**Term Paper Manuscript**

MEGAN ESCOBAR

ALEXANDER BUGARIN

DAVID HA

ALEJANDRO CANO

INTRODUCTION TO BIG DATA CIS 4560-01

California State University, Los Angeles

E-mail : mescob33@calstatela.edu

dha3@calstatela.edu

abugari3@calstatela.edu

acano19@calstatela.edu

**Abstract:** This research focuses on the crimes reported by the police department of the City of Chicago, Illinois [1]. The research intends to find crime trends from their publicly published dataset using Pig and Beeline to do the analyzation. We find the area of focus should be concentrated on simple battery and simple theft. Although most crimes committed are nonviolent, it is still important to compare these crimes to other neighborhoods. This is because only about 30% of all reported crimes results in an arrest. This research aims to provide useful information to the Chicago police department so as to decrease criminal activities by effectively putting more resources, such as officers at certain areas at certain times, where it matters most.

**1. Introduction**

This paper will demonstrate how to download the dataset “Crimes - 2001 to present” from the Chicago Data Portal [2], process it in Beeline CLI using HiveQL, analyze and visualize it using Excel, and lastly, filter it in Pig using HCatalog.

This dataset is from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. There are many columns to analyze, but the ones focused on are case date, block, primary type (of crime), description (of crime), location description, arrest, and more.

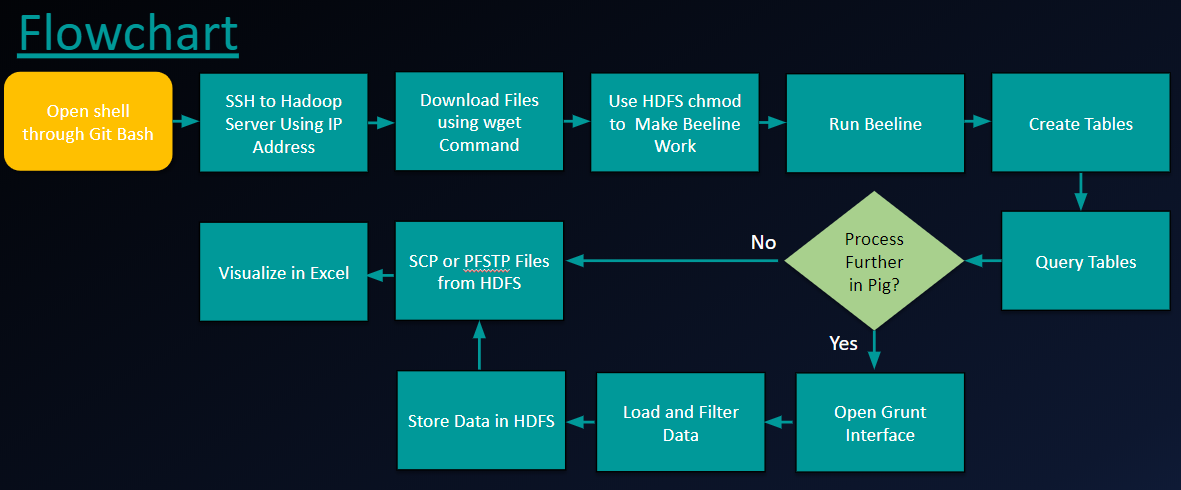
Figure 1.1 shows the research process and the steps taken to reach our conclusion. 

Figure 1.1 Research process

**2. General Instruction**

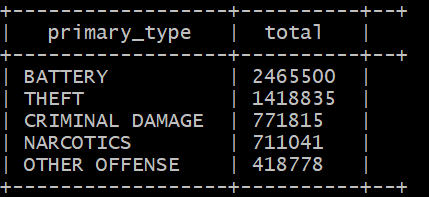
Learning how to download the dataset, extracting information from it, and creating visualizations to make it more understandable.

