BIG DATA ECOSYSTEMS

MSBA - 321

**CHICAGO POLICE DEPARTMENT-ILLINOIS UNIFORM CRIME REPORTING (IUCR) CODES**

BY

PORRSELVI

**CHICAGO POLICE DEPARTMENT-ILLINOIS UNIFORM CRIME REPORTING (IUCR) CODES**

**INTRODUCTION**

Illinois Uniform Crime Reporting (IUCR) codes are four digit codes that law enforcement agencies use to classify criminal incidents when taking individual reports. These codes are also used to aggregate types of cases for statistical purposes. In Illinois, the Illinois State Police establish IUCR codes, but the agencies can add codes to suit their individual needs. The Chicago Police Department currently uses more than 350 IUCR codes to classify criminal offenses, divided into “Index” and “Non-Index” offenses.

USING HIVE TECHNOLOGY, I HAVE ANALYZED CHICAGO POLICE DEPARTMENT DATA OF ILLINOIS UNIFORM CRIME REPORTING (IUCR).

HIVE QUERY LANGUAGE:

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

The term ‘Big Data’ is used for collections of large datasets that include huge volume, high velocity, and a variety of data that is increasing day by day. Using traditional data management systems, it is difficult to process Big Data. Therefore, the Apache Software Foundation introduced a framework called Hadoop to solve Big Data management and processing challenges.

Hadoop:

Hadoop is an open-source framework to store and process Big Data in a distributed environment. It contains two modules, one is MapReduce and another is Hadoop Distributed File System (HDFS).

* **MapReduce:** It is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware.
* **HDFS:**Hadoop Distributed File System is a part of Hadoop framework, used to store and process the datasets. It provides a fault-tolerant file system to run on commodity hardware.

The Hadoop ecosystem contains different sub-projects (tools) such as Sqoop, Pig, and Hive that are used to help Hadoop modules.

