Bookmarks

☐ Bookmark

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 populous 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 in R to understand the motor vehicle thefts in Chicago.

Please download the file mvtWeek1.csv for this problem (do not open this file in any spreadsheet software before completing this problem because it might change the format of the Date field). Here is 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 occured. Each district is composed of many beats, and are defined by the Chicago Police Department.

- **CommunityArea**: 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.

Problem 1.1 - Loading the Data

 (1 point possible)
Read the dataset mvtWeek1.csv into R, using the read.csv function, and call the data frame "mvt". Remember to navigate to the directory on your computer containing the file mvtWeek1.csv first. It may take a few minutes to read in the data, since it is pretty large. Then, use the str and summary functions to answer the following questions.

How many rows of data (observations) are in this dataset?

[   ] ?

You have used 0 of 3 submissions

Problem 1.2 - Loading the Data

(1 point possible)
How many variables are in this dataset?

[   ] ?

You have used 0 of 3 submissions

Problem 1.3 - Loading the Data

(1 point possible)
Using the "max" function, what is the maximum value of the variable "ID"?

[   ] ?

You have used 0 of 3 submissions

Problem 1.4 - Loading the Data

(1 point possible)
What is the minimum value of the variable "Beat"?

[   ] ?

You have used 0 of 3 submissions

Problem 1.5 - Loading the Data

(1 point possible)
How many observations have value TRUE in the Arrest variable (this is the number of crimes for which an arrest was made)?

[   ]   **?**

You have used 0 of 3 submissions

Problem 1.6 - Loading the Data

(1 point possible)

How many observations have a LocationDescription value of ALLEY?

[   ]   **?**

You have used 0 of 3 submissions

Problem 2.1 - Understanding Dates in R

(1 point possible)

In many datasets, like this one, you have a date field. Unfortunately, R does not automatically recognize entries that look like dates. We need to use a function in R to extract the date and time. Take a look at the first entry of Date (remember to use square brackets when looking at a certain entry of a variable).

In what format are the entries in the variable Date?

○ Month/Day/Year Hour:Minute

○ Day/Month/Year Hour:Minute

○ Hour:Minute Month/Day/Year

○ Hour:Minute Day/Month/Year

?

You have used 0 of 1 submissions

Problem 2.2 - Understanding Dates in R

(1 point possible)

Now, let's convert these characters into a Date object in R. In your R console, type

DateConvert = as.Date(strptime(mvt$Date, "%m/%d/%y %H:%M"))

This converts the variable "Date" into a Date object in R. Take a look at the variable DateConvert using the summary function.

What is the month and year of the median date in our dataset? Enter your answer as "Month Year", without the quotes. (Ex: if the answer was 2008-03-28, you would give the answer "March 2008", without the quotes.)

<div style="border:1px solid #ccc; width:300px; height:80px;"></div>   **?**

You have used 0 of 3 submissions

Problem 2.3 - Understanding Dates in R

(1 point possible)

Now, let's extract the month and the day of the week, and add these variables to our data frame mvt. We can do this with two simple functions. Type the following commands in R:

mvt$Month = months(DateConvert)

mvt$Weekday = weekdays(DateConvert)

This creates two new variables in our data frame, Month and Weekday, and sets them equal to the month and weekday values that we can extract from the Date object. Lastly, replace the old Date variable with DateConvert by typing:

mvt$Date = DateConvert

Using the table command, answer the following questions.

In which month did the fewest motor vehicle thefts occur?

[dropdown ▲▼]   **?**

You have used 0 of 2 submissions

Problem 2.4 - Understanding Dates in R

(1 point possible)
On which weekday did the most motor vehicle thefts occur?

[▲▼]   **?**

You have used 0 of 2 submissions

Problem 2.5 - Understanding Dates in R

(1 point possible)
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?

[▲▼]   **?**

You have used 0 of 2 submissions

Problem 3.1 - Visualizing Crime Trends

(3 points possible)
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 to a file. For more information, this website very clearly explains the process.

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. In your R console, type

hist(mvt$Date, breaks=100)

Looking at the histogram, answer the following questions.

In general, does it look like crime increases or decreases from 2002 - 2012?

○ Increases

○ Decreases

?

In general, does it look like crime increases or decreases from 2005 - 2008?

○ Increases

○ Decreases

?

In general, does it look like crime increases or decreases from 2009 - 2011?

○ Increases

○ Decreases

?

You have used 0 of 1 submissions

Problem 3.2 - Visualizing Crime Trends

(1 point possible)
Now, let's see how arrests have changed over time. Create a boxplot of the variable "Date", sorted by the variable "Arrest" (if you are not familiar with boxplots and would like to learn more, check out this tutorial). 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 (IQR). 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.

Does it look like there were 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

○ Second half

?

You have used 0 of 1 submissions

Problem 3.3 - Visualizing Crime Trends

(2 points possible)
Let's investigate this further. Use the table function for the next few questions.

For what proportion of motor vehicle thefts in 2001 was an arrest made?

Note: in this question and many others in the course, we are asking for an answer as a proportion. Therefore, your answer should take a value between 0 and 1.

?

You have used 0 of 5 submissions

Problem 3.4 - Visualizing Crime Trends

(1 point possible)
For what proportion of motor vehicle thefts in 2007 was an arrest made?

?

You have used 0 of 3 submissions

Problem 3.5 - Visualizing Crime Trends

(1 point possible)

For what proportion of motor vehicle thefts in 2012 was an arrest made?

[   ]   ?

Since there may still be open investigations for recent crimes, this could explain the trend we are seeing in the data. There could also be other factors at play, and this trend should be investigated further. However, since we don't know when the arrests were actually made, our detective work in this area has reached a dead end.

You have used 0 of 3 submissions

Problem 4.1 - Popular Locations

(1 point possible)

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?

We want to find the top five locations 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. In your R console, type:

sort(table(mvt$LocationDescription))

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

?

You have used 0 of 2 submissions

Problem 4.2 - Popular Locations

(1 point possible)

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 "Top5". To do this, you can use the | symbol. In lecture, we used the & symbol to use two criteria to make a subset of the data. To only take observations that have a certain value in one variable or the other, the | character can be used in place of the & symbol. This is also called a logical "or" operation.

Alternately, you could create five different subsets, and then merge them together into one data frame using rbind.

How many observations are in Top5?

	?

You have used 0 of 3 submissions

Problem 4.3 - Popular Locations

(2 points possible)

R will remember the other categories of the LocationDescription variable from the original dataset, so running table(Top5$LocationDescription) will have a lot of unnecessary output. To make our tables a bit nicer to read, we

can refresh this factor variable. In your R console, type:

Top5$LocationDescription = factor(Top5$LocationDescription)

If you run the str or table function on Top5 now, you should see that LocationDescription now only has 5 values, as we expect.

Use the Top5 data frame to answer the remaining questions.

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 Problem 4.1.

<div style="border:1px solid #ccc; width:300px; height:80px;"></div> ?

You have used 0 of 3 submissions

Problem 4.4 - Popular Locations

(1 point possible)
On which day of the week do the most motor vehicle thefts at gas stations happen?

[dropdown ▼]   ?

You have used 0 of 2 submissions

Problem 4.5 - Popular Locations

(1 point possible)
On which day of the week do the fewest motor vehicle thefts in residential driveways happen?

[dropdown ▼]   ?

You have used 0 of 2 submissions

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

POWERED BY
OPENedX®