**2.1 Get Data from Chicago Data Portal**

Remotely access Oracle Big Data data by using SSH by opening Git Bash. Using your account to connect to the Hadoop server via IP address. By using the bash command to download the dataset from the city of Chicago’s website, we now have access to the dataset to analyze.

**2.2 Using Beeline CLI**

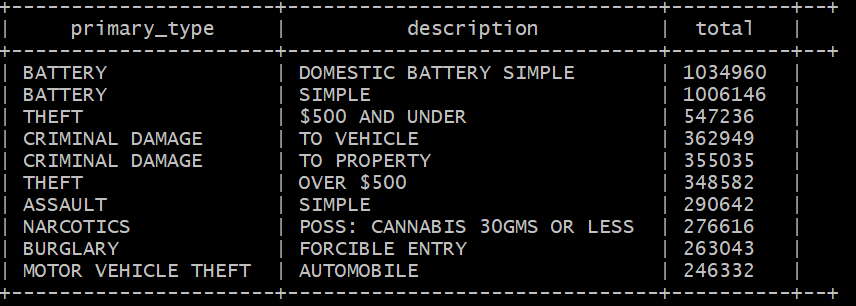
Connecting to Beeline gives the ability to create the table to be loaded and thus fetched and queried.

Figure 2.2.1 Determining which crime is committed most

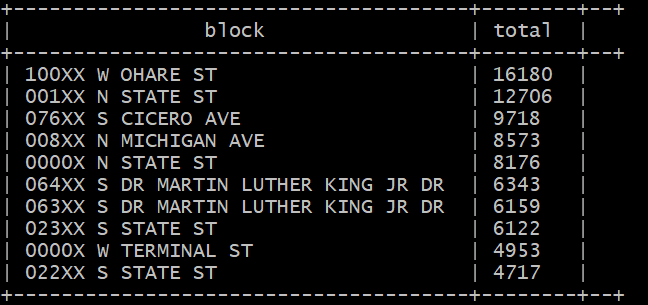
A simple query such as figure 2.2.1 shows the power of Hive. It is able to analyze all the data and retrieve information showing the top five types of crimes. In addition, it has the ability to download data from anywhere in the world and the ability to store that data in a decentralized way, giving us access to a more feasible way of storing, analyzing, and retrieving data.

Using these techniques, it can be seen that the point of interest the City of Chicago has is with battery and theft. Thus, the city should have plans to somehow decrease the cause for battery and theft. By doing so, it can decrease a huge chunk of crime by focusing on only two areas instead on multiple areas.

By then furthering analyzing the description of the crime, Chicago police can easily determine which crime is most prevalent.

Figure 2.2.2 Retrieving the description of the top crimes committed

Chicago police can also determine where crimes are currently most common at and also at what time. Using data such as this is a huge advantage on preventing crime by being able to be ahead of the curve by being able to decipher criminal activities.

Figure 2.2.3 Determining which street has the most crime

Using Beeline, it is also easy to determine the seriousness of the current crime. We are then able to answer the question: “How many reports result in an arrest?” If an arrest is required, it means the crime has a certain level of seriousness. If the city uses this type of analysis, they are then able to determine if the arrest is warranted, or if there is an increase in arrest. With Beeline, it can be analyzed to show that about 30% of all crimes results in an arrest.

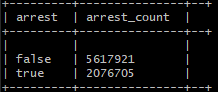


Figure 2.2.4 Figuring out how many arrests were made in Chicago

**2.3 Analyzing violent crimes**

With almost two and a half million entries for battery as one of the crimes committed it becomes difficult for Microsoft Excel to handle. Using Hadoop we can drill down our data and also analyze its content with more scrutiny.

There are various descriptions that fall into the primary\_type battery and with this we can narrow our results and look for interesting visualizations.

