Overview

Which car makes and models are the most popular in the US?

Goals

By the end of this case, you will have learned how to interpret tables, bar charts, and pie charts. You will also be introduced to the concepts of absolute and relative frequency.

Introduction

Business Context. You work as an analyst at a small car dealership. The dealership is planning to expand its marketing campaigns to give better coverage to car makes and models that have proven to be more popular with consumers.

Business Problem. Your job is to *write a short report that summarizes car sales data from the first half of 2020 in the United States*. Though car sales were abnormally down during this period, you can still compare what models were popular to what models were unpopular. Your boss will use your report to make decisions about what makes and models to promote via advertising in the coming months.

Analytical Context. You have found this article in Forbes Wheels magazine. It contains information about how many cars were sold in the United States between January and June 2020. You will use the car sales table in the article as your data source to write your report. You have asked one of your colleagues to take the data and prepare some simple visualizations to help you and your boss make sense of the numbers.

Reading a Table

How to read a table

Tables are made up of two things - columns and rows. **Rows** typically represent people, transactions, items, or any kind of “thing” in general. **Columns** generally represent characteristics or attributes that those “things” have. For instance, in this table:

First name	Last name	Favorite food
Fred	Johnson	Burger
Anna	Stevens	Pasta
Jean	Fischer	Broccoli

Rows represent people (so the “things” are people), and columns represent the information we have about these people, namely their first names, last names, and favorite foods. Each one of the boxes where a column and a row meet is called a **cell**.

To read a table, you usually start by locating the row you are interested in and then the column. Or, if you know the attribute and want to find out more about the “thing,” you first locate the column and then the row.

Example 1

What is the first name of the person whose last name is Fischer?

▼ **Click for the Answer to Example 1**

We know the attribute (the last name) but don't have more information about the element. So, let's locate the column with the attribute:

First name	Last name	Favorite food
Fred	**Johnson**	Burger
Anna	**Stevens**	Pasta
Jean	**Fischer**	Broccoli

Next, we locate the row that corresponds to the last name Fischer:

First name	Last name	Favorite food
Fred	Johnson	Burger
Anna	Stevens	Pasta
Jean	**Fischer**	Broccoli

Now that we know the column and the row, we can go ahead and find the attribute of this row that represents the first name of the person:

First name	Last name	Favorite food
Fred	Johnson	Burger
Anna	Stevens	Pasta
Jean	Fischer	Broccoli

The answer is, therefore, “Jean.”

Exercise 1

What is Anna Stevens’s favorite food?

▼ **Click for the Answer to Exercise 1**

Her favorite food is Pasta. You can find this by first going to the **row** that corresponds to Anna Stevens:

First name	Last name	Favorite food
Fred	Johnson	Burger
Anna	**Stevens**	**Pasta**
Jean	Fischer	Broccoli

and then narrow down your search to the column that contains the favorite food attribute:

First name	Last name	Favorite food
Fred	Johnson	Burger
Anna	Stevens	**Pasta**
Jean	Fischer	Broccoli

Frequency Tables

Frequency Tables are extremely useful for representing virtually any kind of data - from a person's medical records in a hospital database to their entire network of friends on Facebook.

One of the most frequently used kinds of tables are **frequency tables** (pun intended). These are tables that list how often a particular element appears in a dataset - i.e., its frequency in that dataset.

Here is our car sales table (we have extracted only the 10 highest-selling car models):

Make	Model	Cars sold from Jan 2020 - Jun 2020
Ford	F-Series	367,387
Chevrolet	Silverado	264,442
Ram	Pickup	246,253
Toyota	RAV4	183,360
Honda	CR-V	138,898
Honda	Civic	127,858
Toyota	Camry	125,899
Chevrolet	Equinox	124,251
Nissan	Rogue	106,965
GMC	Sierra	106,833

Source: Motor Intelligence (June 2020) via Forbes Wheels.

