**QUESTION 1**

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

* 340159

**QUESTION 2**

How many variables are in this dataset?

* 10

**QUESTION 3**

Using the "max" function, what is the maximum value of the variable "ID"?

* 12457898

**QUESTION 4**

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

* 2001

**QUESTION 5**

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

* 29960

**QUESTION 6**

How many observations have a LocationDescription value of “ALLEY”?

* 5732

**QUESTION 7**

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, in the CrimeData, are the entries in the variable Date?

1. Month/Day/Year Hour:Minute
2. Day/Month/Year Hour:Minute
3. Hour:Minute Month/Day/Year
4. Hour:Minute Day/Month/Year

**QUESTION 8**

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

DateConvert = as.Date(strptime(CrimeData$Date, "%m/%d/%Y %I:%M:%S %p"))

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 2012-03-28, you would give the answer "March 2012", without the quotes.)

* June 2008

**QUESTION 9**

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

CrimeData$Month = months(DateConvert)

CrimeData$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:

CrimeData$Date = DateConvert

Now answer this question: In which month did the fewest crime occur?

* February

**QUESTION 10**

On which weekday did the most crime occur?

* Friday

**QUESTION 11**

Each observation in the dataset represents a theft, and the Arrest variable indicates whether an arrest was later made for this theft. Which month has the largest number of thefts that arrests were made?

* January

**QUESTION 12**

Now find out how many of those thefts are related to AUTOMOBILE?

* 1963

**QUESTION 13**

What is the percentage of ARRESTS for the crime of AUTOMOBILE theft? Use up to one decimal point, e.g., use 12.3 for 0.123.

* 70.7

**QUESTION 14**

Now, let's make some plots to help us better understand how crime has changed over time in Chicago. In this problem, you need to answer three questions.

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(CrimeData$Date, breaks=30)

Looking at the histogram, answer Question (14a) here.

In general, does it look like the crime of thefts increases or decreases from 2002 - 2009?

1. Decrease
2. Increase

**QUESTION 15**

In general, does it look like the crime of thefts increases or decreases from 2009 - 2010?

1. Increase
2. Decrease

**QUESTION 16**

In general, does it look like the crime of thefts increases or decreases from 2013 - 2020?

1. flat
2. Increase
3. Decrease

**QUESTION 17**

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 of thefts 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 2021, so the middle of the time period is the beginning of 2011.)

1. Second Half
2. First Half

**QUESTION 18**

Let's investigate this further. Use the table function for the next few questions.

For what proportion of the thefts in 2020 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. Use up to 4 digits after the decimal point.

* .0470

**QUESTION 19**

For what proportion of the theft crimes in 2007 was an arrest made?

Again, use 4 digits after the decimal point.

* .0957

**QUESTION 20**

For what proportion of AUTOMOBILE thefts in 2020 was an arrest made?

Use 4 digits after the decimal point.

* .0400

**QUESTION 21**

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 AUTOMOBILE 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(CrimeData[CrimeData$Description == "AUTOMOBILE", "LocationDescription"]), decreasing = T)

Which locations are the top five locations for motor vehicle thefts, excluding the "Other" category? You should select 5 of the following options.

RESTAURANT

GAS STATION

CAR WASH

STREET

HOTEL/MOTEL

BANK

OTHER

ALLEY

VACANT LOT/LAND

PARKING LOT/GARAGE (NON. RESID.)

**QUESTION 22**

First, change the LocationDescription variable as a factor variable by using

CrimeData$LocationDescription = as.factor(CrimeData$LocationDescription)

Then try to 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 may use the | symbol, which means "or" or "union", when you add your conditions.

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

How many observations are in those top 5 locations?

Keep in mind that you are still investigating for AUTOMOBILE only.

* 248043

**QUESTION 23**

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 levels of that variable. 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.

* GAS STATION

**QUESTION 24**

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

* Sunday

**QUESTION 25**

On which day of the week do the fewest AUTOMOBILE thefts in STREET happen?

* Tuesday