# [C:\Users\jwoo5\AppData\Local\Temp\templateTermTutorial.html](http://www.calstatela.edu/centers/hipic)

# CIS4560-01 Term Project Tutorial

California State University, Los Angeles

#### Authors: Megan Escobar, Alexander Bugarin, David Ha, Alejandro Cano

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 11/17/2018

**Lab Tutorial**

Megan Escobar (mescob33@calstatela.edu)

David Ha (dha3@calstatela.edu)

Alexander Bugarin (abugari3@calstatela.edu)

Alejandro Cano (acano19@calstatela.edu)

11/17/2018

**Visualizing Crime Statistics using Beeline CLI, HCatalog, and Pig**

In this tutorial you will learn how to process data in Beeline CLI using HiveQL as well as filter large datasets in Pig using HCatalog. You will first download the file “Crimes - 2001 to present” from the Chicago Data Portal, then use this file to perform a number of Hive queries and Pig operations to process and analyze the data, and lastly, download the manipulated data to visualize it using Microsoft Excel.

This dataset reflects reported incidents of crime that have occurred in the city of Chicago from 2001 to this present year, minus the most recent seven days. The data is extracted from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system, and it includes several characteristics of the crime committed including its case number, location, primary type of crime, whether an arrest occurred, and more. Some of the queries you will perform will include analyzing the number of crimes committed per type of crime, determining which area of Chicago is most prone to crime, and discerning which crimes are committed in that area.

**Objectives**

In this hands-on lab, you will learn how to:

* Get data manually using HDFS commands
* Perform analysis using HiveQL commands
* Visualize the data in Microsoft Excel

**Platform Spec**

* IBM Bluemix BigInsights
* CPU Speed: 2195
* # of CPU cores: 4
* # of nodes: 5
* Total Memory Size: 150 GB

Step 1: Get Data from Chicago Data Portal

In this step, you will remotely access your Oracle Big Data that you executed in your Oracle Cloud account using SSH.

1. SSH to the Hadoop server using the ip address given by the instructor.

$ ssh mescob33@129.150.205.28

1. Use the following command to see a list of your directories:

-bash-4.1$ hdfs dfs –ls

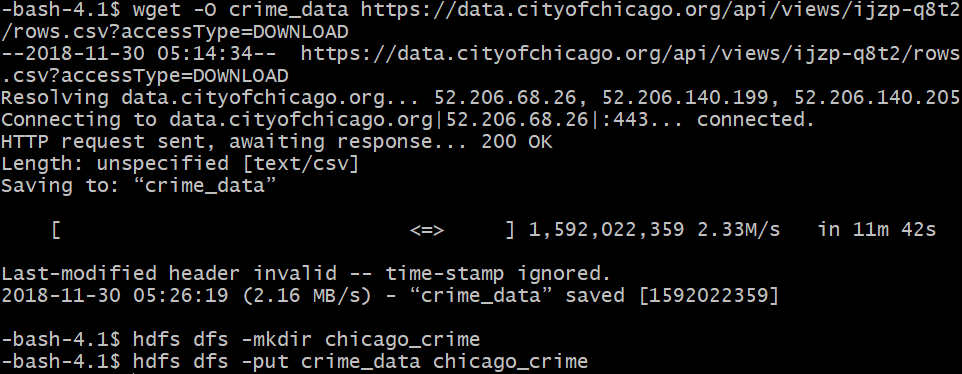
1. Download the csv file “Crimes - 2001 to present.”

-bash-4.1$ wget -O crime\_data https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD

1. Now make a directory to contain the data and put the downloaded file “crime\_data” into it.

-bash-4.1$ hdfs dfs -mkdir chicago\_crime

-bash-4.1$ hdfs dfs -put crime\_data chicago\_crime

1. Remove the file crime\_data from the local server to maintain storage.

-bash-4.1$ rm crime\_data

1. Use the following command to check that “crime\_data” is in the “chicago\_crimes” directory:

-bash-4.1$ hdfs dfs -ls chicago\_crime

1. Run the following HDFS command to make your beeline command works:

-bash-4.1$ hdfs dfs -chmod -R o+w .

Step 2: Beeline CLI

You now need to open another terminal to connect to beeline using the connect URL provided by the instructor.

1. Commands to connect to beeline:

$ ssh mescob33[@129.150.205.28](mailto:mescob33@129.150.205.28)

-bash-4.1$ beeline

beeline>!connect url

1. Be sure to use your database to execute the following commands

use mescob33;

1. Create an external table to contain the “crime\_data” file in a readable format.

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> CREATE EXTERNAL TABLE IF NOT EXISTS crimes(ID STRING, Case\_Number STRING, Case\_Date STRING, Block STRING, IUCR STRING, Primary\_Type STRING, Description STRING, Location\_Description STRING, Arrest BOOLEAN, Domestic BOOLEAN, Beat STRING, District STRING, Ward STRING, Community\_Area STRING, FBI\_Code STRING, X\_Coordinate STRING, Y\_Coordinate STRING, Year STRING, Updated\_On STRING, Latitude STRING, Longitude STRING, Location STRING)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/mescob33/chicago\_crime' TBLPROPERTIES

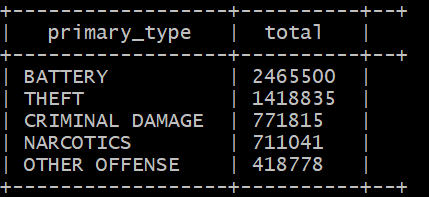
('skip.header.line.count'='2');

1. Verify that the table was created using the SHOW command.

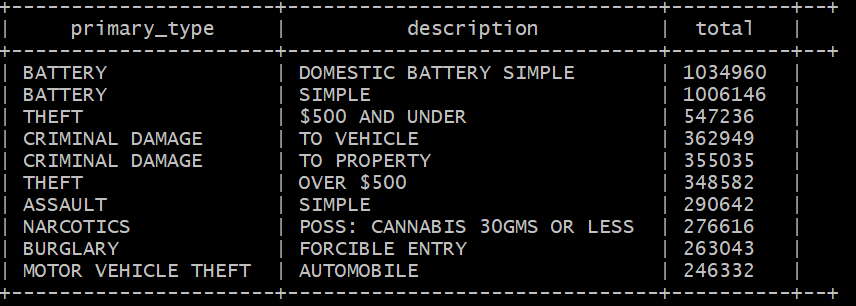
0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> show tables;

We can analyze the data in several ways such as on which blocks the most crimes occur, which crimes are the most committed, how often arrests occur, etc. To do so, we can create queries in Hive to return specified records.

1. To determine which type of crime is committed the most we can use the following query:

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT primary\_type, COUNT(primary\_type) as total from crimes GROUP BY primary\_type ORDER BY total DESC limit 5;

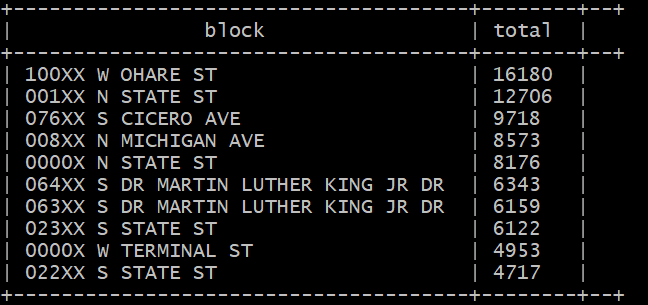
1. Each primary type has an accompanying description which describes in detail the level of the crimes such as if the theft is under $500, aggravated, if a weapon is involved, etc.

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT primary\_type, description, COUNT(description) as total from crimes GROUP BY primary\_type, description ORDER BY total DESC limit 10;

As you can see the total number of the type of crime is comprised of different degrees to the crime such as theft over $500 dollars and theft under $500.

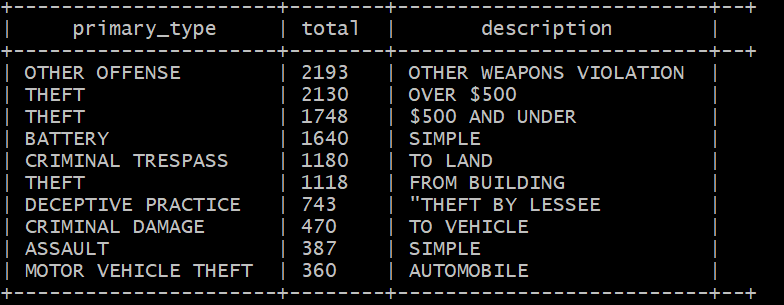
We can also analyze which areas are the most prone to crime. The dataset has limited the location of crimes committed to block streets for the sake of privacy of the victims but we can still query which blocks have the highest rates of crime.

1. To query which blocks have the highest rates of crime, use the following query:

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT block, COUNT(block) as total from crimes GROUP BY block ORDER BY total DESC limit 10;

This gives us a rough estimate on which street has the most crime. As you can see, “100XX W OHARE ST” has the highest frequency of crime occurrence.

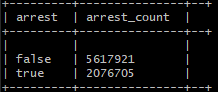
1. Analyzing further, we can see what crimes are being committed in this area.

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> SELECT primary\_type, COUNT (primary\_type) AS total, description from crimes WHERE block == '100XX W OHARE ST' GROUP BY primary\_type, description ORDER BY total DESC limit 10;

From this query, we can determine that the most common crime on 100xx W OHARE ST is weapons violation followed by theft.