Question 1

Exercise 2

Answer the following two questions:

1. What was the model with the highest sales?
2. How many Honda Civic units were sold?

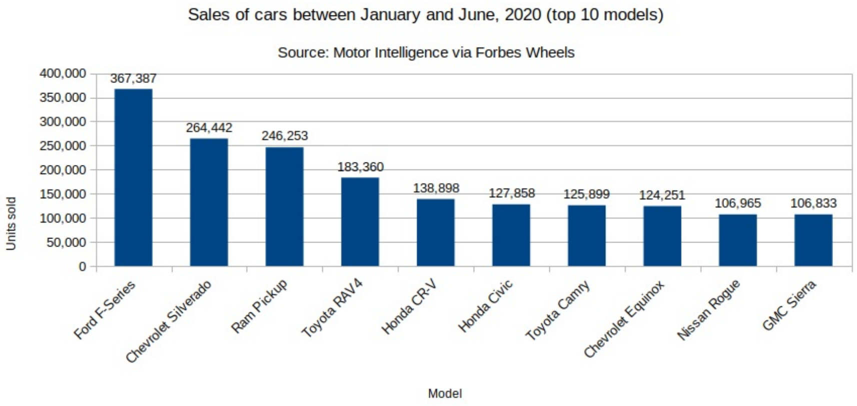
▼ **Click for the Answer to Exercise 2**

(1) The model with the highest sales was the Ford F-Series (367,387).

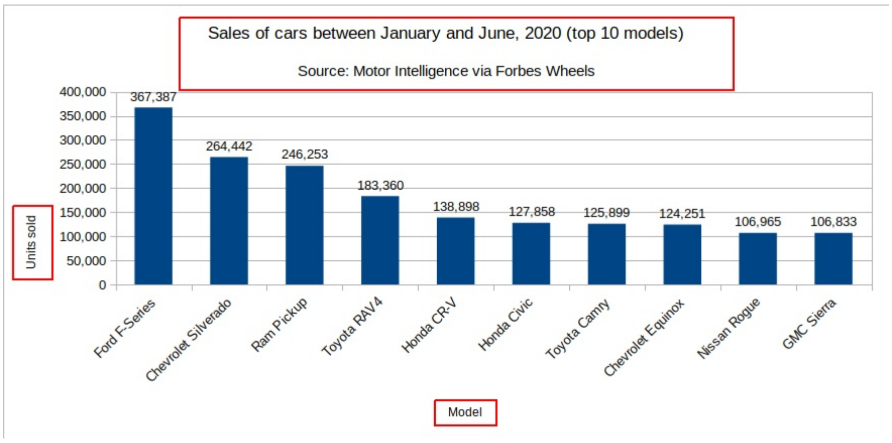
(2) 127,858 Honda Civics were sold in this period.

Bar charts

The car sales frequency table can also be represented graphically via a **bar chart**. Bar charts are one of the most popular ways to represent data. In a bar chart, you have a set of bars (one bar for each row in the table) whose heights represent the frequency that is associated with that row:

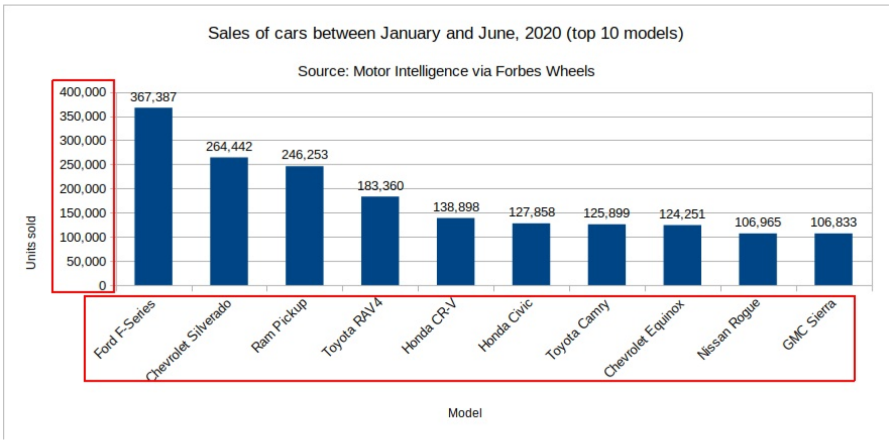


Let's discuss each of the major components of this bar chart. First, you have your *titles* and *axis labels*. These provide key information because they tell you what the chart is trying to represent:

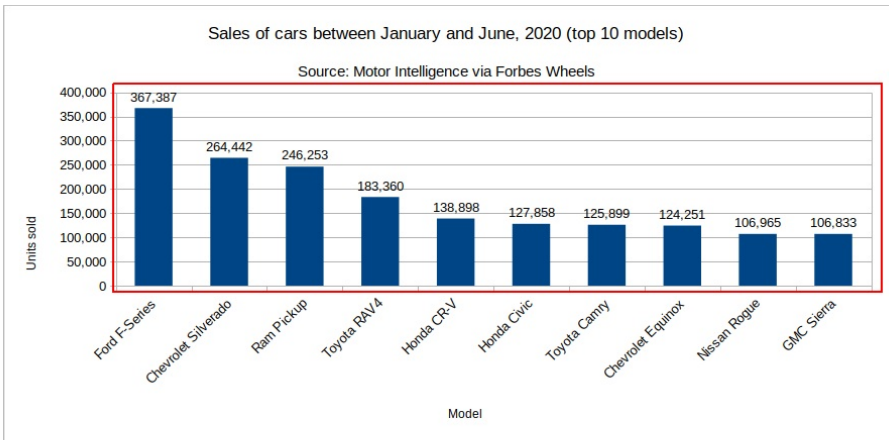


The vertical axis (`Units sold`) is commonly known as the y - *axis*. The horizontal axis (`Model`) is known as the x - *axis*. As you may have noticed, this plot has a subtitle that tells us the source of the data. This is generally good practice because it helps readers understand where the data came from, which in turn can help them assess the quality and trustworthiness of the information conveyed in the plot.

Next, you have *tick labels* on each axis. These tell you what each bar represents (in the case of the x - axis tick labels) and what magnitudes the y - axis represents (in the case of the y - axis tick labels):

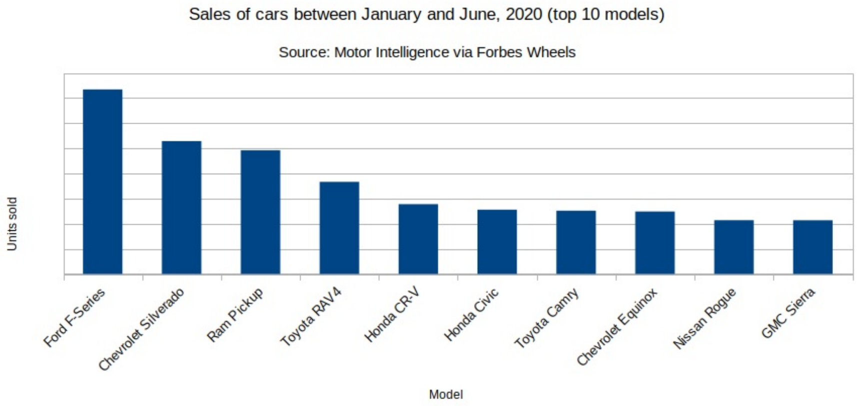


Finally, we have the *bars*, the *bar labels*, and the *grid*. The height of each bar is proportional to the frequency associated with the respective model; that is, the taller the bar, the more units sold of that particular model. We sometimes see labels on top of each bar that tell us the exact frequency of that model. They are often unnecessary, but they can be useful in many situations. We also have the grid (the gray horizontal lines), which helps us compare the heights of different bars at a glance:



Exercise 3

One of the main advantages of bar charts is that they allow you to quickly compare different bars without having to look at the actual numbers. Here's the same plot without the bar and y - axis tick labels:



Just by looking at the grid (don't look at the numbers in the previous plots), pick two models so that one has sales that are approximately twice those of the other. Your answer should be in the form: "Model A had approximately twice the sales of Model B".

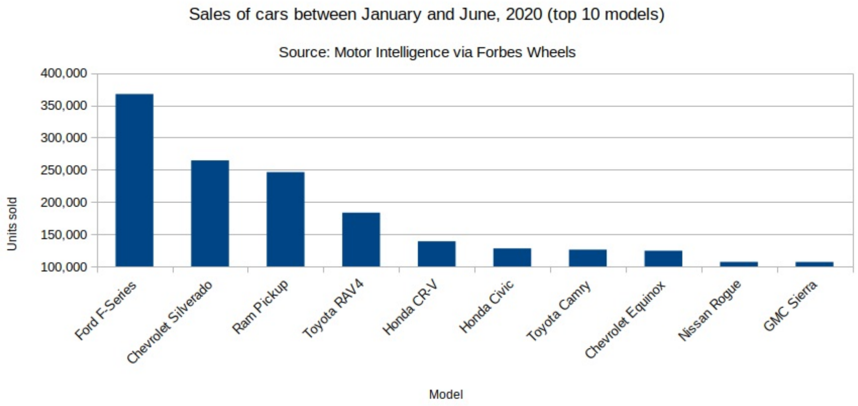
▼ **Click here to see Possible Answers to Exercise 3.**

