Data Science Visualization

The textbook for the Data Science course series is freely available online.

Learning Objectives

- Data visualization principles to better communicate data-driven findings
- How to use ggplot2 to create custom plots
- The weaknesses of several widely used plots and why you should avoid them

Course Overview

Section 1: Introduction to Data Visualization and Distributions

You will get started with data visualization and distributions in R.

Section 2: Introduction to ggplot2

You will learn how to use ggplot2 to create plots.

Section 3: Summarizing with dplyr

You will learn how to summarize data using dplyr.

Section 4: Gapminder

You will see examples of ggplot2 and dplyr in action with the Gapminder dataset.

Section 5: Data Visualization Principles

You will learn general principles to guide you in developing effective data visualizations.

Section 1 Overview

Section 1 introduces you to Data Visualization and Distributions.

After completing Section 1, you will:

- understand the importance of data visualization for communicating data-driven findings.
- be able to use distributions to summarize data.
- be able to use the average and the standard deviation to understand the normal distribution.
- be able to assess how well a normal distribution fits the data using a quantile-quantile plot.
- be able to interpret data from a boxplot.

Introduction to Data Visualization

The textbook for this section is available here

Key points

- Plots of data easily communicate information that is difficult to extract from tables of raw values.
- Data visualization is a key component of exploratory data analysis (EDA), in which the properties of data are explored through visualization and summarization techniques.
- Data visualization can help discover biases, systematic errors, mistakes and other unexpected problems in data before those data are incorporated into potentially flawed analysis.
- This course covers the basics of data visualization and EDA in R using the **ggplot2** package and motivating examples from world health, economics and infectious disease.

Code

```
library(dslabs)
data(murders)
head(murders)
```

```
##        state abb region population total
## 1    Alabama  AL  South    4779736   135
## 2     Alaska  AK   West     710231    19
## 3    Arizona  AZ   West    6392017   232
## 4   Arkansas  AR  South    2915918    93
## 5 California  CA   West   37253956  1257
## 6   Colorado  CO   West    5029196    65
```

Introduction to Distributions

The textbook for this section is available here

Key points

- The most basic statistical summary of a list of objects is its distribution.
- We will learn ways to visualize and analyze distributions in the upcoming videos.
- In some cases, data can be summarized by a two-number summary: the average and standard deviation. We will learn to use data visualization to determine when that is appropriate.

Data Types

The textbook for this section is available here

Key points

- Categorical data are variables that are defined by a small number of groups.
  - Ordinal categorical data have an inherent order to the categories (mild/medium/hot, for example).
  - Non-ordinal categorical data have no order to the categories.
- Numerical data take a variety of numeric values.
  - Continuous variables can take any value.
  - Discrete variables are limited to sets of specific values.

Assessment - Data Types

1. The type of data we are working with will often influence the data visualization technique we use.

We will be working with two types of variables: categorical and numeric. Each can be divided into two other groups: categorical can be ordinal or not, whereas numerical variables can be discrete or continuous.

We will review data types using some of the examples provided in the `dslabs` package. For example, the `heights` dataset.

```
library(dslabs)
data(heights)
```

```
data(heights)
names(heights)
```

```
## [1] "sex"    "height"
```

2. We saw that `sex` is the first variable. We know what values are represented by this variable and can confirm this by looking at the first few entires:

```
head(heights)
```

```
##      sex height
## 1   Male     75
## 2   Male     70
## 3   Male     68
## 4   Male     74
## 5   Male     61
## 6 Female     65
```

What data type is the `sex` variable?

- ☐ A. Continuous
- ☒ B. Categorical
- ☐ C. Ordinal
- ☐ D. None of the above

3. Keep in mind that discrete numeric data can be considered ordinal.

Although this is technically true, we usually reserve the term ordinal data for variables belonging to a small number of different groups, with each group having many members.

The `height` variable could be ordinal if, for example, we report a small number of values such as short, medium, and tall. Let's explore how many unique values are used by the heights variable. For this we can use the `unique` function:

```
x <- c(3, 3, 3, 3, 4, 4, 2)
unique(x)
```

```
x <- heights$height
length(unique(x))
```

```
## [1] 139
```

4. One of the useful outputs of data visualization is that we can learn about the distribution of variables.

For categorical data we can construct this distribution by simply computing the frequency of each unique value. This can be done with the function `table`. Here is an example:

```
x <- c(3, 3, 3, 3, 4, 4, 2)
table(x)
```

```
x <- heights$height
tab <- table(x)
```

5. To see why treating the reported heights as an ordinal value is not useful in practice we note how many values are reported only once.

In the previous exercise we computed the variable tab which reports the number of times each unique value appears. For values reported only once tab will be 1. Use logicals and the function sum to count the number of times this happens.

```
tab <- table(heights$height)
sum(tab==1)
```

```
## [1] 63
```

6. Since there are a finite number of reported heights and technically the height can be considered ordinal, which of the following is true:

☒ A. It is more effective to consider heights to be numerical given the number of unique values we observe and the fact that if we keep collecting data even more will be observed.
☐ B. It is actually preferable to consider heights ordinal since on a computer there are only a finite number of possibilities.
☐ C. This is actually a categorical variable: tall, medium or short.
☐ D. This is a numerical variable because numbers are used to represent it.