9. We can also find the arrest rate ratio;

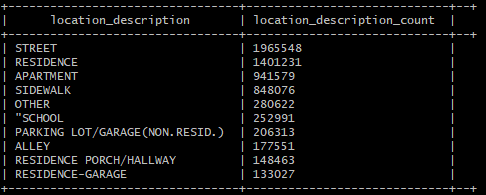
0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> select arrest, count(arrest) as (arrest\_count) from crimes group by arrest;



From this query, we can see that about 30% of all incidents involve the criminal being arrested.

10. We can also find the location where crime tends to happen the most at.

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> select location\_description, count(location\_description) as Location\_Description\_Count from crimes group by location\_description order by location\_description\_count desc limit 10;



As shown, the top three location where crime occur is at “Street”, “Residence”, and “Apartment”.

Step 3: Visualization

We can visualize the data as well as process it further in Microsoft Excel. Excel enables us to create detailed maps of the data which can make it easier for us to compare differences in the severity of the crimes committed, the frequency at which they are committed, and even allows us to process the data in ways which we are limited in the Hive environment with HiveQL. For the following example, we will analyze and later visualize the different categories of the primary type battery.

1. To focus on the primary type “battery”, we will first need to create a table that is filtered to only contain that primary type.

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> CREATE TABLE IF NOT EXISTS battery\_occurence ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' AS select \* from crimes where primary\_type = 'BATTERY';

All occurrences of the primary\_type should now be BATTERY.

1. Now we want to create a new column to rank the types of battery based on its description. The descriptions can range from simple, to domestic battery simple, to several variations of aggravated battery using different weapons. We want to further categorize these as mild if the description is simple, moderate if it is domestic battery simple, and severe if it is anything else.

0: jdbc:hive2://cis4560-bdcsce-4.compute-6082> DROP TABLE IF EXISTS severity;

--create the severity table by selecting from the battery\_occurence table

CREATE TABLE severity

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

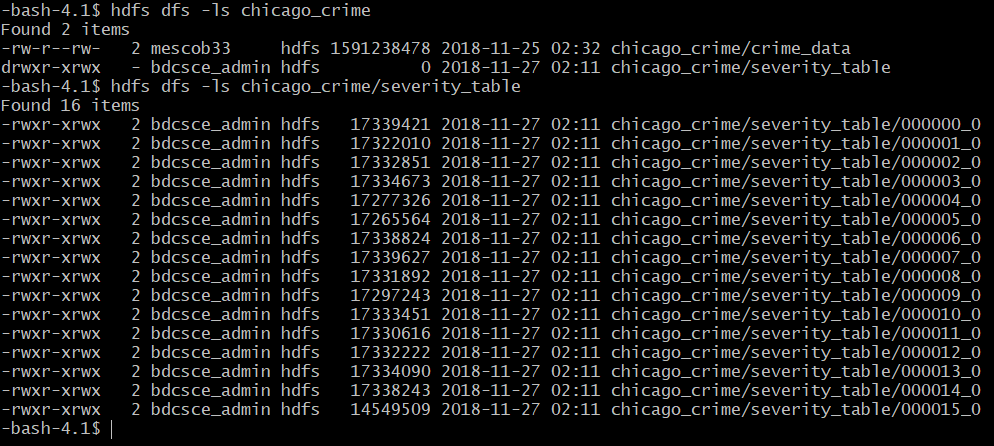
STORED AS TEXTFILE LOCATION '/user/mescob33/chicago\_crime/severity\_table'

AS

SELECT \*, IF((description) == 'SIMPLE', 'Mild', IF((description) == 'DOMESTIC BATTERY SIMPLE', 'Moderate', 'Severe')) AS crime\_severity

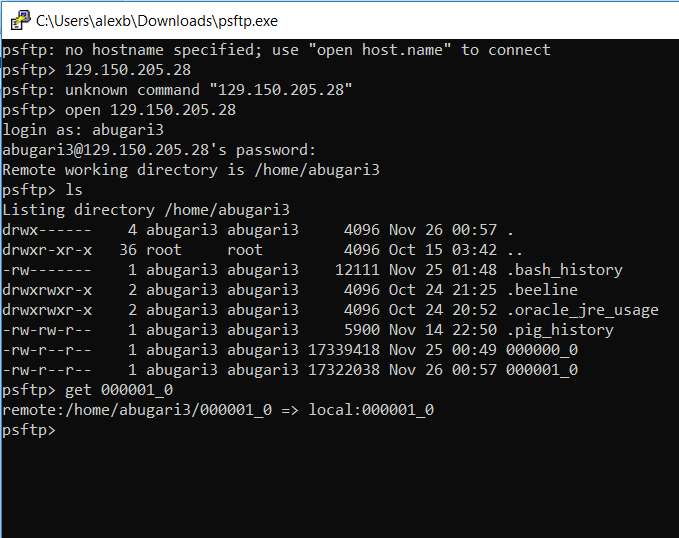
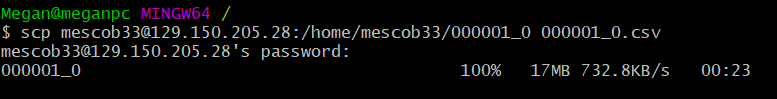
FROM battery\_occurence;

1. Open your HDFS shell and access your chicago\_crime directory.

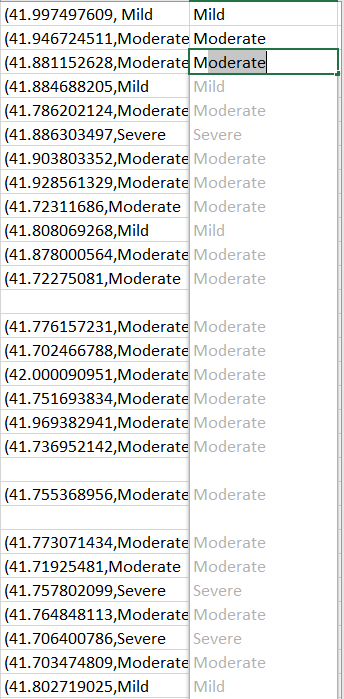
-bash-4.1$ hdfs dfs -ls chicago\_crime/severity\_table

1. Download one file as a sample file to visualize using –get.

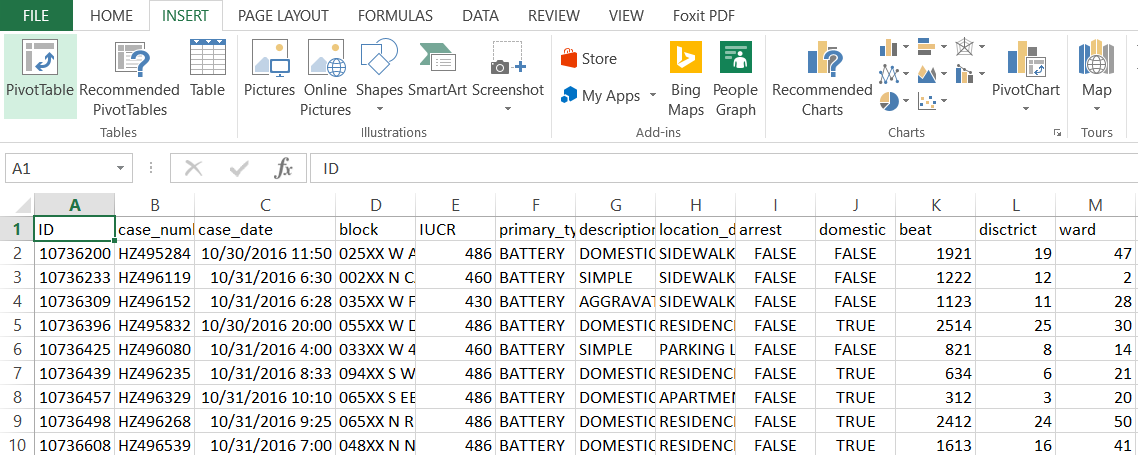
-bash-4.1$ hdfs dfs -get /user/mescob33/chicago\_crime/severity\_table/000001\_0

