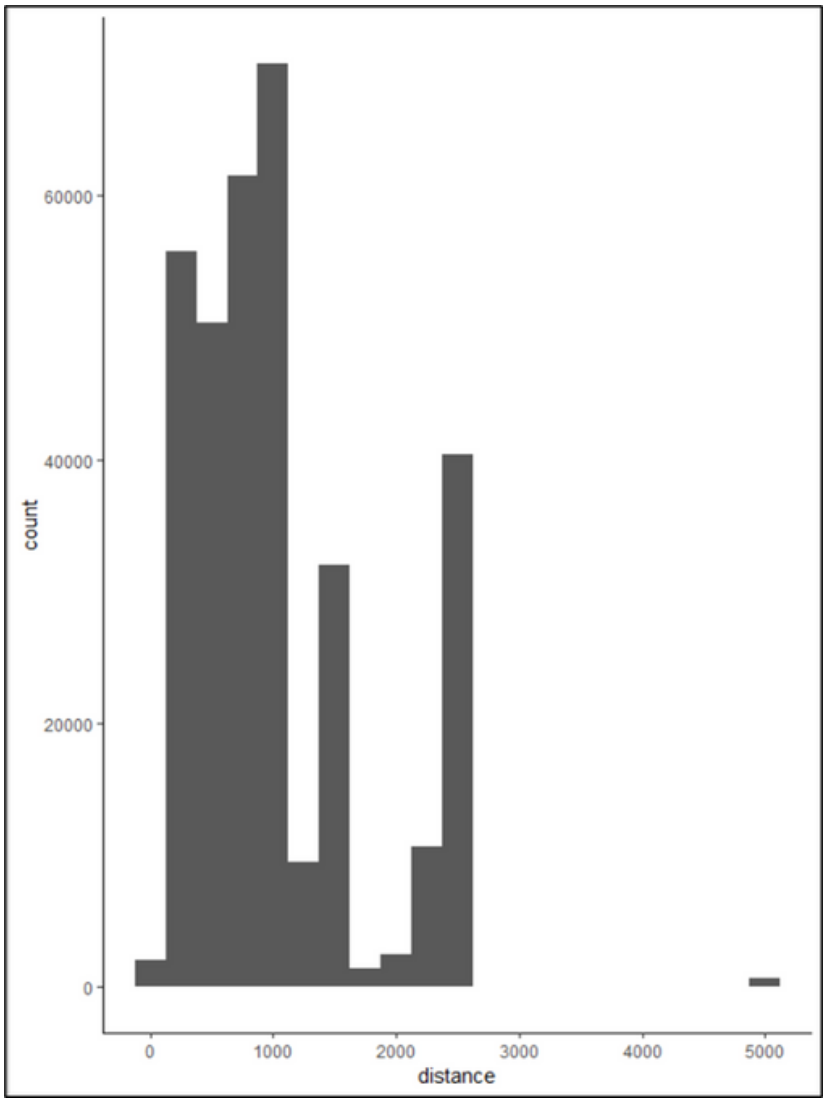
VISUAL DATA ANALYSIS WITH PYTHON



HOW IS THIS USEFUL?

DISTRIBUTION ANALYSIS

Dave ON DATA

Introduction

I have a confession to make. The first time I was introduced to visual data analysis, I didn't get it. It took me many years to understand why I didn't get it. In the end, it boiled down to a simple idea.

When first taught, the visual analysis techniques were not put in the context of answering questions that resonated with me.

For professionals learning any data analysis technique, context is everything. The best way to provide context is to map an analysis technique to the business questions it can answer.

As this document is intended for a broad audience, I will use an example that should resonate broadly - commercial airline travel.

The *nycflights13* dataset used in this document (i.e., the *flights* dataset) provides data regarding 336,776 departing flights from the three New York City area airports (JFK, EWR, and LGA) in 2013.

More info about the dataset can be found here:

- https://github.com/tidyverse/nycflights13

Business Questions

Insightful data analyses always start with a business question. Doesn't matter if your analysis uses visualizations, statistics, or machine learning.