**Hive:** It is a platform used to develop SQL type scripts to do MapReduce operations.

**DESCRIPTION OF THE DATA:**

The data type for all the columns are **STRING.**

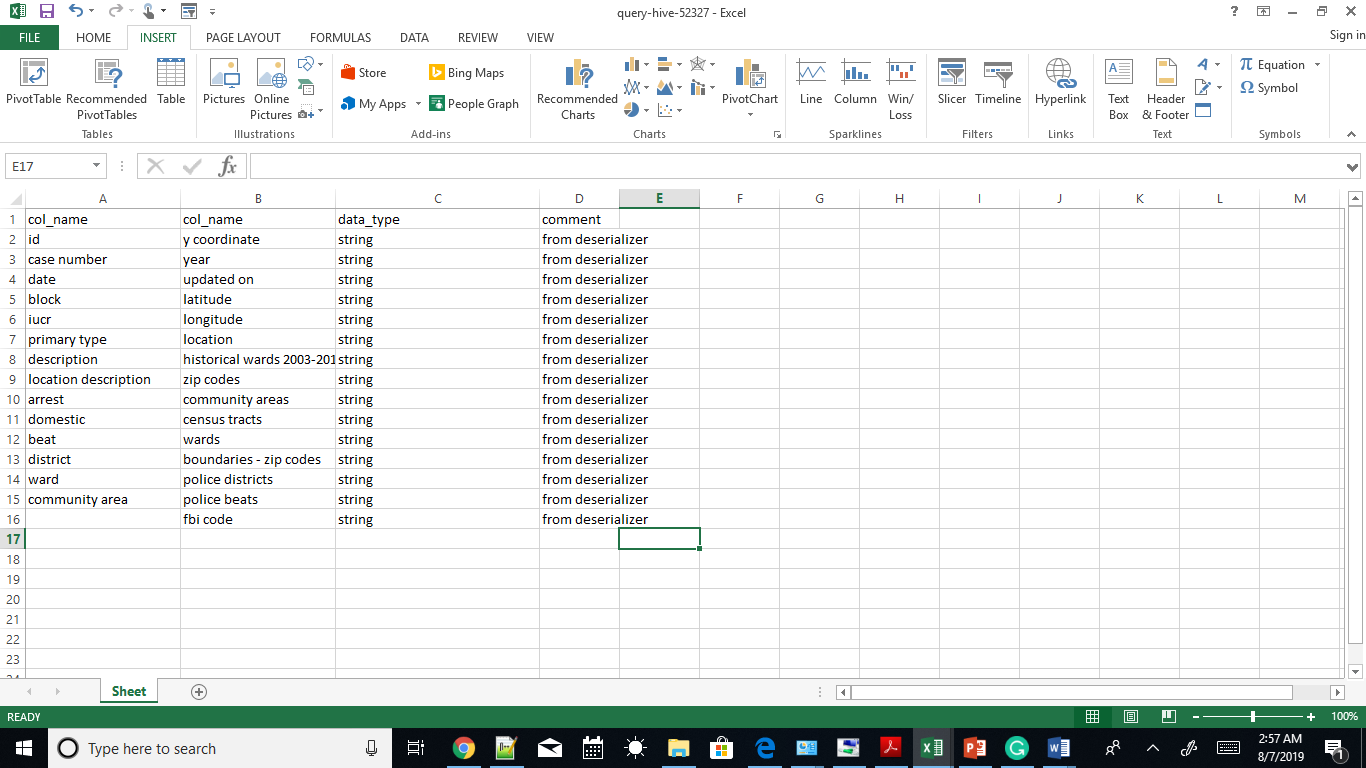
COLUMN NAMES

1. crimes\_2001\_to\_present.id

1. crimes\_2001\_to\_present.case number
2. crimes\_2001\_to\_present.date
3. crimes\_2001\_to\_present.block
4. crimes\_2001\_to\_present.iucr
5. crimes\_2001\_to\_present.primary type
6. crimes\_2001\_to\_present.description
7. crimes\_2001\_to\_present.location description
8. crimes\_2001\_to\_present.arrest
9. crimes\_2001\_to\_present.domestic
10. crimes\_2001\_to\_present.beat
11. crimes\_2001\_to\_present.district
12. crimes\_2001\_to\_present.ward
13. crimes\_2001\_to\_present.community area
14. crimes\_2001\_to\_present.fbi code
15. crimes\_2001\_to\_present.x coordinate
16. crimes\_2001\_to\_present.y coordinate
17. crimes\_2001\_to\_present.year
18. crimes\_2001\_to\_present.updated on
19. crimes\_2001\_to\_present.latitude
20. crimes\_2001\_to\_present.longitude
21. crimes\_2001\_to\_present.location
22. crimes\_2001\_to\_present.historical wards 2003-2015
23. crimes\_2001\_to\_present.zip codes
24. crimes\_2001\_to\_present.community areas
25. crimes\_2001\_to\_present.census tracts
26. crimes\_2001\_to\_present.wards
27. crimes\_2001\_to\_present.boundaries - zip codes
28. crimes\_2001\_to\_present.police districts
29. crimes\_2001\_to\_present.police beats
30. crimes\_2001\_to\_present. Beat

HIVESQL TO KNOW ABOUT ALL THE COLUMNS:

**desc crimes\_2001\_to\_present;**

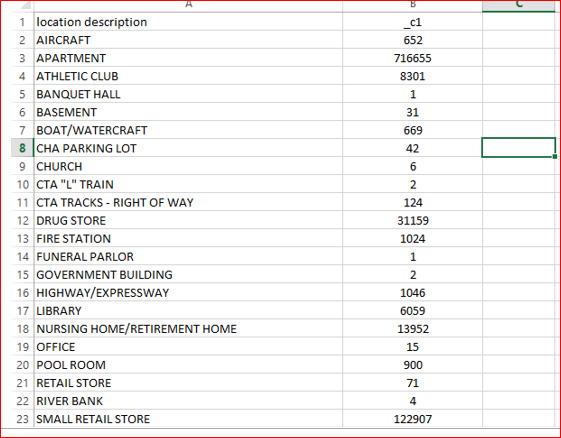


**USING COUNT FUNCTION AND GROUP BY STATEMENT TO RETRIEVE THE DETAILS FOR ‘LOCATION DESCRIPTION’:**

**SELECT `location description`, COUNT(\*)**

**FROM crimes\_2001\_to\_present**

**GROUP BY `location description`;**

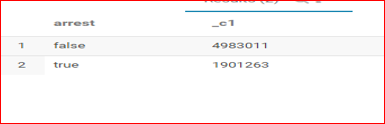


**USING COUNT FUNCTION I DETERMINED TOTAL NUMBER OF ARREST BASED ON TRUE AND FALSE**

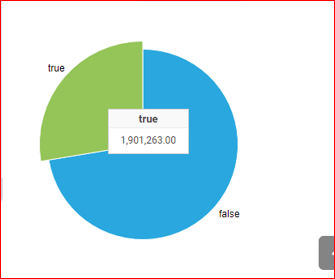
**select arrest, count(True)**

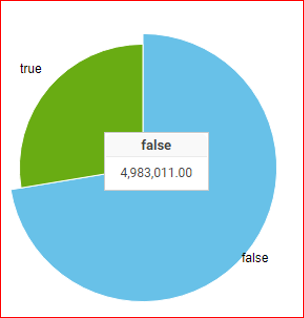
**from crimes\_2001\_to\_present**

**GROUP BY arrest;**



THE FLLOWING CHART SHOWING TOTAL NUMBER OF ARREST BASED ON TRUE AND FALSE





I HAVE SELELCTED ONLY APARTMENT FROM THE LOCATION DESCRIPTION COLUMN BASED ON GAMBLING FROM THE PRIMARY TYPE COLUMN:

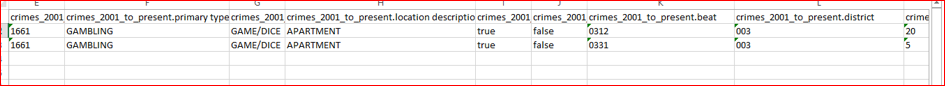
**select id, `primary type`,`location description` from crimes\_2001\_to\_present**

**where `primary type`='GAMBLING'**

**and**

**`location description`='APARTMENT'**

**and year= 2018;**



HERE I HAVE USED COUNT FUNCTION, ORDER BY, AND GROUP BY STATEMENT AS WELL:

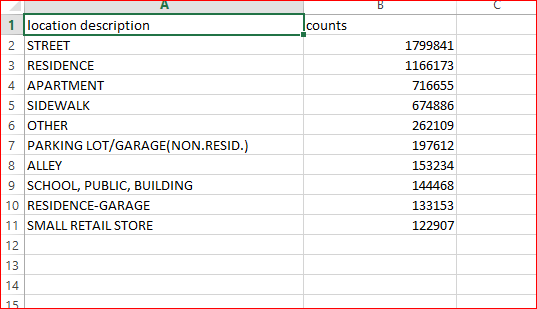
**select `location description`, count(\*) as counts from**