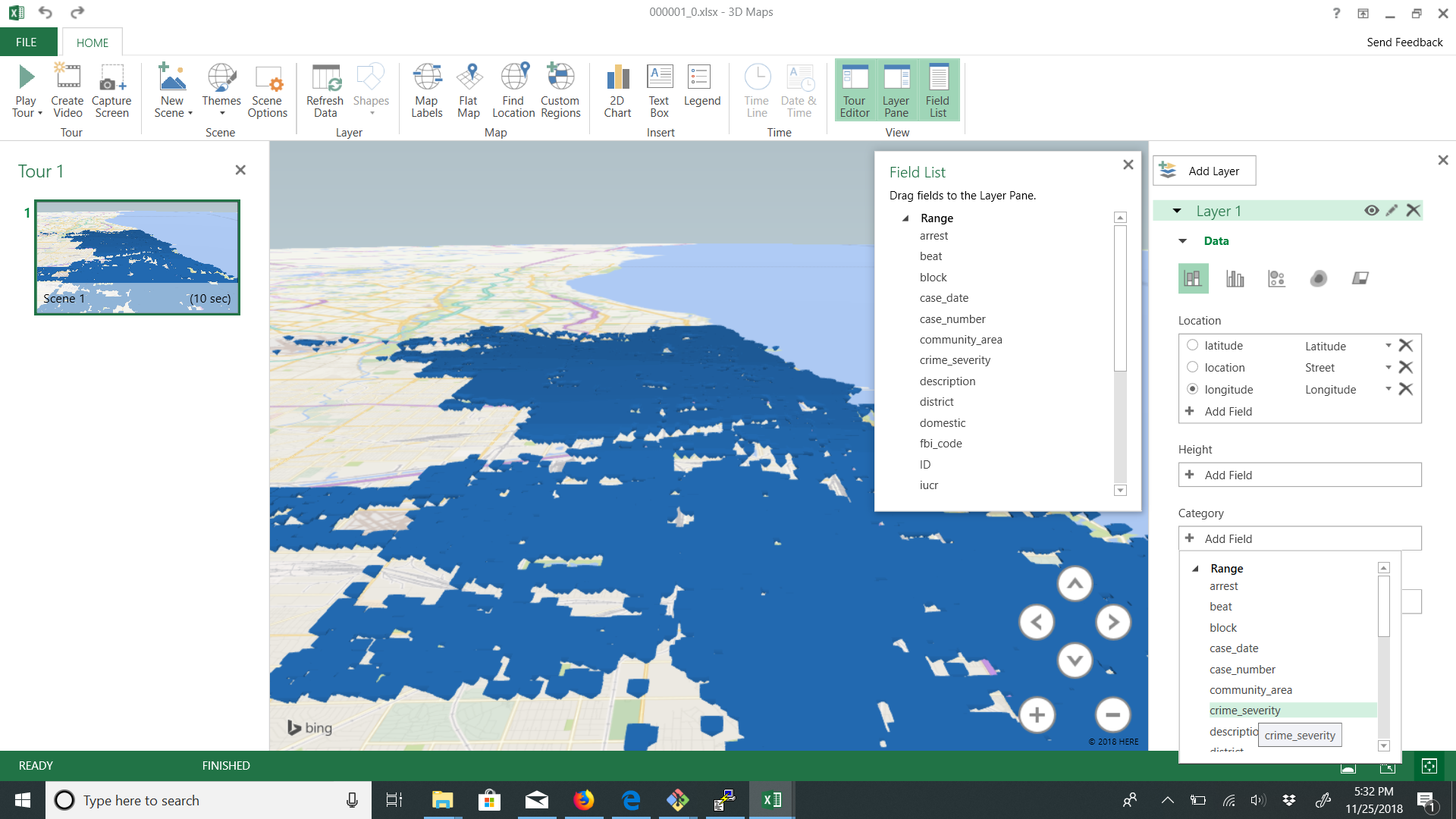
1. You may use pfstp to download this file.
2. Or you may open another terminal and use scp to download the file.

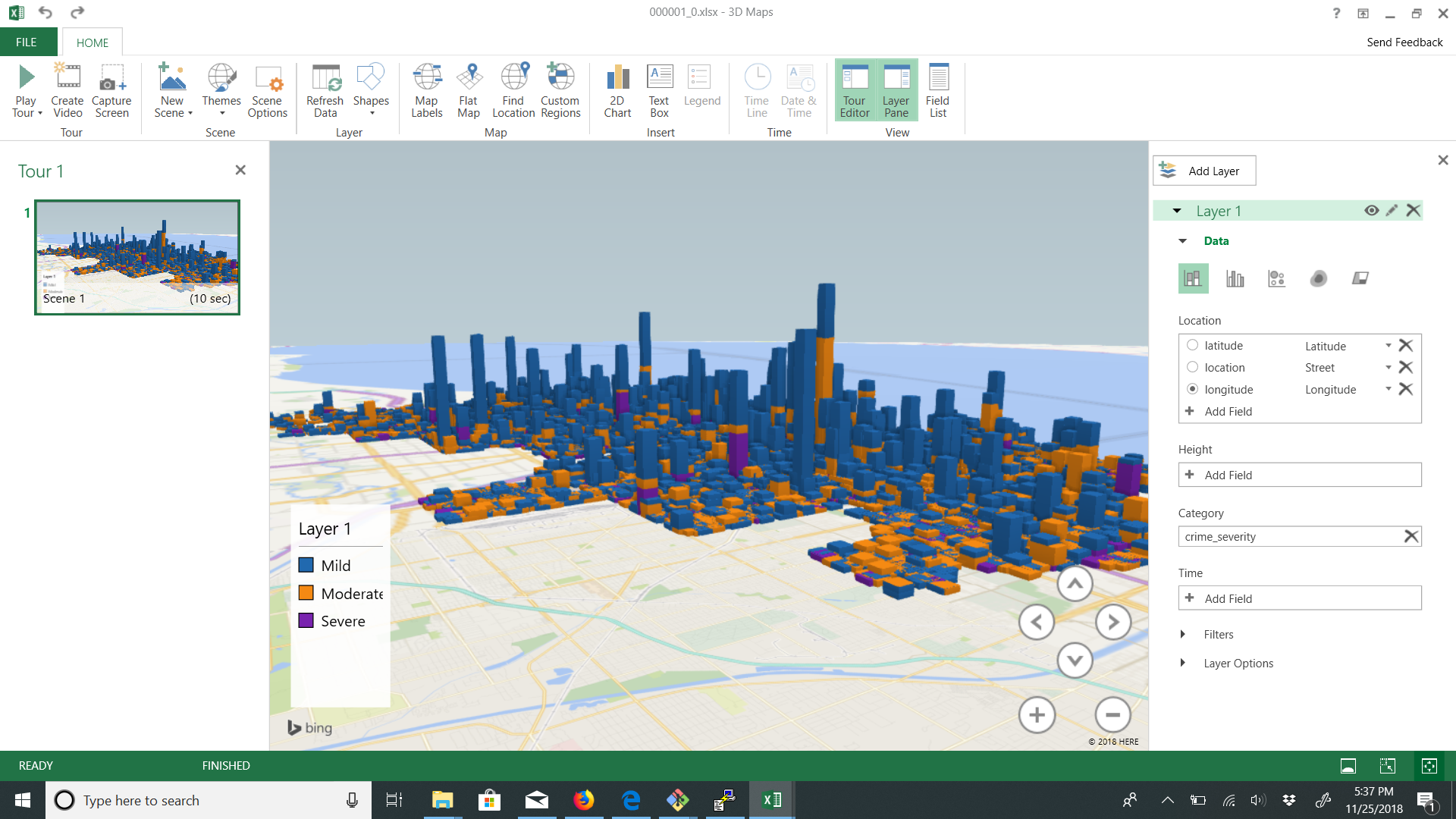
Note: For the Windows user, in order to use psftp you need to download the file at: <https://the.earth.li/~sgtatham/putty/latest/w64/psftp.exe>.

1. Open the file in Excel and choose delimited by comma.
2. If the final column is not separate add two or three cells and Excel will auto generate the rest.
3. Add the columns names.

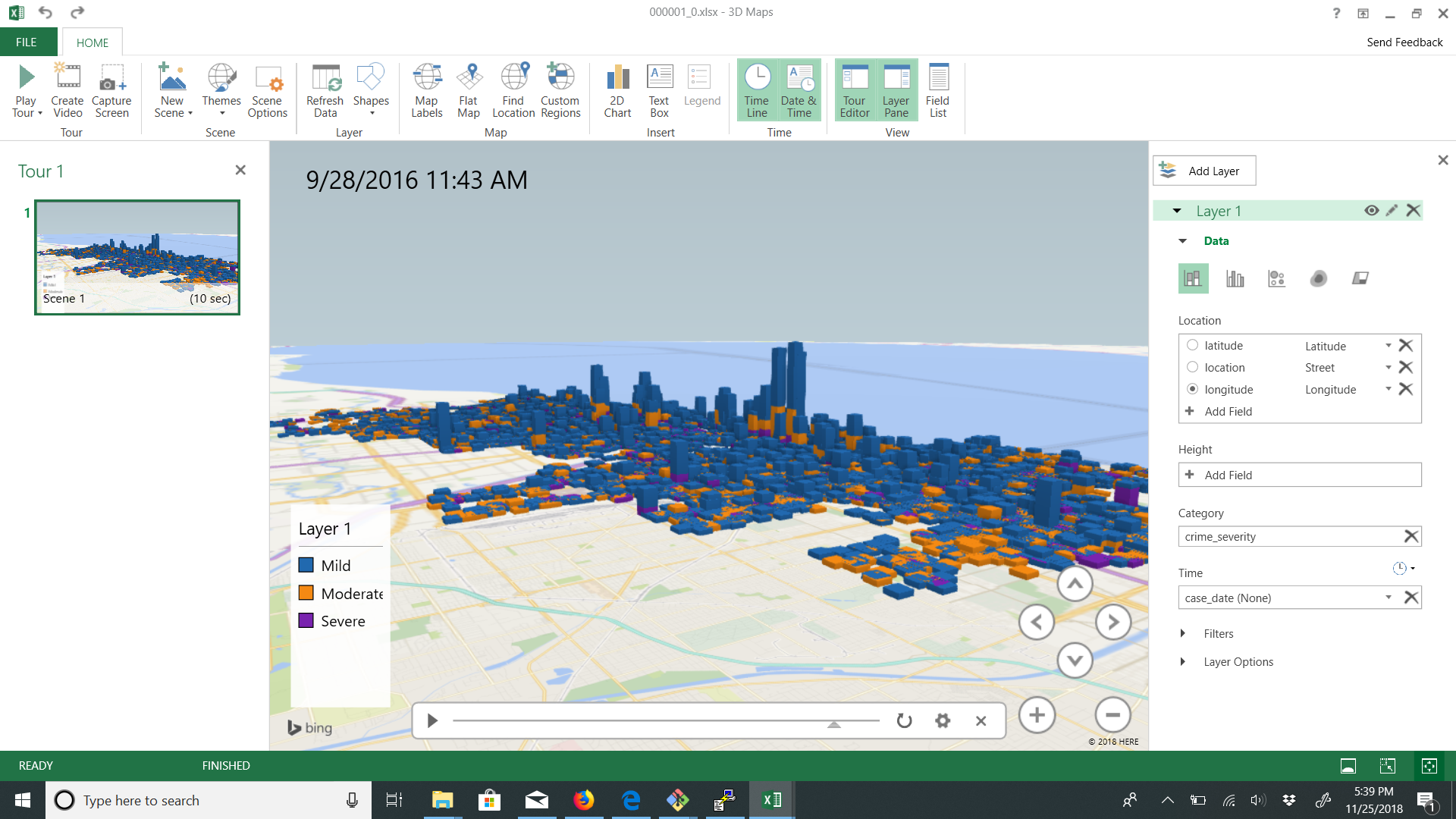
ID, case\_number, case\_date, block, iucr, primary\_type, description, location\_description, arrest, domestic, beat, district, ward, community\_area , fbi\_code, x\_coordinate, y\_coordinate, year, updated\_on, latitude, longitude, location, crime\_severity

