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.

Help

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 (1/1 point)

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?

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.

Check Save 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

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.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.3 - LOADING THE DATA (1/1 point)

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

9181151

Answer: 9181151

EXPLANATION

You can compute the maximum value of the ID variable with max(mvt$ID).

Check Save Hide Answer *You have used 1 of 3 submissions*

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

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

111

Answer: 111

EXPLANATION

If you type summary(mvt) in your R console, you can see the summary statistics for each variable. This shows that the minimum value of Beat is 111. Alternatively, you could use the min function by typing min(mvt$Beat).

Check Save 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

Answer: 15536

EXPLANATION

If you type summary(mvt) in your R console, you can see the summary statistics for each variable. This shows that 15,536 observations fall under the category TRUE for the variable Arrest.

Check Save Hide Answer *You have used 1 of 3 submissions*

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

How many observations have a LocationDescription value of ALLEY?

2308

2308

Answer: 2308

EXPLANATION

If you type summary(mvt) in your R console, you can see the summary statistics for each variable. This shows that 2,308 observations fall under the category ALLEY for the variable LocationDescription. You can also read this from table(mvt$LocationDescription).

Check Save 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.

Check Save Hide Answer *You have used 1 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

EXPLANATION

If you type table(mvt$Month), you can see that the month with the smallest number of observations is February.

Hide Answer *You have used 2 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

EXPLANATION

If you type table(mvt$Weekday), you can see that the weekday with the largest number of observations is Friday.

Final Check | Save | 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

EXPLANATION

If you type table(mvt$Arrest,mvt$Month), you can see that the largest number of observations with Arrest=TRUE occurs in the month of January.

Final Check | Save | 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 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 ✓

EXPLANATION

While there is not a clear trend, it looks like crime generally decreases.

In general, 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.

In general, 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 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 ✗

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 (2/2 points)

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.

Check Save Hide Answer *You have used 3 of 5 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 **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.

Check Save 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 **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.

Check Save 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

☑ Gas Station ✓

□ Hotel/Motel

☑ Street ✓

□ Car Wash

□ Restaurant

☑ Parking Lot/Garage (Non-Residential) ✓

☑ Alley ✓

☑ 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.

Final Check | Save | 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

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)
```

Check | Save | Hide Answer | *You have used 1 of 3 submissions*

PROBLEM 4.3 - POPULAR LOCATIONS (2/2 points)

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.

Final Check Save Hide Answer *You have used 2 of 3 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).

Final Check Save 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).

Final Check Save 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

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

About edX

About

News

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Follow Us

Twitter

Facebook

Meetup

LinkedIn

Google+

MITx: 15.071x The Analytics Edge

parmarmanojkumar

Courseware Course Info Discussion Progress Syllabus Schedule Files Wiki

STOCK DYNAMICS

A stock market is where buyers and sellers trade shares of a company, and is one of the most popular ways for individuals and companies to invest money. The size of the world stock market is now estimated to be in the trillions. The largest stock market in the world is the New York Stock Exchange (NYSE), located in New York City. About 2,800 companies are listed on the NSYE. In this problem, we'll look at the monthly stock prices of five of these companies: IBM, General Electric (GE), Procter and Gamble, Coca Cola, and Boeing. The data used in this problem comes from Infochimps.

Download and read the following files into R, using the read.csv function: IBMStock.csv, GEStock.csv, ProcterGambleStock.csv, CocaColaStock.csv, and 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.

PROBLEM 1.1 - SUMMARY STATISTICS (1/1 point)

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?

480

Answer: 480

EXPLANATION

Using the str function, we can see that each data set has 480 observations. We have monthly data for 40 years, so there are 12*40 = 480 observations.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.2 - SUMMARY STATISTICS (1/1 point)

What is the earliest year in our datasets?

1970

Answer: 1970

EXPLANATION

Using the summary function, the minimum value of the Date variable is January 1, 1970 for any dataset.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.3 - SUMMARY STATISTICS (1/1 point)

What is the latest year in our datasets?

2009

Answer: 2009

EXPLANATION

Using the summary function, the maximum value of the Date variable is December 1, 2009 for any dataset.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.4 - SUMMARY STATISTICS (1/1 point)

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

144.38

Answer: 144.38

EXPLANATION

By typing summary(IBM), we can see that the mean value of the IBM StockPrice is 144.38.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.5 - SUMMARY STATISTICS (1/1 point)

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

9.294

Answer: 9.294

EXPLANATION

By typing summary(GE), we can see that the minimum value of the GE StockPrice is 9.294.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.6 - SUMMARY STATISTICS (1/1 point)

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

146.58

Answer: 146.58

EXPLANATION

By typing summary(CocaCola), we can see that the maximum value of the Coca-Cola StockPrice is 146.58.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.7 - SUMMARY STATISTICS (1/1 point)

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

44.88

Answer: 44.88

EXPLANATION

By typing summary(Boeing), we can see that the median value of the Boeing StockPrice is 44.88.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.8 - SUMMARY STATISTICS (1/1 point)

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

18.19414

Answer: 18.19414

EXPLANATION

By typing sd(ProcterGamble$StockPrice), we can see that the standard deviation of the Procter & Gamble StockPrice is 18.19414.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 2.1 - VISUALIZING STOCK DYNAMICS (2/2 points)

Let's plot the stock prices to see if we can visualize trends in stock prices during this time period. Using the plot function, plot the Date on the x-axis and the StockPrice on the y-axis, for Coca-Cola.

This plots our observations as points, but we would really like to see a line instead, since this is a continuous time period. To do this, add the argument type="l" to your plot command, and re-generate the plot (the character is quotes is the letter l, for line). You should now see a line plot of the Coca-Cola stock price.

Around what year did Coca-Cola has its highest stock price in this time period?

● 1973 ✓
○ 1980
○ 1985
○ 1995
○ 2008

Around what year did Coca-Cola has its lowest stock price in this time period?

○ 1973
● 1980 ✓
○ 1985
○ 1995
○ 2008

EXPLANATION

You can generate the plot using the command plot(CocaCola$Date, CocaCola$StockPrice, type="l"). Looking at the plot, the maximum value of the StockPrice is around 1973, and the minimum value of the StockPrice is around 1980.

Final Check | Save | Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.2 - VISUALIZING STOCK DYNAMICS (1 point possible)

Now, let's add the line for Procter & Gamble too. You can add a line to a plot in R by using the lines function instead of the plot function. Keeping the plot for Coca-Cola open, type in your R console:

lines(ProcterGamble$Date, ProcterGamble$StockPrice)

Unfortunately, it's hard to tell which line is which. Let's fix this by giving each line a color. First, re-run the plot command for Coca-Cola, but add the argument col="red". You should see the plot for Coca-Cola show up again, but this time in red. Now, let's add the Procter & Gamble line (using the lines function like we did before), adding the argument col="blue". You should now see in your plot the Coca-Cola stock price in red, and the Procter & Gamble stock price in blue.

As an alternative choice to changing the colors, you could instead change the line type of the Procter & Gamble line by adding the argument lty=2. This will make the Procter & Gamble line dashed.

Using this plot, answer the following questions.

In March of 2000, the technology bubble burst, and a stock market crash occurred. According to this plot, which company's stock dropped more?

● Coca-Cola ✗
○ Procter and Gamble ✓

EXPLANATION

You can generate the combined plot for Coca-Cola and Procter & Gamble by using the following commands in R:

