Problem 1.1 - Loading the Data

Read the dataset [mvtWeek1.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/96f9b8f751467da3a4b8a5be33e32905/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/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?

Problem 1.2 - Loading the Data

How many variables are in this dataset?

Problem 1.3 - Loading the Data

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

Problem 1.4 - Loading the Data

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

Problem 1.5 - Loading the Data

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

Problem 1.6 - Loading the Data

How many observations have a LocationDescription value of ALLEY?

Problem 2.1 - Understanding Dates in R

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

Top of Form

Bottom of Form

Problem 2.2 - Understanding Dates in R

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.)

Problem 2.3 - Understanding Dates in R

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?

Top of Form

Problem 2.4 - Understanding Dates in R

On which weekday did the most motor vehicle thefts occur?

Top of Form

Problem 2.5 - Understanding Dates in R

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?

Problem 1.1 - Summary Statistics

Download and read the following files into R, using the read.csv function: [IBMStock.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/4fc08d10f171aacf2ef61c6b4b5bb4d8/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/IBMStock.csv), [GEStock.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/448b8be4693d913c2b5153be0c0e25d6/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/GEStock.csv),[ProcterGambleStock.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/bb6ed54230b5b2e29fb66819ed535da0/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/ProcterGambleStock.csv), [CocaColaStock.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/4c30fd7f4f55e537989ca13a6db36289/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/CocaColaStock.csv), and [BoeingStock.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/2e8c9fb294db48e5a999c747b317722d/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/BoeingStock.csv). (Do not open these files in any spreadsheet software before completing this problem because it might change the format of the Date field.)

Call the data frames "IBM", "GE", "ProcterGamble", "CocaCola", and "Boeing", respectively. Each data frame has two variables, described as follows:

* **Date**: the date of the stock price, always given as the first of the month.
* **StockPrice**: the average stock price of the company in the given month.

In this problem, we'll take a look at how the stock dynamics of these companies have changed over time.

Before working with these data sets, we need to convert the dates into a format that R can understand. Take a look at the structure of one of the datasets using the str function. Right now, the date variable is stored as a factor. We can convert this to a "Date" object in R by using the following five commands (one for each data set):

IBM$Date = as.Date(IBM$Date, "%m/%d/%y")

GE$Date = as.Date(GE$Date, "%m/%d/%y")

CocaCola$Date = as.Date(CocaCola$Date, "%m/%d/%y")

ProcterGamble$Date = as.Date(ProcterGamble$Date, "%m/%d/%y")

Boeing$Date = as.Date(Boeing$Date, "%m/%d/%y")

The first argument to the as.Date function is the variable we want to convert, and the second argument is the format of the Date variable. We can just overwrite the original Date variable values with the output of this function. Now, answer the following questions using the str and summary functions.

Our five datasets all have the same number of observations. How many observations are there in each data set?

Problem 1.2 - Summary Statistics

What is the earliest year in our datasets?

Problem 1.3 - Summary Statistics

What is the latest year in our datasets?

Problem 1.4 - Summary Statistics

What is the mean stock price of IBM over this time period?

Problem 1.5 - Summary Statistics

What is the minimum stock price of General Electric (GE) over this time period?

Problem 1.6 - Summary Statistics

What is the maximum stock price of Coca-Cola over this time period?

Problem 1.7 - Summary Statistics

What is the median stock price of Boeing over this time period?

Problem 1.8 - Summary Statistics

What is the standard deviation of the stock price of Procter & Gamble over this time period?

## DEMOGRAPHICS AND EMPLOYMENT IN THE UNITED STATES

The observations in the dataset represent people surveyed in the September 2013 CPS who actually completed a survey. While the full dataset has 385 variables, in this exercise we will use a more compact version of the dataset, [CPSData.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/f041f6c100061fc06bc3b6320e6512fa/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/CPSData.csv), which has the following variables:

**PeopleInHousehold**: The number of people in the interviewee's household.

**Region**: The census region where the interviewee lives.

**State**: The state where the interviewee lives.

**MetroAreaCode**: A code that identifies the metropolitan area in which the interviewee lives (missing if the interviewee does not live in a metropolitan area). The mapping from codes to names of metropolitan areas is provided in the file [MetroAreaCodes.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/fd88455abc1b5b69112daf70f3bb0c77/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/MetroAreaCodes.csv).

**Age**: The age, in years, of the interviewee. 80 represents people aged 80-84, and 85 represents people aged 85 and higher.

**Married**: The marriage status of the interviewee.

**Sex**: The sex of the interviewee.

**Education**: The maximum level of education obtained by the interviewee.

**Race**: The race of the interviewee.

**Hispanic**: Whether the interviewee is of Hispanic ethnicity.

**CountryOfBirthCode**: A code identifying the country of birth of the interviewee. The mapping from codes to names of countries is provided in the file [CountryCodes.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/763710fa6703caea1cf9c708e31e99a3/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/CountryCodes.csv).