For the *flights* dataset, **distribution analysis** can help answer questions like:

- What is the range of arrival delay times?
- Are most arrival delays relatively small? The opposite? Neither?
- Are longer arrival delays more concentrated at one airport versus other airports?
- What is the range of the distances traveled by departing flights?
- Do more short-distance flights depart from one airport versus other airports?

The questions above are specific to the *flights* dataset, but the pattern of questions applies to all domains!

The Average Arrival Delay

The world's most common predictive model is the average.

We use the average all the time to summarize historical data and use the summary to predict the future.

For example, let's say the average flight arrival delay is 7 minutes.

Most people would use this number to schedule connecting flights, assuming most flight delays are close to the average.

What if it was known that 10% of flights arrive 52+ minutes late?

Do you think that would change people's plans for scheduling connecting flights?

While averages are immensely useful, they obscure the data's underlying **distribution**.

Frequency Distributions

Given a data collection, the frequency distribution is a count of the individual values within the data.

Take the following collection of data as an example.

arr_delay
11
20
12
31
7
12
10
17
5
23

Tally the individual values...

The frequencies

value	count
5	1
7	1
10	1
11	1
12	2
17	1
20	1
23	1
31	1

Notice the following:

1. The average value is 14.8
2. 60% of the values are below 14.8
3. 10% of values are double the average

Frequency distributions give us a more thorough understanding of the data!

Frequency Distribution Bins

There can be many, many distinct values in a collection of data. Creating a frequency distribution based on distinct values quickly becomes unwieldy.

To accommodate large datasets, **bins** are used to group the data.

Define some bins and tally the values that fall in each bin...

arr_delay
11
20
12
31
7
12
10
17
5
23

bin	count
1 to 5	1
6 to 10	2
11 to 15	3
16 to 20	2
21 to 25	1
26 to 30	0
31 to 35	1

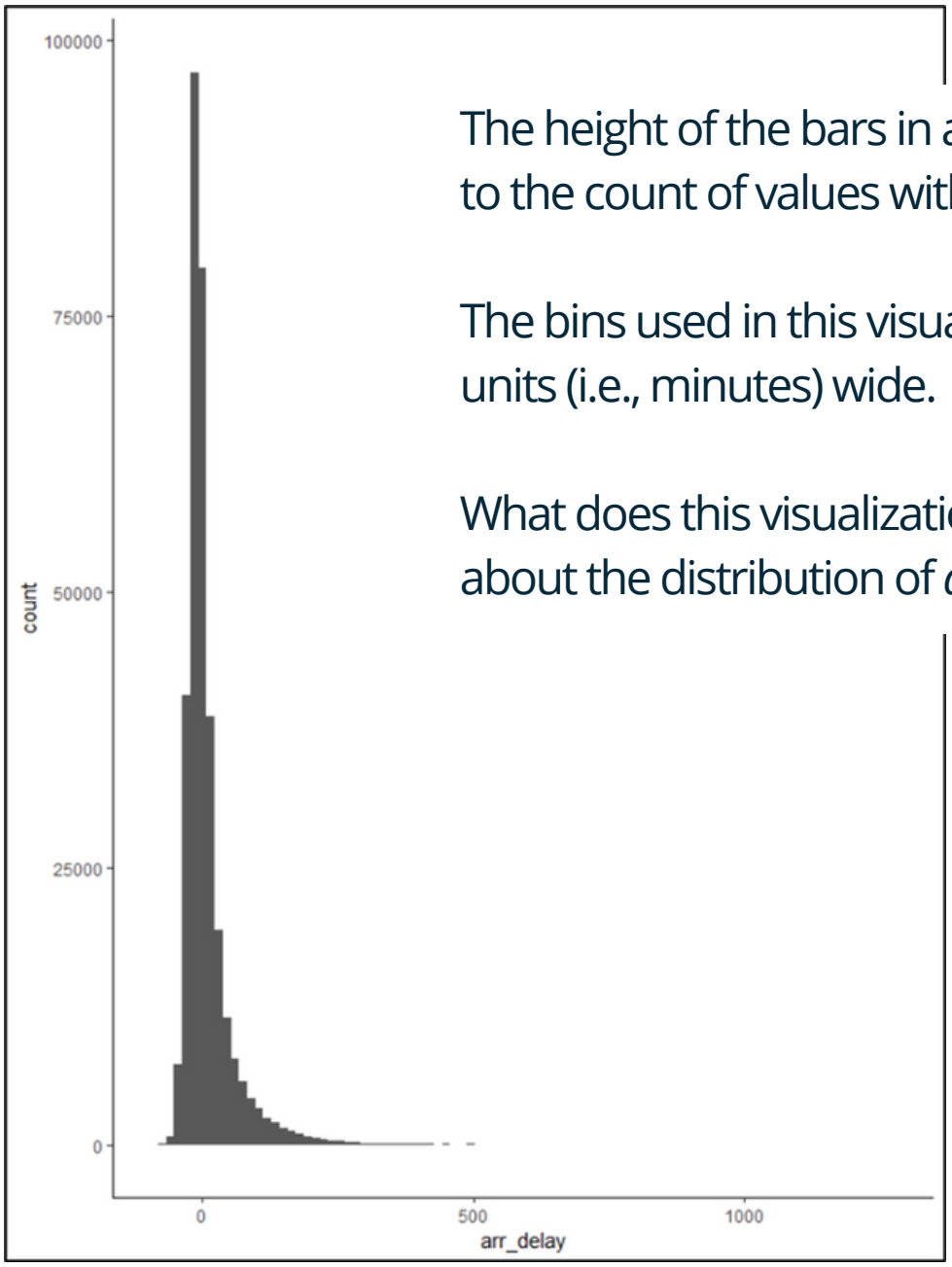
Notice the following:

1. 60% of the data falls in the first three bins
2. There are no values in the "26 to 30" bin
3. Only 10% of values fall in the last two bins

As the data analyst, you define the bins. It is common to experiment with different bin widths while exploring your data.

Histograms

A **histogram** is a frequency distribution visualization. Here's a histogram of the *arr_delay* feature of the *flights* dataset.



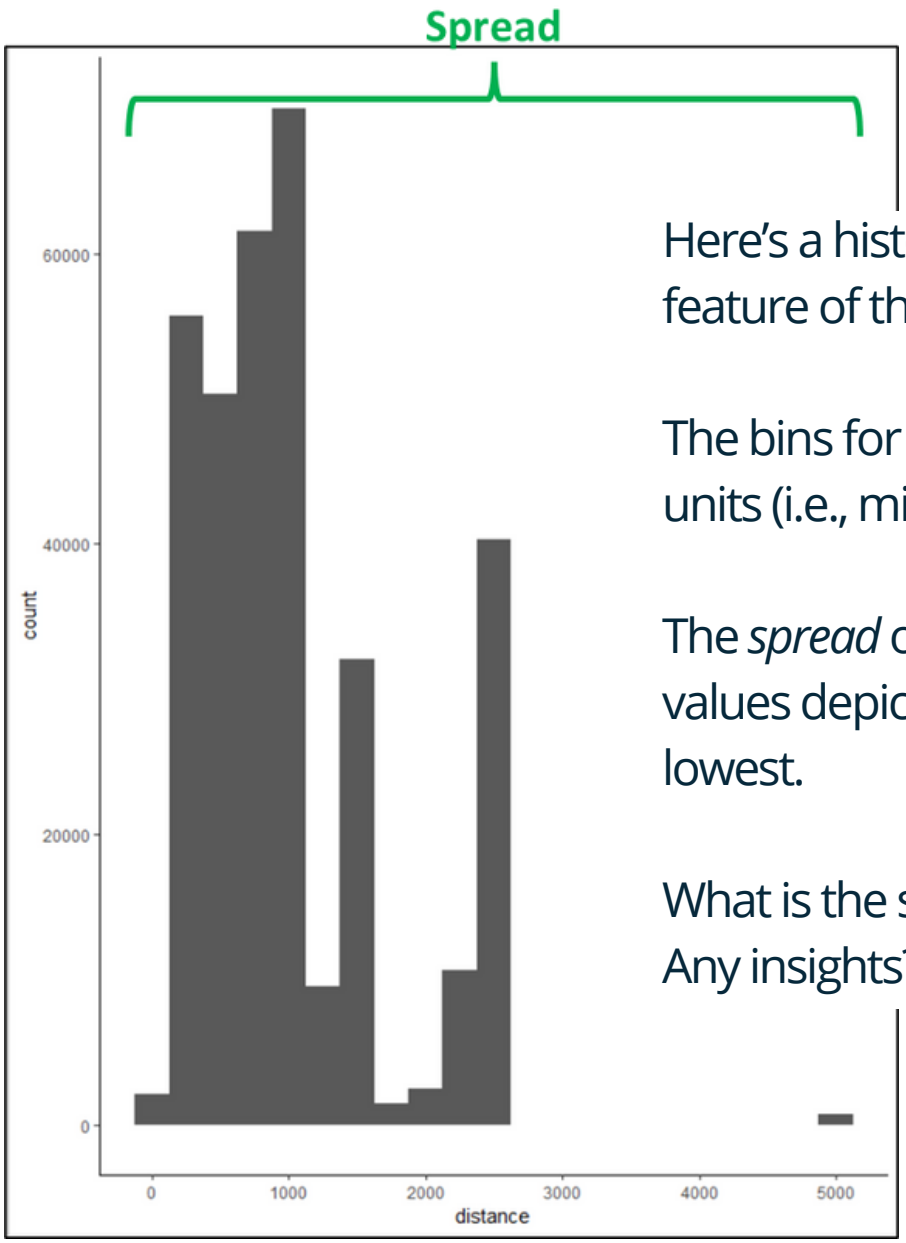
The height of the bars in a histogram corresponds to the count of values within each bin.

