



Bookmarks

▼ [Unit 1: An Introduction to Analytics](#)

[Welcome to Unit 1](#)

[Initial Evaluation](#)

[Evaluations due Apr 26, 2016 02:00 CEST](#)

[The Analytics Edge: Intelligence, Happiness, and Health \(Lecture Sequence\)](#)

[Working with Data: An Introduction to R Lecture Sequence Quick Questions](#)

[Understanding Food: Nutritional Education with Data \(Recitation\)](#)

[Assignment 1](#)

[Homework due Apr 28, 2016 02:00 CEST](#)

► [Entrance Survey](#)

► [Unit 2: Linear Regression](#)

► [Unit 3: Logistic Regression](#)

► [Unit 4: Trees](#)

► [Unit 5: Text Analytics](#)

Unit 1: An Introduction to Analytics > Assignment 1 > Internet Privacy Poll (OPTIONAL)

## Internet Privacy Poll (OPTIONAL)

[Bookmark this page](#)

**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).

- ▶ [Unit 6: Clustering](#)
- ▶ [Kaggle Competition](#)
- ▶ [Unit 7: Visualization](#)
- ▶ [Unit 8: Linear Optimization](#)
- ▶ [Exit Survey](#)
- ▶ [Unit 9: Integer Optimization](#)
- ▶ [Final Exam](#)

- **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

0 points possible (ungraded)

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?

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)`.

You have used 0 of 3 attempts

### Problem 1.2 - Loading and Summarizing the Dataset

0 points possible (ungraded)

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?

**Answer: 487**

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

**Answer: 472**

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

**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.

You have used 0 of 3 attempts

### Problem 1.3 - Loading and Summarizing the Dataset

0 points possible (ungraded)

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?

Select an option ▾

Answer: 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)
```

Submit

You have used 0 of 2 attempts

## Problem 2.1 - Internet and Smartphone Users

0 points possible (ungraded)

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 not having used the Internet and not having used a smartphone?

Answer: 186

How many interviewees reported having used the Internet and having used a smartphone?

**Answer: 470**

How many interviewees reported having used the Internet but not having used a smartphone?

**Answer: 285**

How many interviewees reported having used a smartphone but not having used the Internet?

**Answer: 17****Explanation**

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

**Submit**

You have used 0 of 3 attempts

**Problem 2.2 - Internet and Smartphone Users**

0 points possible (ungraded)

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

**Answer: 1**

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

**Answer: 43****Explanation**

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

**Submit**

You have used 0 of 3 attempts

**Problem 2.3 - Internet and Smartphone Users**

0 points possible (ungraded)

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?

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)`.

Submit

You have used 0 of 3 attempts

**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

0 points possible (ungraded)

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)`

You have used 0 of 2 attempts

**Problem 3.2 - Summarizing Opinions about Internet Privacy**

0 points possible (ungraded)

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

**Answer:** 3.795**Explanation**

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

You have used 0 of 3 attempts

**Problem 3.3 - Summarizing Opinions about Internet Privacy**



0 points possible (ungraded)

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

**Answer:** 105

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

**Answer:** 8

### Explanation

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

Submit

You have used 0 of 3 attempts

### Problem 3.4 - Summarizing Opinions about Internet Privacy

0 points possible (ungraded)

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.

Submit

You have used 0 of 3 attempts

### Problem 3.5 - Summarizing Opinions about Internet Privacy

0 points possible (ungraded)

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

**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)`.

You have used 0 of 3 attempts

### Problem 3.6 - Summarizing Opinions about Internet Privacy

0 points possible (ungraded)

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

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.

You have used 0 of 3 attempts

### Problem 3.7 - Summarizing Opinions about Internet Privacy

0 points possible (ungraded)

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

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.

**Submit**

You have used 0 of 3 attempts

### Problem 4.1 - Relating Demographics to Polling Results

0 points possible (ungraded)

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.

**Submit**

You have used 0 of 1 attempt

### Problem 4.2 - Relating Demographics to Polling Results

0 points possible (ungraded)

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.)

**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))`

**Submit**

You have used 0 of 3 attempts

**Problem 4.3 - Relating Demographics to Polling Results**

0 points possible (ungraded)

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.

You have used 0 of 1 attempt

#### Problem 4.4 - Relating Demographics to Polling Results

0 points possible (ungraded)

Now, plot Age against Info.On.Internet with `plot(jitter(limited$Age), jitter(limited$Info.On.Internet))`. What relationship do 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 people 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.

You have used 0 of 2 attempts

**Problem 4.5 - Relating Demographics to Polling Results**

0 points possible (ungraded)

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?

**Answer:** 4.368

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

**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**.

You have used 0 of 3 attempts

**Problem 4.6 - Relating Demographics to Polling Results**

0 points possible (ungraded)

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?

**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?

**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.

You have used 0 of 3 attempts

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

### Discussion

**Topic:** Unit 1 / Unit 1, Homework: Internet Privacy Poll

© All Rights Reserved



© 2012-2017 edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.

POWERED BY  
**OPENedX**