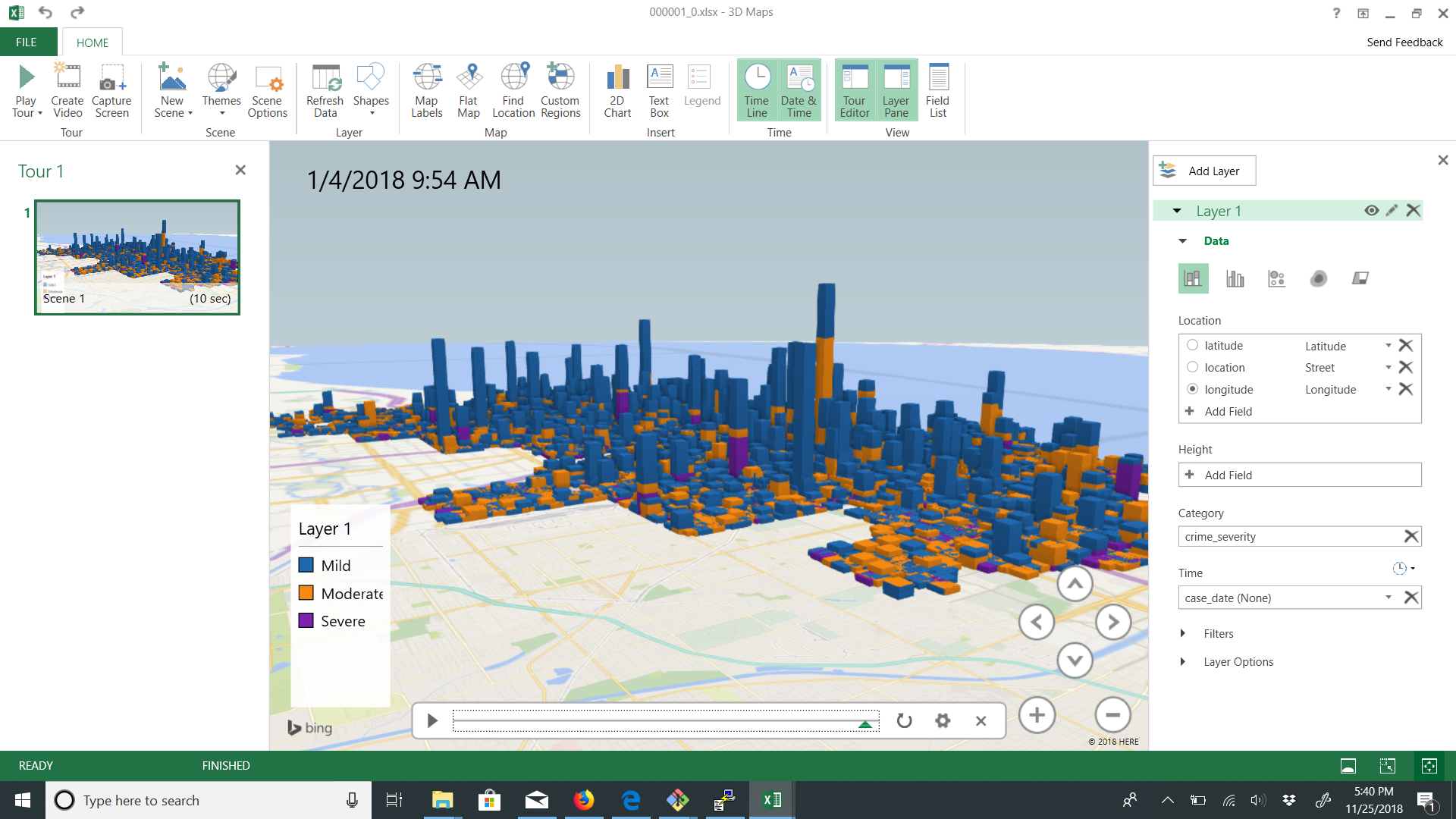
1. Save the file as an excel workbook and add a 3d chart.
2. Select either latitude or longitude for location then for category select crime\_severity.





1. We can then also add a time dimension with the case\_date category. Here we case see a massive spike in activity in between the years 2016 and 2018.





Step 4: Data Processing

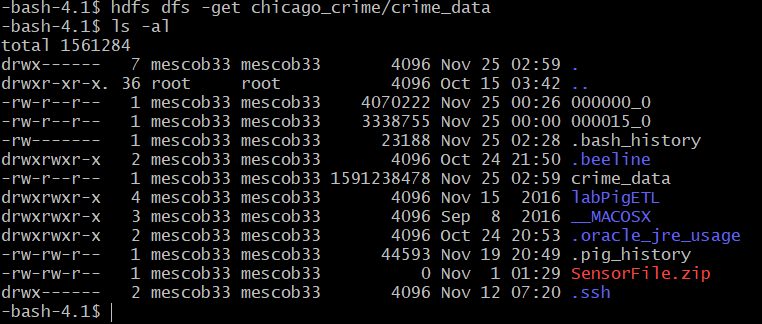
Using Excel, we can also process and visualize how often crime happens with respect to the date, month, and year. For example, how many crime incidents are reported within a month or year, if crime has been going down per year, or which month consistently has the most crime per year.

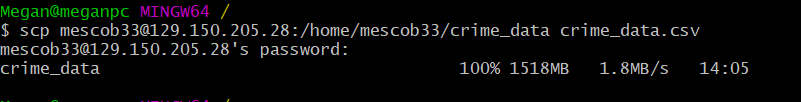
**Step 4.1:** Number of Crime Occurrences per Year and Month

1. Download the entire crime\_data.csv file to the local file systems using the following commands:

-bash-4.1$ hdfs dfs -get chicago\_crime/crime\_data

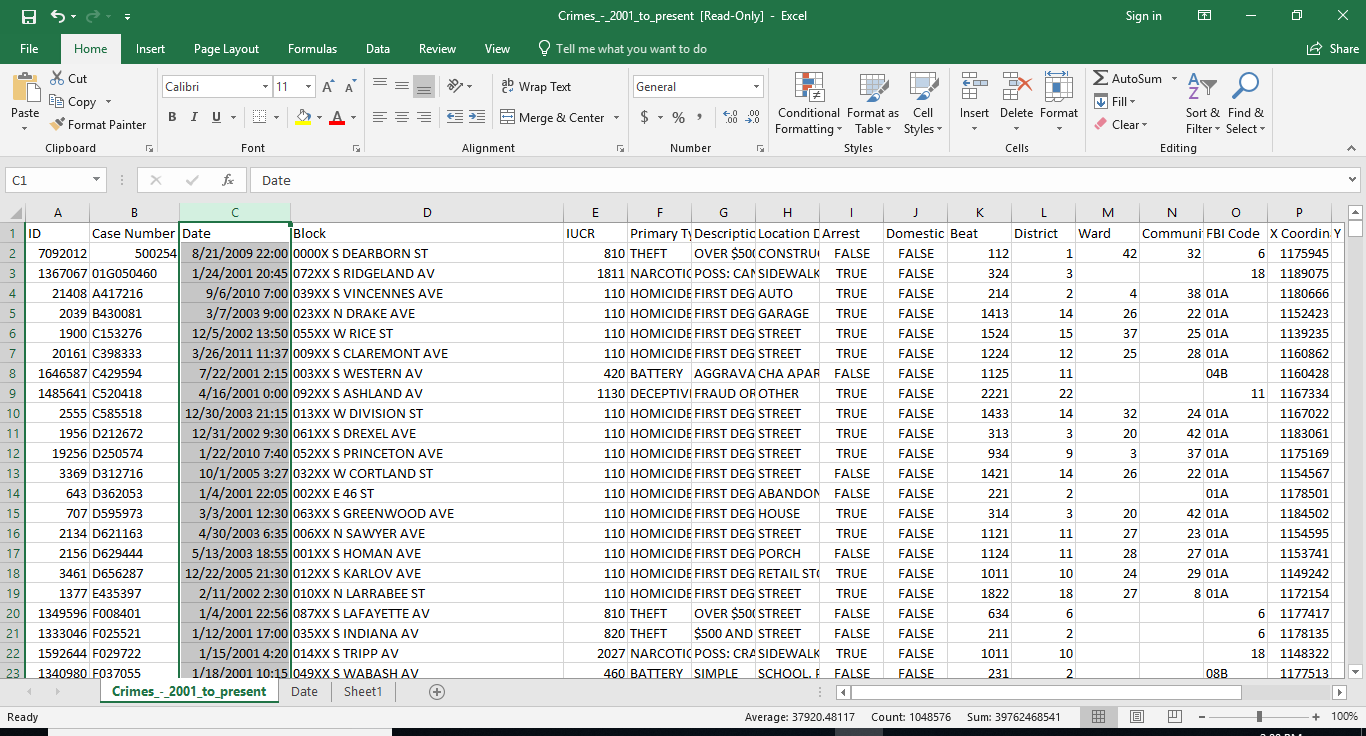
1. Then view the files in your local server to verify that you downloaded the file correctly.

