1.5.1

Measures of Central Tendency

Adding statistical components will deepen our analysis for Louise. Statistics provide an unbiased view of the data and make conclusions based on calculations rather than gut feelings. With the addition of statistics, Louise will be able to make decisions about her campaign with confidence. We'll start with the measures of central tendency: mean, median, and mode.

So far, we've compiled some solid information to present to Louise to help inform her campaign strategy. We've organized and sorted the data, as well as created visualizations that lend strength to our analysis. Now we'll use statistics to beef up our report even more.

When we talk about the statistics of a data set, we're generally concerned with how the data is **distributed**. Are data points clustered around one value, or is the data more spread out? Statistical measures distill information about the distribution of data into a single number. The first measures we're going to look at are **measures of central tendency. Central tendency** refers to the tendency of data to be toward the middle of the dataset. The three key measures of central tendency are the mean, median, and mode.

Mean, Median, and Mode

The **mean** is the sum of the data divided by the number of data points. You can think of the mean as answering the question "If every data point contributed the same amount, what would that amount be?" For example, if you and two friends all chipped in to buy a pizza, and you put in $12, one friend put in $7, and the other friend put in $5, the mean cost would be calculated this way:

(12 + 7 + 5) / 3 = 24 / 3 = $8.

The **median** answers the question "Where is the midpoint of the data?" Also known as the 50th percentile, the median is the value that splits the data into two equal halves: 50% of the data is lower than the median, and 50% of the data is higher. To calculate the median, sort the data points in order, and then locate the point in the middle. For example, if the grades on a quiz are 82, 79, 79, 77, 70, 90, 71, 86, 83, first we would put them in order:

70, 71, 77, 79, 79, 82, 83, 86, 90

Since there are 9 scores, the midpoint will be the 5th score after sorting. There are 4 scores above it and 4 scores below it:

70, 71, 77, 79, **79**, 82, 83, 86, 90

The median score on the quizzes is **79**.

SKILL DRILL

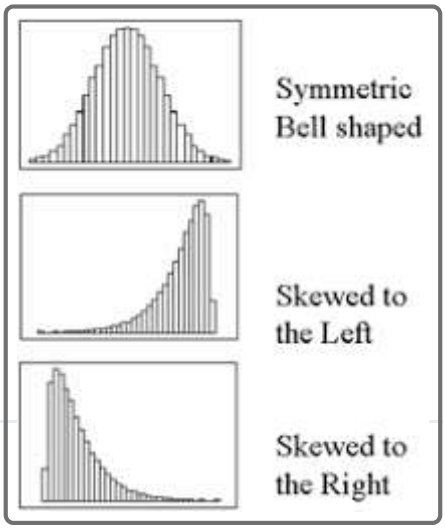
Make a small dataset (fewer than 10 data points) with a median of 50 and a mean of 55.

The **mode** answers the question "What value shows up the most?" This question can be trickier than it seems. Take our quiz scores example. In this example, 79 shows up more than any other score, so 79 is the mode. Let's say there is a student who was absent that day and then takes a makeup quiz, scoring 82. Then there would be two students who scored 82, and two students who scored 79. In this case, the data has two modes: 79 and 82. A dataset can have any number of modes—or even no mode!

IMPORTANT

When the mean and median are close to each other, the data is roughly **symmetric**: half the data is above the mean, half the data is below. If the mean is significantly different than the median, the data is **skewed,** meaning that some number of extreme values are pulling the mean higher or lower. If the mean is much higher than the median, the data is **skewed to the right**. If the mean is much lower than the median, the data is **skewed to the left**.

Skewness is a statistic that quantifies how skewed, or asymmetrical, a distribution is.



Use Measures of Central Tendency with Crowdfunding Data

Let's see how measures of central tendency work in practice. We'll consider Kickstarter campaigns for plays in the U.S. and compare the statistics for the campaigns that succeeded versus those that failed. Follow these steps:

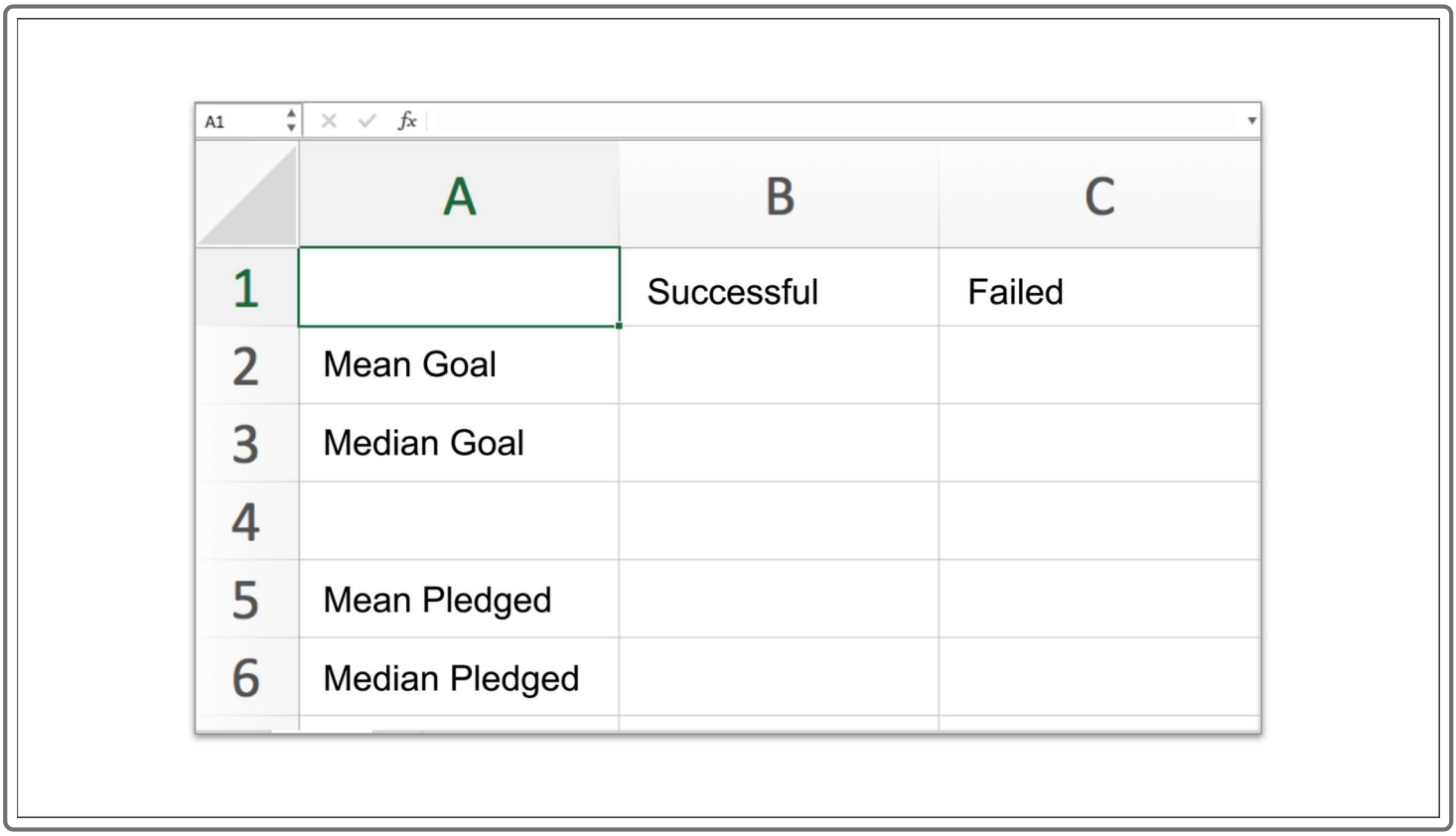
1. Clear any filters on the dataset.
2. For some versions of Excel, the entire header row must be selected first so multiple filters can be applied. Depending on your version, you may need to select the entire row, or only the header you want to filter. Then, apply the following filters:
    - Filter on subcategory for "plays."
    - Filter on country for "US"
    - Filter on outcome for "successful."
3. Copy the filtered dataset and paste it into a new worksheet named "Successful US Kickstarters."

SHOW PRO TIP

4. Return to the Kickstarter worksheet and change the filter on outcome to "failed."
5. Copy and paste this dataset into a new worksheet and name it "Failed US Kickstarters."
6. Create another worksheet and name it "Descriptive Statistics."

We'll be pulling data from each of these new worksheets and performing a few measures of central tendency on them—which is just a fancy way of saying that we'll be finding the mean and median for each dataset's (the failed and successful "US" campaigns) goal and pledged columns.

In the new worksheet, create a table like the one below to hold our results. This way, we'll be able to easily compare the goals and pledges for failed and successful campaigns alike. By comparing the two, we'll be able to determine whether there are any trends between the goals and pledges in successful or failed campaigns.



Failed Campaigns

In B2, enter the formula used to find the mean of a dataset:

```
=AVERAGE('Successful US Kickstarters'!D:D)
```

IMPORTANT

Excel doesn't have a MEAN function; it uses the less precise AVERAGE function to calculate the mean. Statisticians use the term "average" in many contexts, but prefer to be precise in their calculations. You can tell when a statistician is using Excel by the grumbling noise they make when they have to type AVERAGE instead of MEAN.

This formula tells Excel that we're looking for the average number in a dataset, but we're only looking for the average amount of "Successful US Kickstarters." By adding (D:D), we're pinpointing which column we're applying the formula to. The colon indicates a range of data, so by adding D:D to the formula, we're specifying the entire column.

Let's add a few more to our new worksheet.

- In C2, enter the formula `=AVERAGE('Failed US Kickstarters'!D:D)`.
- In B3, enter the formula `=MEDIAN('Successful US Kickstarters'!D:D)`.
- In C3, enter the formula `=MEDIAN('Failed US Kickstarters'!D:D)`.