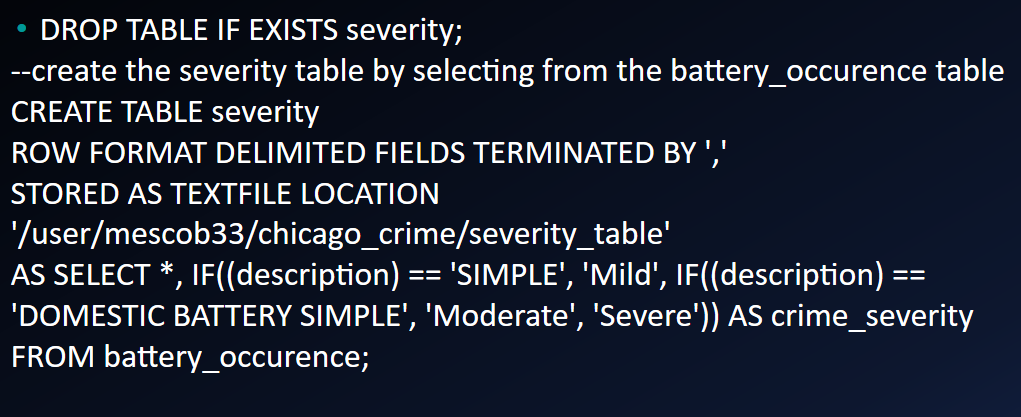


Figure 2.3.1 This is a beeline table that classifies descriptions into a severity that can be visualized.

Using Beeline, we can classify these millions of instances of battery on a basis of either mild, moderate, or a severe type of crime. This is based on if the assault happened with a deadly weapon or not. It would be useful for the police department to know where the most violent crimes are happening so that they can take measures to protect these areas.

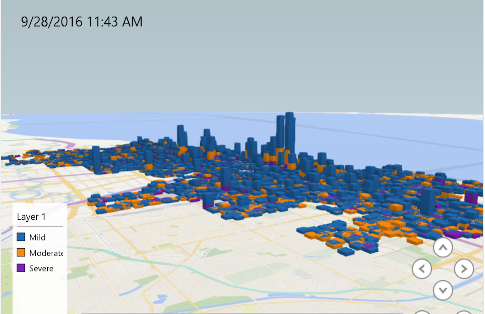


Figure 2.3.2 Importing our Beeline table into Excel can create a very use 3d map tour

In the photo above we can see the severity of crimes being committed within the city. It is easy to see that there are various points in the city with significantly more crime and with the ability to differentiate severity it is easier to see where the most violent crimes happen. Since the data also has a column for date and time we can use this to visualize the increase of crimes throughout the years.

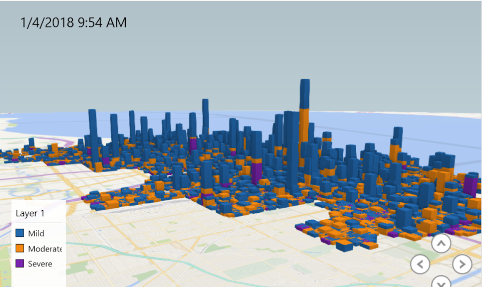


Figure 2.3.3 Using the time element we can see a very significant increase in violent crime between 2016-2018

**2.4 Pig HCatalog**

Beeline isn’t the only method to manipulate our data. Another method of manipulating our data is by using Pig. Pig is a high level scripting language that is used with Apache Hadoop. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark. We can use Pig to filter out some data so we can focus on more specific data.

First the data from the hive table has to be imported into Pig which can be done easily with HCatalog. With narcotics being one of the crimes with the largest amounts of entries, we can try and visualize where in the city these crimes happen most.



Figure 2.4.1 A heatmap of narcotics activity created by filtering Chicago crimes data with Pig.

As can be seen in figure 2.4.2 and 2.4.3, marijuana possession has been on the increase. With this knowledge and analyzation skill that can be applied to more than just marijuana, Chicago police can better enforce rules and regulations, and as well as find correlations.