```
plot(CocaCola$Date, CocaCola$StockPrice, type="l", col="red")
```

```
lines(ProcterGamble$Date, ProcterGamble$StockPrice, col="blue")
```

Looking at the plot, around 2000 both stocks drop, but Procter and Gamble's stock drops more.

To answer this question and the ones that follow, you may find it useful to draw a vertical line at a certain date. To do this, type the command

```
abline(v=as.Date(c("2000-03-01")), lwd=2)
```

in your R console, with the plot still open. This generates a vertical line at the date March 1, 2000. The argument lwd=2 makes the line a little thicker. You can change the date in this command to generate the vertical line in different locations.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 2.3 - VISUALIZING STOCK DYNAMICS (2/2 points)

Answer these questions using the plot you generated in the previous problem.

Around 1983, the stock for one of these companies (Coca-Cola or Procter and Gamble) was going up, while the other was going down. Which one was going up?

● Coca-Cola ✓
○ Procter and Gamble

EXPLANATION

Looking at the plot generated by the commands:

```
plot(CocaCola$Date, CocaCola$StockPrice, type="l", col="red")
```

```
lines(ProcterGamble$Date, ProcterGamble$StockPrice, col="blue")
```

we can see that around 1983 the stock for Coca-Cola has an upward trend.

In the time period shown in the plot, which stock generally has lower values?

● Coca-Cola ✓

○ Procter and Gamble

EXPLANATION

Looking at the plot, the red line (for Coca-Cola) is generally lower than the blue line.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 3.1 - VISUALIZING STOCK DYNAMICS 1995-2005 (1/1 point)

Let's take a look at how the stock prices changed from 1995-2005 for all five companies. In your R console, start by typing the following plot command:

```
plot(CocaCola$Date[301:432], CocaCola$StockPrice[301:432], type="l", col="red", ylim=c(0,210))
```

This will plot the CocaCola stock prices from 1995 through 2005, which are the observations numbered from 301 to 432. The additional argument, ylim=c(0,210), makes the y-axis range from 0 to 210. This will allow us to see all of the stock values when we add in the other companies.

Now, use the lines function to add in the other four companies, remembering to only plot the observations from 1995 to 2005, or [301:432]. You don't need the "type" or "ylim" arguments for the lines function, but remember to make each company a different color so that you can tell them apart. Some color options are "red", "blue", "green", "purple", "orange", and "black". To see all of the color options in R, type colors() in your R console.

(If you prefer to change the type of the line instead of the color, here are some options for changing the line type: lty=2 will make the line dashed, lty=3 will make the line dotted, lty=4 will make the line alternate between dashes and dots, and lty=5 will make the line long-dashed.)

Use this plot to answer the following four questions.

Which stock fell the most right after the technology bubble burst in March 2000?

○ Coca-Cola

○ Procter and Gamble

○ IBM

● General Electric (GE) ✓

○ Boeing

EXPLANATION

You can create the plot needed to answer the questions in this problem by typing the following commands into your R console:

```
plot(CocaCola$Date[301:432], CocaCola$StockPrice[301:432], type="l", col="red", ylim=c(0,210))
```

