The Analytics Edge

Test your knowledge of R

The purpose of this set of exercises is to help build your familiarity with R. My goal throughout the course will be to try and assist you to become comfortable in being able to use tools such as R for analytics. My goal will not be to make you the most proficient R programmer but rather focus on how to use R to help in analytics.

1. Define:
   > x <− c(4,2,6)
   > y <− c(1,0,-1)
   Determine the result of the following:

   - length(x)
   - sum(x)
   - sum(x^2)
   - x+y
   - x∗y
   - x-2
   - x^2
   - x∗y[1:2]

   Use R to check the results.

2. Decide what the following sequences are and use R to check your answers:

   - 7:11
   - seq(2,9)
   - seq(4,10,by=2)
   - seq(3,30,length=10)
   - seq(6,-4,by=-2)

3. Determine what the result will be of the following R expressions, and then use R to check you are right:

   - rep(2,4)
   - rep(c(1,5),4)
   - rep(c(1,2),c(4,4))

4. Define:
   > x <− c(5,9,2,3,4,6,7,0,8,12,2,9)
   Decide what each of the following is and use R to check your answers:

   - x[2]
   - x[2:4]
   - x[c(2,3,6)]
   - x[c(1:5,10:12)]
   - x[-(10:12)]

5. Create in R the matrices
$$x = \begin{bmatrix} 3 & 2 \\ -1 & -1 \end{bmatrix}$$
   and
$$y = \begin{bmatrix} 1 & 4 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$
   Calculate the following and check your answers in R:

   - 2∗x
   - x∗x
   - x%∗%x
   - x%∗%y
   - t(y)
   - solve(x)

   With x and y as above, calculate the effect of the following subscript operations and check your answers in R.

   - x[1,]
   - x[2,]
   - x[,2]
   - y[1,2]
   - y[,2:3]

6. 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 question, 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. The dataset AnonymityPoll.csv 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).

(a) 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?

(b) Look at the breakdown of the number of people with smartphones using the table() command on the Smartphone variable.
- How many interviewees responded that they use a smartphone?
- How many interviewees responded that they don't use a smartphone?
- How many interviewees did not respond to the question, resulting in a missing value, or NA, in the summary() output?

(c) Look at the breakdown of the number of people with smartphones and Internet use using the table() command.
- How many interviewees reported not having used the Internet and not having used a smartphone?
- How many interviewees reported having used the Internet and having used a smartphone?
- How many interviewees reported having used the Internet but not having used a smartphone?
- How many interviewees reported having used a smartphone but not having used the Internet?

(d) 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 have a missing value for their Internet use?
- How many interviewees have a missing value for their smartphone use?

(e) 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.
How many interviewees are in the new data frame?

(f) For all the remaining questions use the limited data frame you have created.
Which variables have missing values in the limited data frame?

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

(h) How many interviewees reported a value of 0 for Info.On.Internet?
How many interviewees reported the maximum value of 11 for Info.On.Internet?

(i) What proportion of interviewees who answered the Worry.About.Info question worry about how much information is available about them on the Internet?

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

(k) Build a histogram of the age of interviewees.
What is the best represented age group in the population - people aged around 20, people aged around 40, people aged around 60, people aged around 80?

(l) 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. 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 using the plot() function.
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?

(m) 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?

(n) Now, plot Age against Info.On.Internet with plot(jitter(limited$Age), jitter(limited$Info.On.Internet)). Comment on the relationship you observe between Age and Info.On.Internet?

(o) Use the tapply() function to find the average of the Info.On.Internet value, depending on whether an interviewee is a smartphone user or not?

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

7. In this question, we will investigate graphically the R internal dataset swiss using a different visualization tool. The data contains the variables:

   - Fertility - common standardized fertility measure
   - Catholic - % of catholics
   - Agriculture - % of men working in agriculture environment
   - Examination - % draftees receiving highest mark on army examination
   - Education - % education beyond primary school for draftees
   - Infant.Mortality - % of live births who live less than 1 year

   of 47 counties in the west of Switzerland dated at 1888. With ?swiss, you can get more information on the meaning of the variables.

   (a) Read the help file of stars(). Make a star plot of all variables. What can you say about Sierre?
   R-Hint: stars(as.matrix(swiss), ...)

   (b) We are interested in the relation between Fertility and Education. Therefore we would like to make a scatter-plot of Fertility against Education whose points are stars with the information of the other variables. In addition we need the argument location.
   R-Hint: stars(as.matrix(swiss[, c(2,3,5,6)]),location = as.matrix(swiss[, c(4,1)]),axes = T, ...)

   (c) Set the argument draw.segments to TRUE to get segments instead of stars. Place a legend with key.loc.

   (d) What relation do you get from the plots?

8. In this question, we will visualize the attributes of parole violators from a dataset. In many criminal justice systems around the world, inmates deemed not to be a threat to society are released from prison under the parole system prior to completing their sentences. They are still considered to be serving their sentences while on parole and they can be returned to prison if they violate the terms of their parole. Parole boards use data on parole violators to better understand whether to approve or deny an application for parole. The dataset Parole.csv has the following fields:

   - Male = 1 if the parolee is male, 0 if female
   - Racewhite = 1 if the parolee is white, 0 otherwise
   - Age = The parolee's age in years at the time of release from prison
   - State = The parolee's state (Kentucky, Louisiana, Virginia, and Other). The first three states were selected due to having a high representation in the dataset.
   - TimeServed = The number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
   - MaxSentence = The maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
   - MultipleOffenses = 1 if the parolee was incarcerated for multiple offenses, 0 otherwise.
   - Crime = The parolee's main crime leading to incarceration (Larceny, Drugs, Driving, and Other).
   - Violator = 1 if the parolee violated their parole, and 0 if the parolee completed the parole without violation.

