HiveQL – Chicago Crime Data

This data shows the crime incidents happened in the City of Chicago from 2001 till date. We are using HiveQL to run queries on this dataset and understand the trend and pattern of the crimes from the query result. The objective of this project is to identify the common type of Crime, risky locations and year and overall what the data can say about Chicago's crime rate.

We can start by querying the **top 3 seen types of crimes i.e. Crimes that are mostly common. (Q1)**

Results (3)

	primary type	crimecount
1	THEFT	1450019
2	BATTERY	1257663
3	CRIMINAL DAMAGE	785088

The result indicates that Theft being the highest followed by Battery and Criminal Damage as the 3 most common type of crime committed in the City of Chicago. We can also identify the **least seen types of crimes i.e. Crimes that are the least common. (Q2)**

Results (3)

	primary type	crimecount
1	DOMESTIC VIOLENCE	1
2	NON-CRIMINAL (SUBJECT SPECIFIED)	9
3	RITUALISM	23

As per the result Domestic Violence, Non-Criminal and Ritualism are the least common crimes identified. We should be able to identify **the top 3 safest and riskiest neighborhoods (Q3)** as well along with the crime. I'm considering zip code and block as neighborhood as different blocks are coming under one Zip code for riskiest neighborhoods. Since there are multiple blocks with 1 crime count, I'm including only zip codes for the safest neighborhoods.

- Top 3 **safest** neighborhoods

Hive Execute and watch *Add a description...*

38.66s default text

```
SELECT `zip codes`, count(`primary type`) AS crimecount
FROM `msba_db2`.`crimes_2001_to_present`
WHERE `zip codes` IS NOT NULL
GROUP BY `zip codes`
ORDER BY crimecount LIMIT 3;
```

Results (3)

	zip codes	crimecount
1	4094	1
2	9458	5
3	26912	11

- Top 3 **riskiest** neighborhoods

```
SELECT block,`zip codes`, COUNT('primary type') AS crimecount
FROM `msba_db2`.`crimes_2001_to_present`
GROUP BY `zip codes`, block
ORDER by crimecount DESC LIMIT 10;
```

	block	zip codes	crimecount
1	100XX W OHARE ST	16197	15112
	001XX N STATE ST	14310	9407
3	076XX S CICERO AVE	4300	9294

We can find identifying which where **the safest and Riskier 3 years (Q4).**

- **Riskier** 3 years

36.22s

```
SELECT year, count(`primary type`) AS crimecount
FROM `msba_db2`.`crimes_2001_to_present`
GROUP BY year
ORDER BY crimecount DESC LIMIT 3;
```

	year	crimecount
1	2002	486757
2	2001	485754
3	2003	475946

* Limiting to 4 for safest 3 years, as assuming 2019 is still running and this count is not complete.

- **Safest** 3 years

37.49s de

```
SELECT year, count(`primary type`) AS crimecount
FROM `msba_db2`.`crimes_2001_to_present`
GROUP BY year
ORDER BY crimecount LIMIT 4;
```

Results (4)

	year	crimecount
1	2019	98989
2	2015	264143
3	2018	266849
4	2017	268215

Identifying the safest and riskier consecutive 3 years? (ex: 1996-1997-1998) (Q5)

41.62s msba_db2 text ?

```
SELECT year AS startyear,crimecount AS crimecountcurrentyear,
crimecount + lead(crimecount,1) OVER (ORDER BY year) + lead(crimecount,2)
OVER (ORDER BY year) AS crimecountover3years
FROM
(SELECT year,
count(`primary type`) as crimecount
FROM `msba_db2`.`crimes_2001_to_present`
GROUp BY year)A
ORDER BY crimecountover3years;
```

There is no data for 2019 and 2020 because of which we have NULL values for the 3 years from 2018. Lowest 3years value is for 2017 – 634,053 however since we only do not have full year data for 2019, I'm choosing value 801424 i.e. 264143 + 269066 + 268215 for safest 3 years 2015, 2016 and 2017

Results (19)

	startyear	crimecountcurrentyear	crimecountover3years
1	2019	98989	NULL
2	2018	266849	NULL
3	2017	268215	634053
4	2015	264143	801424
5	2016	269066	804130
6	2014	275324	808533
7	2013	307135	846602
8	2012	335987	918446
9	2011	351794	994916
10	2010	370321	1058102
11	2009	392698	1114813
12	2008	427065	1190084
13	2007	437016	1256779
14	2006	448114	1312195

Results (19)

	startyear	crimecountcurrentyear	crimecountover3years
1	2019	98989	NULL
2	2018	266849	NULL
3	2017	268215	634053
4	2015	264143	801424
5	2016	269066	804130
6	2014	275324	808533