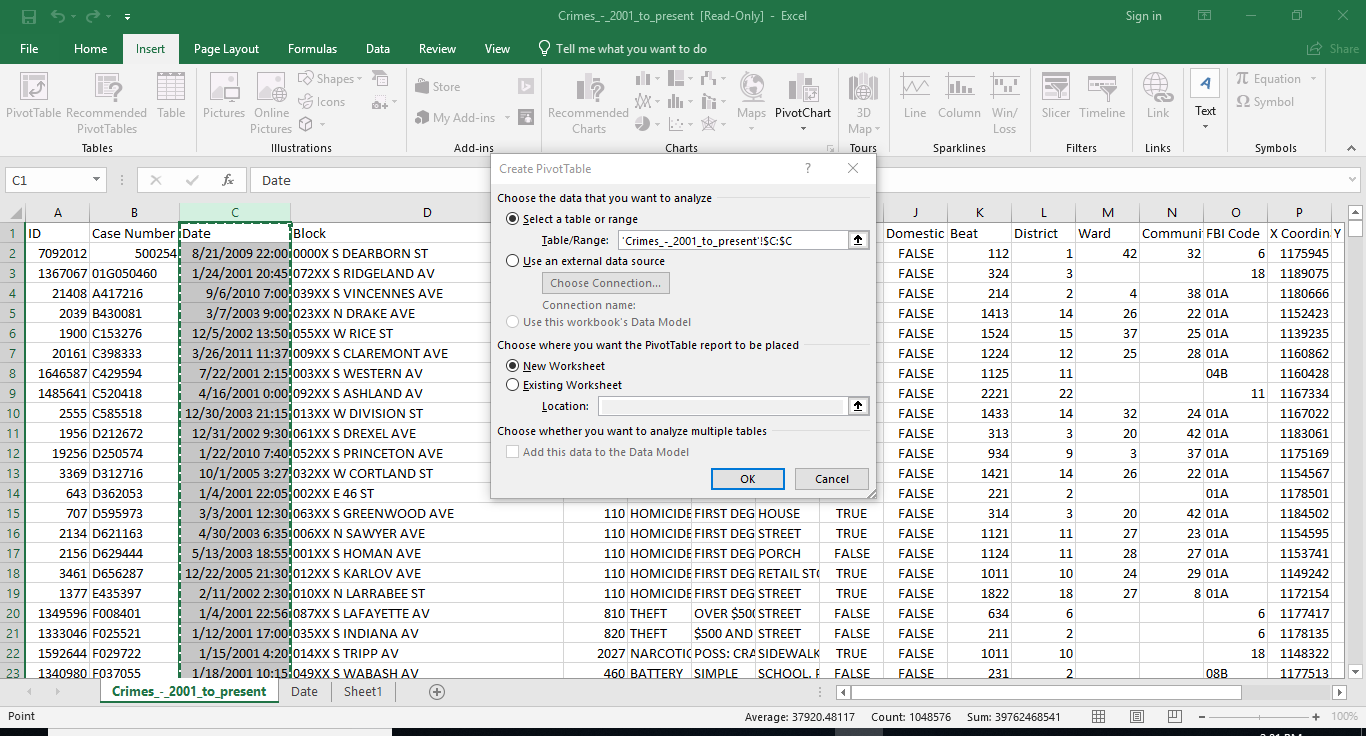
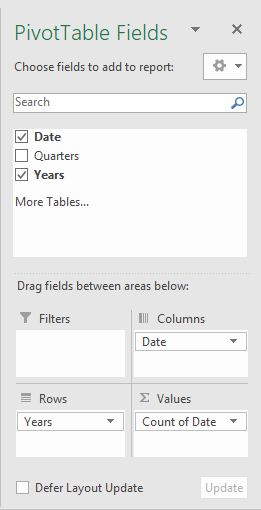
-bash-4.1$ ls -al

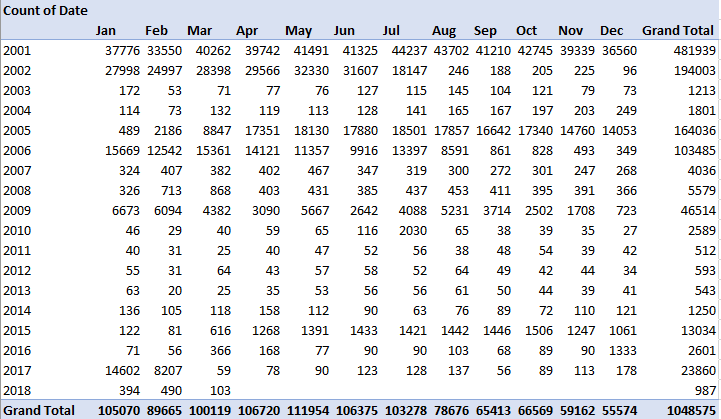
1. Open another terminal with git bash or putty to import the crime\_data file using pfstp or scp as demonstrated in Step 3: Visualization.
2. After downloading the csv file, you may delete the file from your local server once again to conserve space.

-bash-4.1$ rm crime\_data

1. In order to visualize the data, open Excel and your file.
2. Highlight **Date** column.



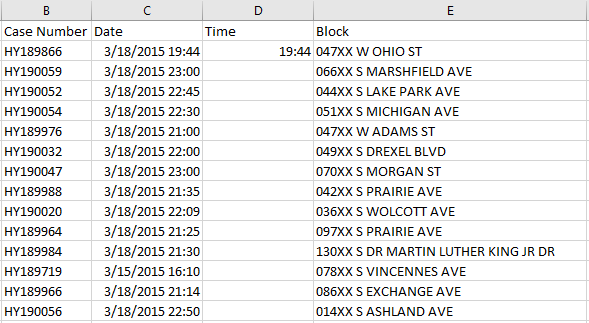
1. Go to Insert > Pivot Table > OK
2. Drag Date field in Columns and Values area. Drag year field in Rows area.
3. Occurrences are then displayed.



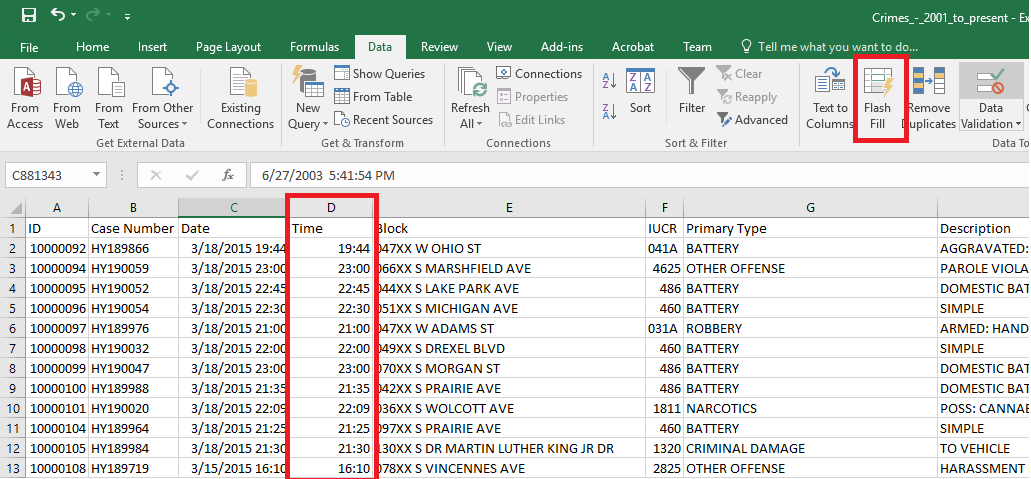
**Step 4.2:** Prevalence of Crime with Respect to Time of Day

Similarly, we can also analyze what time crime is most prevalent.

