**An Analytical Detective**

Crime is an international concern, but it is documented and handled in very different ways in different countries. In the United States, violent crimes and property crimes are recorded by the Federal Bureau of Investigation (FBI). Additionally, each city documents crime, and some cities release data regarding crime rates. The city of Chicago, Illinois releases crime data from 2001 onward online.

Chicago is the third most popular city in the United States, with a population of over 2.7 million people. The city of Chicago is shown in the map below, with the state of Illinois highlighted in red.



There are two main types of crimes: violent crimes, and property crimes. In this problem, we’ll focus on one specific type of property crime, called “motor vehicle theft” (sometimes referred to as grand theft auto). This is the act of stealing, or attempting to steal, a car. In this problem, we’ll use some basic data analysis to understand the motor vehicle thefts in Chicago.

**Data:** download mvtWeek1.csv (CSV)

Here’s a list of descriptions of the variables:

**ID:** a unique identifier for each observation

**Date:** the date the crime occurred

**LocationDescription:** the location where the crime occurred.

**Arrest:** whether or not an arrest was made for the crime (TRUE if an arrest was made, and FALSE if an arrest was not made).

**Domestic:** whether or not the crime was a domestic crime, meaning that it was committed against a family member (TRUE if it was domestic, and FALSE if it was not domestic).

**Beat:** the area, or “beat” in which the crime occurred. This is the smallest regional division defined by the Chicago police department.

**District:** the police district in which the crime occurred. Each district is composed of many beats, and are defined by the Chicago Police Department.

**Community Area:** the community area in which the crime occurred. Since the 1920s, Chicago has been divided into what are called “community areas”, of which there are now 77. The community areas were devised in an attempt to create socially homogeneous regions.

**Year:** the year in which the crime occurred.

**Latitude:** the latitude of the location at which the crime occurred.

**Longitude:** the longitude of the location at which the crime occurred.

**Q&A**

**The instructions assume you are using Microsoft Excel, but you can use any tool you are comfortable with for generating calculation and visualizations.**

Loading the Data **(6 pts)**

* How many rows of data (observations) are in this dataset?
  + There are 191, 642 rows of data in this dataset
* How many variables are in this dataset?
  + There are 11 variables in this dataset
* Using the “max” function, what is the maximum value of the variable “ID”?
  + 9181151
* What is the minimum value of the variable “Beat”?
  + 111
* How many observations have the value TRUE in the Arrest variable (this is the number of crimes for which an arrest was made)?
  + 15536
* How many observations have a LocationDescription value of ALLEY?
  + 2308

**Understanding the date variables (12 pts)**

* In many datasets, like this one, you have a date field. There are many ways to represent the date. In what format are the entries in the variable Date?
  + MM/DD/YYYY HH:MM
* Insert three new columns after the date column. Label the columns: Month, Day, and Year. Use the following Excel functions to extract each date element from the date variable in column B.
  + In cell c2 type =text(B2,”MM”) this should extract the month from cell B2
  + In cell d2 type =text(B2,”dd”) this should extract the day from cell B2
  + In cell e2 type =text(B2,”yyyy”) this should extract the year from cell B2



<https://www.extendoffice.com/documents/excel/3118-excel-extract-month-and-year-from-date.html>

* What is the month and year of the median date in our dataset? Enter your answer as “Month Year”, without the quotes. (Example: if the answer was 2008-03-28, you would give the answer “March 2008”, without quotes.)
  + May 2006
* In which month did the fewest motor vehicle thefts occur?
  + February
* On which weekday did the most motor vehicle thefts occur?
  + Friday
* Each observation in the dataset represents a motor vehicle theft, and the Arrest variable indicates whether an arrest was later made for this theft. Which month has the largest number of motor vehicle thefts for which an arrest was made?
  + January

**Visualizing Crime Trends I (2 pts)**

Now let’s make some plots to help us better understand how crime has changed over time in Chicago. Throughout this problem, and in general you can save your plot file. You can use any visualization tool you choose.

First, let’s make a histogram of the variable Date. We’ll add an extra argument, to specify the number of bars we want in our histogram. We want 100 bins.

* How to generate a histogram using Excel: <https://www.youtube.com/watch?v=xekiDJzajYk>

Looking at the histogram, answer the following questions:

* In general, does it look like crime increases or decreases from 2002 – 2012?
  + In general, crime looks like it decreases during this time period
* In general, does it look like crime increases or decreases from 2009 – 2011?
  + In general, the crime increases during this time period

**Visualizing Crime Trends II (10 pts)**

Now, let’s see how arrests have changed over time. Create a boxplot of the variable “Date”, sorted by the variable “Arrest”.

