Set 1: Summary Statistics

Skill 1.1: Describe the purpose of summary statistics
Skill 1.2: Describe categorical variables
Skill 1.3: Describe numerical variables
Skill 1.4: Interpret the mean and standard deviation of a dataset
Skill 1.5: Interpret skewed distributions
Skill 1.6: Calculate the median and IQR for a dataset
Skill 1.7: Identify outliers and *robust* summary statistics
Skill 1.8: Aggregate a dataset
Skill 1.9: Explore relationships between numerical variables

Skill 1.1: Describe the purpose of summary statistics

Skill 1.1 Concepts

Skill 1.1 Exercise 1

In the above exercise, the best we could do was generalize the data by looking over the rows and columns in the data set. To answer precise questions about the data, we need some kind of "data vocabulary" that can help us measure and describe the variables. *Summary statistics* can be used for this purpose!

Summary statistics is the science that is concerned with methods for collecting, organizing, analyzing, and interpreting data. And, with a basic understanding, we can communicate and understand a lot more specific information about the musicians in the city.

But learning statistics is often associated with a lot of negativity,

- Memorization of lots of math formulas
- Long calculations done by hand
- Confusing or meaningless interpretations

None of these struggles need to be part of learning to use statistics. In this lesson, we'll gain a conceptual understanding of how summary statistics can easily help us communicate and interpret our dataset.

Skill 1.2: Describe categorical variables

Skill 1.2 Concepts

To start our summary for the data, let's describe some of the *categorical variables* in the musician dataset — those variables that contain qualitative information on the city's musicians. First, let's look at information about the title variable, which tells us the job title each musician holds.

The following table shows:

- **frequency**: the count of musicians for each job title
- **proportion**: the frequency divided by the total number of musicians
- **percentage**: the proportion converted from a decimal to a percentage

Job Title Categories

TITLE	FREQUENCY	PROPORTION	PERCENTAGE
performer	333	0.35	35%
manager	113	0.12	12%
producer	87	0.09	9%
educator	239	0.25	25%
composer	186	0.19	19%
TOTAL	958	1.00	100%

From the table, we can learn about the different job titles of musicians in the city. There are 333 performers out of a total of 958 musicians. The proportion of performers is 333 ÷ 958 = 0.35. To make this even easier to understand, we can convert the proportion to a percentage by multiplying it by 100: 35% of musicians in the city are performers.

We can also compare one category to another by checking the ratio of their frequencies. For example, there are far fewer managers than performers. Their ratio is 333 performers to 113 managers, which can be simplified by dividing: 333 ÷ 113 = 2.95. This means there are almost 3 performers for every manager in the city.

Skill 1.2 Exercise 1

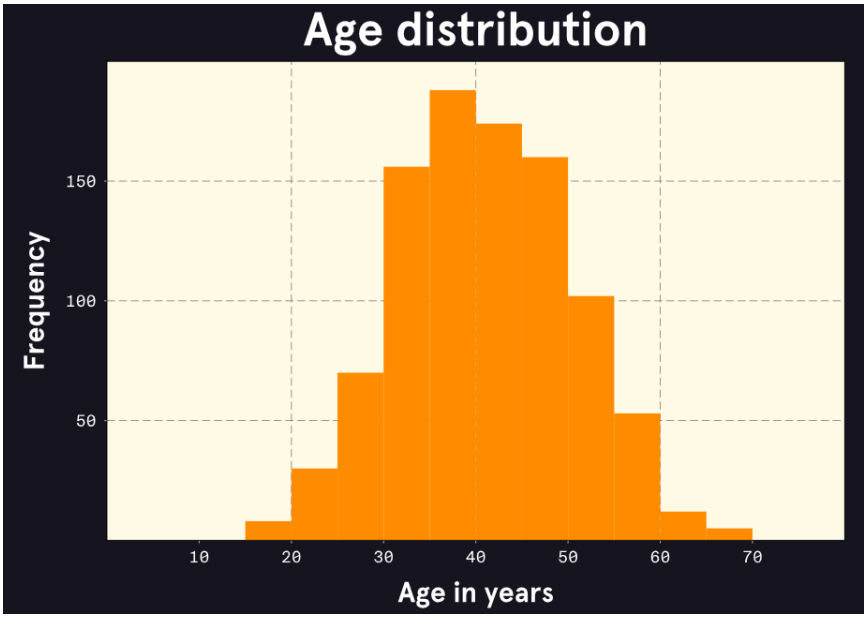
Skill 1.3: Describe numerical variables

Skill 1.3 Concepts

Now that we've learned about some of the categorical variables in our musician dataset, it's time to explore some numeric variables - those with quantitative data. There are a lot of ways we can describe the distribution of a numeric variable.

