edX (https://www.edx.org)

MITx: 15.071x The Analytics Edge

sandipan_dey (/dashboard) ▼

Courseware (/courses/MITx/15.071x/1T2014/courseware)   Course Info (/courses/MITx/15.071x/1T2014/info)

Discussion (/courses/MITx/15.071x/1T2014/discussion/forum)   Progress (/courses/MITx/15.071x/1T2014/progress)

Syllabus (/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)

Schedule (/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

Help

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 (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2).

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 (/c4x/MITx/15.071x/asset/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 (1/1 point)

Read the dataset mvtWeek1.csv (/c4x/MITx/15.071x/asset/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?

191641

191641

Answer: 191641

> **EXPLANATION**
>
> If you type str(mvt) in the R console, the first row of output says that this is a data frame with 191,641 observations.

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 1.2 - LOADING THE DATA (1/1 point)

How many variables are in this dataset?

11

11

Answer: 11

> **EXPLANATION**
>
> If you type str(mvt) in the R console, the first row of output says that this is a data frame with 11 variables.

Hide Answer   *You have used 1 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"?

9181151

9181151

Answer: 9181151

Hide Answer   *You have used 0 of 3 submissions*

PROBLEM 1.4 - LOADING THE DATA (1/1 point)

What is the minimum value of the variable "Beat"?

111

111

Answer: 111

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 1.5 - LOADING THE DATA (1/1 point)

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

15536

15536

Answer: 15536

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 1.6 - LOADING THE DATA (1/1 point)

How many observations have value a LocationDescription value of ALLEY?

2308

2308

Answer: 2308

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 2.1 - UNDERSTANDING DATES IN R (1/1 point)

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

EXPLANATION

If you type mvt$Date[1] in your R console, you can see that the first entry is 12/31/12 23:15. This must be in the format Month/Day/Year Hour:Minute.

Hide Answer   *You have used 1 of 1 submissions*

PROBLEM 2.2 - UNDERSTANDING DATES IN R (1/1 point)

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.)

| May 2006 | **Answer:** May 2006

EXPLANATION

If you type summary(DateConvert), you can see that the median date is 2006-05-21.

Hide Answer   *You have used 2 of 3 submissions*

PROBLEM 2.3 - UNDERSTANDING DATES IN R (1/1 point)

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?

[February ▼]   February

Hide Answer   *You have used 1 of 2 submissions*

PROBLEM 2.4 - UNDERSTANDING DATES IN R (1/1 point)

On which weekday did the most motor vehicle thefts occur?

Friday ▼   Friday

Hide Answer   *You have used 1 of 2 submissions*

PROBLEM 2.5 - UNDERSTANDING DATES IN R (1/1 point)

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 ▼   January

Hide Answer   *You have used 1 of 2 submissions*

PROBLEM 3.1 - VISUALIZING CRIME TRENDS (3/3 points)

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 (http://www.stat.berkeley.edu/classes/s133/saving.html) 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.

Does it look like crime increases or decreases from 2002 - 2012?

○ Increases

◉ Decreases ✔

Does it look like crime increases or decreases from 2005 - 2008?

○ Increases

○ Decreases ✔

> **EXPLANATION**
>
> In this time period, there is a clear downward trend in crime.

Does it look like crime increases or decreases from 2009 - 2011?

◉ Increases ✔
○ Decreases

> **EXPLANATION**
>
> In this time period, there is a clear upward trend in crime.

Hide Answer   *You have used 1 of 1 submissions*

PROBLEM 3.2 - VISUALIZING CRIME TRENDS (1/1 point)

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 (http://msenux.redwoods.edu/math/R/boxplot.php)). 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.

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

> **EXPLANATION**
>
> You can create the boxplot with the command boxplot(mvt$Date ~ mvt$Arrest). If you look at the boxplot, the one for Arrest=TRUE is definitely skewed towards the bottom of the plot, meaning that there were more crimes for which arrests were made in the first half of the time period.

Hide Answer   *You have used 1 of 1 submissions*

PROBLEM 3.3 - VISUALIZING CRIME TRENDS (1/1 point)

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.