For Riskier 3 years, 2001, 2002 and 2003: 485754+486757+475946 =1448457

52.36s msba_db2 text

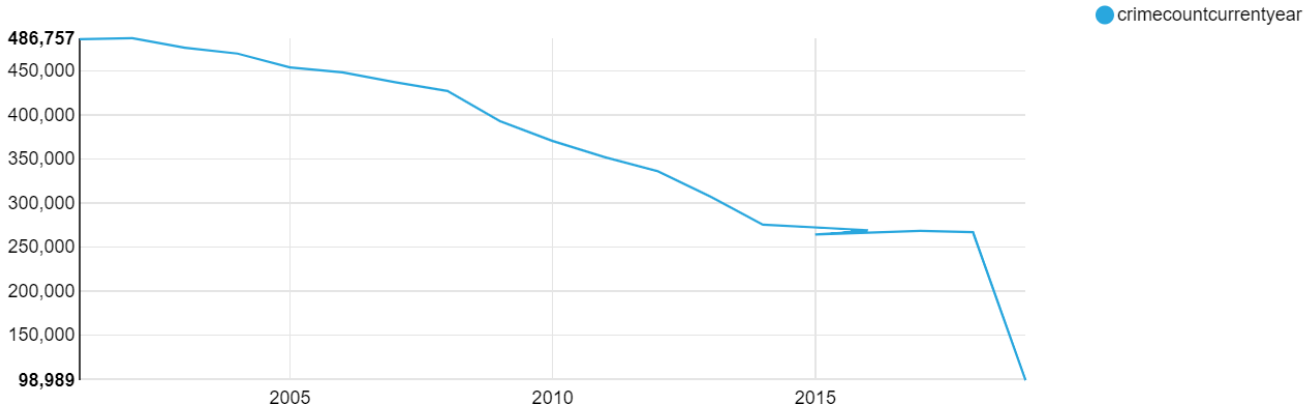
```
SELECT year AS startyear,crimecount AS crimecountcurrentyear,
crimecount + lead(crimecount,1) OVER (ORDER BY year) + lead(crimecount,
OVER (ORDER BY year) AS crimecountover3years
FROM
(SELECT year,
count(`primary type`) as crimecount
FROM `msba_db2`.`crimes_2001_to_present`
GROUp BY year)A
ORDER BY crimecountover3years DESC;
```

Results (19)

	startyear	crimecountcurrentyear	crimecountover3years
1	2001	485754	1448457
2	2002	486757	1432087
3	2003	475946	1399047
4	2004	469384	1371215
5	2005	453717	1338847
6	2006	448114	1312195

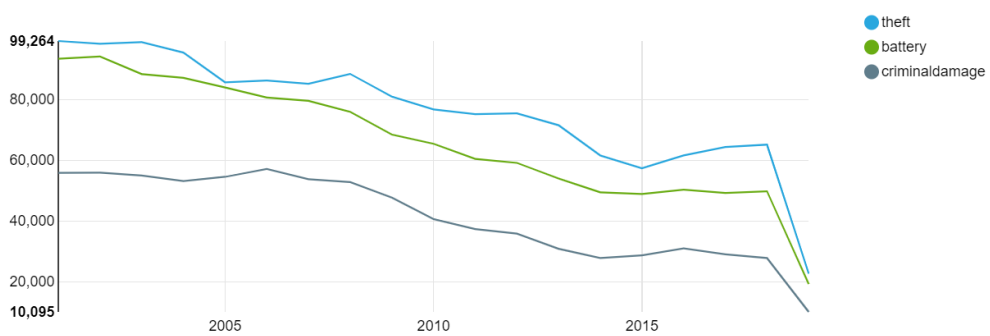
Now we can see how crime has been evolving over time for all the data points we just saw.

The below chart indicates crime count over time, we can see it has been lowest in 2019 and highest in 2001.



Below chart shows top 3 crimes (Theft, Battery and Criminal Damage) over time.

```
SELECT year,
count(CASE WHEN `primary type` = 'THEFT' THEN 1 END) AS Theft,
count(case WHEN `primary type` = 'BATTERY' THEN 1 end) AS Battery,
count(case WHEN `primary type` = 'CRIMINAL DAMAGE' THEN 1 end) AS CriminalDamage
FROM `msba_db2`.`crimes_2001_to_present`
GROUP BY year
ORDER BY year;
```



Reference

Mode (n.d.) SQL Window Functions. Retrieved from https://mode.com/resources/sql-tutorial/sql-window-functions/

Microsoft (2017). Retrieved from https://docs.microsoft.com/en-us/sql/t-sql/functions/lead-transact-sql?view=sql-server-2017