Project Overview & Data  
The Business Problem  
You’re working for a marketing consultancy firm (congrats on the new job!) One of your  
clients is a company that designs operating systems, and they want to build a major  
apps store into their user interface. To this end, they want to know whether Google Play  
apps have higher reviews on average than Apple Store apps (or vice versa), as they’re  
intending to strike a deal with just one of these companies.  
To complete this project, you’ll work through a series of steps known as the Data  
Science Pipeline, or DSP. The DSP is a general and fairly broad sequence of steps data  
scientists use to tackle and solve business problems. While working on this project,  
you’ll complete each step of the DSP, including:  
1. Sourcing and loading data  
2. Cleaning and transforming data  
3. Visualizing and modeling data  
4. Evaluating, presenting findings, and concluding your work  
Download the Data  
And now, for the first step! Download both the Google and Apple data from here and  
here, respectively. You may need to make a free Kaggle account in order to do this; this  
is a useful thing to do in any case, as you’re likely to use Kaggle during your career as a  
data scientist!

**Part 1: Sourcing and Loading**

Save

20 - 30 Minutes

7 Points

Sourcing and loading is the first stage of the Data Science Pipeline. Sourcing means finding the data itself, which can be trickier than it first seems. Happily, for this project, we’ve sourced the data from Kaggle and know that it’s a good, reliable source of relatively clean data.

A picture containing clipart

Description automatically generatedFor each step of this project, **work through the tier you feel comfortable with: tier 1 is the easiest, and tier 3 is the hardest.** The only difference between the tiers is that there are more blanks to fill in as the tier increases.

**Project Files** *(Please note that each of these tiered files contain separate Jupyter notebooks for part 1, 2, and 3 of this project. Please only work on part 1 right now.)*

* [**Tier 1**](https://www.springboard.com/archeio/download/27e436773c40449fb6189da70a153a61/)
* [**Tier 2**](https://www.springboard.com/archeio/download/5114aae4936e4d419ea0d1e34c69c05f/)
* [**Tier 3**](https://www.springboard.com/archeio/download/8223c619283642bbacbb403a179c8750/)

A note about sourcing data after you've completed this course: when sourcing data for future projects, it’s important to consider:

* What available sources of data there are
* How reliable those sources are — that is, the accuracy of the information captured in that data
* How clean that data is — how much work will it be to transform the data into a form where it can be accurately modeled
* Whether there are any relevant data protection regulations we have to abide by

Loading is typically a little easier if we’ve done the sourcing right. We just have to read the data into our development environment. If the dataset is very large, we can use big data tools like Apache Spark.

# Springboard Apps project - Tier 1 - Sourcing and Loading

Welcome to the final project of this Springboard prep course! To give you a taste of your future career, we're going to walk through exactly the kind of notebook that you'd write as a data scientist. In the process, we'll be sure to signpost the general framework for our investigation - the Data Science Pipeline - as well as give reasons for why we're doing what we're doing.

**Brief**

Did Apple Store apps receive better reviews than Google Play apps?

## Stages of the project

1. Sourcing and loading
   * Load the two datasets
   * Pick the columns that we are going to work with
   * Subsetting the data on this basis
2. Cleaning, transforming and visualizing
   * Check the data types and fix them
   * Add a platform column to both the Apple and the Google dataframes
   * Changing the column names to prepare for a join
   * Join the two data sets
   * Eliminate the NaN values
   * Filter only those apps that have been reviewed at least once
   * Summarize the data visually and analytically (by the column platform)
3. Modelling
   * Hypothesis formulation
   * Getting the distribution of the data
   * Permutation test
4. Evaluating and concluding
   * What is our conclusion?
   * What is our decision?
   * Other models we could have used.

## Importing the libraries

In this case we are going to import pandas, numpy, scipy, random and matplotlib.pyplot

import pandas as pd

import numpy as np

import matplotlib as plt

# scipi is a library for statistical tests and visualizations

from scipy import stats

# random enables us to generate random numbers

import random

## Stage 1 - Sourcing and loading data

### 1a. Source and load the data

Let's download the data from Kaggle. Kaggle is a fantastic resource: a kind of social medium for data scientists, it boasts projects, datasets and news on the freshest libraries and technologies all in one place. The data from the Apple Store can be found [here](https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps) and the data from Google Store can be found [here](https://www.kaggle.com/lava18/google-play-store-apps). Download the datasets and save them in your working directory.