A *distribution* is a function that shows all possible values of a variable and how frequently each value occurs. This may sound pretty technical, but visualizing the distribution can make it easy to understand.

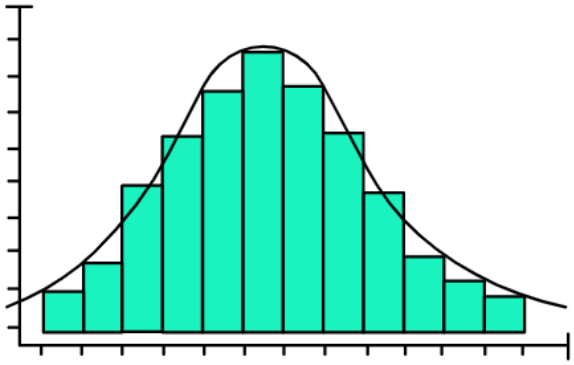
In the visualization below, the distribution of musician ages is plotted with age on the x-axis and frequency on the y-axis.



From this plot, we can see:

- Ages range from about 15 to 70.
- There are few musicians under 30 or over 50 years old.
- There are a lot of musicians between the ages of 30 and 50.

This distribution might be considered bell-shaped or hill-shaped and symmetrical. This is actually a very common pattern and is called a *normal distribution*.



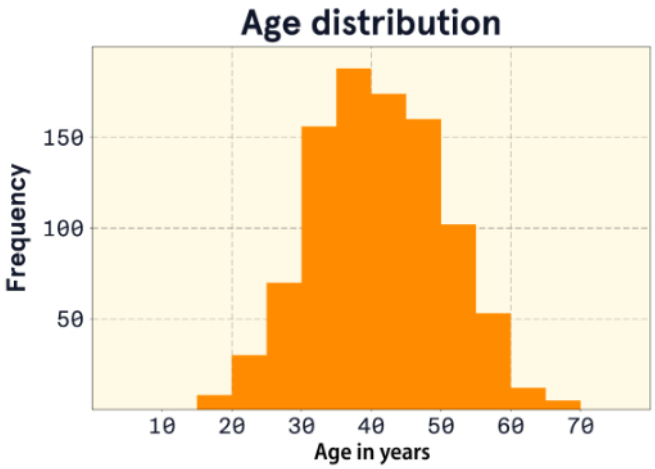
Viewing a plot or knowing a variable is normally distributed gives us some general information, but still nothing specific. We need exact measurements to describe where the center of the distribution is and how wide the values are spread away from that center. There are several sets of statistics we may use for these measurements, and we will need to know when to use which combination.

Skill 1.3 Exercise 1

Skill 1.4: Interpret the mean and standard deviation of a dataset

Skill 1.4 Concepts

Let's return to our visualization of the distribution of musician ages.

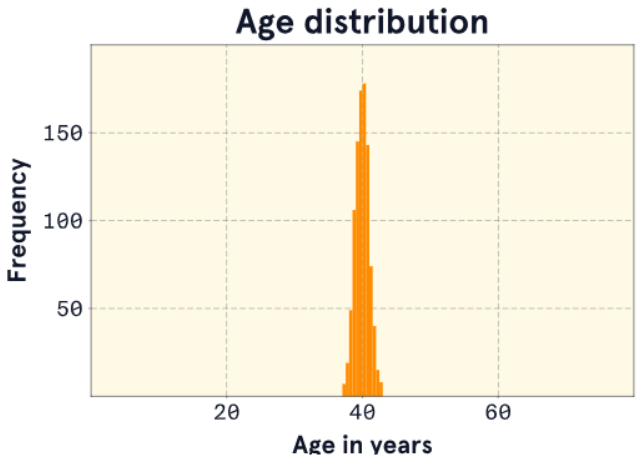


- What would you say is the typical age of a musician in Melody Metropolis?
- Are most musicians about this age, or are there lots of musicians of many different ages?

