Weapon Description(corrected)
Strong-Arm (Hands, Fist, Feet Or Bodily Force)
Unknown Weapon/Other Weapon
Unidentified Weapon
Unidentified Weapon
Unidentified Weapon
Unknown Weapon/Other Weapon
Unidentified Weapon
Unidentified Weapon
Unidentified Weapon
Rock/Thrown Object
Verbal Threat
Unidentified Weapon
Unknown Weapon/Other Weapon
Unidentified Weapon
Strong-Arm (Hands, Fist, Feet Or Bodily Force)
Folding Knife
Unidentified Weapon
Unidentified Weapon
Unknown Weapon/Other Weapon
Unknown Weapon/Other Weapon
Strong-Arm (Hands, Fist, Feet Or Bodily Force)
Unidentified Weapon
Unidentified Weapon
Unknown Weapon/Other Weapon
Unidentified Weapon
Unidentified Weapon

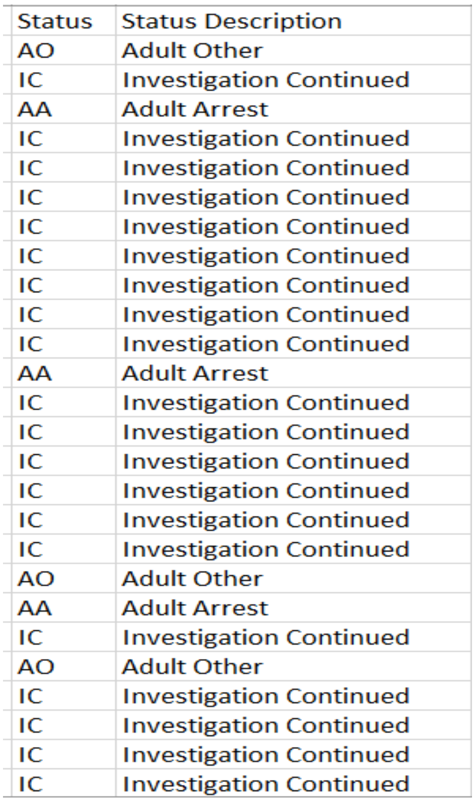
Status Desc

For this column, I corrected the Column Title from **Status Desc -> Status Description.**

To replace rows that contained **“Invest Cont”** within the column, I selected the Column **Status Description**, used **CTRL F** to find all rows that contained that phrase, then replaced all with **“Investigation Continued”.**



	Status	Status Desc	Crm
)	AO	Adult Other	
	IC	Invest Cont	
	AA	Adult Arrest	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	AA	Adult Arrest	
	IC	Invest Cont	
	IC	Invest Cont	
)	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	AO	Adult Other	
	AA	Adult Arrest	
)	IC	Invest Cont	
	AO	Adult Other	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	
	IC	Invest Cont	



Status	Status Description
AO	Adult Other
IC	Investigation Continued
AA	Adult Arrest
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
AA	Adult Arrest
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
AO	Adult Other
AA	Adult Arrest
IC	Investigation Continued
AO	Adult Other
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued
IC	Investigation Continued