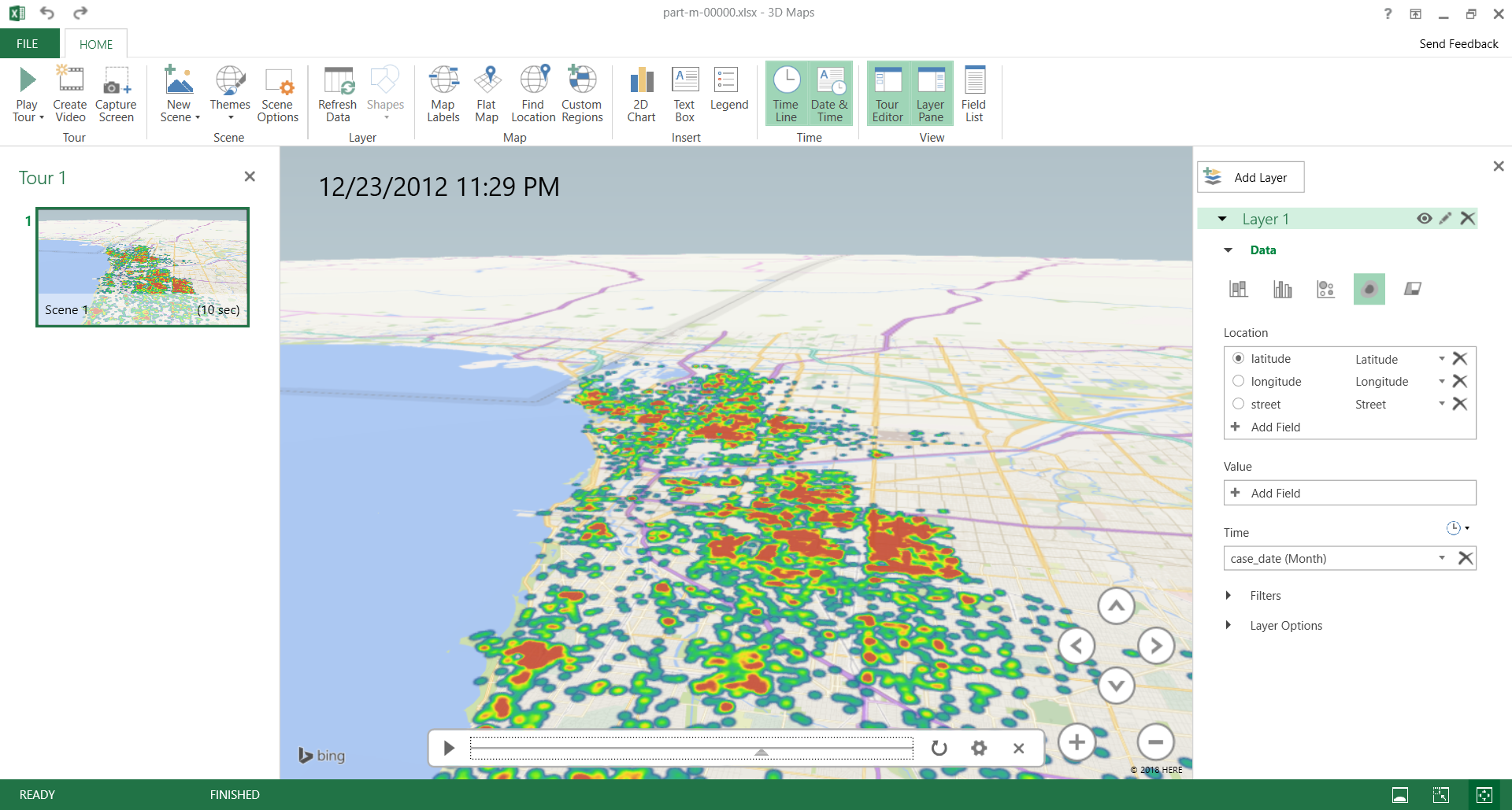


Figure 2.4.2 Marijuana Possession in 12/23/2012

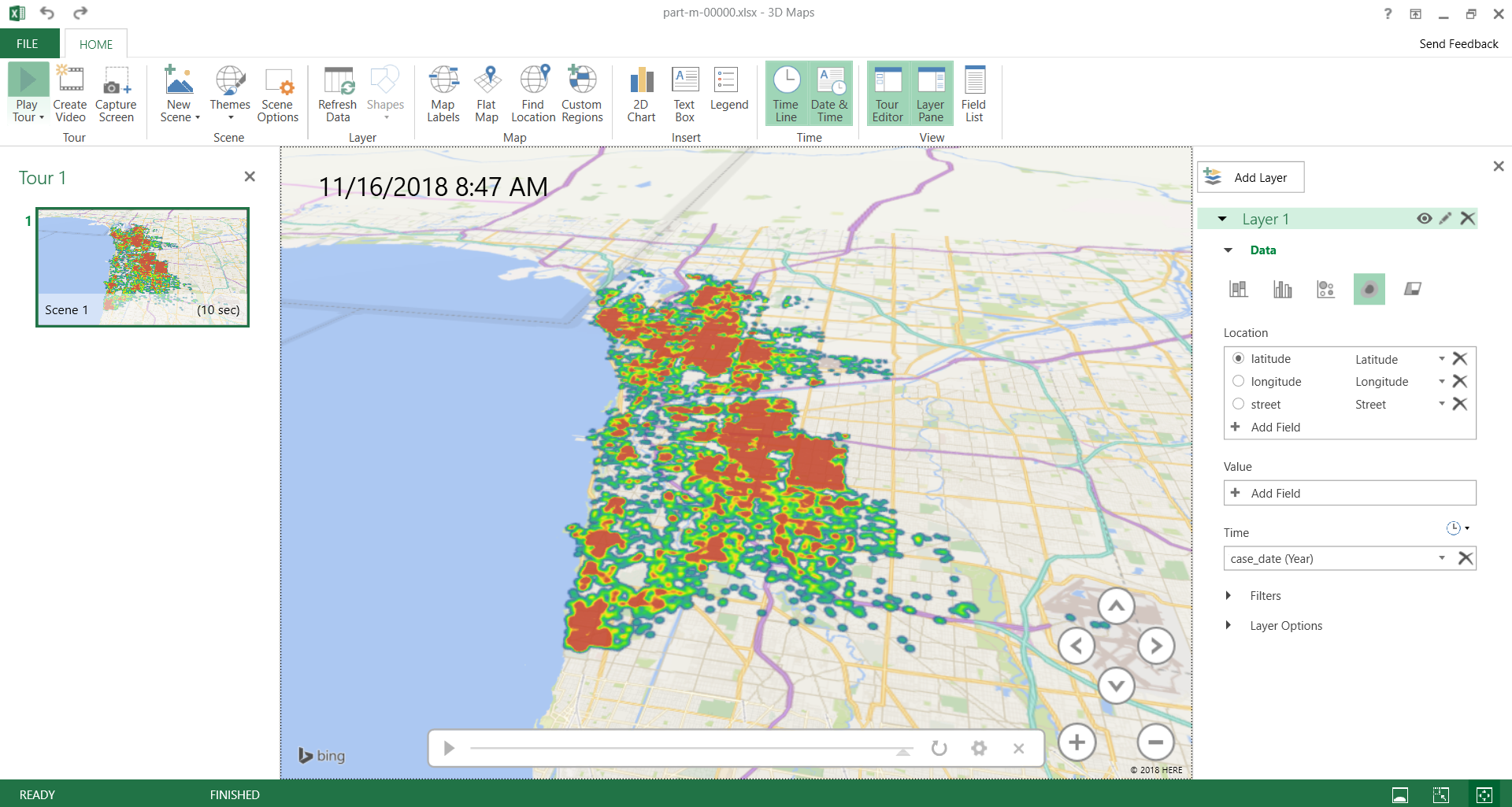


Figure 2.4.3 Marijuana possession in 11/16/2018

**2.5 Data Processing in Excel**

Analyzing data in Microsoft