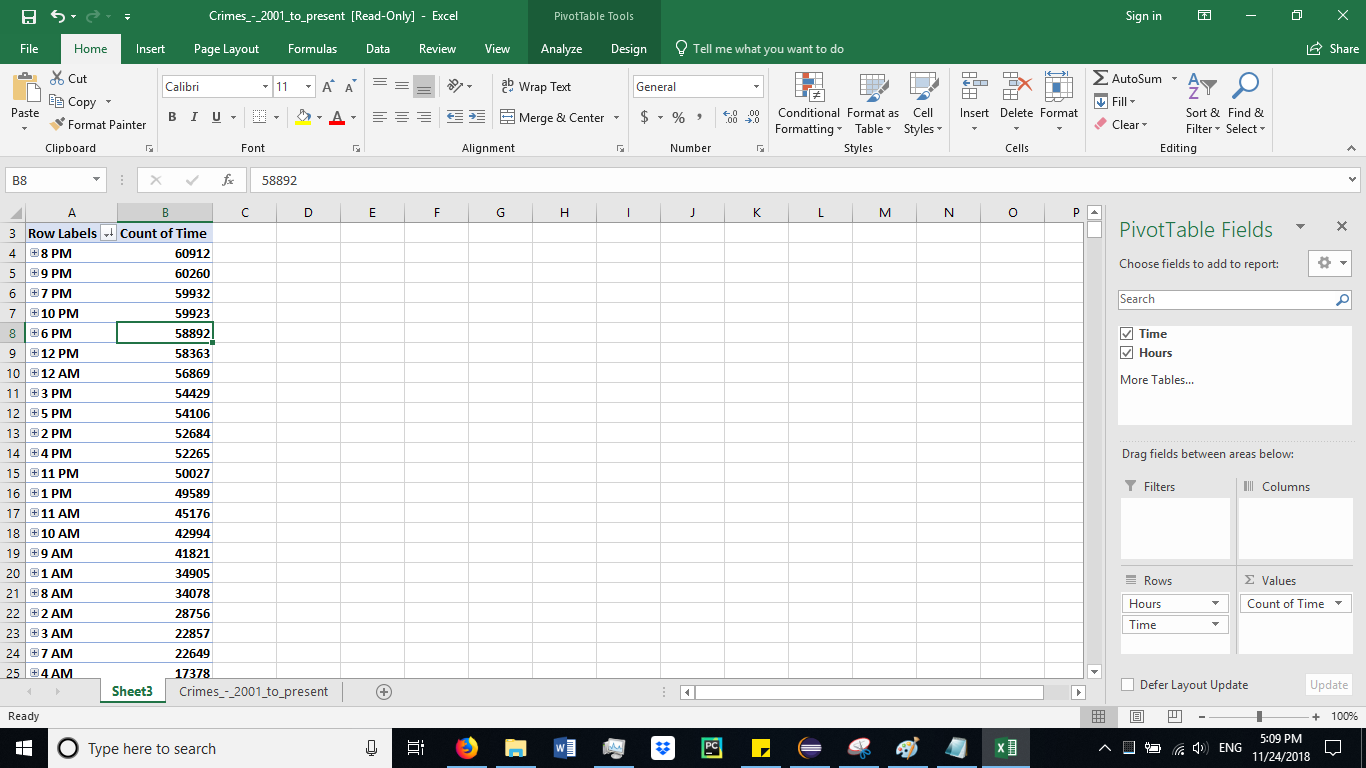
1. Insert a new column right of Date and name it Time. In the first column of Time, put the time from the Date column that is just left of it. In this case, it will be “19:44”



2. Then, go to Data > Flash Fill, while focus is still on, in this case, “19:44”. As you can see, the Time column gets automatically filled.

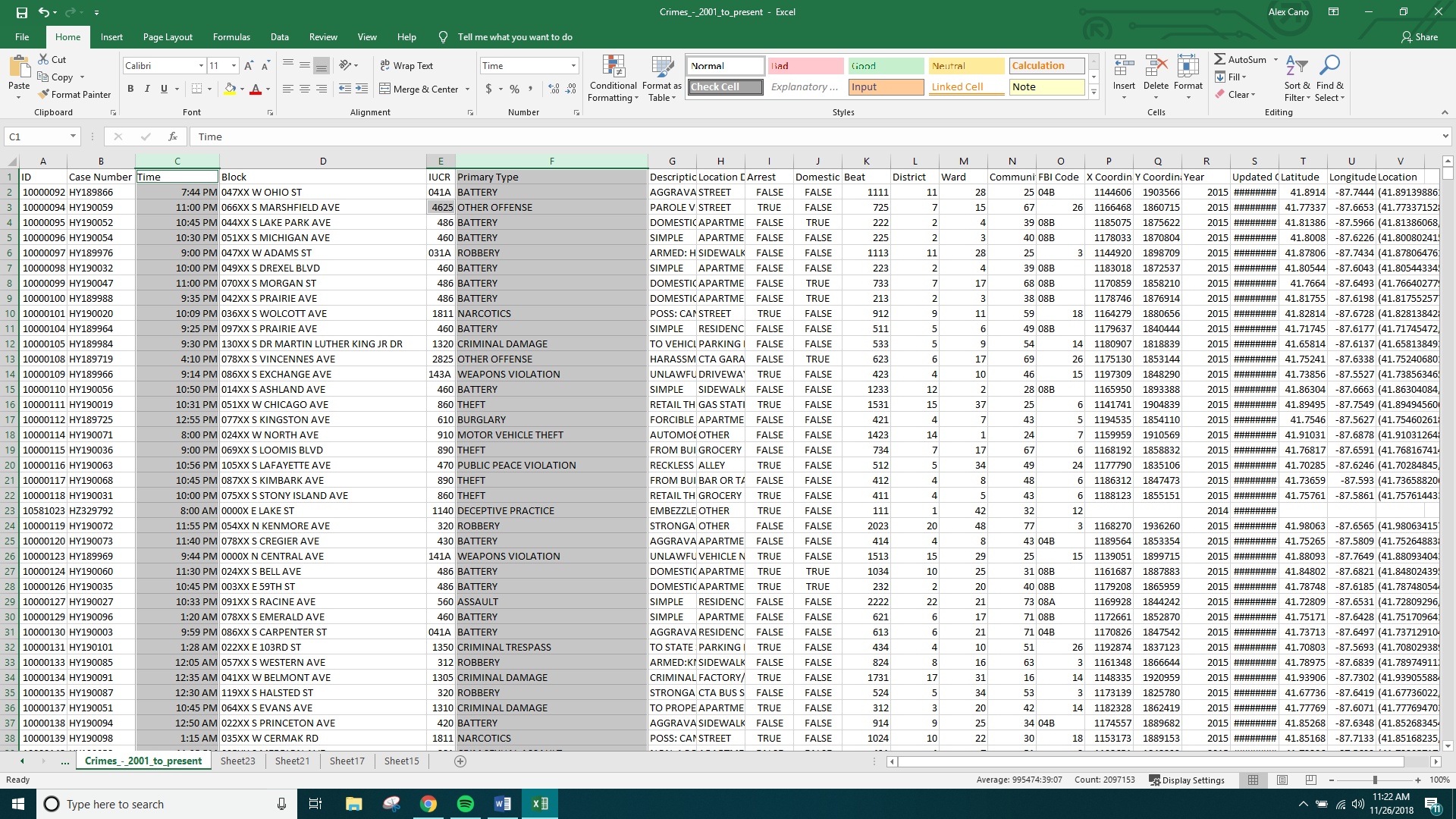
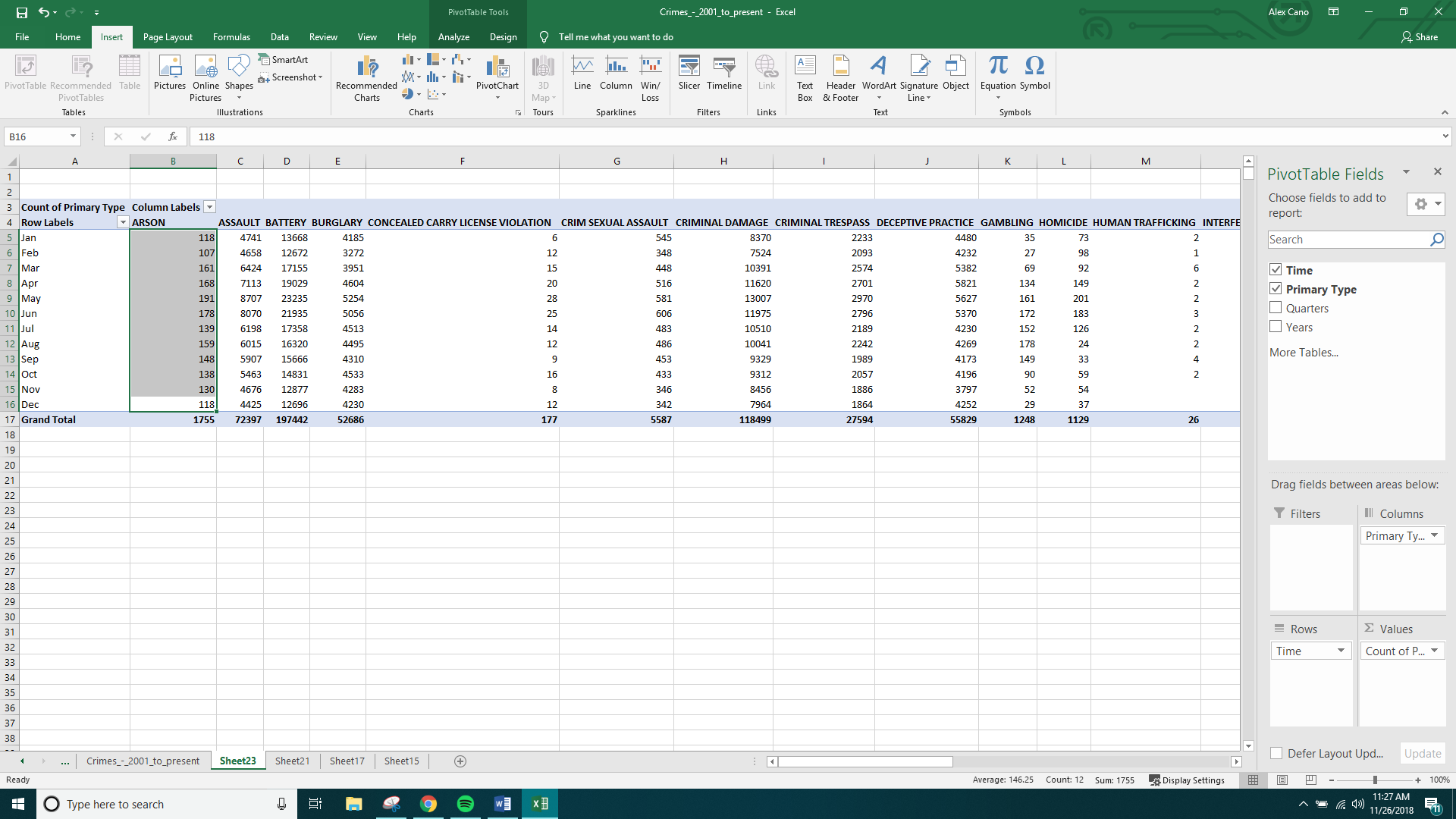
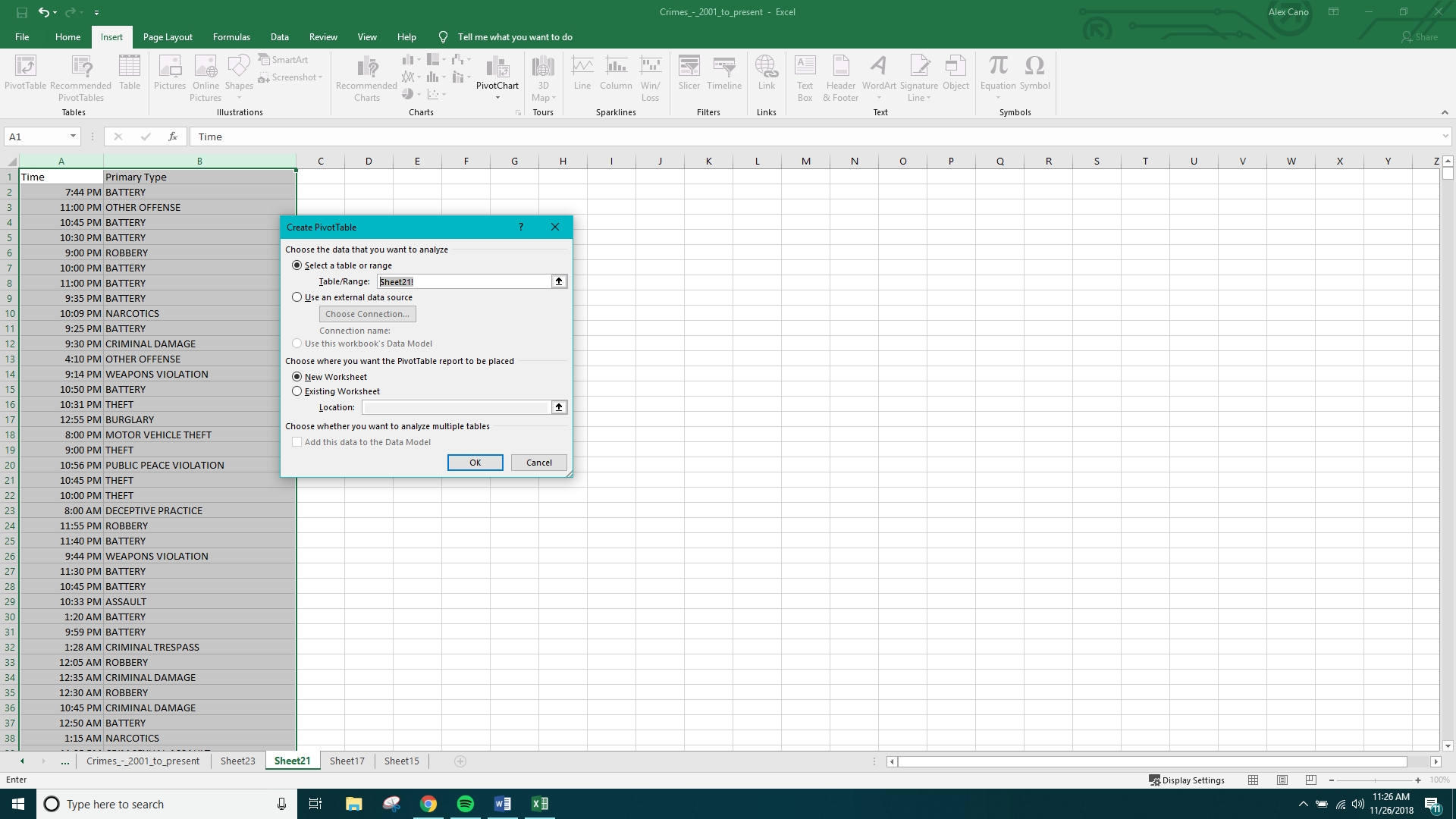