# Now that the files are saved, we want to load them into Python using read\_csv and pandas.

# Create a variable called google, and store in it the path of the csv file that contains your google dataset.

# If your dataset is in the same folder as this notebook, the path will simply be the name of the file.

google = \_ \_ \_

# Read the csv file into a data frame called Google using the read\_csv() pandas method.

Google = pd.read\_csv(\_ \_ \_)

# Using the head() pandas method, observe the first three entries.

Google.\_ \_

# Create a variable called apple, and store in it the path of the csv file that contains your apple dataset.

apple = \_ \_ \_

# Read the csv file into a pandas DataFrame object called Apple.

Apple = \_ \_ \_

# Observe the first three entries like you did with your other data.

Apple.\_ \_ \_

**1b. Pick the columns we'll work with**

From the documentation of these datasets, we can infer that the most appropriate columns to answer the brief are:

1. Google:
   * Category # Do we need this?
   * Rating
   * Reviews
   * Price (maybe)
2. Apple:
   * prime\_genre # Do we need this?
   * user\_rating
   * rating\_count\_tot
   * price (maybe)

**1c. Subsetting accordingly**

Let's select only those columns that we want to work with from both datasets. We'll overwrite the subsets in the original variables.

# Subset our DataFrame object Google by selecting just the variables ['Category', 'Rating', 'Reviews', 'Price']

\_ \_ \_ = Google[['Category', 'Rating', 'Reviews', 'Price']]

# Check the first three entries

\_ \_ \_

# Do the same with our Apple object, selecting just the variables ['prime\_genre', 'user\_rating', 'rating\_count\_tot', 'price']

Apple = \_ \_ \_

# Let's check the first three entries

\_ \_ \_

1

**Part 2: Cleaning, Transforming, and Visualizing**

Saved

50 Minutes - 1 Hour 15 Minutes

19 Points

For this part of the project, you’ll:

* Check data types and fix them
* Add a `platform` column to both the `Apple` and the `Google` dataframes
* Change the column names to prepare for a join
* Join the two data sets
* Eliminate the `NaN` values
* Filter only those apps that have been reviewed at least once
* Summarize the data visually and analytically (by the column `platform`)

As always, if you have any questions, please reach out to your mentor or online community.

**Project Files***(Please note that each of these tiered files contain separate Jupyter notebooks for part 1, 2, and 3 of this project. Please only work on part 1 right now.)*

* [**Tier 1**](https://www.springboard.com/archeio/download/27e436773c40449fb6189da70a153a61/)
* [**Tier 2**](https://www.springboard.com/archeio/download/5114aae4936e4d419ea0d1e34c69c05f/)
* [**Tier 3**](https://www.springboard.com/archeio/download/8223c619283642bbacbb403a179c8750/)

'Course Logo

Icon

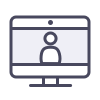
Description automatically generatedCharito

[Community](https://springboard-2784014.hs-sites.com/community?__hstc=267612241.29d59d13a224641ba418b0e0a4fba7e0.1659832822290.1668741940537.1669063883515.27&__hssc=267612241.1.1669063883515&__hsfp=2923744165)

How was your call on Sunday, November 20?

Search

My Progress

[Office Hour Videos](https://www.youtube.com/playlist?list=PLjKxfZpleE6ARDh12WsI12JjdKUN8J2nb)

scroll icon

* 1

[Program overview](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/44839)

  2

[Intro to Data Science & Python](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10671)

  3

[Intermediate Python](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10690)

  4

[Foundations of Probability](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10700)

  5

[Python: Data Structures & Algorithms](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10703)

  6

[Technical Skills Survey](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10712)

  7

[Your Data Science Toolbox](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10720)

* 8

[The Apps Project](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10727)



* + 8.1

[Context and Data](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10728) 2/2

* + 8.2

[Cleaning, Transforming, & Visualizing Data](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10731) 1/1

* + 8.3

[Statistics](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10733) 0/2

* + 8.4

[Data Storytelling](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10736) 0/4

* + 8.5

[Prework: Python](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18552) 0/6

* 9

[Congratulations!](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/10742)

* 10

[What is Data Science?](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18565)



* 11

[Problem Identification](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18576)



* 12

[The Python Data Science Stack](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18594)



* 13

[Creating Your Job Search Strategy](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18610)



* 14

[Applying the Data Science Method](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18619)



* 15

[Data Wrangling](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18648)



* 16

[SQL & Databases](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18674)



* 17

[Your Elevator Pitch and LinkedIn Profile](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18694)



* 18

[Statistics for Exploratory Data Analysis](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18708)



* 19

[Python Statistics in EDA](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18751)



* 20

[Effective Networking](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18769)



* 21

[Machine Learning Overview](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18786)



* 22

[Supervised Learning](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18803)



* 23

[Unsupervised Learning](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18856)



* 24

[Feature Engineering](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18892)



* 25

[Informational Interviews](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18911)



* 26

[Machine Learning Applications](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18921)



* 27

[Find the Right Job Title and Companies](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18940)



* 28

[Data Storytelling](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18954)



* [Specializations](https://www.springboard.com/workshops/data-science-career-track-f2c/learn#/curriculum/18978)

scroll icon

**Statistics**

All data science projects entail the use of at least some statistics, and this one is no exception. We’ve made the statistics as light and intuitive as possible however, and have justified the use of statistical techniques with reference to our specific problem. There’s plenty of time for you to become more comfortable with statistics while working through the Data Science Career Track course itself. Let's dive in!