In a boxplot, the bold horizontal line is the median value of the data, the box shows the range of values between the first quartile and third quartile, and the whiskers (the dotted lines extending outside the box) show the minimum and maximum values, excluding any outliers (which are plotted as circles). Outliers are defined by first computing the difference between the first and third quartile values, or the height of the box. This number is called the Inter-Quartile Range (QR). Any point that is greater than the third quartile plus the IQR or less than the first quartile minus the IQR is considered an outlier.

Creating a Boxplot in Excel:

<https://www.youtube.com/watch?v=EiS-q6euyf8>

<https://www.youtube.com/watch?v=0eZ5G06TKfE>

Does it look like there are more crimes for which arrests were made in the first half of the time period or the second half of the time period? (Note that the time period is from 2001 to 2012, so the middle of the time period is the beginning of 2007)

* First half
  + The average mean of the number of arrests made looks higher in the first half than the second
* Second half

**Visualizing Crime Trends III (10 pts)**

* For what proportion of motor vehicle thefts in 2001 was an arrest made? Note, we are asking for an answer as a proportion. Therefore, your answer should take a value between 0 and 1.
  + 0.104117
* How to calculate proportions using Excel: <https://www.youtube.com/watch?v=ZGUTAxdQ7yw>

**Visualizing Crime Trends IV (5 pts)**

* For what proportion of motor vehicle thefts in 2007 was an arrest made?
  + 0.084874

**Visualizing Crime Trends V (10 pts)**

* For what proportion of motor vehicle thefts in 2012 was an arrest made?
  + 0.039029
* Analyzing this data could be useful to the Chicago Police Department when deciding where to allocate resources. If they want to increase the number of arrests that are made for motor vehicle thefts, where should they focus their efforts?
  + They should focus their efforts into analyzing which time of the year the most motor vehicle threats occur and be more alert during those times.

We want to find the top five locates where motor vehicle thefts occur. If you create a table of the LocationDescription variable, it is unfortunately very hard to read since there are 78 different locations in the data set. By using the sort function, we can view this same table, but sorted by the number of observations in each category.

* Which locations are the top five locations for motor vehicle thefts, excluding the “Other” category? You should select 5 of the following options:
  + Bank
  + Gas station
  + Hotel/Motel
  + Street
  + Car wash
  + Restaurant
  + Parking lot/Garage (non-residential)
  + Alley
  + Driveway (residential)
  + Vacant lot/land

**Popular Locations I (5 pts)**

Create a subset of your data, only taking observations for which, the theft happened in one of these five locations, and call this new data set “Top 5”. To only take observations that have a certain value in one variable or the other, the logical OR operation is used. Alternatively, you could create five different subsets, and then merge them together into one dataset.

* How many observations are in the Top 5?
  + 177510

**Popular Locations II**

Use the Top 5 subset data to answer the remining questions:

You should see the LocationDescription now only has 5 values, as we expect.

* One of the locations has a much higher arrest rate than the other locations. Which is it? Please enter the text in exactly the same way as how it looks in the answer options for **Visualizing Crime Trends V**.
  + Gas station

On which day of the week do the fewest motor vehicle thefts in residential driveways happen?

* Saturday

Choose one aspect of the data to create a visualization to support one or more of your findings. Use best practices for data visualization.

You can edit and save your answers in this document, but include visualizations in separate files (make sure you label the visualizations accordingly). Create a 1-2-page summary of your findings, make sure you refer to your visualizations. This means you need to label the visualizations (Fig 1, for example). Your summary should tell a story. **Upload your data file showing your calculations.** Combine all of the files (this file + all visualizations) into one zip file and upload in Brightspace.

Grading

Loading the Data (6 pts)

Understanding the data variables (12 pts)

Visualizing Crime Trends I (2 pts)

Visualizing Crime Trends II (10 pts)

Visualizing Crime Trends III (10 pts)

Visualizing Crime Trends IV (5 pts)

Visualizing Crime Trends V (10 pts)

Popular Locations I (5 pts)

Popular Locations II (20 pts)

Summary (20 pts)

**Total: 100 points**

**Save the zip file as:**

LastnameFirstInitial\_CGT275S22\_A1.zip