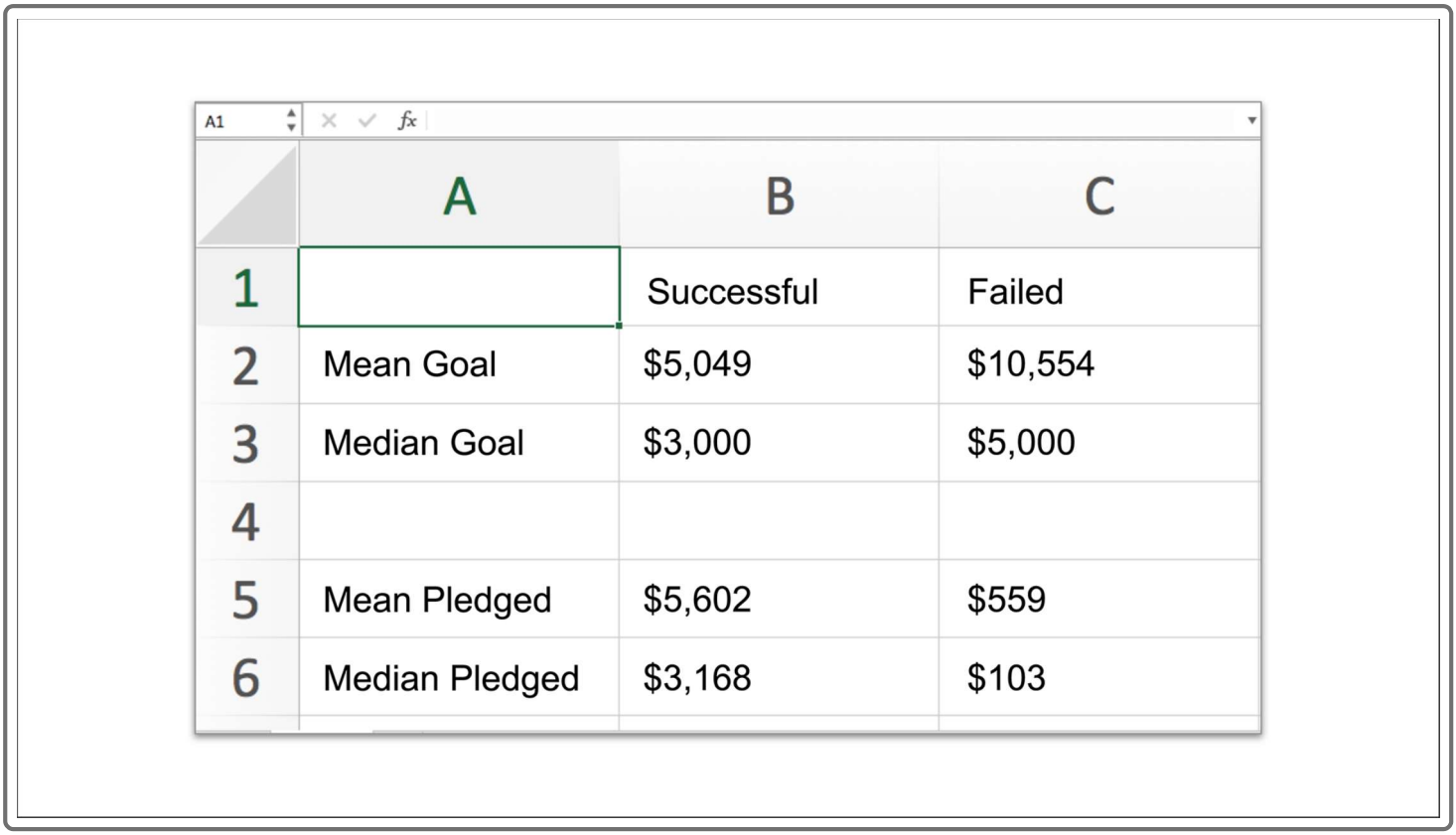
SKILL DRILL

In cells B5 and C5, enter the AVERAGE formula. Remember to use the correct columns for each: B5 is finding the average pledged for successful campaigns, and C5 is finding the average pledged for failed campaigns.

In cells B6 and C6, enter the MEDIAN formula. Be sure to pull data from the Pledged column for these as well and filter them as either successful or failed.

Format the cells B2:C6 to Currency with no decimals.

You should now see a table that looks like this:



	A	B	C
1		Successful	Failed
2	Mean Goal	$5,049	$10,554
3	Median Goal	$3,000	$5,000
4			
5	Mean Pledged	$5,602	$559
6	Median Pledged	$3,168	$103

This simple table allows us to determine a few things. For one, failed Kickstarter campaigns have much higher fundraising goals than successful Kickstarter campaigns. Louise is asking for more than twice the average successful Kickstarter goal, so this isn't great news for her campaign. In addition, the mean and median pledged

amounts are much lower than the successful pledges, which indicates that failed Kickstarter campaigns are unsuccessful for reasons other than asking for too much money. In other words, if the failed projects were also getting a median pledge amount of around $3,000, it's possible that those that failed just asked for too high of a price. Since the median is much lower, there must be another factor keeping people from pledging to those unsuccessful projects.

SKILL DRILL

Make the same table for two other subcategories. (They don't need to be in the Theater category.) How do the distributions differ, if at all?