1

**Part 3: Modeling**

Save

40 Minutes - 1 Hour

16 Points

Data scientists use a variety of models to draw insights from datasets; the model they choose to apply to their specific dataset depends on the kind of problem they’re trying to solve. For this project, you’ll apply a basic permutation test that requires very few assumptions for it to be properly applied.

**Project Files***(Please note that each of these tiered files contain separate Jupyter notebooks for part 1, 2, and 3 of this project. Please only work on part 1 right now.)*

* [**Tier 1**](https://www.springboard.com/archeio/download/27e436773c40449fb6189da70a153a61/)
* [**Tier 2**](https://www.springboard.com/archeio/download/5114aae4936e4d419ea0d1e34c69c05f/)
* [**Tier 3**](https://www.springboard.com/archeio/download/8223c619283642bbacbb403a179c8750/)

Other examples of models you may find yourself using in the future include:

* An unsupervised learning model will help you find patterns in data when you don’t know in advance what it is that you’re looking for.
* A supervised learning model will help you classify unseen examples into various categories after having presented your machine with many examples of correctly categorized examples.
* A time series model will help you predict the values of a given variable based on its historical values.

Have you ever been inspired to action by a speech or presentation? Learning how to share your insights in an engaging way will help to instill those same feelings of inspiration in others.

How you express a message is often just as important — if not more important — than the message itself. Even if you find a really valuable insight through data analysis, you’re likely to have a hard time putting that insight into action if you don’t communicate your thoughts about it properly. That’s where storytelling comes in — when you organize your insights into a good story and tell that story effectively, you’ll be much more likely to have a positive impact on the project you're working on. Data scientists frequently need to communicate what they've learned about a problem through their work. This subunit will teach you some techniques that will help you to communicate what you've discovered about the Yelp dataset.

# Data Storytelling: The Essential Data Science Skill Everyone Needs

[Brent Dykes](https://www.forbes.com/sites/brentdykes/)

Contributor

I write about how to drive more value with data and analytics.

Mar 31, 2016,11:26am EDT

##### Tweet This

 [People hear statistics, but they feel stories](https://twitter.com/intent/tweet?url=http%3A%2F%2Fwww.forbes.com%2Fsites%2Fbrentdykes%2F2016%2F03%2F31%2Fdata-storytelling-the-essential-data-science-skill-everyone-needs%2F&text=People%20hear%20statistics%2C%20but%20they%20feel%20stories)

*  [People hear statistics, but they feel stories](https://twitter.com/intent/tweet?url=http%3A%2F%2Fwww.forbes.com%2Fsites%2Fbrentdykes%2F2016%2F03%2F31%2Fdata-storytelling-the-essential-data-science-skill-everyone-needs%2F&text=People%20hear%20statistics%2C%20but%20they%20feel%20stories)

  Once your business has started collecting and combining all kinds of data, the next elusive step is to extract value from it. Your data may hold tremendous amounts of potential value, but not an ounce of value can be created unless insights are uncovered and translated into actions or business outcomes. During a [2009 interview](http://www.mckinsey.com/industries/high-tech/our-insights/hal-varian-on-how-the-web-challenges-managers), Google’s Chief Economist Dr. Hal R.Varian stated, "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades." Fast forward to 2016 and many businesses would agree with Varian’s astute assessment.



Google's Chief Economist, Dr. Hal R. Varian predicted the growing importance of data skills back in... [+]

As data becomes increasingly ubiquitous, companies are desperately searching for talent with these data skills. LinkedIn recently reported data analysis is [one of the hottest skill categories over the past two years](http://blog.linkedin.com/2016/01/12/the-25-skills-that-can-get-you-hired-in-2016/) for recruiters, and it was the only category that consistently ranked in the top 4 across all of the countries they analyzed. Interestingly, much of the current hiring emphasis has centered on the data preparation and analysis skills—not the "last mile" skills that help convert insights into actions. Many of the heavily-recruited individuals with advanced degrees in economics, mathematics, or statistics struggle with communicating their insights to others effectively—essentially, telling the story of their numbers.

