Chicago Motor Vehicle Theft Analysis

Overview

- This project involves the analysis of vehicle thefts in the city of Chicago from 2001 until 2022
- We will analyze the worst hit neighborhoods within the city and analyze the overall trends
- Dataset: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
- Portfolio: https://muthalibabdul.github.io/Main.Portfolio/
- Software: SAS Studio 9.4 for Academics

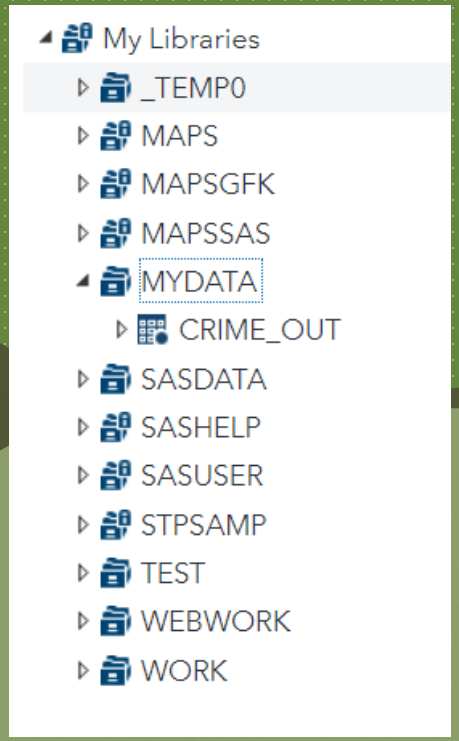
STEP 1: DEFINE THE PROBLEM

Crimes are taking over the second most beautiful city of Chicago. In this project we will analyze real world crimes data from Chicago Data Portal and try to answer some of the simple questions like

1. How many types of crimes have been reported and what are their frequency during the study?
2. How many Car Theft cases were reported in the city of Chicago during the year 2001 to 2022?
3. How many arrests were made of those reported cases? what is the percentage of arrests made?
4. Which year has reported the highest number of car theft in the city?
5. What is the worst hit community area in Chicago affected car thefts?
6. Which months of the year are more vulnerable to car thefts?

STEP 2: ACCESSING DATA / DATA IMPORT

```
1 /* Using proc import to import xlsx file initially, Do not run, takes 10 mins to load */
2 options validvarname=v7;
3 proc import datafile="/home/u60639771/Chicago_Crime/Chicago_Crimes_2001-2022.xlsx" dbms=xlsx
4 out=mydata.crime_out;
```



- Using Proc Import function, we first pull the xlsx file into our library.
- Options Validvarname enables SAS to specify the variables follow SAS standards
- The dbms function lets us chose the format our data is in.
- Out function creates a library called "mydata" with the data set named "crime_out"
- Lastly, we can see our data set appears in the library

Step 3 : Data Exploration

```
/* Using proc contents to understand the data set */
proc contents data=mydata.crime_out;
run;
```

The CONTENTS Procedure

Data Set Name	MYDATA.CRIME_OUT	Observations	1048575
Member Type	DATA	Variables	22
Engine	V9	Indexes	0
Created	05/27/2023 21:59:45	Observation Length	344
Last Modified	05/27/2023 21:59:45	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

- Using proc contents, we have a glimpse of what is in the data set. For example, we can see the total observations (rows), variables (columns), data set name
- One of the crucial aspect of proc contents is the being able to see the variables, its type, length of the variable, format, labels
- Contents is one of the crucial steps before we start doing some data analysis/cleaning

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
9	Arrest	Num	8	BEST.		Arrest
11	Beat	Num	8	BEST.		Beat
4	Block	Char	35	$35.	$35.	Block
2	Case_Number	Char	9	$9.	$9.	Case Number
14	Community_Area	Num	8	BEST.		Community Area
3	Date	Num	8	DATETIME16.		Date
7	Description	Char	60	$60.	$60.	Description
12	District	Num	8	BEST.		District
10	Domestic	Num	8	BEST.		Domestic
15	FBI_Code	Char	3	$3.	$3.	FBI Code
1	ID	Num	8	BEST.		ID
5	IUCR	Char	4	$4.	$4.	IUCR
20	Latitude	Num	8	BEST.		Latitude
22	Location	Char	29	$29.	$29.	Location
8	Location_Description	Char	53	$53.	$53.	Location Description
21	Longitude	Num	8	BEST.		Longitude
6	Primary_Type	Char	33	$33.	$33.	Primary Type
19	Updated_On	Num	8	DATETIME16.		Updated On
13	Ward	Num	8	BEST.		Ward
16	X_Coordinate	Num	8	BEST.		X Coordinate
17	Y_Coordinate	Num	8	BEST.		Y Coordinate
18	Year	Num	8	BEST.		Year

```
/* Diff data values in a single column */
proc freq data=mydata.crime_out;
tables Primary_Type;
run;
```