```
lines(ProcterGamble$Date[301:432], ProcterGamble$StockPrice[301:432], col="blue")
lines(IBM$Date[301:432], IBM$StockPrice[301:432], col="green")

lines(GE$Date[301:432], GE$StockPrice[301:432], col="purple")

lines(Boeing$Date[301:432], Boeing$StockPrice[301:432], col="orange")
```

You can add a vertical line to the plot at March 2000 by typing the following command:

```
abline(v=as.Date(c("2000-03-01")), lwd=2)
```

By looking at this plot, you can see that the stock for General Electric falls significantly more than the other stocks after the technology bubble burst.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 3.2 - VISUALIZING STOCK DYNAMICS 1995-2005 (1/1 point)

Which stock reaches the highest value in the time period 1995-2005?

○ Coca-Cola
○ Procter and Gamble
● IBM ✓
○ General Electric (GE)
○ Boeing

EXPLANATION

Looking at the plot (see the previous explanation for how to create the plot), you can see that IBM has the highest value, around 1999.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 3.3 - VISUALIZING STOCK DYNAMICS 1995-2005 (1 point possible)

In October of 1997, there was a global stock market crash that was caused by an economic crisis in Asia. Comparing September 1997 to November 1997, which companies saw a decreasing trend in their stock price? (Select all that apply.)

□ Coca-Cola
□ Procter and Gamble ✓
□ IBM
□ General Electric (GE)
☑ Boeing ✓

EXPLANATION

You can create vertical lines at September 1997 and November 1997 with the following commands:

abline(v=as.Date(c("1997-09-01")), lwd=2)

abline(v=as.Date(c("1997-11-01")), lwd=2)

Looking at the plot, two companies had a decreasing trend in stock prices from September 1997 to November 1997: Boeing and Procter & Gamble.

Hide Answer *You have used 2 of 2 submissions*

PROBLEM 3.4 - VISUALIZING STOCK DYNAMICS 1995-2005 (1/1 point)

In the last two years of this time period (2004 and 2005) which stock seems to be performing the best, in terms of increasing stock price?

○ Coca-Cola
○ Procter and Gamble
○ IBM
○ General Electric (GE)
● Boeing ✓

EXPLANATION

Looking at the plot, you can see that Boeing is steadily increasing from 2004 to the beginning of 2006.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4.1 - MONTHLY TRENDS (1/1 point)

Lastly, let's see if stocks tend to be higher or lower during certain months. Use the tapply command to calculate the mean stock price of IBM, sorted by months. To sort by months, use

months(IBM$Date)

as the second argument of the tapply function.

For IBM, compare the monthly averages to the overall average stock price. In which months has IBM historically had a higher stock price (on average)? Select all that apply.

☑ January ✓
☑ February ✓
☑ March ✓
☑ April ✓
☑ May ✓
□ June
□ July
□ August

☐ September
☐ October
☐ November
☐ December

EXPLANATION

The overall average stock price for IBM is 144.375, which can be computed using the command mean(IBM$StockPrice). Comparing the monthly averages to this, using the command tapply(IBM$StockPrice, months(IBM$Date), mean), we can see that the price has historically been higher than average January - May, and lower than average during the remaining months.

Final Check | Save | Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4.2 - MONTHLY TRENDS (1/1 point)

Repeat the tapply function from the previous problem for each of the other four companies, and use the output to answer the remaining questions.

General Electric and Coca-Cola both have their highest average stock price in the same month. Which month is this?

○ January
○ February
○ March
● April ✓
○ May
○ June
○ July
○ August
○ September
○ October
○ November
○ December

EXPLANATION

You can see the monthly average stock prices for GE and Coca-Cola by using the commands:

```
tapply(GE$StockPrice, months(GE$Date), mean)
```

```
tapply(CocaCola$StockPrice, months(CocaCola$Date), mean)
```

General Electric has an average stock price of 64.48 in April, which is higher than any other month. Coca-Cola has an average stock price of 62.69 in April, which is higher than any other month.

Final Check | Save | Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4.3 - MONTHLY TRENDS (1/1 point)

For the months of December and January, every company's average stock is higher in one month and lower in the other. In which month are the stock prices lower?

● December ✓

○ January

EXPLANATION

IBM has an average price of 140.76 in December, and 150.24 in January, which can be seen with the command:

tapply(IBM$StockPrice, months(IBM$Date), mean)

Having lower stock prices in December is a trend that holds for all five companies.

After seeing these trends, we are ready to buy stock in certain months and sell it in others! But, we should be careful, because one really good or really bad year could skew the average to show a trend that is not really there in general.

Hide Answer *You have used 1 of 1 submissions*

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

Show Discussion

New Post

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law,

About edX

About

News

Contact

FAQ

Follow Us

Twitter

Facebook

Meetup

literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

edX Blog

Donate to edX

Jobs at edX

LinkedIn

Google+

DEMOGRAPHICS AND EMPLOYMENT IN THE UNITED STATES

In the wake of the Great Recession of 2009, there has been a good deal of focus on employment statistics, one of the most important metrics policymakers use to gauge the overall strength of the economy. In the United States, the government measures unemployment using the Current Population Survey (CPS), which collects demographic and employment information from a wide range of Americans each month. In this exercise, we will employ the topics reviewed in the lectures as well as a few new techniques using the September 2013 version of this rich, nationally representative dataset (available online).

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, 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.

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.

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 (1/1 point)

Load the dataset from CPSData.csv into a data frame called CPS, and view the dataset with the summary() and str() commands.

EXPLANATION

You can load the data with:

```
CPS = read.csv("CPSData.csv")
```

How many interviewees are in the dataset?

131302

131302

Answer: 131302

EXPLANATION

From str(CPS), we can read that there are 131302 interviewees.

Check | Save | Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.2 - LOADING AND SUMMARIZING THE DATASET (1/1 point)

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.

Educational and health se

Answer: Educational and health services

EXPLANATION

The output of summary(CPS) orders the levels of a factor variable like Industry from largest to smallest, so we can see that "Educational and health services" is the most common Industry. table(CPS$Industry) would have provided the breakdown across all industries.

Final Check | Save | Hide Answer *You have used 1 of 2 submissions*

PROBLEM 1.3 - LOADING AND SUMMARIZING THE DATASET (2/2 points)

Recall from the homework assignment "The Analytical Detective" that 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?

New Mexico

Answer: New Mexico

Which state has the largest number of interviewees?

California

Answer: California

EXPLANATION

These can be read from sort(table(CPS$State))

Check | Save | Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.4 - LOADING AND SUMMARIZING THE DATASET (1/1 point)

What proportion of interviewees are citizens of the United States?

0.9421943

Answer: 0.942

EXPLANATION

From table(CPS$Citizenship), we see that 123,712 of the 131,302 interviewees are citizens of the United States (either native or naturalized). This is a proportion of 123712/131302=0.942.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.5 - LOADING AND SUMMARIZING THE DATASET (1/1 point)

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.)

☑ American Indian ✓
□ Asian
☑ Black ✓
☑ Multiracial ✓
□ Pacific Islander
☑ White ✓

EXPLANATION

The breakdown of race and Hispanic ethnicity can be obtained with table(CPS$Race, CPS$Hispanic).

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.1 - EVALUATING MISSING VALUES (1/1 point)

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

□ PeopleInHousehold
□ Region
□ State
☑ MetroAreaCode ✓
□ Age
☑ Married ✓
□ Sex
☑ Education ✓
□ Race
□ Hispanic
□ CountryOfBirthCode
□ Citizenship
☑ EmploymentStatus ✓
☑ Industry ✓

EXPLANATION

This can be read from the output of summary(CPS).

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.2 - EVALUATING MISSING VALUES (1/1 point)

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:

○ 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. ✓

○ 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.

EXPLANATION

We can test the relationship between these four variable values and whether the Married variable is missing with the following commands:

```
table(CPS$Region, is.na(CPS$Married))

table(CPS$Sex, is.na(CPS$Married))

table(CPS$Age, is.na(CPS$Married))