Excel is a simple, yet powerful way in quickly generating insights from data. The original crime\_data file contains column headers that can cause problems within excel when opening it as a CSV file.[[1]](#footnote-1) In order to bypass this problem when downloading the data, the first line within the file must be removed and be saved as “new\_crime.csv” in the local server.[[2]](#footnote-2)

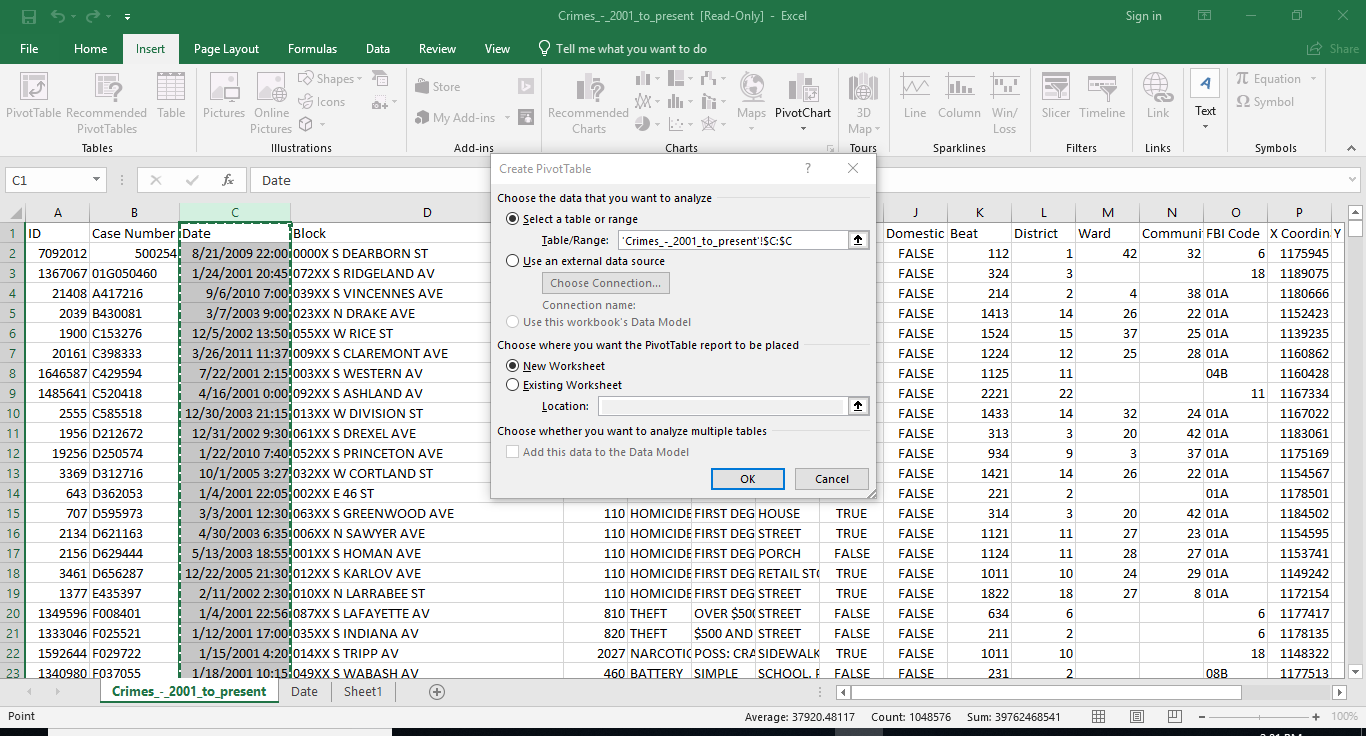


Figure 2.5.1 Showing how to create a pivot table.