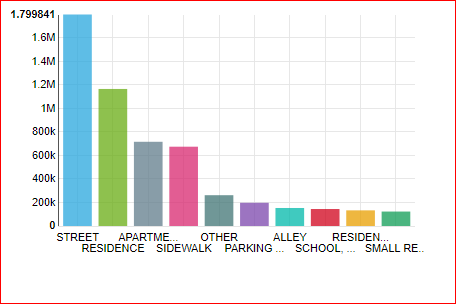
**crimes\_2001\_to\_present group by `location description`**

**order by counts desc**

**limit 10;**

**THE FOLLOWING DATA SHOWS ALL THE DISTINCT LOCATIONS FROM THE LOCATION DESCRIPTION COLUMN AND TOTAL NUMBER OF LOCATIONS AS WELL.**





**HERE, ALL THE ATRIBUTES ARE DECLARED AS A STRING EVEN INTEGER VALUE ALSO. TO USE MATHEMATICAL FUNCTION WE NEED TO CAST FROM STRING INTO INTEGER. USING SUB QUERY I SELECTED THE DATA FOR ONLY AT WARD NUMBER 10 AND THE DISTRICT NUMBER 4. THE ABOVE DETAILS DESCRIBED IN THE CHART.**

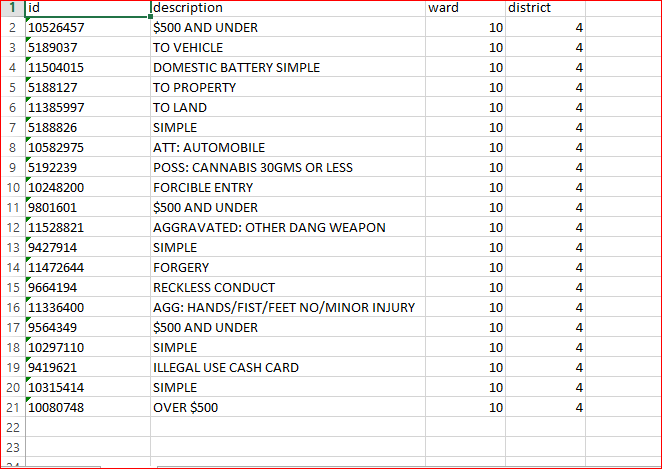
**select id, description, CAST(ward as INT),CAST(district as INT) from crimes\_2001\_to\_present**

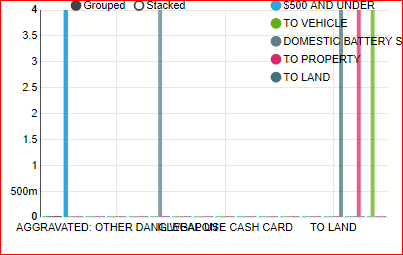
**where ward IN (select ward from crimes\_2001\_to\_present**

**where district > 003**

**group by ward)**

**limit 20;**





**HERE I HAVE USED MANY AND CONDITIONS AND SUBQUERY TO DETERMIND ONLY AUTOMOBILE THEFT HAPPENED IN THE STREET WITH THE CONDITIONS OF ARREST AND DOMESTIC ARE TRUE AT THE YEAR OF 2018.THE FOLLOWING TABLE SHOWS THE DATA AND THE CHAART DESCRIBES IN DETAILD VISUALIZATION.**

**select id, description, iucr, arrest,domestic, year, `location description` from crimes\_2001\_to\_present**

**where year= '2018'**

**and `location description`="STREET"**

**and arrest= "true"**

**and domestic="true"**

**and description**

**IN**

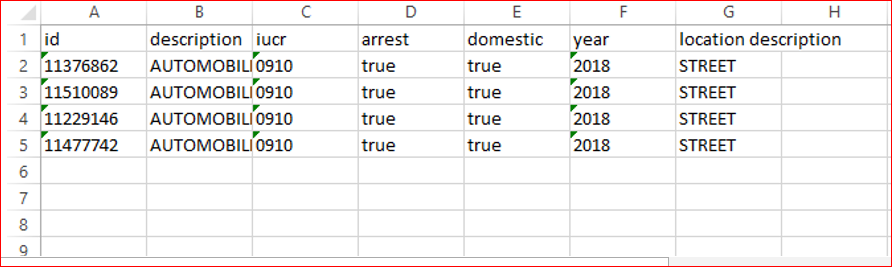
**(select description from crimes\_2001\_to\_present**

**where iucr ='0910'**

**and description ="AUTOMOBILE"**

**group by description**

**);**



**SELECT COUNT(\*) as iucrdet,iucr,**

**CASE WHEN iucr <910 THEN 'not arrest'**

**WHEN iucr >910 THEN 'arrest'**

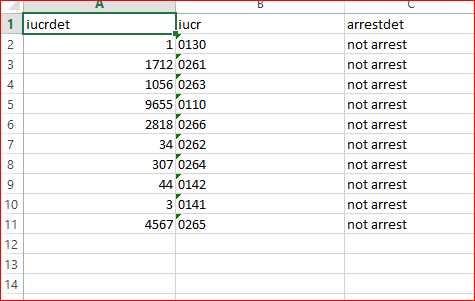
**ELSE 'case'**

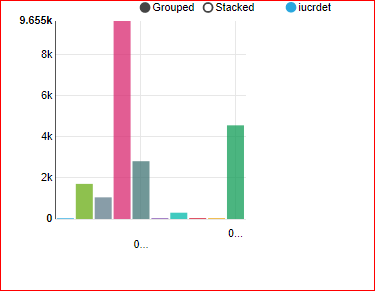
**END as arrestdet**

**FROM crimes\_2001\_to\_present**

**GROUP BY iucr**

**LIMIT 10;**





**USING PIE CHART I HAVE DESCRIBED THE NUMBER OF BEAT. IN 2018 AT THE STREET SOME THEFT HAPPENED ESPECIALY I HAVE DRILLED TO TAKE OUT THE DETAILS OF CHICAGO POLICE DEPARTMENT DATA.**