table(CPS$Citizenship, is.na(CPS$Married))
```

For each possible value of Region, Sex, and Citizenship, there are both interviewees with missing and non-missing Married values. However, Married is missing for all interviewees Aged 0-14 and is present for all interviewees aged 15 and older. This is because the CPS does not ask about marriage status for interviewees 14 and younger.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.3 - EVALUATING MISSING VALUES (2/2 points)

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).

2 **Answer:** 2

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

3 **Answer:** 3

EXPLANATION

The breakdown of missing MetroAreaCode by State can be obtained with table(CPS$State, is.na(CPS$MetroAreaCode)). Alaska and Wyoming have no interviewees living in a metropolitan area, and the District of Columbia, New Jersey, and Rhode Island have all interviewees living in a metro area.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 2.4 - EVALUATING MISSING VALUES (1/1 point)

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

● Midwest ✓
○ Northeast
○ South
○ West

EXPLANATION

To evaluate the number of interviewees not living in a metropolitan area, broken down by region, we can run table(CPS$Region, is.na(CPS$MetroAreaCode)). We can then compute the proportion of interviewees in each region that live in a non-metropolitan area: 34.8% in the Midwest, 21.6% in the Northeast, 23.8% in the South, and 24.4% in the West.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 2.5 - EVALUATING MISSING VALUES (4/4 points)

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%?

Wisconsin **Answer:** Wisconsin

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

Montana **Answer:** Montana

EXPLANATION

The correct way to invoke tapply to answer these questions is:

tapply(is.na(CPS$MetroAreaCode), CPS$State, mean)

It is actually easier to answer this question if the proportions are sorted, which can be accomplished with:

sort(tapply(is.na(CPS$MetroAreaCode), CPS$State, mean))

From this output, we can see that Wisconsin is the state closest to having 30% of its interviewees from a non-metropolitan area (it has 29.933% non-metropolitan interviewees) and Montana is the state with highest proportion of non-metropolitan interviewees without them all being non-metropolitan, at 83.608%.

Answering each of these questions would have been tedious using the table() output.

Check Save Hide Answer *You have used 1 of 4 submissions*

PROBLEM 3.1 - INTEGRATING METROPOLITAN AREA DATA (2/2 points)

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 and 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?

271

271

Answer: 271

EXPLANATION

This can be read from str(MetroAreaMap) or nrow(MetroAreaMap).

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

149

149

Answer: 149

EXPLANATION

This can be read from str(CountryMap) or nrow(CountryMap).

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.2 - INTEGRATING METROPOLITAN AREA DATA (2/2 points)

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?

MetroArea

Answer: MetroArea

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.

34238

Answer: 34238

EXPLANATION

From summary(CPS), we see that the variable MetroArea was added to the CPS data frame, and that it is missing 34238 values.

Check | Save | Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.3 - INTEGRATING METROPOLITAN AREA DATA (1/1 point)

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

- ○ Atlanta-Sandy Springs-Marietta, GA
- ○ Baltimore-Towson, MD
- ● Boston-Cambridge-Quincy, MA-NH ✓
- ○ San Francisco-Oakland-Fremont, CA

EXPLANATION

From table(CPS$MetroArea), we can read that Boston-Cambridge-Quincy, MA-NH has the largest number of interviewees of these options, with 2229.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 3.4 - INTEGRATING METROPOLITAN AREA DATA (2/2 points)

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.

Laredo, TX

Answer: Laredo, TX

EXPLANATION

The correct application of tapply here is

tapply(CPS$Hispanic, CPS$MetroArea, mean)

It will be easiest to obtain the maximum by actually using the sorted output:

sort(tapply(CPS$Hispanic, CPS$MetroArea, mean))

As we can see, 96.6% of the interviewees from Laredo, TX, are of Hispanic ethnicity, the highest proportion among metropolitan areas in the United States.

Check | Save | Hide Answer *You have used 1 of 5 submissions*

PROBLEM 3.5 - INTEGRATING METROPOLITAN AREA DATA (2/2 points)

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.

4

4

Answer: 4

EXPLANATION

As in the previous problem, we want the following command:

sort(tapply(CPS$Race == "Asian", CPS$MetroArea, mean))

We can read from the sorted output that Honolulu, HI; San Francisco-Oakland-Fremont, CA; San Jose-Sunnyvale-Santa Clara, CA; and Vallejo-Fairfield, CA had at least 20% of their interviewees of the Asian race.

Check Save Hide Answer *You have used 1 of 5 submissions*

PROBLEM 3.6 - INTEGRATING METROPOLITAN AREA DATA (1/1 point)

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.

Iowa City, IA **Answer:** Iowa City, IA

EXPLANATION

To obtain the sorted list of proportions by metropolitan area, we run:

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

We can see that Iowa City, IA had 2.9% of interviewees not finish high school, the smallest value of any metropolitan area.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 4.1 - INTEGRATING COUNTRY OF BIRTH DATA (2/2 points)

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?

Country **Answer:** Country

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

176 **Answer:** 176

EXPLANATION

The merge operation in this case is

CPS = merge(CPS, CountryMap, by.x="CountryOfBirthCode", by.y="Code", all.x=TRUE)

From summary(CPS), we can read that Country is the name of the added variable, and that it has 176 missing values.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 4.2 - INTEGRATING COUNTRY OF BIRTH DATA (2/2 points)

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

Philippines

Answer: Philippines

EXPLANATION

From the summary(CPS) output, or alternately sort(table(CPS$Country)), we see that the top two countries of birth were United States and Mexico, both of which are in North America. The third highest value, 839, was for the Philippines.

Check Save Hide Answer *You have used 1 of 5 submissions*

PROBLEM 4.3 - INTEGRATING COUNTRY OF BIRTH DATA (2/2 points)

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.

0.3083749

0.3083749

Answer: 0.309

EXPLANATION

From table(CPS$MetroArea == "New York-Northern New Jersey-Long Island, NY-NJ-PA", CPS$Country != "United States"), we can see that 1668 of interviewees from this metropolitan area were born outside the United States and 3736 were born in the United States (it turns out an additional 5 have a missing country of origin). Therefore, the proportion is 1668/(1668+3736)=0.309.

Check Save Hide Answer *You have used 2 of 5 submissions*

PROBLEM 4.4 - INTEGRATING COUNTRY OF BIRTH DATA (3/3 points)

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.

○ 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

In Brazil?

- ● 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

In Somalia?

- ○ 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

EXPLANATION

To obtain the number of TRUE values in a vector of TRUE/FALSE values, you can use the sum() function. For instance, sum(c(TRUE, FALSE, TRUE, TRUE)) is 3. Therefore, we can obtain counts of people born in a particular country living in a particular metropolitan area with:

```
sort(tapply(CPS$Country == "India", CPS$MetroArea, sum, na.rm=TRUE))
```

```
sort(tapply(CPS$Country == "Brazil", CPS$MetroArea, sum, na.rm=TRUE))
```

```
sort(tapply(CPS$Country == "Somalia", CPS$MetroArea, sum, na.rm=TRUE))
```

We see that New York has the most interviewees born in India (96), Boston has the most born in Brazil (18), and Minneapolis has the most born in Somalia (17).

Help

Hide Answer *You have used 1 of 1 submissions*

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

Show Discussion New Post

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Meetup

LinkedIn

Google+

MITx: 15.071x The Analytics Edge

parmarmanojkumar

Courseware Course Info Discussion Progress Syllabus Schedule Files Wiki

IMPORTANT NOTE: This problem is optional, and will not count towards your grade. We have created this problem to give you extra practice with the topics covered in this unit.

INTERNET PRIVACY POLL (OPTIONAL)

Internet privacy has gained widespread attention in recent years. To measure the degree to which people are concerned about hot-button issues like Internet privacy, social scientists conduct polls in which they interview a large number of people about the topic. In this assignment, we will analyze data from a July 2013 Pew Internet and American Life Project poll on Internet anonymity and privacy, which involved interviews across the United States. While the full polling data can be found here, we will use a more limited version of the results, available in AnonymityPoll.csv. The dataset has the following fields (all Internet use-related fields were only collected from interviewees who either use the Internet or have a smartphone):

- **Internet.Use**: A binary variable indicating if the interviewee uses the Internet, at least occasionally (equals 1 if the interviewee uses the Internet, and equals 0 if the interviewee does not use the Internet).
- **Smartphone**: A binary variable indicating if the interviewee has a smartphone (equals 1 if they do have a smartphone, and equals 0 if they don't have a smartphone).
- **Sex**: Male or Female.
- **Age**: Age in years.
- **State**: State of residence of the interviewee.
- **Region**: Census region of the interviewee (Midwest, Northeast, South, or West).
- **Conservativeness**: Self-described level of conservativeness of interviewee, from 1 (very liberal) to 5 (very conservative).
- **Info.On.Internet**: Number of the following items this interviewee believes to be available on the Internet for others to see: (1) Their email address; (2) Their home address; (3) Their home phone number; (4) Their cell phone number; (5) The employer/company they work for; (6) Their political party or political affiliation; (7) Things they've written that have their name on it; (8) A photo of them; (9) A video of them; (10) Which groups or organizations they belong to; and (11) Their birth date.
- **Worry.About.Info**: A binary variable indicating if the interviewee worries about how much information is available about them on the Internet (equals 1 if they worry, and equals 0 if they don't worry).
- **Privacy.Importance**: A score from 0 (privacy is not too important) to 100 (privacy is very important), which combines the degree to which they find privacy important in the following: (1) The websites they browse; (2) Knowledge of the place they are located when they use the Internet; (3) The content and files they download; (4) The times of day they are online; (5) The applications or programs they use; (6) The searches they perform; (7) The content of their email; (8) The people they exchange email with; and (9) The content of their online chats or hangouts with others.
- **Anonymity.Possible**: A binary variable indicating if the interviewee thinks it's possible to use the Internet anonymously, meaning in such a way that online activities can't be traced back to them (equals 1 if he/she believes you can, and equals 0 if he/she believes you can't).
- **Tried.Masking.Identity**: A binary variable indicating if the interviewee has ever tried to mask his/her identity when using the Internet (equals 1 if he/she has tried to mask his/her identity, and equals 0 if he/she has not tried to mask his/her identity).
- **Privacy.Laws.Effective**: A binary variable indicating if the interviewee believes United States law provides reasonable privacy protection for Internet users (equals 1 if he/she believes it does, and equals 0 if he/she believes it doesn't).

PROBLEM 1.1 - LOADING AND SUMMARIZING THE DATASET

Using read.csv(), load the dataset from AnonymityPoll.csv into a data frame called poll and summarize it with the summary() and str() functions.

How many people participated in the poll?

1002

1002

Answer: 1002

EXPLANATION

The number of people who took the poll is equal to the number of rows of the data frame, and can be obtained with nrow(poll) or from the output of str(poll).

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.2 - LOADING AND SUMMARIZING THE DATASET

Let's look at the breakdown of the number of people with smartphones using the table() and summary() commands on the Smartphone variable. (HINT: These three numbers should sum to 1002.)

How many interviewees responded that they use a smartphone?

487

487

Answer: 487

How many interviewees responded that they don't use a smartphone?

472

472

Answer: 472

How many interviewees did not respond to the question, resulting in a missing value, or NA, in the summary() output?

43

43

Answer: 43

EXPLANATION

From the output of table(poll$Smartphone), we can read that 487 interviewees use a smartphone and 472 do not. From the summary(poll$Smartphone) output, we see that another 43 had missing values. As a sanity check, 487+472+43=1002, the total number of interviewees.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 1.3 - LOADING AND SUMMARIZING THE DATASET

By using the table() function on two variables, we can tell how they are related. To use the table() function on two variables, just put the two variable names inside the parentheses, separated by a comma (don't forget to add poll$ before each variable name). In the output, the possible values of the first variable will be listed in the left, and the possible values of the second variable will be listed on the top. Each entry of the table counts the number of observations in the data set that have the value of the first value in that row, and the value of the second variable in that column. For example, suppose we want to create a table of the variables "Sex" and "Region". We would type

table(poll$Sex, poll$Region)

in our R Console, and we would get as output

Midwest Northeast South West

Female 123 90 176 116

Male 116 76 183 122

This table tells us that we have 123 people in our dataset who are female and from the Midwest, 116 people in our dataset who are male and from the Midwest, 90 people in our dataset who are female and from the Northeast, etc.