Describe Heights to ET

The textbook for this section is available:

- Case Study describing student heights
- Distribution Function
- CDF Intro
- Histograms

Key points

- A distribution is a function or description that shows the possible values of a variable and how often those values occur.
- For categorical variables, the distribution describes the proportions of each category.
- A *frequency table* is the simplest way to show a categorical distribution. Use `prop.table` to convert a table of counts to a frequency table. *Barplots* display the distribution of categorical variables and are a way to visualize the information in frequency tables.
- For continuous numerical data, reporting the frequency of each unique entry is not an effective summary as many or most values are unique. Instead, a distribution function is required.
- The *cumulative distribution function (CDF)* is a function that reports the proportion of data below a value a for all values of a: $F(a) = Pr(x \leq a)$.
- The proportion of observations between any two values a and b can be computed from the CDF as $F(b) - F(a)$.
- A *histogram* divides data into non-overlapping bins of the same size and plots the counts of number of values that fall in that interval.

Code

```
# load the dataset
library(dslabs)
data(heights)
```

```
# make a table of category proportions
prop.table(table(heights$sex))
```

```
##
##    Female      Male
## 0.2266667 0.7733333
```

Smooth Density Plots

The textbook for this section is available here

Key points

- *Smooth density plots* can be thought of as histograms where the bin width is extremely or infinitely small. The smoothing function makes estimates of the true continuous trend of the data given the available sample of data points.
- The degree of smoothness can be controlled by an argument in the plotting function. (We will learn functions for plotting later.)
- While the histogram is an assumption-free summary, the smooth density plot is shaped by assumptions and choices you make as a data analyst.
- The y-axis is scaled so that the area under the density curve sums to 1. This means that interpreting values on the y-axis is not straightforward. To determine the proportion of data in between two values, compute the area under the smooth density curve in the region between those values.
- An advantage of smooth densities over histograms is that densities are easier to compare visually.

A further note on histograms: note that the choice of binwidth has a determinative effect on shape. There is no "true" choice for binwidth, and you can sometimes gain insights into the data by experimenting with binwidths.

Assessment - Distributions

1. You may have noticed that numerical data is often summarized with the average value.

For example, the quality of a high school is sometimes summarized with one number: the average score on a standardized test. Occasionally, a second number is reported: the standard deviation. So, for example, you might read a report stating that scores were 680 plus or minus 50 (the standard deviation). The report has summarized an entire vector of scores with with just two numbers. Is this appropriate? Is there any important piece of information that we are missing by only looking at this summary rather than the entire list? We are going to learn when these 2 numbers are enough and when we need more elaborate summaries and plots to describe the data.

Our first data visualization building block is learning to summarize lists of factors or numeric vectors. The most basic statistical summary of a list of objects or numbers is its distribution. Once a vector has been summarized as distribution, there are several data visualization techniques to effectively relay this information. In later assessments we will practice to write code for data visualization. Here we start with some multiple choice questions to test your understanding of distributions and related basic plots.

In the murders dataset, the region is a categorical variable and on the right you can see its distribution. To the closet 5%, what proportion of the states are in the North Central region?

In the murders dataset, the region is a categorical variable and the following is its distribution.

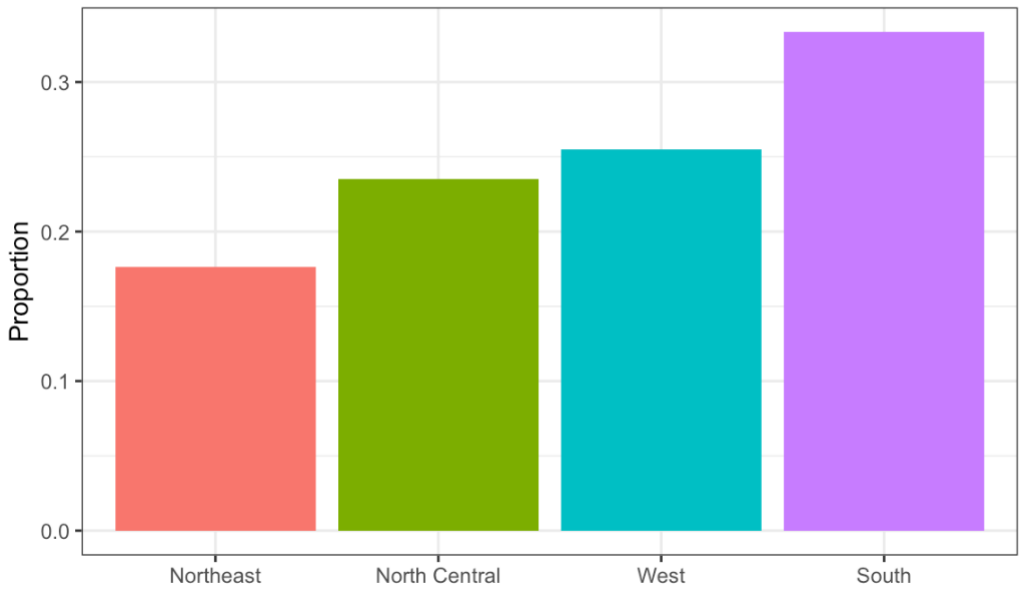

Figure 1: Region vs. Proportion

To the closet 5%, what proportion of the states are in the North Central region?

☐ A. 75%
☐ B. 50%
☒ C. 20%
☐ D. 5%

2. Distributions - 2