Location and Street

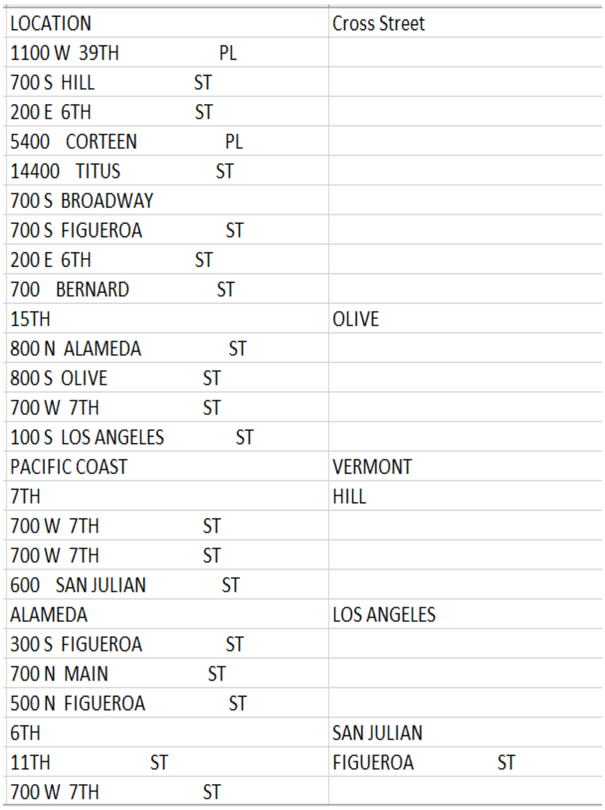
I merged these 2 columns to form the **Address** column using the **CONCATENATE** function.

I used the **TRIM** function normalize the spacings which were irregular as observed in the **Location** and **Cross Street** Columns.

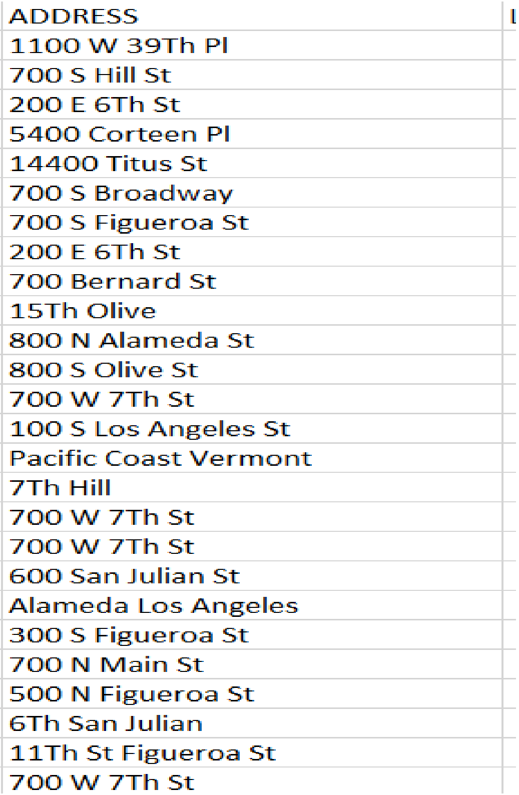
I also employed the PROPER function to normalize the capitalized fonts used in entry of inputs in the Location and Cross Street columns.

To achieve this, I used the formula below to populate the column via auto fill.

=TRIM(PROPER(CONCATENATE(AG2, " ", AH2)))



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	
15TH	OLIVE
800 N ALAMEDA ST	
800 S OLIVE ST	
700 W 7TH ST	
100 S LOS ANGELES ST	
PACIFIC COAST	VERMONT
7TH	HILL
700 W 7TH ST	
700 W 7TH ST	
600 SAN JULIAN ST	
ALAMEDA	LOS ANGELES
300 S FIGUEROA ST	
700 N MAIN ST	
500 N FIGUEROA ST	
6TH	SAN JULIAN
11TH ST	FIGUEROA ST
700 W 7TH ST	



ADDRESS
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
15Th Olive
800 N Alameda St
800 S Olive St
700 W 7Th St
100 S Los Angeles St
Pacific Coast Vermont
7Th Hill
700 W 7Th St
700 W 7Th St
600 San Julian St
Alameda Los Angeles
300 S Figueroa St
700 N Main St
500 N Figueroa St
6Th San Julian
11Th St Figueroa St
700 W 7Th St

CONCLUSION:

In conclusion, the data cleaning task successfully addressed inconsistencies, errors, and missing values in the dataset, enhancing its overall quality and reliability. The refined data now serves as a solid foundation for accurate analysis and informed decision-making.