You might find it helpful to use the table() function to answer the following questions:

Which of the following are states in the Midwest census region? (Select all that apply.)

- □ Colorado
- ☑ Kansas ✓
- □ Kentucky
- ☑ Missouri ✓
- ☑ Ohio ✓
- □ Pennsylvania

Which was the state in the South census region with the largest number of interviewees?

Texas ▾ Texas

EXPLANATION

From table(poll$State, poll$Region), we can identify the census region of a particular state by looking at the region associated with all its interviewees. We can read that Colorado is in the West region, Kentucky is in the South region, Pennsylvania is in the Northeast region, but the other three states are all in the Midwest region. From the same chart we can read that Texas is the state in the South region with the largest number of interviewees, 72.

Another way to approach these problems would have been to subset the data frame and then use table on the limited data frame. For instance, to find which states are in the Midwest region we could have used:

MidwestInterviewees = subset(poll, Region=="Midwest")

table(MidwestInterviewees$State)

and to find the number of interviewees from each South region state we could have used:

SouthInterviewees = subset(poll, Region=="South")

table(SouthInterviewees$State)

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2.1 - INTERNET AND SMARTPHONE USERS

As mentioned in the introduction to this problem, many of the response variables (Info.On.Internet, Worry.About.Info, Privacy.Importance, Anonymity.Possible, and Tried.Masking.Identity) were not collected if an interviewee does not use the Internet or a smartphone, meaning the variables will have missing values for these interviewees.

How many interviewees reported neither Internet use nor smartphone use?

186

186

Answer: 186

How many interviewees reported both Internet use and smartphone use?

470

470

Answer: 470

How many interviewees reported Internet use but no smartphone use?

285

285

Answer: 285

How many interviewees reported smartphone use but no Internet use?

17

17

Answer: 17

EXPLANATION

These four values can be read from table(poll$Internet.Use, poll$Smartphone)

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 2.2 - INTERNET AND SMARTPHONE USERS

How many interviewees have a missing value for their Internet use?

1

1

Answer: 1

How many interviewees have a missing value for their smartphone use?

43

43

Answer: 43

EXPLANATION

The number of missing values can be read from summary(poll)

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 2.3 - INTERNET AND SMARTPHONE USERS

Use the subset function to obtain a data frame called "limited", which is limited to interviewees who reported Internet use or who reported smartphone use. 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.

How many interviewees are in the new data frame?

792

792

Answer: 792

EXPLANATION

The new data frame can be constructed with:

limited = subset(poll, Internet.Use == 1 | Smartphone == 1)

The number of rows can be computed with nrow(limited).

Check Save Hide Answer *You have used 1 of 3 submissions*

Important: For all remaining questions in this assignment please use the limited data frame you created in Problem 2.3.

PROBLEM 3.1 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

Which variables have missing values in the limited data frame? (Select all that apply.)

☐ Internet.Use
☑ Smartphone ✓
☐ Sex
☑ Age ✓
☐ State
☐ Region
☑ Conservativeness ✓
☐ Info.On.Internet
☑ Worry.About.Info ✓

☑ Privacy.Importance ✓
☑ Anonymity.Possible ✓
☑ Tried.Masking.Identity ✓
☑ Privacy.Laws.Effective ✓

EXPLANATION

You can read the number of missing values for each variable from summary(limited)

Final Check | Save | Hide Answer | *You have used 1 of 2 submissions*

PROBLEM 3.2 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

What is the average number of pieces of personal information on the Internet, according to the Info.On.Internet variable?

3.795

3.795

Answer: 3.795

EXPLANATION

This can be obtained with mean(limited$Info.On.Internet) or summary(limited$Info.On.Internet)

Check | Save | Hide Answer | *You have used 1 of 3 submissions*

PROBLEM 3.3 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

How many interviewees reported a value of 0 for Info.On.Internet?

105

105

Answer: 105

How many interviewees reported the maximum value of 11 for Info.On.Internet?

8

8

Answer: 8

EXPLANATION

These can be read from table(limited$Info.On.Internet)

Check | Save | Hide Answer | *You have used 1 of 3 submissions*

PROBLEM 3.4 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

What proportion of interviewees who answered the Worry.About.Info question worry about how much information is available about them on the Internet? Note that to compute this proportion you will be dividing by the number of people who answered the Worry.About.Info question, not the total number of people in the data frame.

0.4886076

0.4886076

Answer: 0.4886

EXPLANATION

From table(limited$Worry.About.Info), we see that 386 of interviewees worry about their info, and 404 do not. Therefore, there were 386+404=790 people who answered the question, and the proportion of them who worry about their info is 386/790=0.4886. Note that we did not divide by 792 (the total number of people in the data frame) to compute this proportion.

An easier way to compute this value is from the summary(limited) output. The mean value of a variable that has values 1 and 0 will be the proportion of the values that are a 1.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.5 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

What proportion of interviewees who answered the Anonymity.Possible question think it is possible to be completely anonymous on the Internet?

0.3691899

0.3691899

Answer: 0.3692

EXPLANATION

From table(limited$Anonymity.Possible), 278 respondents said anonymity is possible and 475 said it is not. Therefore, the desired proportion is 278/(278+475)=0.3692. This can also be read from summary(limited$Anonymity.Possible).

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.6 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

What proportion of interviewees who answered the Tried.Masking.Identity question have tried masking their identity on the Internet?

0.1632653

0.1632653

Answer: 0.1632653

EXPLANATION

This can be computed with the command table(limited$Tried.Masking.Identity). The output tells us that of all the respondents who answered the Tried.Masking.Identity question, 128 out of (128+656) have tried masking their identity on the internet.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3.7 - SUMMARIZING OPINIONS ABOUT INTERNET PRIVACY