To answer these questions more specifically, we should take some measurements of our variable.

- The **mean**, also called the average, describes the center of a numeric distribution by adding all values and dividing by the count.
- The **standard deviation** describes the spread of values in a numeric distribution relative to the mean. It is calculated by finding the average squared distance from each data point to the mean and square-rooting the result.

The mean age of musicians is 40.6 years and the standard deviation is about 9.3 years. We might interpret this standard deviation as moderate variability in age. Had the standard deviation been 1 year, we might say there's hardly any variability in age. The plot of this narrow distribution might look something like the following:

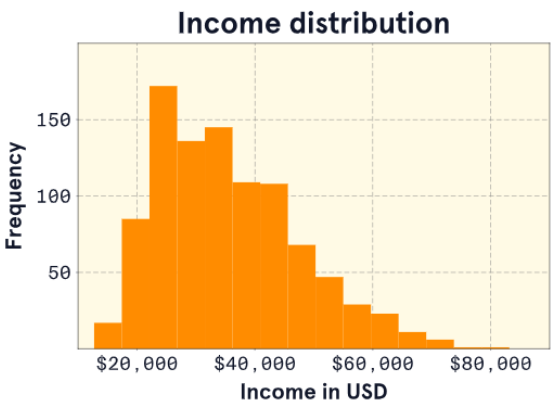


Skill 1.4 Exercise

Skill 1.5: Interpret skewed distributions

Skill 1.5 Concepts

As we're moving through the numeric variables in our musician dataset, we come across some interesting details when we inspect the income variable.



MEAN	STANDARD DEVIATION
$34,795	$11,971

- We notice that the shape of the distribution is different than the shape of the age distribution. There are quite a few musicians with higher incomes that are creating a longer tail on the right side.
- We also notice that the mean indicates that the typical income is $34,795. This value seems a little high since most of the incomes seem to be between $15,000 and $40,000.

What we have learned is that the income distribution is skewed. A *skewed* distribution is asymmetrical with a steep change in frequency on one side and a flatter, trailing change in frequency on the other. Specifically, the income distribution is right-skewed (also called positively-skewed) because the tail is on the right side.

So why does the mean seem wrong? Remember, the mean is the sum of all the values in the dataset divided by the total count. That sum is made very large by all the higher incomes in that right tail. This makes the mean a greater number than we would like it be. When the data are skewed, the mean may not be the best measure of a typical observation.

There are a number of ways to deal with this issue. We will handle the problem with the income data by taking some alternative measurements.

Skill 1.5 Exercise 1

Skill 1.6: Calculate the median and IQR for a dataset

Skill 1.6 Concepts

Let's find an alternative measure to the mean. We want to find a value that represents the typical musician income, but we don't want to use the actual values in the computation because the data are skewed.

One method would be to find the middle value when all values are arranged from smallest to largest. This value is called the **median**, but it's also referred to as the 50th percentile or the second quartile (Q2).

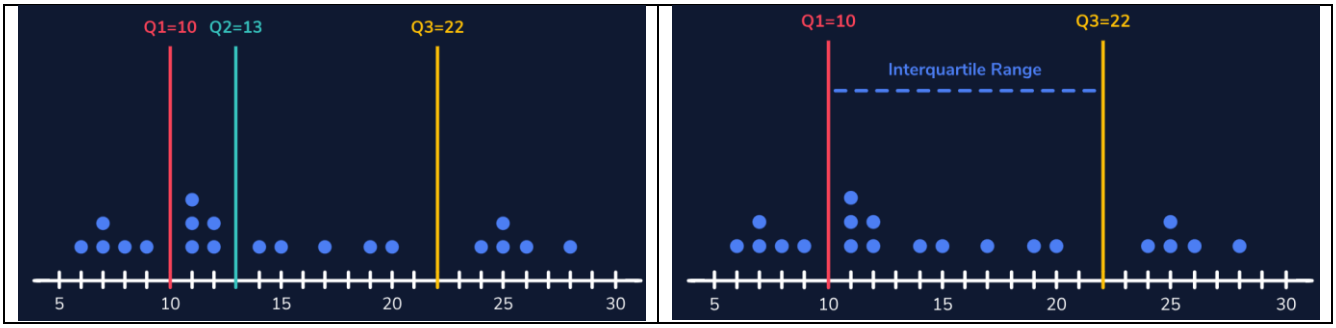
Let's consider a simple data set where the median (Q2) is 13. Half the data points are less than 13, and half are greater than 13.