After downloading the file, it can then be opened in excel. The column headers all provide data which can be used for better analyzing when converted into a pivot table. In this case the first pivot table will show how many crimes occur over the years per month, but it will be complex since the date column of the crimes table contains both the date and time as seen In figure 2.5.2. In order to solve this problem, a new column must be made to contain the time, so the month and year function can be displayed on the two new columns for precise processing.

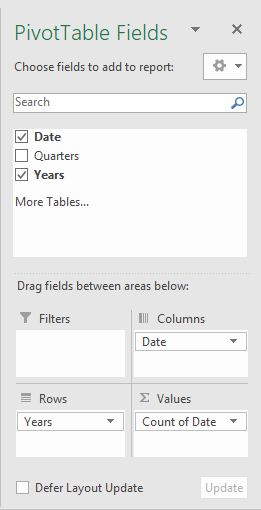


Figure 2.5.2 shows the fields that must be dragged into the correct section.

In order to create the correct table, the fields must be dragged into their correct sections, as seen in Figure 2.5.2. This figure shows that date needs to be dragged into the columns section, years into the rows section, and date into the values section. After doing so, it will generate the correct pivot table needed to analyze the data, which is seen below in Figure 2.5.3. From this we can ultimately see that the month in which crime occurs most was in May.

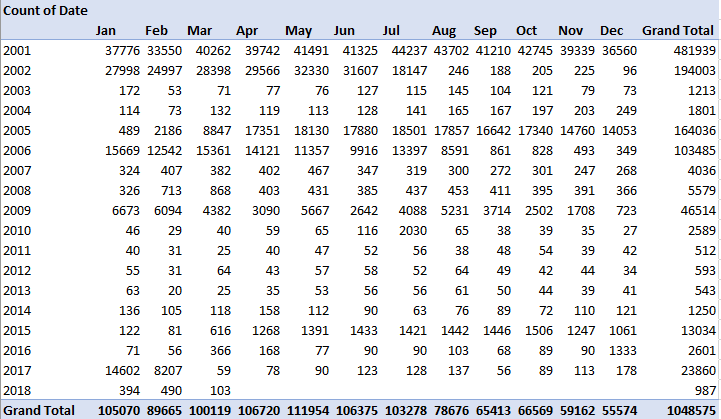


Figure 2.5.3 The final pivot table created to determine what crime occurs most in 2001.