**Citizenship**: The United States citizenship status of the interviewee.

**EmploymentStatus**: The status of employment of the interviewee.

**Industry**: The industry of employment of the interviewee (only available if they are employed).

### Problem 1.1 - Loading and Summarizing the Dataset

Load the dataset from [CPSData.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/f041f6c100061fc06bc3b6320e6512fa/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/CPSData.csv) into a data frame called CPS, and view the dataset with the summary() and str() commands.

How many interviewees are in the dataset?

### Problem 1.2 - Loading and Summarizing the Dataset

Among the interviewees with a value reported for the Industry variable, what is the most common industry of employment? Please enter the name exactly how you see it.

### Problem 1.3 - Loading and Summarizing the Dataset

You can call the sort() function on the output of the table() function to obtain a sorted breakdown of a variable. For instance, sort(table(CPS$Region)) sorts the regions by the number of interviewees from that region.

Which state has the fewest interviewees?

Which state has the largest number of interviewees?

### Problem 1.4 - Loading and Summarizing the Dataset

What proportion of interviewees are citizens of the United States?

### Problem 1.5 - Loading and Summarizing the Dataset

The CPS differentiates between race (with possible values American Indian, Asian, Black, Pacific Islander, White, or Multiracial) and ethnicity. A number of interviewees are of Hispanic ethnicity, as captured by the Hispanic variable. For which races are there at least 250 interviewees in the CPS dataset of Hispanic ethnicity? (Select all that apply.)

Top of Form

American Indian

Asian

Black

Multiracial

Pacific Islander

White

### Problem 2.1 - Evaluating Missing Values

Which variables have at least one interviewee with a missing (NA) value? (Select all that apply.)

Top of Form

PeopleInHousehold

Region

State

MetroAreaCode

Age

Married

Sex

Education

Race

Hispanic

CountryOfBirthCode

Citizenship

EmploymentStatus

Industry

### Problem 2.2 - Evaluating Missing Values

Often when evaluating a new dataset, we try to identify if there is a pattern in the missing values in the dataset. We will try to determine if there is a pattern in the missing values of the Married variable. The function is.na(CPS$Married) returns a vector of TRUE/FALSE values for whether the Married variable is missing. We can see the breakdown of whether Married is missing based on the reported value of the Region variable with the function table(CPS$Region, is.na(CPS$Married)). Which is the most accurate:

Top of Form

The Married variable being missing is related to the Region value for the interviewee.

The Married variable being missing is related to the Sex value for the interviewee.

The Married variable being missing is related to the Age value for the interviewee. correct

The Married variable being missing is related to the Citizenship value for the interviewee.

The Married variable being missing is not related to the Region, Sex, Age, or Citizenship value for the interviewee.Bottom of Form

### Problem 2.3 - Evaluating Missing Values

As mentioned in the variable descriptions, MetroAreaCode is missing if an interviewee does not live in a metropolitan area. Using the same technique as in the previous question, answer the following questions about people who live in non-metropolitan areas.

How many states had all interviewees living in a non-metropolitan area (aka they have a missing MetroAreaCode value)? For this question, treat the District of Columbia as a state (even though it is not technically a state).

How many states had all interviewees living in a metropolitan area? Again, treat the District of Columbia as a state.

### Problem 2.4 - Evaluating Missing Values

Which region of the United States has the largest proportion of interviewees living in a non-metropolitan area?

Top of Form

Midwest correct

Northeast

South

WestBottom of Form

### Problem 2.5 - Evaluating Missing Values

While we were able to use the table() command to compute the proportion of interviewees from each region not living in a metropolitan area, it was somewhat tedious (it involved manually computing the proportion for each region) and isn't something you would want to do if there were a larger number of options. It turns out there is a less tedious way to compute the proportion of values that are TRUE. The mean() function, which takes the average of the values passed to it, will treat TRUE as 1 and FALSE as 0, meaning it returns the proportion of values that are true. For instance, mean(c(TRUE, FALSE, TRUE, TRUE)) returns 0.75. Knowing this, use tapply() with the mean function to answer the following questions:

Which state has a proportion of interviewees living in a non-metropolitan area closest to 30%?

Which state has the largest proportion of non-metropolitan interviewees, ignoring states where all interviewees were non-metropolitan?

### Problem 3.1 - Integrating Metropolitan Area Data

Codes like MetroAreaCode and CountryOfBirthCode are a compact way to encode factor variables with text as their possible values, and they are therefore quite common in survey datasets. In fact, all but one of the variables in this dataset were actually stored by a numeric code in the original CPS datafile.

When analyzing a variable stored by a numeric code, we will often want to convert it into the values the codes represent. To do this, we will use a dictionary, which maps the the code to the actual value of the variable. We have provided dictionaries [MetroAreaCodes.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/fd88455abc1b5b69112daf70f3bb0c77/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/MetroAreaCodes.csv) and [CountryCodes.csv](https://d37djvu3ytnwxt.cloudfront.net/assets/courseware/v1/763710fa6703caea1cf9c708e31e99a3/asset-v1:MITx+15.071x_3+1T2016+type@asset+block/CountryCodes.csv), which respectively map MetroAreaCode and CountryOfBirthCode into their true values. Read these two dictionaries into data frames MetroAreaMap and CountryMap.