These data span 22 values, ranging from 6 to 28. We could use this as our measure of spread, but what if the highest number wasn't 28 but 280? The median would still be 13, but now the range is 274 (280-6), which doesn't tell us a lot about the bulk of the data.

A better measurement might be the **interquartile range (IQR)**. A quartile is simply a marker for a quarter (25%) of the data.

- The first quartile marks 25% (Q1 = 10).
- The second quartile marks 50% (Q2 = 13 — the median)
- The third quartile marks 75% (Q3 = 22)

The IQR is the difference between Q3 and Q1 (22-10 = 12), marking the range for just the middle 50% of the data.



Let's find out how the median and IQR work out for our income data.

MEDIAN	IQR
$32,978	$17,150

This looks better — the IQR of $17, 150 is lower than the median of $32,978 and seems more typical.

Skill 1.6 Exercise 1

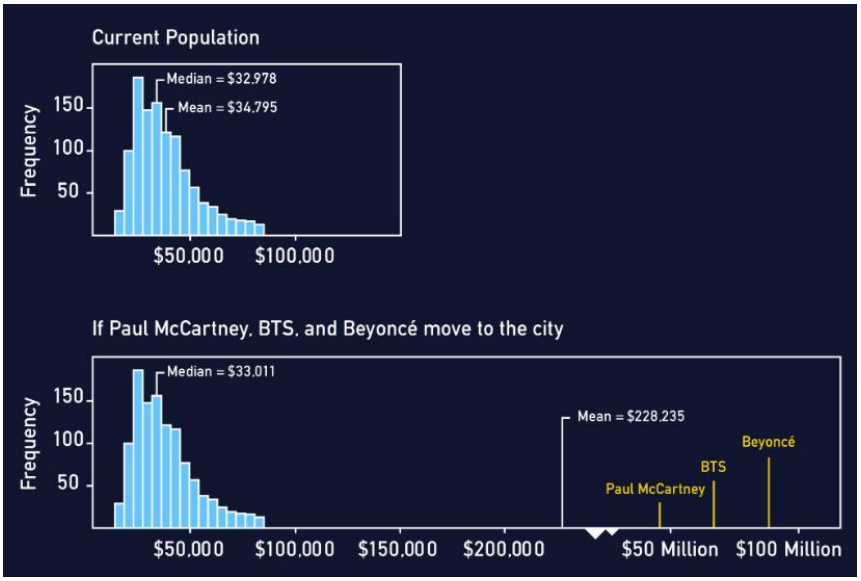
Skill 1.7: Identify outliers and *robust* summary statistics

Skill 1.7 Concepts

For our income data, the difference between the mean ($34,795) and median ($32,978) was only about $2,000. You may be wondering: Is the difference ever larger?

Let's imagine some very famous celebrity musicians have all decided to move to Melody Metropolis. We know celebrities make much more money than the typical musician in our dataset. In the plots below, we've added three new incomes to the distribution:

- **$48 million**: Paul McCartney, British musician of the Beatles
- **$57 million**: BTS, South Korean K-pop band
- **$81 million**: Beyoncé, American singer-songwriter



The second plot shows that the median appears almost unaffected by the addition of these three gigantic incomes: the median moves from $32,978 to $33,011. However, the mean makes a drastic change from $34,795 to $228,235. The mean is now well beyond even the maximum in the original distribution. An income of $228,235 is definitely not a great measure of the center of our income distribution.

These celebrity incomes are examples of *outliers*, extreme values that are distant from the rest of the distribution. Just as with skewness, outliers tend to more heavily influence the mean than the median. This same pattern occurs with measures of spread: the standard deviation is more influenced by outliers and skewness than the interquartile range (IQR).