What proportion of interviewees who answered the Privacy.Laws.Effective question find United States privacy laws effective?

0.2558459

0.2558459

Answer: 0.2558459

EXPLANATION

We can find this number with the command table(limited$Privacy.Laws.Effective). The output tells us that 186 out of (186+541) people who answered the Privacy.Laws.Effective question find US privacy laws effective.

Check Save Hide Answer *You have used 1 of 3 submissions*

PROBLEM 4.1 - RELATING DEMOGRAPHICS TO POLLING RESULTS

Often, we are interested in whether certain characteristics of interviewees (e.g. their age or political opinions) affect their opinions on the topic of the poll (in this case, opinions on privacy). In this section, we will investigate the relationship between the characteristics Age and Smartphone and outcome variables Info.On.Internet and Tried.Masking.Identity, again using the limited data frame we built in an earlier section of this problem.

Build a histogram of the age of interviewees. What is the best represented age group in the population?

○ People aged about 20 years old
○ People aged about 40 years old
● People aged about 60 years old ✓
○ People aged about 80 years old

EXPLANATION

From hist(limited$Age), we see the histogram peaks at around 60 years old.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 4.2 - RELATING DEMOGRAPHICS TO POLLING RESULTS

Both Age and Info.On.Internet are variables that take on many values, so a good way to observe their relationship is through a graph. We learned in lecture that we can plot Age against Info.On.Internet with the command plot(limited$Age, limited$Info.On.Internet). However, because Info.On.Internet takes on a small number of values, multiple points can be plotted in exactly the same location on this graph.

What is the largest number of interviewees that have exactly the same value in their Age variable AND the same value in their Info.On.Internet variable? In other words, what is the largest number of overlapping points in the plot plot(limited$Age, limited$Info.On.Internet)? (HINT: Use the table function to compare the number of observations with different values of Age and Info.On.Internet.)

6

6

Answer: 6

EXPLANATION

By reviewing the output of table(limited$Age, limited$Info.On.Internet), we can see that there are 6 interviewees with age 53 and Info.On.Internet value 0, with age 60 and Info.On.Internet value 0, and with age 60 and Info.On.Internet value 1.

A more efficient way to have obtained the maximum number would have been to run max(table(limited$Age, limited$Info.On.Internet))

Final Check | Save | Hide Answer *You have used 2 of 3 submissions*

PROBLEM 4.3 - RELATING DEMOGRAPHICS TO POLLING RESULTS

To avoid points covering each other up, we can use the jitter() function on the values we pass to the plot function. Experimenting with the command jitter(c(1, 2, 3)), what appears to be the functionality of the jitter command?

- ○ jitter randomly reorders the values passed to it, and two runs will yield the same result
- ○ jitter randomly reorders the values passed to it, and two runs will yield different results
- ○ jitter adds or subtracts a small amount of random noise to the values passed to it, and two runs will yield the same result
- ● jitter adds or subtracts a small amount of random noise to the values passed to it, and two runs will yield different results ✓

EXPLANATION

By running the command jitter(c(1, 2, 3)) multiple times, we can see that the jitter function randomly adds or subtracts a small value from each number, and two runs will yield different results.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 4.4 - RELATING DEMOGRAPHICS TO POLLING RESULTS

Now, plot Age against Info.On.Internet with plot(jitter(limited$Age), jitter(limited$Info.On.Internet)). What relationship to you observe between Age and Info.On.Internet?

- ○ Older age seems strongly associated with a larger value for Info.On.Internet
- ○ Older age seems moderately associated with a larger value for Info.On.Internet
- ○ Older age does not seem associated with a change in the value of Info.On.Internet
- ● Older age seems moderately associated with a smaller value for Info.On.Internet ✓
- ○ Older age seems strongly associated with a smaller value for Info.On.Internet

EXPLANATION

For younger people aged 18-30, the average value of Info.On.Internet appears to be roughly 5, while most peopled aged 60 and older have a value less than 5. Therefore, older age appears to be associated with a smaller value of Info.On.Internet, but from the spread of dots on the image, it's clear the association is not particularly strong.

Final Check | Save | Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4.5 - RELATING DEMOGRAPHICS TO POLLING RESULTS

Use the tapply() function to obtain the summary of the Info.On.Internet value, broken down by whether an interviewee is a smartphone user.

What is the average Info.On.Internet value for smartphone users?

4.367556

4.367556

Answer: 4.368

What is the average Info.On.Internet value for non-smartphone users?

2.922807

2.922807

Answer: 2.923

EXPLANATION

The proper application of tapply here is:

tapply(limited$Info.On.Internet, limited$Smartphone, summary)

We can read the average for non-smartphone users from the summary output labeled with 0 and the average for smartphone users from the summary output labeled with 1.

Check | Save | Hide Answer *You have used 1 of 3 submissions*

PROBLEM 4.6 - RELATING DEMOGRAPHICS TO POLLING RESULTS

Similarly use tapply to break down the Tried.Masking.Identity variable for smartphone and non-smartphone users.

What proportion of smartphone users who answered the Tried.Masking.Identity question have tried masking their identity when using the Internet?

0.1925466

0.1925466

Answer: 0.1925

What proportion of non-smartphone users who answered the Tried.Masking.Identity question have tried masking their identity when using the Internet?

0.1174377

0.1174377

Answer: 0.1174

EXPLANATION

We can get the breakdown for smartphone and non-smartphone users with:

```
tapply(limited$Tried.Masking.Identity, limited$Smartphone, table)
```

Among smartphone users, 93 tried masking their identity and 390 did not, resulting in proportion 93/(93+390)=0.1925. Among non-smartphone users, 33 tried masking their identity and 248 did not, resulting in proportion 33/(33+248)=0.1174.

This could have also been read from tapply(limited$Tried.Masking.Identity, limited$Smartphone, summary).

Next week, we will begin to more formally characterize how an outcome variable like Info.On.Internet can be predicted with a variable like Age or Smartphone.

Final Check | Save | Hide Answer *You have used 2 of 3 submissions*

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

Show Discussion New Post

Help