- Our focus for this project is "Vehicle Theft". Hence, we will use the proc freq procedure to check for different data values within "Primary_Type" column and assess if any data value matches our interest.
- We can see we have "Motor Vehicle Theft" as one of the Primary_Type

Primary Type				
Primary_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ARSON	1702	0.16	1702	0.16
ASSAULT	75056	7.16	76758	7.32
BATTERY	195571	18.65	272329	25.97
BURGLARY	51096	4.87	323425	30.84
CONCEALED CARRY LICENSE VIOLATION	345	0.03	323770	30.88
CRIM SEXUAL ASSAULT	5392	0.51	329162	31.39
CRIMINAL DAMAGE	116324	11.09	445486	42.48
CRIMINAL SEXUAL ASSAULT	179	0.02	445665	42.50
CRIMINAL TRESPASS	26623	2.54	472288	45.04
DECEPTIVE PRACTICE	71363	6.81	543651	51.85
GAMBLING	785	0.07	544436	51.92
HOMICIDE	29	0.00	544465	51.92
HUMAN TRAFFICKING	42	0.00	544507	51.93
INTERFERENCE WITH PUBLIC OFFICER	4544	0.43	549051	52.36
INTIMIDATION	601	0.06	549652	52.42
KIDNAPPING	798	0.08	550450	52.50
LIQUOR LAW VIOLATION	977	0.09	551427	52.59
MOTOR VEHICLE THEFT	42949	4.10	594376	56.68
NARCOTICS	57029	5.44	651405	62.12
NON - CRIMINAL	13	0.00	651418	62.12
NON-CRIMINAL	133	0.01	651551	62.14
NON-CRIMINAL (SUBJECT SPECIFIED)	6	0.00	651557	62.14
OBSCENITY	263	0.03	651820	62.16
OFFENSE INVOLVING CHILDREN	8721	0.83	660541	62.99
OTHER NARCOTIC VIOLATION	20	0.00	660561	63.00
OTHER OFFENSE	69118	6.59	729679	69.59
PROSTITUTION	3441	0.33	733120	69.92
PUBLIC INDECENCY	43	0.00	733163	69.92
PUBLIC PEACE VIOLATION	6379	0.61	739542	70.53
RITUALISM	1	0.00	739543	70.53
ROBBERY	42280	4.03	781823	74.56
SEX OFFENSE	4022	0.38	785845	74.94
STALKING	738	0.07	786583	75.01
THEFT	243834	23.25	1030417	98.27
WEAPONS VIOLATION	18158	1.73	1048575	100.00

STEP 4: DATA CLEANING

```
24
25 /* Using Keep function to keep only a limited number of variables for further analysis */
26 data crime1;
27 set mydata.crime_out;
28 keep ID Date Primary_Type Location_Description Community_Area Arrest;
29 run;
30
```

- Using the keep function, we keep only the variables we would be interested for our further analysis

Total rows: 1048575 Total columns: 6 Rows 1-100

	ID	Date	Primary_Type	Location_Description	Arrest	Community_Area
1	10224738	05SEP15:13:30:00	BATTERY	RESIDENCE	0	61
2	10224739	04SEP15:11:30:00	THEFT	CTA BUS	0	25
3	11646166	01SEP18:00:01:00	THEFT	RESIDENCE	0	44
4	10224740	05SEP15:12:45:00	NARCOTICS	SIDEWALK	1	21
5	10224741	05SEP15:13:00:00	ASSAULT	APARTMENT	0	25
6	10224742	05SEP15:10:55:00	BURGLARY	RESIDENCE	0	71
7	10224743	04SEP15:18:00:00	BURGLARY	RESIDENCE-GARAGE	0	24
8	10224744	05SEP15:13:00:00	THEFT	GROCERY FOOD STORE	1	31
9	10224745	05SEP15:11:30:00	ROBBERY	STREET	0	27
10	11645836	01MAY16:00:25:00	DECEPTIVE PRACTICE		0	63
11	10224746	05SEP15:14:00:00	THEFT	PARKING LOT/GARAGE(NON.RESID.)	0	65
12	10224749	05SEP15:11:00:00	BATTERY	SMALL RETAIL STORE	0	11
13	10224750	05SEP15:03:00:00	OTHER OFFENSE	APARTMENT	0	49