   In this question, we will visualize the attributes of parole violators using histograms with the ggplot2 package. We'll learn how to use histograms to show counts by one variable, and then how to visualize 3 dimensions by creating multiple histograms.

   (a) Read the data into a dataframe called Parole. What fraction of parole violators are female?

   (b) In this dataset, which crime is the most common in Kentucky?

   (c) In the ggplot2 package, we need to specify a dataset, aesthetic, and geometry while creating visualizations. To create a histogram, the geometry will be geom_histogram. Create a histogram to find out the distribution of the age of parolees, by typing the following command in your R console:
   > ggplot(data = Parole, aes(x = Age)) + geom_histogram()

(d) By default, geom_histogram divides the data into 30 bins. Change the width of the bins to 5 years by adding the argument binwidth = 5. Also set the center of one of the bins to 17.5 by adding the argument center = 17.5. Also define the argument closed = c("left") to indicate that left endpoint is included in the bin, but the right endpoint isn't. Which among these age brackets has the most parolees?

- $[20, 25)$
- $[25, 30)$
- $[30, 35)$
- $[35, 40)$

(e) Redo the histogram by adding the argument color = c("blue") to geom_histogram. What does this argument do?

- Changes the fill color of the bars
- Changes the background color of the plot
- Changes the outline color of the bars
- Changes the color of the axis labels

(f) Now suppose we are interested in seeing how the age distribution of male parolees compares to the age distribution of female parolees. One option would be to create a heatmap with Age on one axis and Male (a binary variable in our data set) on the other axis. Another option would be to stick with histograms, but to create a separate histogram for each gender. ggplot has the ability to do this automatically using the facet_grid command. To create separate histograms for male and female, type the following command into your R console:
> ggplot(data = Parole, aes(x = Age)) + geom_histogram(binwidth=5,closed=c("left"), center=17.5,color=c("blue"))+facet_grid(Male∼.)
The histogram for female parolees is on the top and the male parolees is on the bottom. What is the age bracket with the most female parolees?

- $[20, 25)$
- $[25, 30)$
- $[30, 35)$
- $[35, 40)$

(g) Now change the facet_grid argument to facet_grid(.∼Male). What does this do?

- Creates histograms of the Male variable, sorted by the different values of age.
- Puts the histograms side-by-side instead of on top of each other.
- Puts the histogram for male parolees on the top.
- This doesn't change anything - the plot looks exactly the same as it did before.

(h) An alternative choice to creating separate histograms is to color the groups differently. To do this, we need to tell ggplot that a property of the data (male or not male) should be translated to an aesthetic property of the histogram. We can do this with the fill parameter as follows:
> ggplot(data = Parole, aes(x = Age,fill = as.factor(Male))) +
geom_histogram(binwidth=5,closed="left",center=17.5,color=c("blue")))
Here we need to specify the fill argument as a factor for the function to work. Create the new histogram.

(i) Coloring the groups differently is a good way to see the breakdown of age by sex within the single, aggregated histogram. However, the bars here are stacked, meaning that the height of the bars in each age bin represents the total number of parolees in that age bin, not just the number of parolees in that group. An alternative to a single, stacked histogram is to create two histograms and overlay them on top of each other. This is a simple adjustment to our previous command. We just need to 1) Tell ggplot not to stack the histograms by adding the argument position="identity" to the geom_histogram function and 2) Make the bars semi-transparent so we can see both colors by adding the argument alpha=0.5 to the geom_histogram function. The new arguments prevent the bars from being stacked and make them semi-transparent. Redo the plot, making both of these changes.
Which of the following buckets contain no female paroles? Choose all that apply:

- $[15, 20)$
- $[20, 25)$
- $[25, 30)$
- $[30, 35)$
- $[35, 40)$
- $[40, 45)$
- $[45, 50)$
- $[50, 55)$
- $[55, 60)$
- $[60, 65)$
- $[65, 70)$

(j) Which of the histograms (faceting or overlaying) do you think better visualizes the data? Why?

(k) Now let us explore the amount of time served by parolees. Create a basic histogram as in part (c) but with TimeServed on the x-axis. Set the binwidth to 1 month, center to 0.5 and closed to "right". What is the most common length of time served according to this histogram?

(l) Now, suppose we suspect that it is unlikely that each crime has the same distribution of time served. To visualize this change use facet_grid to create a separate histogram of TimeServed for each value of the variable Crime. Which crime type has no observations where time served is less than one month?

- Drug
- Driving
- Larceny
- Other

(m) Now instead of faceting the histogram, overlay them. Remember to set the position and alpha parameters so that histograms are not stacked. Also make sure to indicate the fill aesthetic is Crime. In this case, faceting seems like a better alternative. Why?

- With four different groups, it can be hard to tell them apart when they are overlayed, especially if they have similar values.
- ggplot doesn't let us overlay plots with more than two groups.
- Overlaying the plots doesn't allow us to observe which crime type is the most common