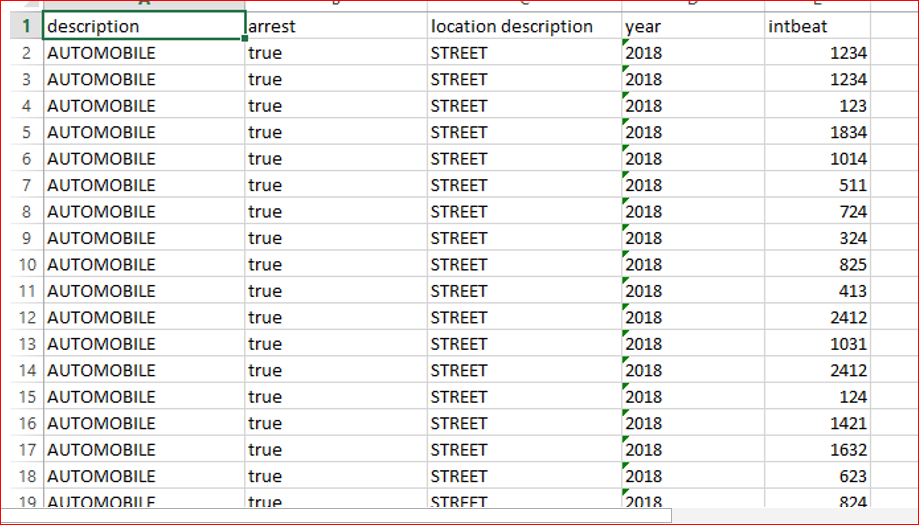
**SELECT description,arrest,`location description`,year ,CAST(beat AS INT) as INTBEAT FROM crimes\_2001\_to\_present**

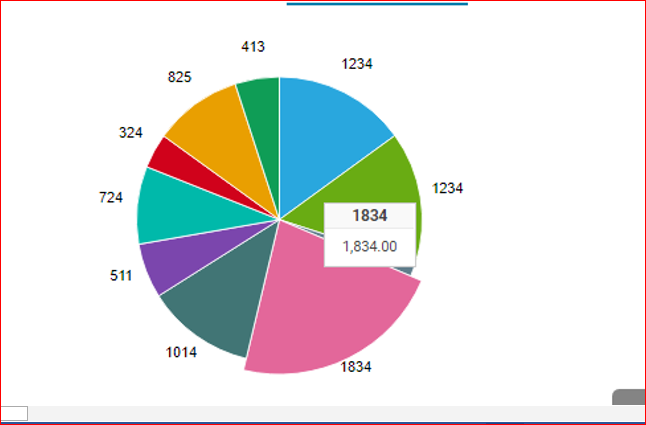
**where description ="AUTOMOBILE"**

**and arrest="true"**

**and `location description`="STREET"**

**and year = '2018';**





**TO SKIP NULL VALUES I USED COUNT FUNCTION TO DETERMIN THE PARTICULAR ID ARRESTED DETAILS. ID NUMBER ‘11704541’ ARRESTED ONLY ONE TIME IN HIS LIFE.**

**select count(block) as COUNTBLOCK,**

**count(arrest) as COUNTARREST**

**from crimes\_2001\_to\_present**

**where id = '11704541';**



**I HAVE DRILLED THE DATA SET TO IDENTIFY THE NUMBER OF DISTRICTS AT THE PARTICULAR WARD USING AVRAGE FUNCTION, GROUP BY, AND ORDER BY STATEMENT.THE FOLLOWING DATA TABLE AND THE DIFFERENT TYPE OF BAR CHARTS DESCRIBE THE DETAILS.**