STEP 5: DATA MANIPULATION

```
/* Subsetting data to MOTOR VEHICLE THEFT and creating new variables*/
data crime2;
set work.crime1;
keep ID NewDate Year Month Day Primary_Type Location_Description Community_Area Arrest;
where Primary_Type="MOTOR VEHICLE THEFT";
/* Date = '01JAN2001'd; */
NewDate=datepart(Date);
format NewDate date9.;
Year = year(NewDate);
Month = month(NewDate);
Day = day(NewDate);
run;
```

- Here we use the where statement to subset the data based on our interest "MOTOR VEHICLE THEFT"
- Then we extract the date portion using datepart function and format the date to date9. format
- Lastly we segregate the month, day and year using simple functions for individual analyses

- We can see the changes reflecting in the Primary_Type column
- We can also see the newly created variables Year, Month and Day

Total rows: 42949 Total columns: 9

	ID	Primary_Type	Location_Description	Arrest	Community_Area	NewDate	Year	Month	Day
1	10224800	MOTOR VEHICLE THEFT	STREET	0	6	05SEP2015	2015	9	5
2	10224846	MOTOR VEHICLE THEFT	DRIVEWAY - RESIDENTIAL	0	29	05SEP2015	2015	9	5
3	10224876	MOTOR VEHICLE THEFT	STREET	0	22	05SEP2015	2015	9	5
4	10224893	MOTOR VEHICLE THEFT	PARKING LOT/GARAGE(NON.RESID.)	0	1	05SEP2015	2015	9	5
5	10224905	MOTOR VEHICLE THEFT	STREET	0	71	05SEP2015	2015	9	5
6	10224915	MOTOR VEHICLE THEFT	STREET	0	25	04SEP2015	2015	9	4
7	10224926	MOTOR VEHICLE THEFT	STREET	0	41	05SEP2015	2015	9	5
8	10225067	MOTOR VEHICLE THEFT	STREET	1	34	05SEP2015	2015	9	5
9	10225087	MOTOR VEHICLE THEFT	ALLEY	0	25	05SEP2015	2015	9	5
10	10225123	MOTOR VEHICLE THEFT	STREET	0	67	05SEP2015	2015	9	5
11	10225144	MOTOR VEHICLE THEFT	STREET	0	67	03SEP2015	2015	9	3
12	10225160	MOTOR VEHICLE THEFT	STREET	0	77	05SEP2015	2015	9	5

- Next we use proc format to create a format to make the Arrest column eligible for presentation.
- We can also use a simple IF-THEN statement
- On the left top portion of the code, we first create the format
- Next, in the code section below, we use the format arr
- Lastly, on the right we can see the changes reflecting accordingly

```
/* Creating format for Arrest */
proc format;
value arr
        0='No'
        1='Yes';
run;
```

```
/* Using arr format */
data crime3;
set crime2;
format arrest arr.;
run;
```

Total rows: 42949 Total columns: 9

	ID	Primary_Type	Location_Description	Arrest	Community_Area	NewDate	Year	Month	Day
1	10224800	MOTOR VEHICLE THEFT	STREET	No	6	05SEP2015	2015	9	5
2	10224846	MOTOR VEHICLE THEFT	DRIVEWAY - RESIDENTIAL	No	29	05SEP2015	2015	9	5
3	10224876	MOTOR VEHICLE THEFT	STREET	No	22	05SEP2015	2015	9	5
4	10224893	MOTOR VEHICLE THEFT	PARKING LOT/GARAGE(NON.RESID.)	No	1	05SEP2015	2015	9	5
5	10224905	MOTOR VEHICLE THEFT	STREET	No	71	05SEP2015	2015	9	5
6	10224915	MOTOR VEHICLE THEFT	STREET	No	25	04SEP2015	2015	9	4
7	10224926	MOTOR VEHICLE THEFT	STREET	No	41	05SEP2015	2015	9	5
8	10225067	MOTOR VEHICLE THEFT	STREET	Yes	34	05SEP2015	2015	9	5
9	10225087	MOTOR VEHICLE THEFT	ALLEY	No	25	05SEP2015	2015	9	5
10	10225123	MOTOR VEHICLE THEFT	STREET	No	67	05SEP2015	2015	9	5
11	10225144	MOTOR VEHICLE THEFT	STREET	No	67	03SEP2015	2015	9	3
12	10225160	MOTOR VEHICLE THEFT	STREET	No	77	05SEP2015	2015	9	5