The bins used in this visualization are set to be 15 units (i.e., minutes) wide.

What does this visualization tell you at first glance about the distribution of *arr_delay*?

Interpreting Histograms - Spread

When interpreting histograms, you are analyzing three visual characteristics: 1) Spread. 2) Center. 3) Shape.



Here's a histogram of the *distance* feature of the *flights* dataset.

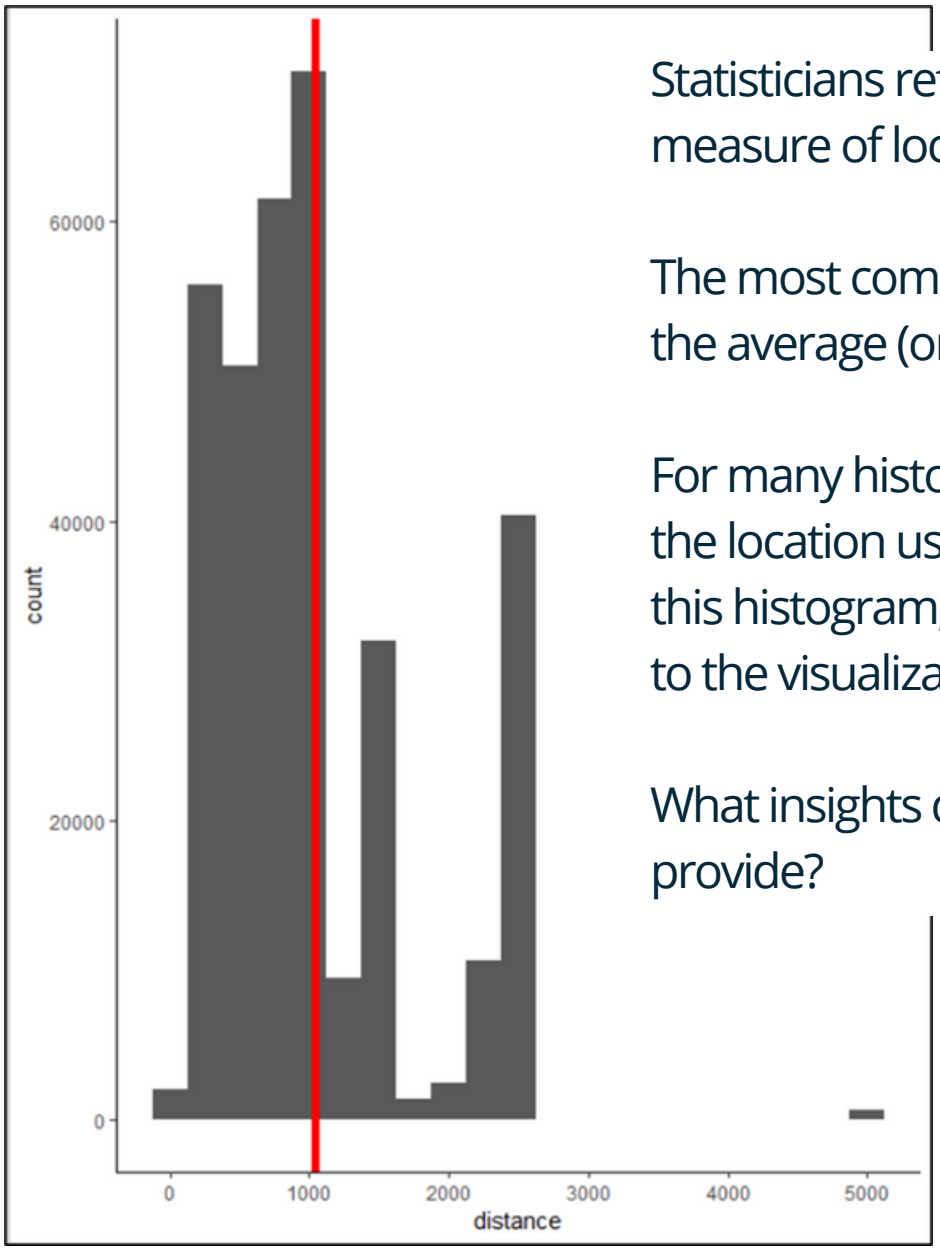
The bins for this histogram are set to 250 units (i.e., miles) wide.

The *spread* of a histogram is the range of values depicted from the highest to the lowest.

What is the spread for this histogram? Any insights?

Interpreting Histograms - Center

When interpreting the *center* of a histogram, you are looking to quantify what a typical value might be.



Statisticians refer to this idea as a measure of location or central tendency.

