Introduction to Data Visualization

1. Looking at the numbers and character strings
2. at the final dataset is really useful.
3. To convince yourself, print and stare at the US murders data table.
4. You can do this with this code.
5. What do you learn from staring at this table?
6. How quickly can you determine which states have the largest populations?
7. Which states have the smallest?
8. How large is a typical state?
9. Is there a relationship between population size and total murders?
10. How do the murder rates vary across regions of the country?
11. For most human brains, it is quite difficult
12. to extract this information just from looking at the numbers.
13. In contrast, the answer to all of these questions
14. are readily available from examining this plot.
15. We are reminded of the saying, "A picture is worth a thousand words."
16. Data visualization provides a powerful way
17. to communicate a data driven finding.
18. In some cases, the visualization is so convincing that no follow up analysis
19. is required.
20. The growing availability of informative data sets and software tools
21. has lead to increased reliance on data of visualization
22. across many industries, academia, and government.
23. A visible example are news organizations that are increasingly
24. embracing data journalism and including effective info-graphics
25. and charts as part of their reporting.
26. A particularly effective example is a Wall Street Journal article
27. showing data related to the impact of vaccines
28. on battling infectious diseases.
29. One of the graphs shows measles cases by US state
30. through the years with a vertical line demonstrating
31. when the vaccine was introduced.
32. Another striking example comes from the New York Times.
33. This article showed data on scores from the New York City regents exam.
34. These scores are collected for several reasons,
35. including to determine if a student graduates from high school.
36. In New York City, you need a 65 to pass.
37. The distribution of the test scores forces
38. us to notice something somewhat problematic.
39. The most common test score is the minimum passing grade.
40. With very few just below that value.
41. This unexpected result is consistent with students
42. close to passing having their scores bumped up.
43. This is an example of how data visualization can
44. lead to discoveries which would otherwise
45. be missed if we simply subject that data to a battery of data analysis, tools,
46. or procedures.
47. Data visualization is the strongest tool of what
48. we call exploratory data analysis.
49. John Tukey, considered the father of exploratory data analysis
50. once said, the greatest value of a picture
51. is when it forces us to notice what we never expected to see.
52. We note that many widely used data analysis
53. tools were initiated by discoveries made with exploratory data analysis.
54. Exploratory data analysis is perhaps the most important part of data analysis,
55. yet it is often overlooked.
56. Data visualization is also now pervasive and philanthropic
57. in educational organizations.
58. One example comes from GAPminder and the talks, New Insights on Poverty
59. and the Best Stats You&#39;ve Ever Seen, Hans Roslings
60. forced us to notice the unexpected with a series of plots
61. related to world health and economics.
62. In his videos, he used animated graphs to show us how the world was changing,
63. and how old narratives are no longer true.
64. It is also important to note that mistakes, biases, systematic errors,
65. and other unexpected problems often lead to data
66. that should be handled with care.
67. Failure to discover these problems often leads to flawed analyzes
68. and false discoveries.
69. As an example, consider that measurement devices sometimes fail
70. and that most data analysis procedures are not designed to detect these yet.
71. These data analysis procedures will still give you an answer.
72. The fact that it can be hard or impossible to notice an error just
73. from the reporter results makes data visualization particularly important.
74. In this course, we will learn the basics of data visualization
75. and exploratory data analysis.
76. We will use motivating examples.
77. We will use the ggplot2 package to code.
78. To learn the basics, we will start with a somewhat artificial example, heights
79. reported by students.
80. Then we will cover two of the examples we mentioned,
81. world health and economics, and infectious diseases
82. trends in the United States.
83. Note that there&#39;s much more to data visualization that
84. will be covered here.
85. But you will get a very good introduction to the topic.
86. You may have noticed that numerical data is often
87. summarized with an average value.
88. For example, the quality of a high school is sometimes
89. summarized with one number--
90. the average score in a standardized test.
91. Occasionally, a second number is reported
92. as well-- the standard deviation.
93. So, for example, you might read a report stating that scores at this high school
94. were 680 plus or minus 50.
95. The last number is the standard deviation.
96. Note that the report has summarized an entire vector of scores
97. with just two numbers.
98. Is this appropriate?
99. Is there any important piece of information
100. we're missing by only looking at this summary rather than the entire list?
101. It turns out that in some cases, these two numbers
102. are pretty much all we need to understand the data.
103. Data visualization techniques will help us determine when
104. this two-number summary is appropriate.
105. These same techniques will serve as alternatives for when
106. these two numbers are not enough.
107. Our first data visualization building block
108. is learning to summarize lists of factors or numeric vectors.
109. The most basic statistical summary of a list of objects or numbers
110. is its distribution.
111. Once a vector has been summarized as a distribution,
112. there are several data visualization techniques to effectively relay
113. this information.