STEP 6: Data Analysis & Interpretation

- Starting with brief analysis, we use the proc freq procedure to assess the highest number of car theft by year (the order function sorts the year by frequency highest to lowest)
- We see years towards the end of the table having unrealistic frequency.
- This might be and error within the data?

```
56
57 /* proc freq */
58 proc freq data=crime3 order=freq;
59 tables Year;
60 run;
61
```

The FREQ Procedure

Year	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2017	11230	26.15	11230	26.15
2016	11127	25.91	22357	52.05
2018	9842	22.92	32199	74.97
2015	5389	12.55	37588	87.52
2019	3454	8.04	41042	95.56
2001	1411	3.29	42453	98.85
2020	365	0.85	42818	99.69
2021	56	0.13	42874	99.83
2002	32	0.07	42906	99.90
2022	32	0.07	42938	99.97
2014	6	0.01	42944	99.99
2013	2	0.00	42946	99.99
2005	1	0.00	42947	100.00
2008	1	0.00	42948	100.00
2011	1	0.00	42949	100.00

```
61
62 /* No Motor Vehicle Theft in 2003 2004? */
63 data crimenew;
64 set mydata.crime_out;
65 where year(Date) in(2003,2004) and Primary_Type="MOTOR VEHICLE THEFT";
66 run;
67
```

```
72
73          data crimenew;
74          set mydata.crime_out;
75          where year(Date) in(2003,2004) and Primary_Type="MOTOR VEHICLE THEFT";
76          run;

NOTE: There were 0 observations read from the data set MYDATA.CRIME_OUT.
      WHERE YEAR(Date) in (2003, 2004) and (Primary_Type='MOTOR VEHICLE THEFT');
NOTE: The data set WORK.CRIMENEW has 0 observations and 22 variables.
```

- We can use the where statement and in function to include all the years we see no crimes occurring.
- Our log output says we have 0 observation.
- We can further check the original data and act as per directions by the manager

```
/* Macros */
%macro tabs (data, vari);
proc freq data=&data order=freq;
tables &vari;
run;
%mend;
```

- Lastly we use a simple macro function to create a macro which makes our code less complicated to execute with different variables.
- First step involve creating the variables for data and variables

- The next step involve calling the macro with data dataset name and the variable of interest. Only changes in the second portion of the macro needs to be changed based on our variable of interest.

```
%tabs (work.crime3, Arrest);
```

The FREQ Procedure

Arrest				
Arrest	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	39542	92.07	39542	92.07
Yes	3407	7.93	42949	100.00

```
%tabs (work.crime3, Community_Area);
```

The FREQ Procedure

Community Area				
Community_Area	Frequency	Percent	Cumulative Frequency	Cumulative Percent
25	2834	6.82	2834	6.82
24	1425	3.43	4259	10.25
28	1395	3.36	5654	13.61
23	1347	3.24	7001	16.85
29	1280	3.08	8281	19.93
43	1166	2.81	9447	22.74
8	1150	2.77	10597	25.51
19	1132	2.72	11729	28.23

```
%tabs (work.crime3, Month);
```

The FREQ Procedure

Month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4991	11.62	4991	11.62
12	4013	9.34	9004	20.96
8	3975	9.26	12979	30.22
10	3818	8.89	16797	39.11
9	3625	8.44	20422	47.55
11	3559	8.29	23981	55.84
7	3387	7.89	27368	63.72
3	3247	7.56	30615	71.28
4	3223	7.50	33838	78.79
2	3201	7.45	37039	86.24
5	3026	7.05	40065	93.29
6	2884	6.71	42949	100.00

- INSGHTS?

- There have been 35 different types of crimes reported
- 42949 cases were reported in the city of Chicago during the year 2001 to 2022
- Only 3407 arrests were made which equates to 7.93%
- 2017 saw the highest number of car thefts with 11,230 case
- Community Area 25 with 2835 cases and frequency percentage of 6.82% is the worst hit neighborhood in the city of Chicago
- December and January are the most vulnerable months for car thefts