How many observations (codes for metropolitan areas) are there in MetroAreaMap?

How many observations (codes for countries) are there in CountryMap?

### Problem 3.2 - Integrating Metropolitan Area Data

To merge in the metropolitan areas, we want to connect the field MetroAreaCode from the CPS data frame with the field Code in MetroAreaMap. The following command merges the two data frames on these columns, overwriting the CPS data frame with the result:

CPS = merge(CPS, MetroAreaMap, by.x="MetroAreaCode", by.y="Code", all.x=TRUE)

The first two arguments determine the data frames to be merged (they are called "x" and "y", respectively, in the subsequent parameters to the merge function). by.x="MetroAreaCode" means we're matching on the MetroAreaCode variable from the "x" data frame (CPS), while by.y="Code" means we're matching on the Code variable from the "y" data frame (MetroAreaMap). Finally, all.x=TRUE means we want to keep all rows from the "x" data frame (CPS), even if some of the rows' MetroAreaCode doesn't match any codes in MetroAreaMap (for those familiar with database terminology, this parameter makes the operation a left outer join instead of an inner join).

Review the new version of the CPS data frame with the summary() and str() functions. What is the name of the variable that was added to the data frame by the merge() operation?

How many interviewees have a missing value for the new metropolitan area variable? Note that all of these interviewees would have been removed from the merged data frame if we did not include the all.x=TRUE parameter.

### Problem 3.3 - Integrating Metropolitan Area Data

Which of the following metropolitan areas has the largest number of interviewees?

Top of Form

Atlanta-Sandy Springs-Marietta, GA

Baltimore-Towson, MD

Boston-Cambridge-Quincy, MA-NH correct

San Francisco-Oakland-Fremont, CABottom of Form

### Problem 3.4 - Integrating Metropolitan Area Data

Which metropolitan area has the highest proportion of interviewees of Hispanic ethnicity? Hint: Use tapply() with mean, as in the previous subproblem. Calling sort() on the output of tapply() could also be helpful here.

### Problem 3.5 - Integrating Metropolitan Area Data

Remembering that CPS$Race == "Asian" returns a TRUE/FALSE vector of whether an interviewee is Asian, determine the number of metropolitan areas in the United States from which at least 20% of interviewees are Asian.

### Problem 3.6 - Integrating Metropolitan Area Data

Normally, we would look at the sorted proportion of interviewees from each metropolitan area who have not received a high school diploma with the command:

sort(tapply(CPS$Education == "No high school diploma", CPS$MetroArea, mean))

However, none of the interviewees aged 14 and younger have an education value reported, so the mean value is reported as NA for each metropolitan area. To get mean (and related functions, like sum) to ignore missing values, you can pass the parameter na.rm=TRUE. Passing na.rm=TRUE to the tapply function, determine which metropolitan area has the smallest proportion of interviewees who have received no high school diploma.

### Problem 4.1 - Integrating Country of Birth Data

Just as we did with the metropolitan area information, merge in the country of birth information from the CountryMap data frame, replacing the CPS data frame with the result. If you accidentally overwrite CPS with the wrong values, remember that you can restore it by re-loading the data frame from CPSData.csv and then merging in the metropolitan area information using the command provided in the previous subproblem.

What is the name of the variable added to the CPS data frame by this merge operation?

.

How many interviewees have a missing value for the new country of birth variable?

### Problem 4.2 - Integrating Country of Birth Data

Among all interviewees born outside of North America, which country was the most common place of birth?

### Problem 4.3 - Integrating Country of Birth Data

What proportion of the interviewees from the "New York-Northern New Jersey-Long Island, NY-NJ-PA" metropolitan area have a country of birth that is not the United States? For this computation, don't include people from this metropolitan area who have a missing country of birth.

### Problem 4.4 - Integrating Country of Birth Data

Which metropolitan area has the largest number (note -- not proportion) of interviewees with a country of birth in India? Hint -- remember to include na.rm=TRUE if you are using tapply() to answer this question.

Top of Form

Boston-Cambridge-Quincy, MA-NH

Minneapolis-St Paul-Bloomington, MN-WI

New York-Northern New Jersey-Long Island, NY-NJ-PA

Washington-Arlington-Alexandria, DC-VA-MD-WVNot yet answered.

Bottom of Form

In Brazil?

Top of Form

Boston-Cambridge-Quincy, MA-NH

Minneapolis-St Paul-Bloomington, MN-WI

New York-Northern New Jersey-Long Island, NY-NJ-PA

Washington-Arlington-Alexandria, DC-VA-MD-WVNot yet answered.

Bottom of Form

In Somalia?

Top of Form

Boston-Cambridge-Quincy, MA-NH

Minneapolis-St Paul-Bloomington, MN-WI

New York-Northern New Jersey-Long Island, NY-NJ-PA

Washington-Arlington-Alexandria, DC-VA-MD-WV

Top of Form