**SELECT**

**ward,avg\_district**

**FROM**

**(SELECT**

**ward,**

**AVG(district) as avg\_district**

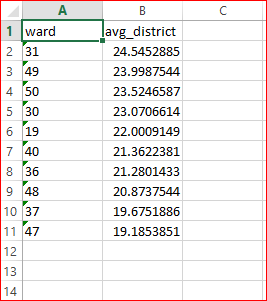
**FROM**

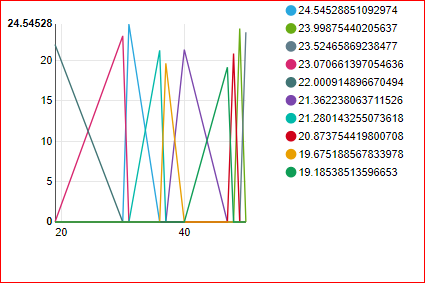
**crimes\_2001\_to\_present**

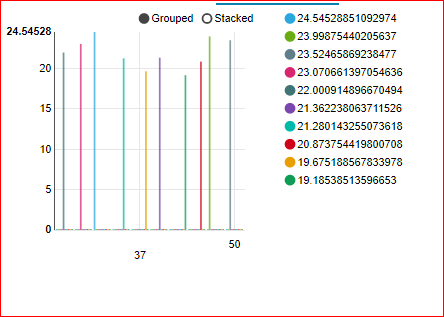
**GROUP BY ward**

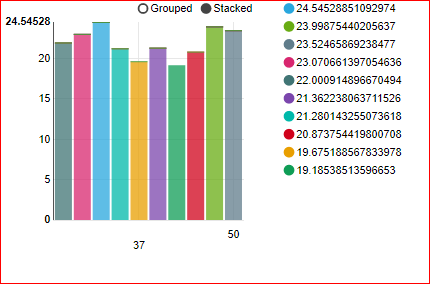
**ORDER BY avg\_district DESC**

**LIMIT 10) as AVGDISRTICTWARD;**







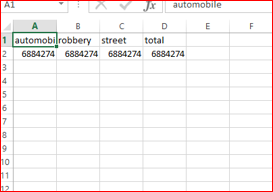


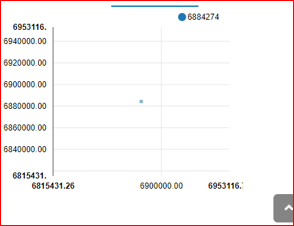
**USING COUNT FUNCTION I HAVE DETERMINED TOTAL NUMBER OF ROBBERY HAPPENED ONLY AT THE STREET. THE KIND OF ROBBERY IS AUTOMOBILE.**

**select count(description='AUTOMOBILE')as AUTOMOBILE, count('primary type'='ROBBERY') as ROBBERY,**

**count('location description'='STREET') as STREET, count(1) as TOTAL**

**from crimes\_2001\_to\_present;**





**THE FOLLOWING TABLE PROVIDES THE DETAILS OF MOTOR VEHICLE THEFT AHPPENED AT THE STREET, AND THE ARREST IS TRUE WITH NUMBER OF BEAT. FOR VISULAIZATION I HAVE USED BAR CHARTS TO PROVIDE MORE DETAILS.**

**SELECT ‘primary type’, `description`,beat ,CAST(beat AS INT) as INTBEAT FROM crimes\_2001\_to\_present**

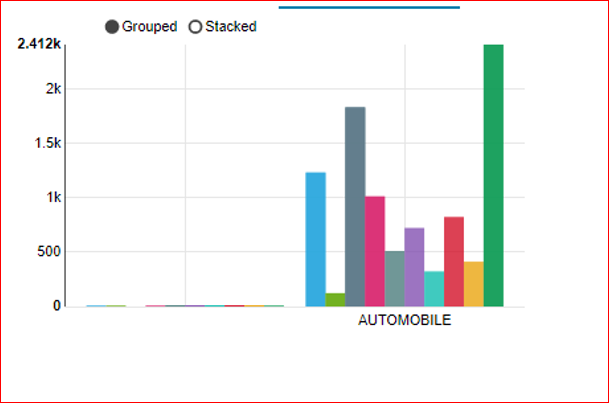
**where description ="AUTOMOBILE"**

**and arrest="true"**

**and ‘primary type’=”MOTOR VEHICLE THEFT”**

**and `location description`="STREET";**





**I HAVE DILLED THE DAT SET TO TAKE OUT THE DETAILS FOR THE PARTICULAR ID, USING ARITHMATIC OPERATOR, SUB QUERY,AND OPERATOR, CAST FUNCTION IDENTIFIED FOR THAT PARTICULE THEIF ATH STEET WHAT HE DID.IS HE ARRESTED OR NOT?**

**select id, description, iucr, arrest,domestic, year, `location description`, cAST(beat as INT) from crimes\_2001\_to\_present**

**where year= '2018'**

**and `location description`="STREET"**

**and arrest= "true"**

**and domestic="true"**

**and beat >='1000'**

**and description**

**IN**

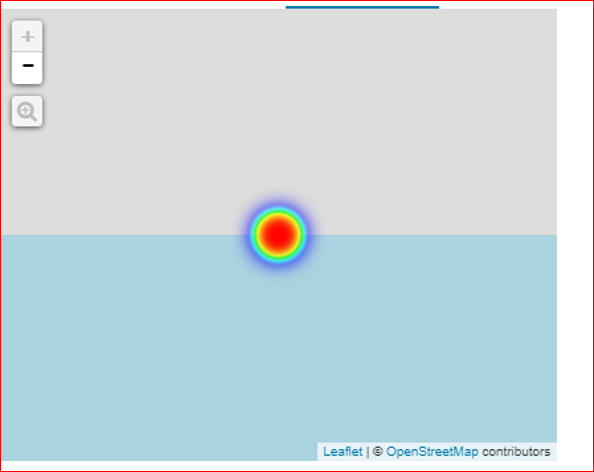
**(select description from crimes\_2001\_to\_present**

**where iucr ='0910'**

**and description ="AUTOMOBILE"**

**group by description**

**);**





**USING SUBQUERY, BETWEEN OPERATOR, AND CAST FUNCTION IDENTIFIED THE DETAILS ID, DESCRIPTION,DISTRICT, WARD NUMBER BETWEEN 2011 TO 2018 AND THE DISTRICT BETWEEN 003 TO 005. THE BAR CHART PROVIDES THE VISUALIZATION OF THE DATA TO UNDERSTAND EASLY.**

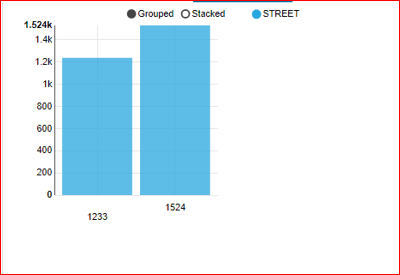
**select id, description, district, CAST(ward as INT) from crimes\_2001\_to\_present**