Some possible answers are:

- Ford F-Series and Toyota RAV4. You see that the sales of the Ford F-Series extended just above seven grid lines, and those of the Toyota RAV4 reached about three and a half.
- For the same reason, the Chevrolet Silverado and Honda Civic.
- Another example is Ram Pickup and Chevrolet Equinox.

Exercise 4

We've added the tick labels back in the chart below. Try to validate your answers from the previous exercise using this bar chart. Can you do it easily? What is wrong with this plot?



▼ **Click here for the Answer to Exercise 4.**

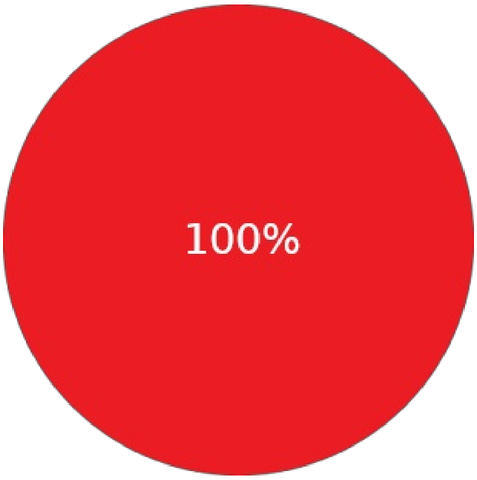
Although this plot and the original one depict the same frequency table, they look very different. That is because this plot's y - axis does *not* start at zero but rather at 100,000. This effectively creates the illusion that all the bars have shrunk, especially those that had the shortest bars to begin with. This is misleading because the relative proportions between bars are lost. For instance, the Ram Pickup sales now look much, much greater than those of the Honda CR-V (certainly larger than just a 2x difference).

It is very important that you are aware of the minimum and maximum values of the y - axis whenever you are reading a bar chart. Bar charts whose vertical axes don’t start at zero should make you immediately suspicious because they can be easily misleading. There are cases in which an axis can legitimately start at a value that is nonzero, but that is the exception, not the norm.

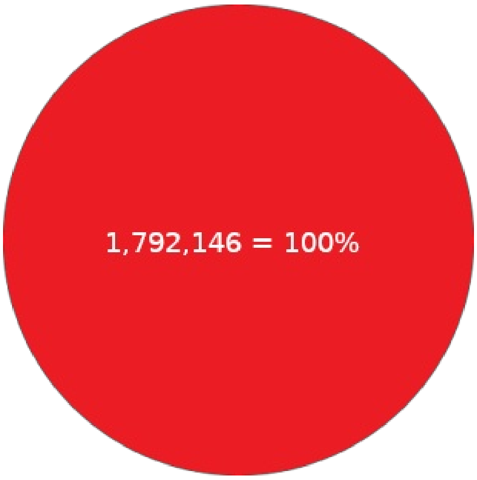
Pie charts

Bar charts are not the only way to represent count data. There is a popular alternative that is sometimes useful - the pie chart. Pie charts look like pies - hence the name. Each category gets a slice of the pie, and the size of the slice is proportional to the frequency of the category in the frequency table, whereby more frequent categories get a larger slice.

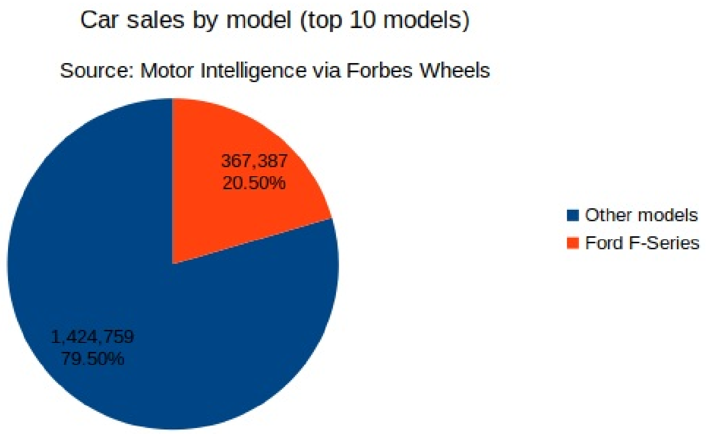
Pie charts are good for representing frequencies not only as absolute numbers but also as percentages, i.e., what we call **relative frequencies** (as opposed to the **absolute frequencies** we have been plotting so far). In a relative frequency pie chart, the total area of the pie is considered to represent the number 100%:



For instance, our car sales table has 1,792,146 units sold. Therefore, the total area of the pie chart would represent that number as 100% of total sales:



Ford F-Series sales amounted to 367,387 units sold, i.e., 20.5% of 1,792,146. It is almost exactly one-fifth of the total sales, which means that one-fifth of the pie chart area should be assigned to the Ford F-Series like this:

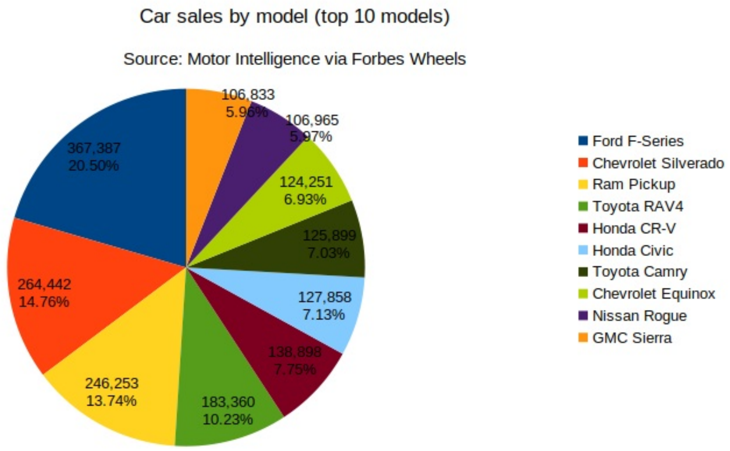


This plot has two categories: `Ford F-Series`, with an absolute frequency of 367,387 and a relative frequency of 20.5%, and `Other models`, with an absolute frequency of 1,424,759 units sold and a relative frequency of 79.5%. If you sum both absolute frequencies, you get the total sales (1,792,146), and if you sum both relative frequencies, you get 100%.

You can also see that this pie chart has a text box to the right. This is the *legend*. It tells us what each color represents in the pie chart - orange represents the Ford F-Series, and blue represents all the other models combined.

Question 2

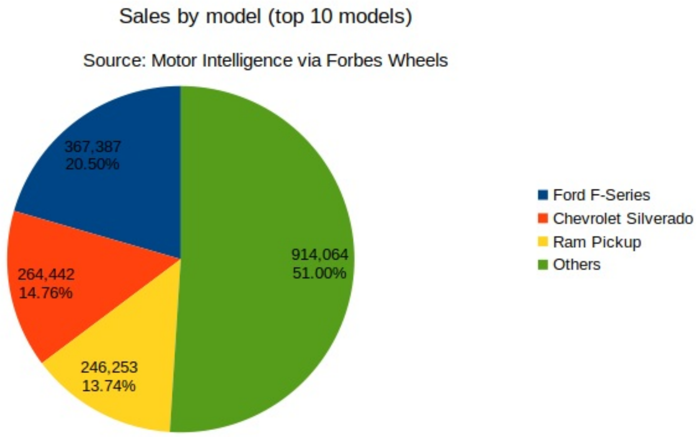
Here's the pie chart equivalent of our bar charts from before:



▼ Click for a Possible Answer to Question 2.

The pie chart version of the frequency table is arguably more difficult to read than the bar chart version. That is because when there are many categories, it becomes difficult to compare the sizes of the slices since they can't easily be measured against an axis and grid lines. This is why many people prefer bar charts over pie charts. Pie charts are still useful, but they work best when the number of categories is reasonably small. In general, bar charts should be your first option unless a pie chart could be more easily understood.