**Data Types**

1. An important first step in deciding how to visualize data
2. is to know what type of data it is.
3. We will be working with two types of variables-- categoricals
4. and numericals.
5. Each can be divided into two further groups.
6. Categoricals can be divided into ordinals and non-ordinals.
7. And numerical variables can be divided into discrete or continuous.
8. Variables that are defined by a small number of groups
9. we call categorical data.
10. Two simple examples are sex, male or female, or regions of the states
11. that we looked at in the first of course--
12. Northeast, South, North Central, West.
13. Some categorical data can be ordered.
14. For example, spiciness can be mild, medium, or hot.
15. Even if they are not numbers per se, they can still be ordered.
16. In statistics textbooks they sometimes refer to these as ordinal data.
17. The other data type are numericals.
18. Examples that we have seen or will see are population sizes, murder rates,
19. and heights.
20. We can further divide numerical data into continuous and indiscreet.
21. Continuous variables are those that can take any value such as heights
22. if measured with enough precision.
23. For example, a pair of twins maybe 68.12 inches and 68.11 inches respectively.
24. Counts such as population sizes are discrete
25. because they have to be round numbers.
26. Note that discrete numeric data can be considered ordinal.
27. An example are heights rounded to the nearest inch.
28. Although this is technically true, we usually
29. reserve the term "ordinal data" for variables
30. belonging to a small number of different groups
31. with each group having many members.
32. In contrast, when we have many groups with few cases in each group,
33. we typically refer to this as discrete numerical variables.
34. So for example, the number of packs of cigarettes a person smokes a day
35. rounded to the closest pack-- so 0, 1, or 2--
36. would be considered ordinal.
37. While the number of cigarettes that we smoke--
38. 0, 1, 2, 3 up to maybe 36--
39. would be considered a numerical variable.
40. But indeed, these examples can be considered both
41. when it comes to data visualization.
42. Now that we've learned about the different data types,
43. we're ready to learn about data visualization techniques.

**DataCamp**

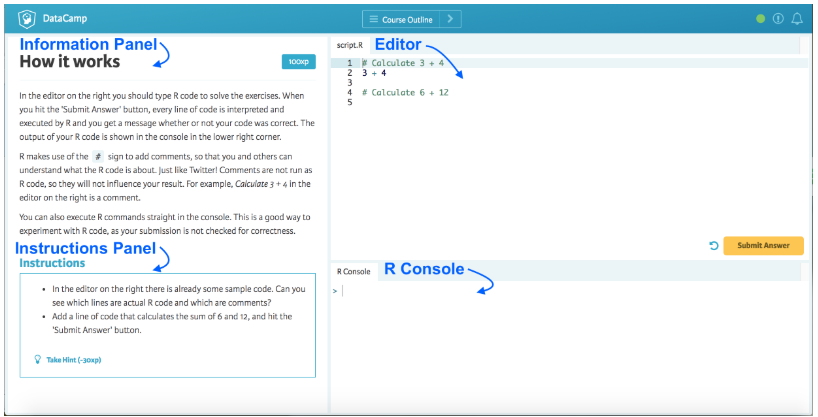
You are about the take the first assessment. In this course we will be using the DataCamp platform for all assessments. DataCamp provides an R console and and a script editor right here on your browser.  This is why **installing R is not required** for this course. Here we give a brief DataCamp tutorial. If you are already familiar with DataCamp you can skip this section and proceed to the next section

To start an DataCamp assessment, you will click on the button that says *Click here to start the assessment*, which looks like this: *.* You will see button like this one in the next section: **Assessment 1.1: Data Types**.

The DataCamp interface has **four** panels. They are:

* + **The Information Panel:** General information about the assessment.
  + **The Instructions Panel:** Exercise instructions. The multiple choice questions appear here when applicable.
  + **The Editor:** Here is where you type and edit your answers in the form of an R script. Example code also appears here. The editor also includes reminders of the instructions.  Note that # denotes comments. These are not run as code, instead, they tell others what your code is about!
  + **R console:** This is where R commands get executed. You can send commands from the editor to the console but you can also type in commands directly to test out code.

Here is a screenshot of what DataCamp looks like:



There are two ways to send commands from the editor to the console:

1) If you hit the  Submit Answer button, the entire code in the editor gets executed and your answer is evaluated. Remember, after you click Submit Answer in an assessment, your code will be evaluated. If you do not take the hint, you get unlimited tries.

2) If your cursor is on the editor and you hit command-return on a Mac or control-return on Windows, that line gets executed in the console. You do not submit an answer when you do this. This is a good way to test your script before you submit.

**Tip:** DataCamp suggests useful keyboard shortcuts after most exercises.