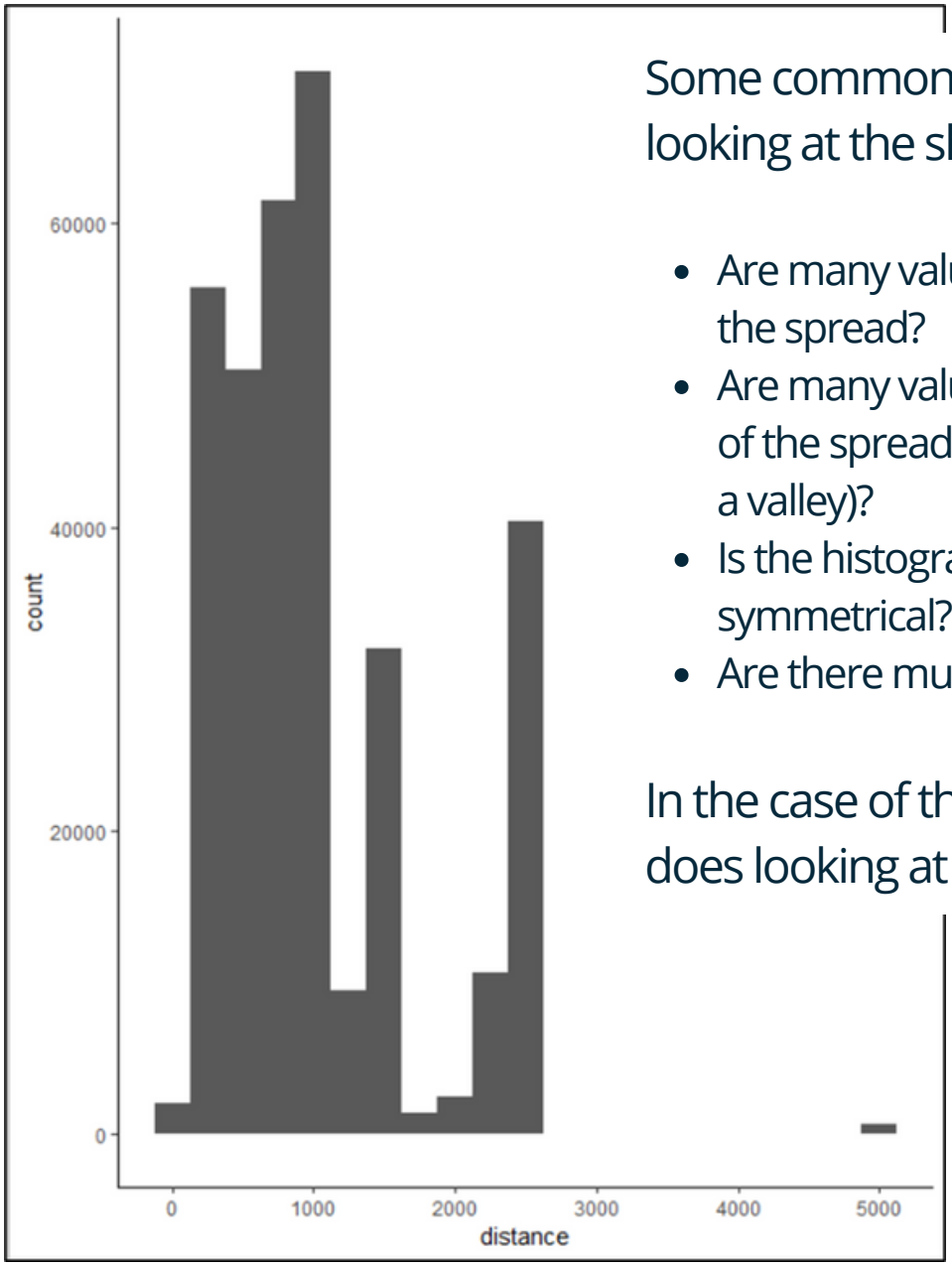
The most common measure of location is the average (or arithmetic mean).

For many histograms, you can estimate the location using your eyes. In cases like this histogram, you can add the average to the visualization.

What insights does the red average line provide?

Interpreting Histograms - Shape

When interpreting the *shape* of a histogram, you are looking to understand the density (i.e., counts) of values throughout the spread.



Some common questions to ask while looking at the shape:

- Are many values bunched up at one end of the spread?
- Are many values bunched up at both ends of the spread (i.e., the histogram looks like a valley)?
- Is the histogram mound-shaped and symmetrical?
- Are there multiple peaks?

In the case of this histogram, what insights does looking at the shape provide?

Histograms with Python

The following Python code creates the histogram on the page 9 using the mighty *plotnine* library:

```
from plotnine import ggplot, theme_classic,
                     geom_histogram, aes,
                     geom_vline

mean_dist = flights['distance'].mean()

(ggplot(flights) +
   theme_classic() +
   geom_histogram(binwidth = 250,
                  mapping = aes(x = 'distance')) +
   geom_vline(mapping = aes(xintercept = mean_dist),
              color = 'red', size = 2))
```

