







# Wentworth Institute of Technology

## School of Computing and Data Science

Big Data Systems– Summer 2022

Dr. Salem Othman

Read each question carefully before answering it. You can use Google cloud or your machine. I may give partial credit based on the code you write down, so if time permits, show your work! Solve these questions using **PIG**.

 Chicago_Crimes_2001_to_2004.csv	6/18/2022 9:53 PM	Microsoft Excel Co...	464,204 KB
 Chicago_Crimes_2005_to_2007.csv	6/18/2022 9:53 PM	Microsoft Excel Co...	460,233 KB
 Chicago_Crimes_2008_to_2011.csv	6/18/2022 9:53 PM	Microsoft Excel Co...	661,811 KB
 Chicago_Crimes_2012_to_2017.csv	6/18/2022 9:45 PM	Microsoft Excel Co...	358,208 KB

Given four datasets of Chicago crime, answer the following questions. A schema might need to be created manually. Also, you will need to Concatenate datasets and filter to remove rows that contains null values of year column.

- Q1. What was the most common type of crime committed?
- Q2. Per distinct district, what was the average time when a crime was committed (give both average time and district per row).
- Q3. How many first-degree murders involved both an arrest and domestic violence?
- Q4. What was the most common day of the week for theft?
- Q5. What block in Chicago had the most deceptive practice, of column primary type, in 2013?
- Q6. What percent of thefts involved an arrest in each Chicago ward?
- Q7. Plot the number of thefts and arsons in Chicago from 2010-2015 in both a bar graph and in a line graph?
- Q8. From 2003 until 2015, did the number of robberies increase or decrease? To support your result, include the total amount of robberies per year, with the percent decrease or increase. Also plot in at least three graphs of your choosing to support the result.
- Q9. Find the amount of Criminal Damage that occurred per hour on Thursday between 2009 and 2012, plot in a line graph.
- Q10. What was the least common day for arsons in 2017?
- Q11. **Write your own 3 questions; they have to be complicated questions with at least two parts (similar to number 7 or number 8) and answer them yourself.**

**Note 1: If you find interpretation of any question difficult, interpret it the way you understood it, and write your interpretation down before you answer the question. So, I will read it and be fair grading you.**

## Dataset

You can use the dataset in Brightspace or download it from the following link.

<https://www.kaggle.com/datasets/currie32/crimes-in-chicago?resource=download>

## Content

**ID** - Unique identifier for the record.

**Case Number** - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.

**Date** - Date when the incident occurred. this is sometimes a best estimate.

**Block** - The partially redacted address where the incident occurred, placing it on the same block as the actual address.

**IUCR** - The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at <https://data.cityofchicago.org/d/c7ck-438e>.

**Primary Type** - The primary description of the IUCR code.

**Description** - The secondary description of the IUCR code, a subcategory of the primary description.

**Location Description** - Description of the location where the incident occurred.

**Arrest** - Indicates whether an arrest was made.

**Domestic** - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.

**Beat** - Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at <https://data.cityofchicago.org/d/aerh-rz74>.

**District** - Indicates the police district where the incident occurred. See the districts at <https://data.cityofchicago.org/d/fthy-xz3r>.

**Ward** - The ward (City Council district) where the incident occurred. See the wards at <https://data.cityofchicago.org/d/sp34-6z76>.

**Community Area** - Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at <https://data.cityofchicago.org/d/cauq-8yn6>.

**FBI Code** - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at [http://gis.chicagopolice.org/clearmap\\_crime\\_sums/crime\\_types.html](http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html).

**X Coordinate** - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

**Y Coordinate** - The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.

**Year** - Year the incident occurred.

**Updated On** - Date and time the record was last updated.

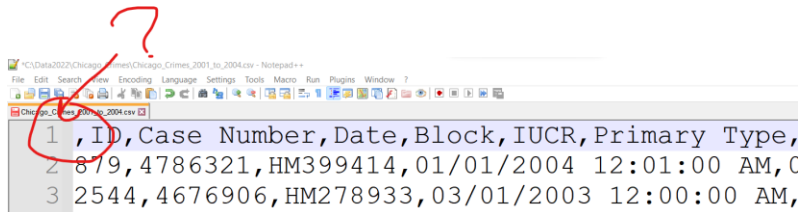
**Latitude** - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

**Longitude** - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.

**Location** - The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

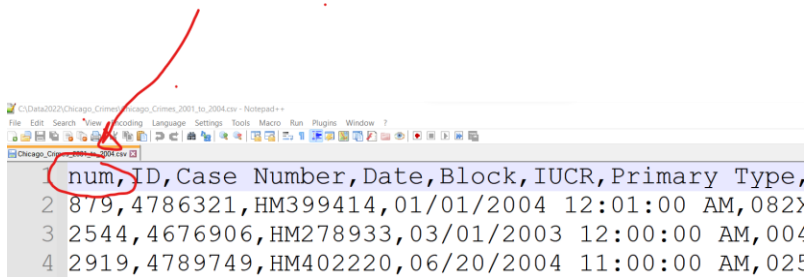
**Note 2:** I added a new column name to each file that uploaded to Brightspace. If you download the files from the link above, then you need to come up with some way to fix this issue. See the pictures below.

Original File:



```
1 , ID, Case Number, Date, Block, IUCR, Primary Type,  
2 879, 4786321, HM399414, 01/01/2004 12:01:00 AM, C  
3 2544, 4676906, HM278933, 03/01/2003 12:00:00 AM,
```

The updated File:



```
1 num, ID, Case Number, Date, Block, IUCR, Primary Type,  
2 879, 4786321, HM399414, 01/01/2004 12:01:00 AM, 082>  
3 2544, 4676906, HM278933, 03/01/2003 12:00:00 AM, 004  
4 2919, 4789749, HM402220, 06/20/2004 11:00:00 AM, 025
```