**where year between 2011 and 2018**

**and ward IN (select ward from crimes\_2001\_to\_present**

**where district between 003 and 005**

**);**



**USING SUBQUERY, BETWEEN OPERATOR, AND CAST FUNCTION IDENTIFIED THE DETAILS ID, DESCRIPTION, DISTRICT, WARD NUMBER BETWEEN 2011 AND 2018.** HERE, **I TOOK OUT THE DATA BETWEEN 2011 AND 2104.**

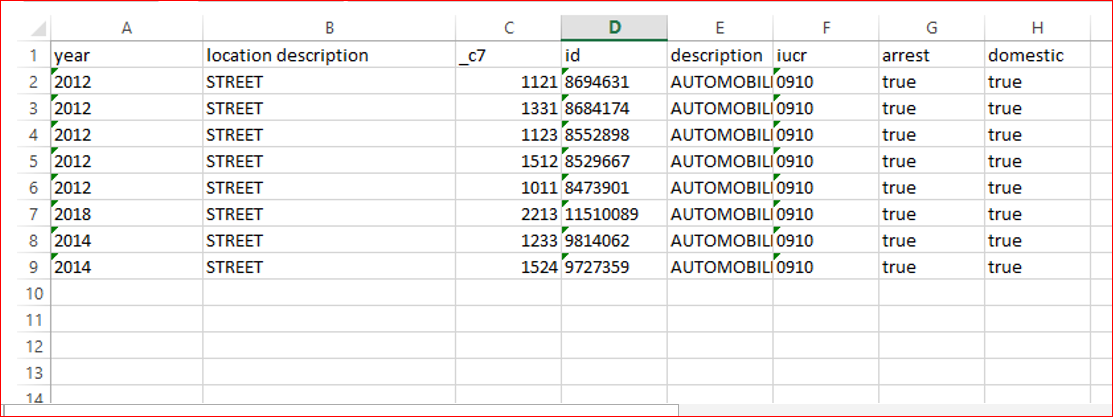
**select id, description, district, CAST(ward as INT) from crimes\_2001\_to\_present**