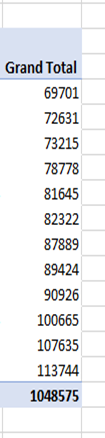
1. Just like Step 4.1 of Analyzing Occurrences of Crime in Excel. Create a pivot table of the column Time. Drag the Time field into the Row and Values Column. Time and Hours should appear in the Rows Field, if it doesn’t, drag Time into the Rows column again and it should appear.



**Step 4.3:** Least/Most Crimes Committed per Month

Analyzing which month had the least/most crimes committed.

1. Follow the steps in “Analyzing the Frequency of Crime in Excel” to get time column. Highlight time and primary type to create pivot table.
2. Continue with default options to create the pivot table.
3. Drag the time and primary type to the column, row and value areas. This will show a count of primary types done in the given months
4. To find the least/most month that given crimes were done do the following:

* Highlight the column Count of Primary Type
* On the Home tab click on Sort & Filter  
  

This shows that February was the month where the least amount of crimes committed compared to May where it was the most.

Step 3: Pig HCatalog

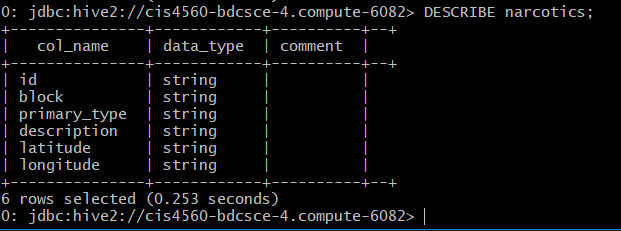
We can analyze specific crimes and store the information in HDFS. In the following example we will create a more detailed table from the crimes table named “narcotics.” From this table we will use the Pig interface Grunt to filter the data by the description “POSS: CANNABIS 30MG OR LESS” and store this data in HDFS.

1. Open another terminal and create a directory to contain your output.

-bash-4.1$ hdfs dfs -mkdir narco

1. In your beeline CLI environment, create a table which will consist of information regarding the primary\_type “narcotics.”

CREATE TABLE IF NOT EXISTS narcotics ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' AS select id, block, primary\_type, description, latitude, longitude, case\_date from crimes where primary\_type = 'NARCOTICS';

DESCRIBE narcotics;

1. Open another terminal and open the Pig interface Grunt by typing in pig -useHCatalog in the terminal, which adopts HCatalog to share Hive tables with Pig

pig -useHCatalog

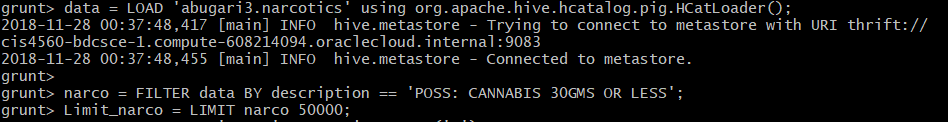
1. To load hive data into pig you must export the data using HCatLoader().

data = LOAD 'mescob33.narcotics' using org.apache.hive.hcatalog.pig.HCatLoader();

1. Filter the data to only contain information where the description is 'POSS: CANNABIS 30GMS OR LESS' and limit the results to 50,000 records.

narco = FILTER data BY description == 'POSS: CANNABIS 30GMS OR LESS';

limit\_narco = LIMIT narco 50000;