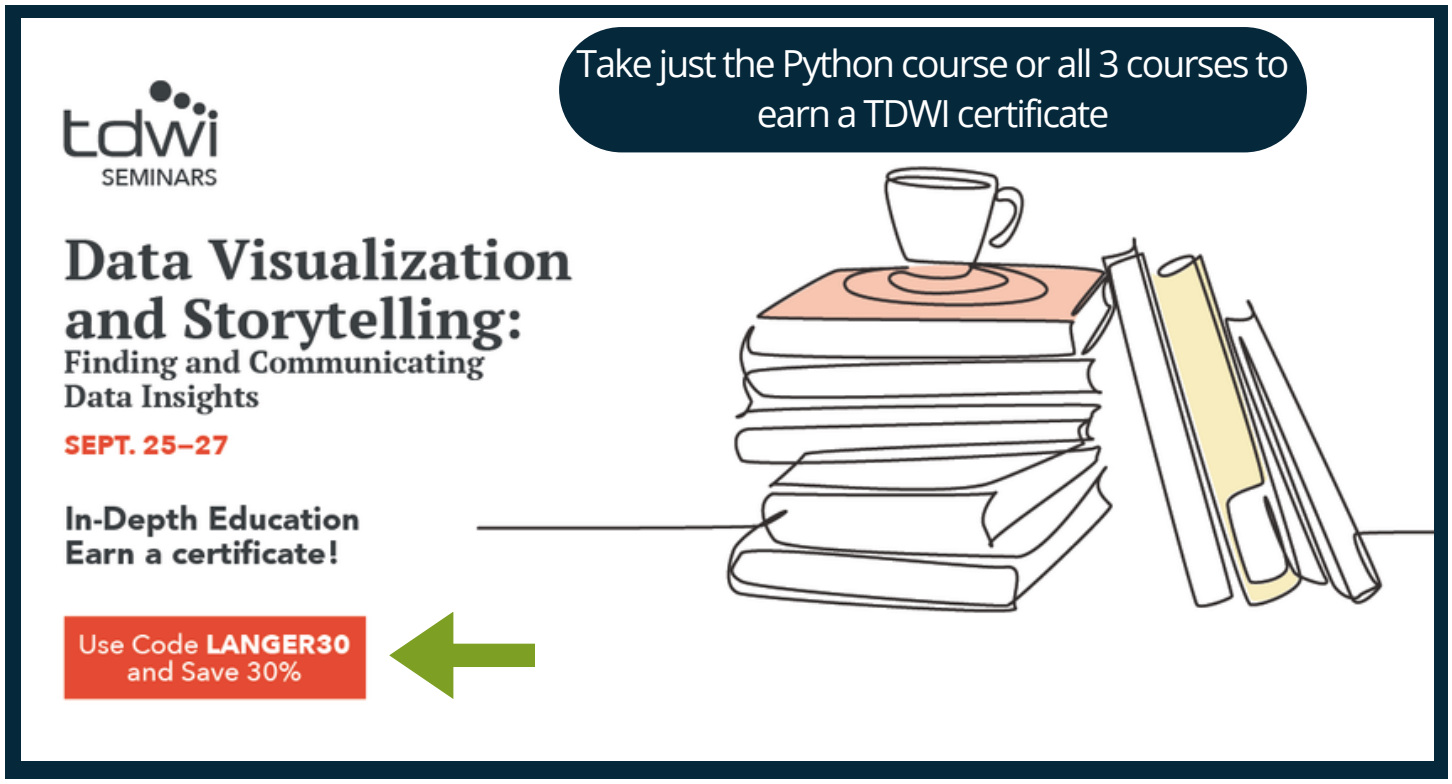
Jumpstart Your Data Analysis Skills

The content in this document comes from the live virtual training course taking place September 26th, 2023 :

- Visual Data Analysis with Python

This course includes 4 hands-on Python labs to jumpstart your data analysis skills.

Be sure to check with your manager. TDWI is an approved training vendor for many organizations.



Top-Rated Live Training

My live teaching is consistently top-rated by attendees. I combine an engaging style with many hands-on labs to build skills. Like Tyler, I can empower you with "go do" tools you can apply at work immediately.

No experience with Python? No worries!

Attendees of my live Python training courses will get **free access to a 4-hour Python online tutorial**. The Python Quick Start gives you the foundation you need to start learning data analysis.

About the Author

My name is Dave Langer and I am the founder of Dave on Data.

I'm a hands-on analytics professional, having used my skills with Excel, SQL, and R/Python to craft insights, advise leaders, and shape company strategy.

I'm also a skilled educator, having trained 100s of working professionals in live in-person classroom settings and 1000s more via live virtual training and online courses.

In the past, I've held analytics leaderships roles at Schedulicity, Data Science Dojo, and Microsoft.

Drop me an email if you have any questions: dave@daveondata.com