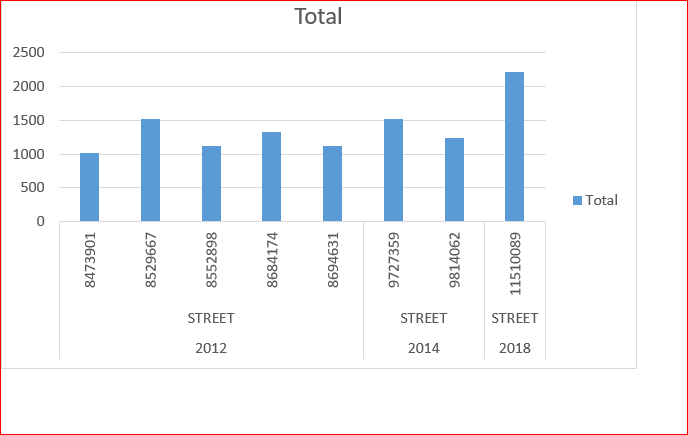
**where year between 2011 and 2014**

**and ward IN (select ward from crimes\_2001\_to\_present**

**where district between 003 and 005**

**);**





**TO VIEW ALL THE DATA IN 2019, THE WARD NUMBER IS 7, COMMUNITY AREA BETWEEN 1 AND 50 AND THE DESCRIPTION IS SIMPLE.**

**SELECT \* FROM crimes\_2001\_to\_present**

**WHERE description = 'SIMPLE'**

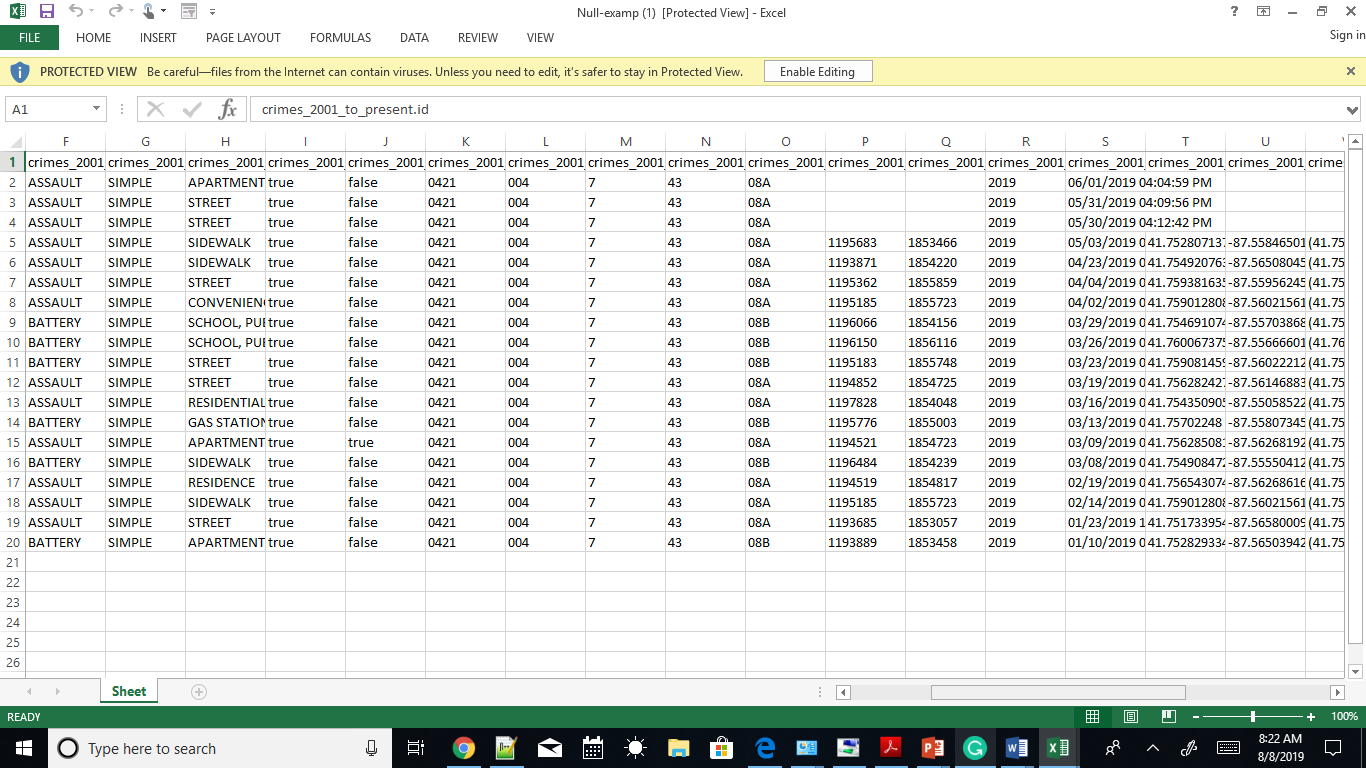
**and year= '2019'**

**and ward='7'**

**and `community area` between 1 and 50**

**and beat= '0421'**

**and arrest='true';**



**WHEN THE BEET EXCEEDS GREATER THAN 1000 IN AUTOMOBILE THEFT AT THE STREET IN 2012 WITH IUCR NUMBER 0910- TO SEE HOW MANY AUTOMOBILE THEFT HAPPENED. THE BAR CHART VISUALIZED IN THE DIFFERENT COLOR TO THE USER.**

**select id, description, iucr, arrest,domestic, year, `location description`, CAST(beat as INT) from crimes\_2001\_to\_present**

**where year= '2012'**

**and `location description`="STREET"**

**and arrest= "true"**

**and domestic="true"**

**and beat >='1000'**

**and description**

**IN**

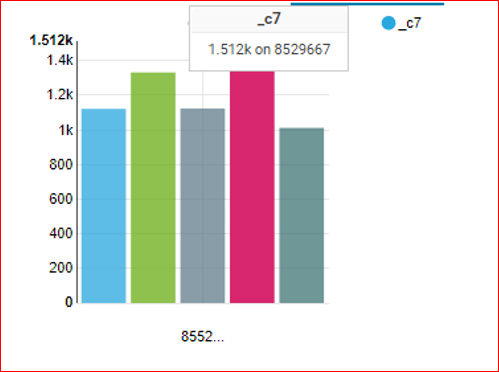
**(select description from crimes\_2001\_to\_present**