0.1041173 ✔

0.1041173

Answer: 0.1041173

> **EXPLANATION**

If you create a table using the command table(mvt$Arrest, mvt$Year), the column for 2001 has 2152 observations with Arrest=TRUE and 18517 observations with Arrest=FALSE. The fraction of motor vehicle thefts in 2001 for which an arrest was made is thus 2152/(2152+18517) = 0.1041173.

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 3.4 - VISUALIZING CRIME TRENDS (1/1 point)

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

0.08487395

0.08487395

Answer: 0.08487395

> **EXPLANATION**
>
> If you create a table using the command table(mvt$Arrest, mvt$Year), the column for 2007 has 1212 observations with Arrest=TRUE and 13068 observations with Arrest=FALSE. The fraction of motor vehicle thefts in 2007 for which an arrest was made is thus 1212/(1212+13068) = 0.08487395.

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 3.5 - VISUALIZING CRIME TRENDS (1/1 point)

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

0.03902924

0.03902924

Answer: 0.03902924

> **EXPLANATION**
>
> If you create a table using the command table(mvt$Arrest, mvt$Year), the column for 2012 has 550 observations with Arrest=TRUE and 13542 observations with Arrest=FALSE. The fraction of motor vehicle thefts in 2012 for which an arrest was made is thus 550/(550+13542) = 0.03902924.

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.

Hide Answer   *You have used 1 of 3 submissions*

PROBLEM 4.1 - POPULAR LOCATIONS (1/1 point)

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
- [x] Gas Station ✔
- [ ] Hotel/Motel
- [x] Street ✔
- [ ] Car Wash
- [ ] Restaurant
- [x] Parking Lot/Garage (Non-Residential) ✔
- [x] Alley ✔
- [x] Driveway (Residential) ✔
- [ ] Vacant Lot/Land

EXPLANATION

If you type sort(table(mvt$LocationDescription)), the locations with the largest number of motor vehicle thefts are listed last. These are Street, Parking Lot/Garage (Non. Resid.), Alley, Gas Station, and Driveway - Residential.

Hide Answer   *You have used 1 of 2 submissions*

PROBLEM 4.2 - POPULAR LOCATIONS (1/1 point)

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?

    177510

    177510

Answer: 177510

EXPLANATION

You can create this subset with the command:

Top5 = subset(mvt, LocationDescription=="STREET" | LocationDescription=="PARKING LOT/GARAGE(NON.RESID.)" | LocationDescription=="ALLEY" | LocationDescription=="GAS STATION" | LocationDescription=="DRIVEWAY - RESIDENTIAL")

If you look at the structure of this data frame with str(Top5), you can see that there are 177510 observations.

Another way of doing this would be to use the %in% operator in R. This operator checks for inclusion in a set. You can create the same subset by typing the following two lines in your R console:

TopLocations = c("STREET", "PARKING LOT/GARAGE(NON.RESID.)", "ALLEY", "GAS STATION", "DRIVEWAY - RESIDENTIAL")

Top5 = subset(mvt, LocationDescription %in% TopLocations)

Hide Answer   *You have used 2 of 3 submissions*

PROBLEM 4.3 - POPULAR LOCATIONS (1/1 point)

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 function again on Top5, 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.

Gas Station

Answer: Gas Station

EXPLANATION

If you create a table of LocationDescription compared to Arrest, table(Top5$LocationDescription, Top5$Arrest), you can then compute the fraction of motor vehicle thefts that resulted in arrests at each location. Gas Station has by far the highest percentage of arrests, with over 20% of motor vehicle thefts resulting in an arrest.

Hide Answer   *You have used 2 of 2 submissions*

PROBLEM 4.4 - POPULAR LOCATIONS (1/1 point)

On which day of the week do the most motor vehicle thefts at gas stations happen?

Saturday ▾   Saturday

EXPLANATION

This can be read from table(Top5$LocationDescription, Top5$Weekday).

Hide Answer   *You have used 1 of 2 submissions*

PROBLEM 4.5 - POPULAR LOCATIONS (1/1 point)

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

Saturday ▾   Saturday

EXPLANATION

This can be read from table(Top5$LocationDescription, Top5$Weekday).

Hide Answer   *You have used 1 of 2 submissions*

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

Show Discussion   ✎ New Post