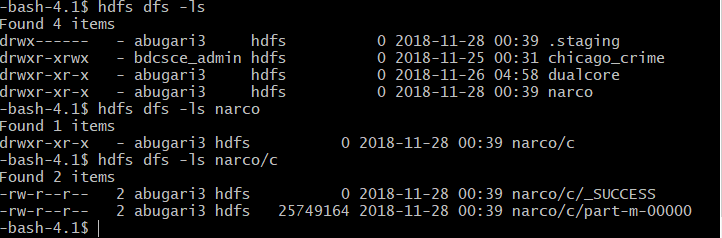


1. With the following command, you will store the output into a new file in HDFS.

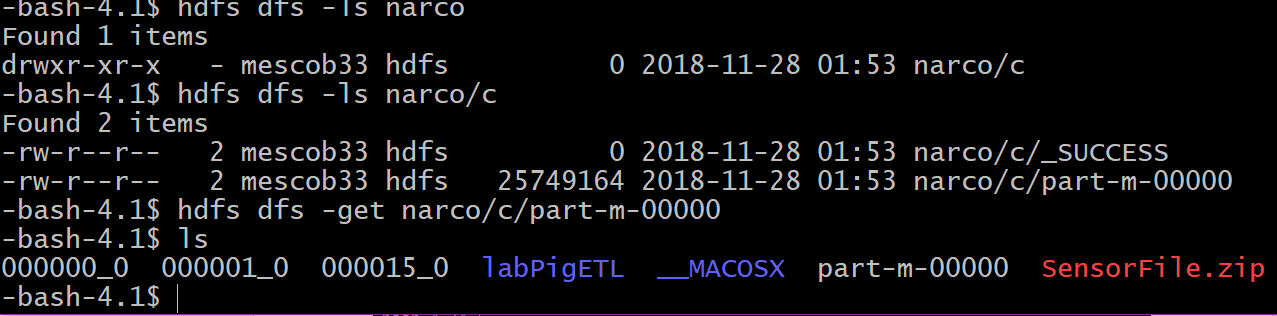
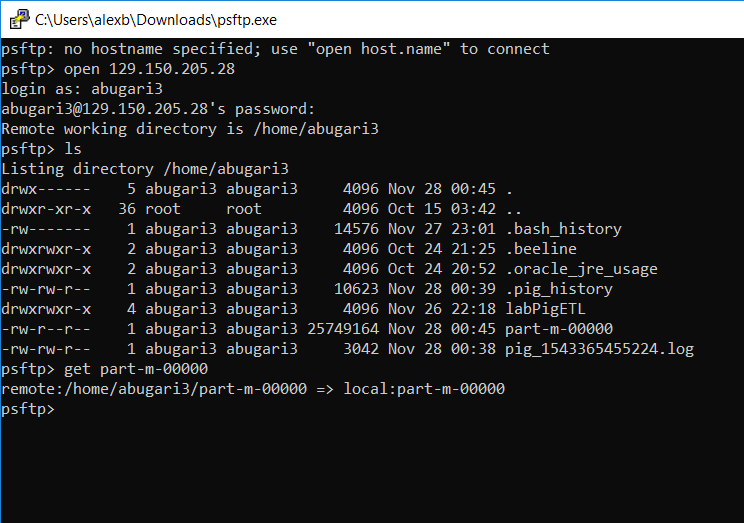
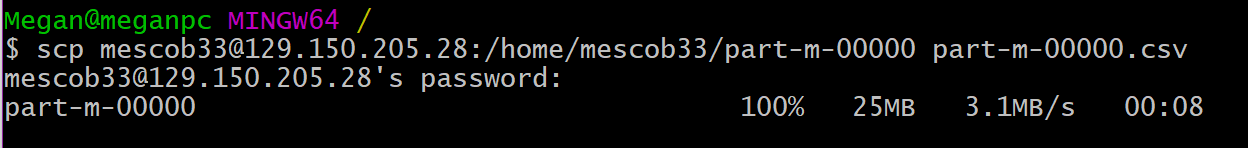
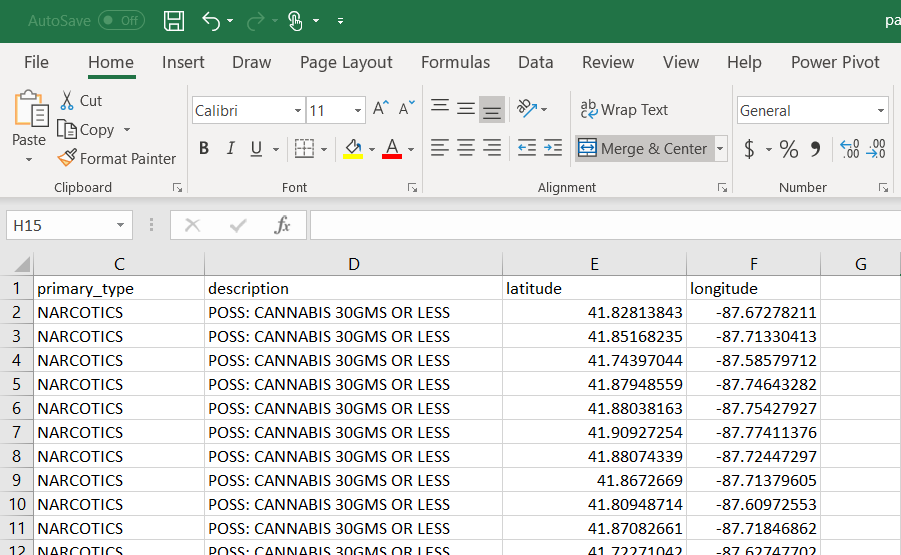
STORE narco INTO 'narco/c' USING PigStorage(',');



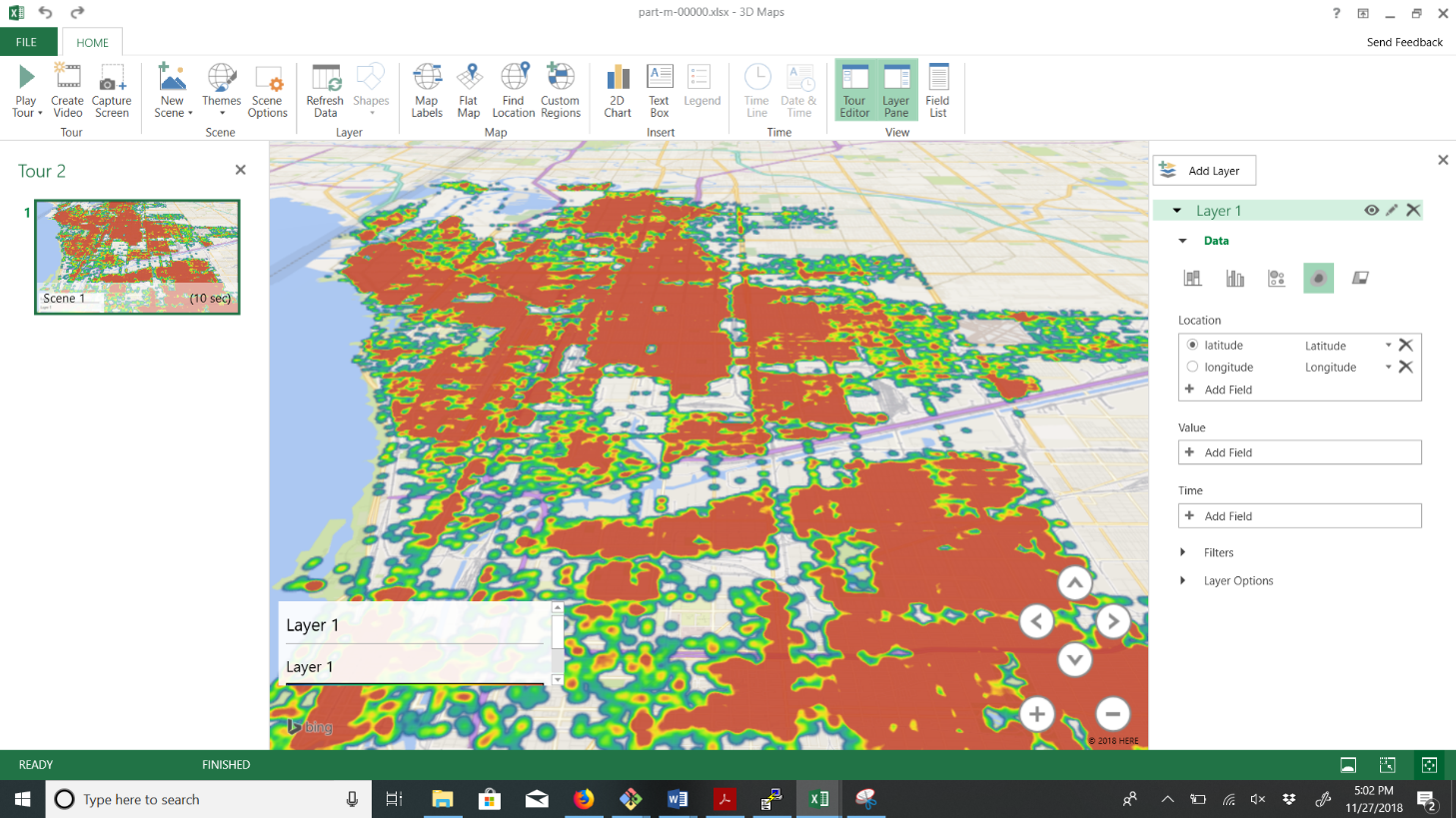
Check the output to see if the file is there



Download the file either using pfstp or scp

1. To download the file, use the –get command to download the file from HDFS to your local server.
2. You can then open another terminal and retrieve the file using scp.
3. Or use the following commands using psftp.
4. Open the file in Excel and add the following column headers.
5. Save the excel as a Workbook and insert a 3d tour



1. Change the data type to a heatmap and here we get a detailed visualization of the activity regarding marijuana possession.

References

* 1. URL of Data Source website: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>
  2. URL of Data Source CSV file: https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD
  3. URL of your Github: https://github.com/DayWait/BigData.git