**where iucr ='0910'**

**and description ="AUTOMOBILE"**

**group by description**

**);**





**TO SEE IN WARD NUMBER 37 AT DISTRICT NUMBER 011WHAT HAPPENED**

**select id, arrest, description, ward,district**

**from crimes\_2001\_to\_present**

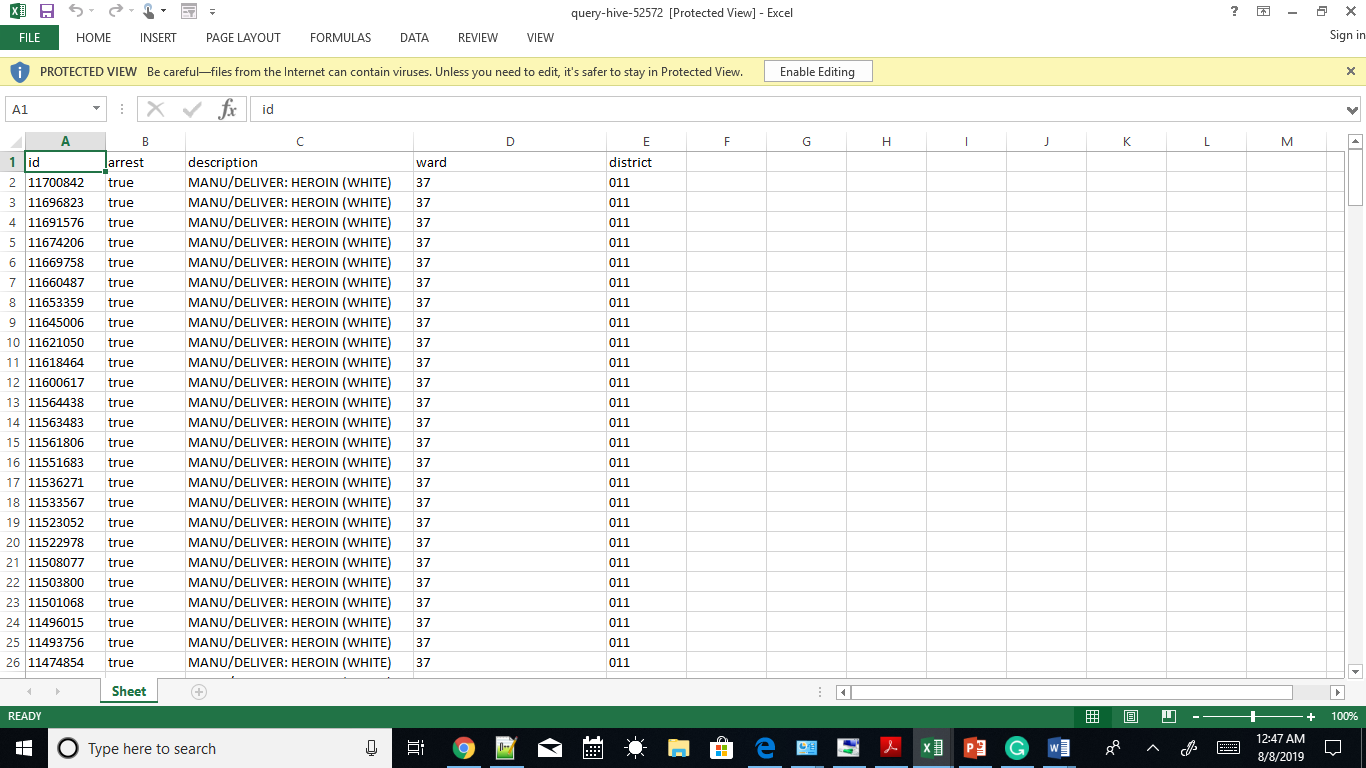
**where iucr='2014'**

**and**

**district='011'**

**and**

**ward = '37';**



**HERE I USED WHERE CLAUSE, COUNT FUNCTION, TWO SUBQUERIES, GROUP BY, AVERAGE FUNCTIONAND ORDER BY STATEMENT WITH LIMIT NUMBER 10. I IDENTIFIED THE AVERAGE OF BEAT WITH PARTICULAR WARD.**

**SELECT COUNT (ward),**

**description**

**FROM crimes\_2001\_to\_present**

**WHERE ward IN**

**(SELECT ward**

**FROM**

**(SELECT ward,**

**AVG(beat) as avg\_beat**

**FROM crimes\_2001\_to\_present**

**GROUP BY ward**

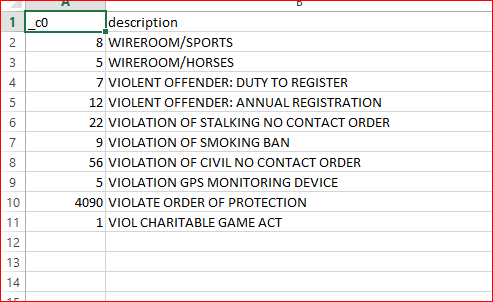
**ORDER BY avg\_beat DESC**

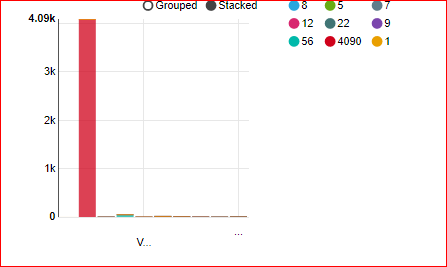
**LIMIT 10) as data\_frmcrime)**

**GROUP BY description**

**ORDER BY description DESC**

**LIMIT 10;**





**TO SELECT AUTOMOBILE FROM THE DESCRIPTION COLUMN WITH CONDITION OF ARREST IS TRUE AND THE DOMESTIC COLUMN IS EQUAL TO TRUE AT THE PARTICULAR YEAR OF 2012. I USED THE FOLLOWING QUERY TO TAKE OUT THE PARTICULAR DATA.**

**select id, description, iucr, arrest, domestic, year, `location description`, CAST(beat as INT) from crimes\_2001\_to\_present**

**where year= '2012'**

**and `location description`="STREET"**

**and arrest= "true"**

**and domestic="true"**

**and beat >='1000'**

**and description**

**IN**

**(select description from crimes\_2001\_to\_present**

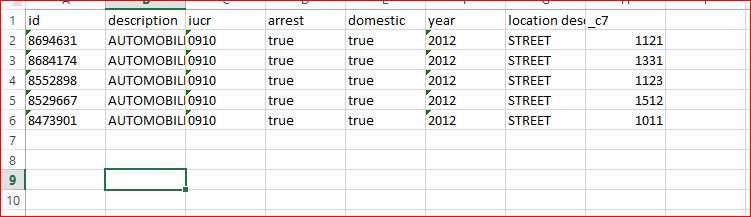
**where iucr ='0910'**

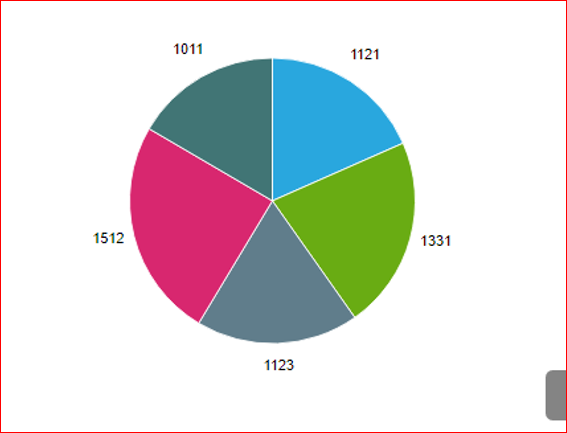
**and description ="AUTOMOBILE"**

**group by description**

**);**

**DATA:**





**THE ABOVE THE CHART SHOWS WHERE BEAT IS GREATER THAN 1000, SO IT SHOWS FROM 1011 TO 1021.**

**CONCLUSION:**

In conclusion, I have used the HIVE environment to analyze Chicago police department data set. Using HIVESQL I have drilled the data set to take out the theft information because the data set contains unorganized data. To get information from the unstructured data can use the SQL language.

I have collected all the data based on the description, particular id details, total number of beat, an average of the ward in the particular district, and based on the years. This Chicago police Department report provides all the information. In my opinion, the HIVE environment tool is very convenient to use the SQL language.