Because the median and IQR are NOT heavily influenced by extreme values, we say they are *robust*. Robust statistics. are often a better choice to measure the center and spread of a distribution that is skewed or has outliers.

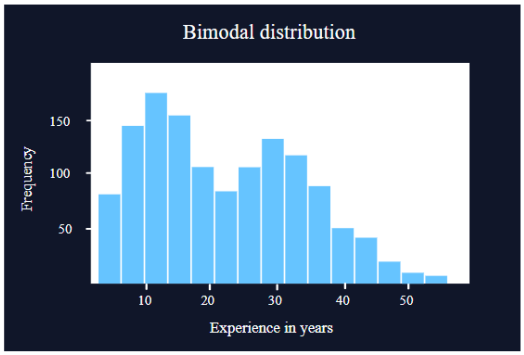
Skill 1.7 Exercise 1

Skill 1.8: Aggregate a dataset

Skill 1.8 Concepts

One measure that we haven't covered that is usually talked about alongside the mean and median is the mode. The mode is defined as the value with the highest frequency, but we can also think of the mode as the value where the peak of the distribution occurs. While not great for computations, the mode can help us identify interesting features in a variable.

For instance, there might be more than one mode, such as in our distribution of years of experience. In the following plot, we can see there's one peak near the 10-year mark and another near the 30-year mark. We would call this distribution *bimodal* because it has two modes.



Sometimes bimodal distributions occur when there are differences across categories of another variable. Given that the city seems to have a lot of young people in bands, let's see if this pattern is reflected when we find the mean of each category of the band variable.

MEAN [IN A BAND]	MEAN [NOT IN A BAND]
14.4 years	26.2 years

These means are very different and very close to the locations of the modes in our plot. This indicates that there may be some differences in experience level between these two groups that are showing up in our distribution plots as two peaks.

By making this separation and then summarizing with the mean, we have **aggregated** our data. In this case, we have aggregated by summarizing a numeric variable (experience) across each value of a categorical variable (band).

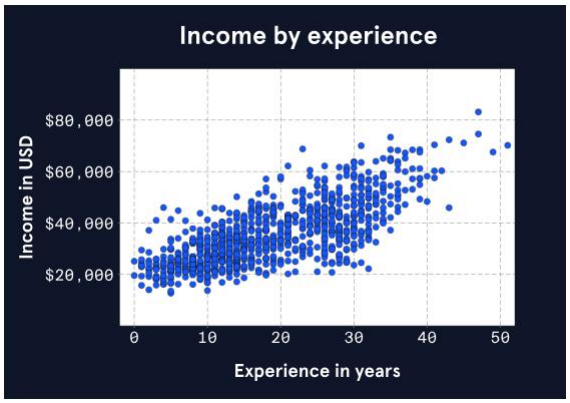
Skill 1.8 Exercise 1

Skill 1.9: Explore relationships between numerical variables

Skill 1.9 Concepts

Aggregating data is a way of exploring variable relationships. We specifically looked at relationships between a numeric variable and a categorical variable, but we should also examine relationships between two numeric variables.

For example, we might wonder: Does musician income vary with years of experience? To start, we can take a look at a *scatter plot* with experience on the x-axis and income on the y-axis. Each point in the plot represents a musician, and the coordinates of that point are the musician's experience (x) and income (y).



The cloud of points in the plot has a pattern. The points move from the lower left to the upper right part of the plot. In other words, lower levels of experience tend to be associated with lower incomes, and higher levels of experience tend to be associated with higher incomes. The points don't form a perfect line though — there is some variation.

We can describe this relationship more precisely by measuring the *correlation coefficient*. This number ranges from -1 to +1 and tells us two things about a linear relationship:

- **Direction**: A positive coefficient means that higher values in one variable are associated with higher values in the other. A negative coefficient means higher values in one variable are associated with lower values of the other.
- **Strength**: The farther the coefficient is from 0, the stronger the relationship and the more the points in a scatter plot look like a line.

The correlation coefficient for income and experience is 0.74 — the relationship is positive and moderately strong.

Skill 1.9 Exercise 1