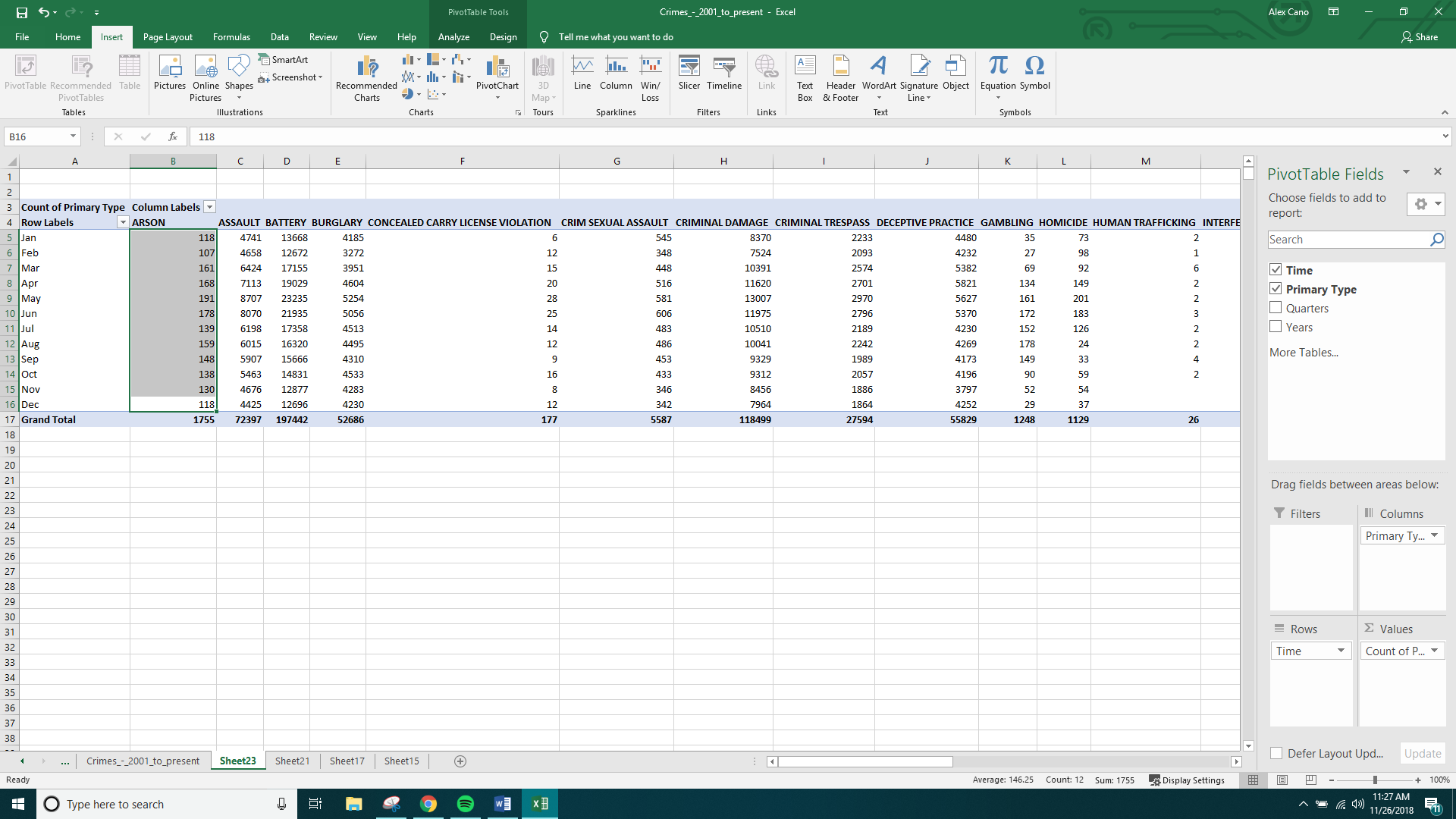


Figure 2.5.4 Shows pivot table created from time and primary type columns.

Pivot tables can also be created with more than one column in excel, as seen in Figure 2.5.4. The columns that were used here were time and primary type. The given columns were then dragged into their corresponding sections for the result that we needed. February was the month that least crimes were committed, in comparison to May, that indicates it was the month with the most committed crimes. Also, can be seen in figure 2.5.5 below.

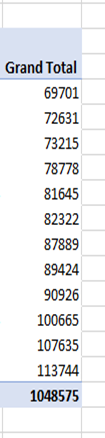


Figure 2.5.5 The months that most (May) and least (Feb.) crimes committed.

**References**

[1] Crimes - 2001 to present | City of Chicago | Data Portal. (n.d.). Retrieved November/December, 2018, from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

*[2] 2001 Chicago Crime Data Set*. (n.d.). Retrieved November/December, 2018, from https://data.cityofchicago.org/api/views/ijzp-q8t2/rows.csv?accessType=DOWNLOAD.

1. Excel interprets CSV as SYLK. (2016, September 19). Retrieved from https://www.drupal.org/project/csv\_serialization/issues/2802081 [↑](#footnote-ref-1)
2. Sed Command in Linux/Unix with examples. (2018, September 06). Retrieved from https://www.geeksforgeeks.org/sed-command-in-unix/ [↑](#footnote-ref-2)