One way to reduce the number of categories is by grouping those with the smallest counts under an `Others` category. For example:



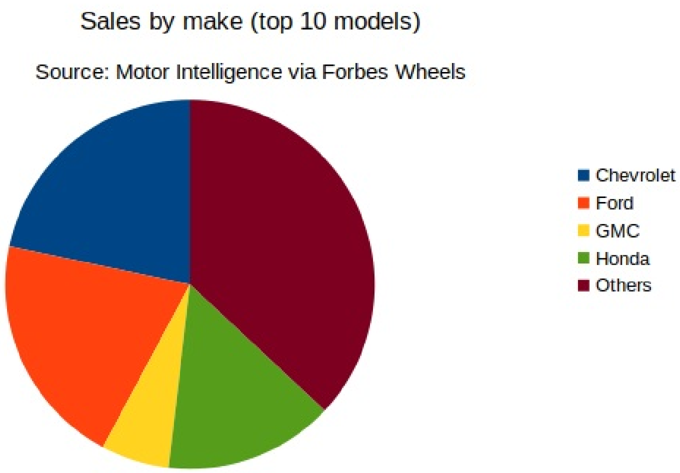
Sales by model (top 10 models)

Source: Motor Intelligence via Forbes Wheels

This is easier to read, but we lost a lot of information about the categories grouped under `Others`. If that information is not critical, we can go with this plot instead of a bar chart. This pie chart gives us additional information that was *not* apparent from the bar chart - it is now much more evident that three models comprise almost half of the market share out of the ten models in this top 10 list. However, if the information about the other models is important, a bar chart would probably be the best option.

Exercise 5

This is a pie chart of total sales by car *make* (previously, we plotted total sales by car *model*).



Sales by make (top 10 models)

Source: Motor Intelligence via Forbes Wheels

Try to estimate the relative frequencies to fill the missing spaces below (you can make approximations):

- Chevrolet: ___
- Ford: ___
- GMC: ___
- Honda: ___
- Others: ___

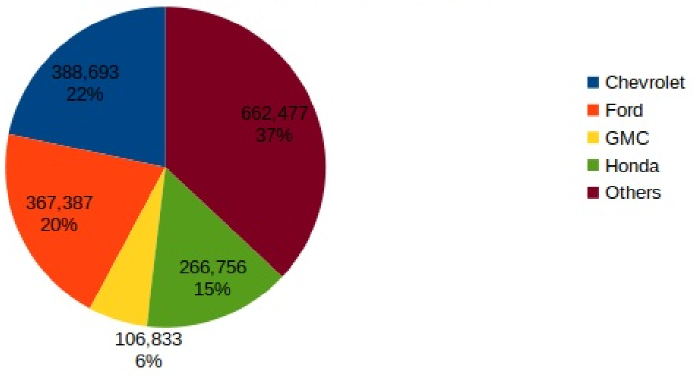
▼ **Click for the Answer to Exercise 5.**
Answer. This is the frequency table:

Make	Absolute frequency	Relative frequency
Chevrolet	388,693	21.69%
Ford	367,387	20.50%
GMC	106,833	5.96%
Honda	266,756	14.88%
Others	662,477	36.97%
TOTAL	1,792,146	100.00%

And this is the pie chart with labels:

Sales by make (top 10 models)

Source: Motor Intelligence via Forbes Wheels



Conclusions & Takeaways

Now that you know how to read frequency tables, bar charts, and pie charts, you are very well-equipped to write your news summary and hand it to your boss. Here are some key ideas to keep in mind as you write it:

- To read a table, you can start by locating the relevant column(s) and then the relevant row(s) or vice versa, depending on which initial information you have.
- Absolute frequencies are counts, and relative frequencies are percentages.
- Bar heights in bar charts are proportional to frequencies.
- y - axes in bar charts should *always* start at zero unless there is a *very* good reason for not doing so.
- Pie charts can show the same information as bar charts but are more difficult to read when there are many categories.

Here is a short summary document for review. You can download by clicking here.

Attribution

“Ford Mustang V6 Coupé (VI) – Frontansicht, 2. Oktober 2016, New York”, M 93, Oct 2, 2016, Creative Commons Attribution-Share Alike 3.0 Germany License, https://commons.wikimedia.org/wiki/File:Ford_Mustang_V6_Coup%C3%A9_(VI)_%E2%80%93_Frontansicht,_2._Oktober_2016,_New_York.jpg