[**Subscribe Now: Forbes Entrepreneurs & Small Business Newsletters**](https://app.e2ma.net/app2/audience/signup/1788435/1752888/?v=a)

PROMOTED

[All the trials and triumphs of building a business – delivered to your inbox.](https://app.e2ma.net/app2/audience/signup/1788435/1752888/?v=a)

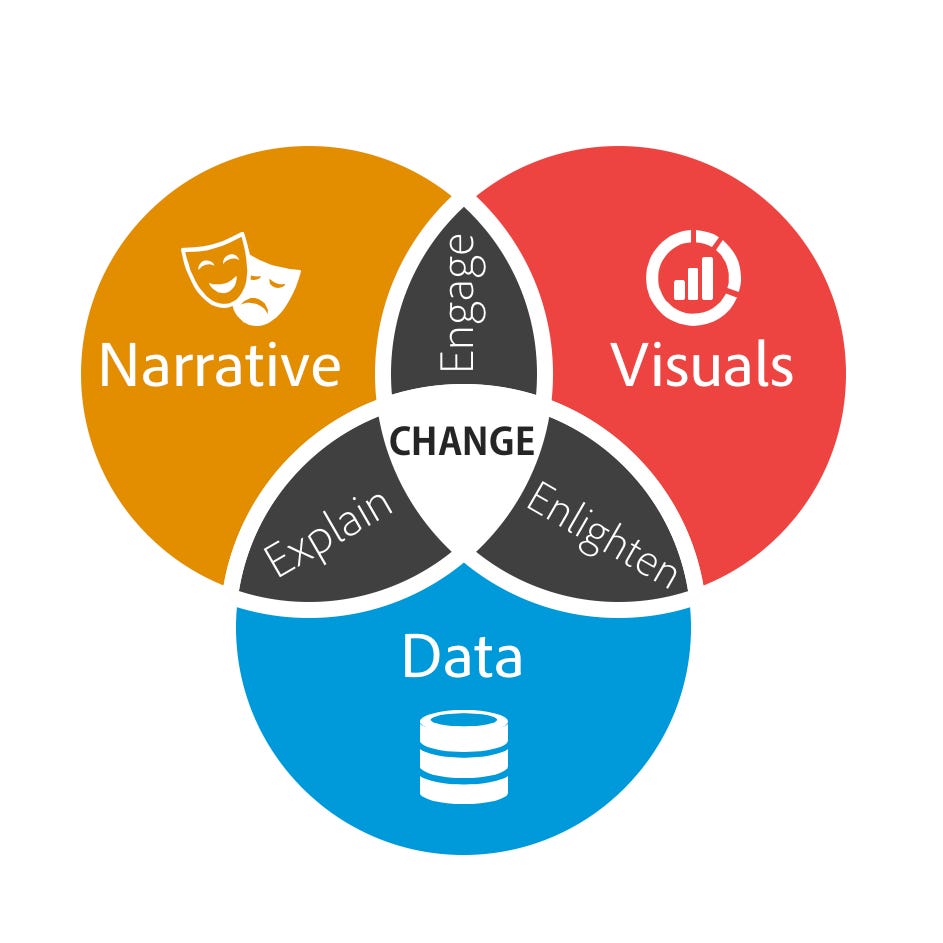
The need for more data storytellers is only going to increase in the future. With the shift towards more self-service capabilities in analytics and business intelligence, the pool of people generating insights will expand beyond just analysts and data scientists. This new breed of data tools will make it easier for people across business functions to access and explore the data on their own. As a result, we’re going to see an unprecedented number of insights being generated within companies than ever before. However, unless we can improve the communication of these insights we will also see a poorer insight-to-value conversion rate. If an insight isn’t understood and isn’t compelling, no one will act on it and no change will occur.

Data visualization expert Stephen Few said, “Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.” Any insight worth sharing is probably best shared as a data story. The phrase “data storytelling” has been associated with many things—data visualizations, infographics, dashboards, data presentations, and so on. Too often data storytelling is interpreted as just visualizing data effectively, however, it is much more than just creating visually-appealing data charts. Data storytelling is a structured approach for communicating data insights, and it involves a combination of three key elements: data, visuals, and narrative.

It’s important to understand how these different elements combine and work together in data storytelling. When narrative is coupled with data, it helps to **explain** to your audience what’s happening in the data and why a particular insight is important. Ample context and commentary are often needed to fully appreciate an insight. When visuals are applied to data, they can **enlighten** the audience to insights that they wouldn’t see without charts or graphs. Many interesting patterns and outliers in the data would remain hidden in the rows and columns of data tables without the help of data visualizations.



Finally, when narrative and visuals are merged together, they can **engage** or even entertain an audience. It’s no surprise we collectively spend billions of dollars each year at the movies to immerse ourselves in different lives, worlds, and adventures. When you combine the right visuals and narrative with the right data, you have a data story that can influence and drive **change**.



**Why data storytelling is essential**

For thousands of years, storytelling has been an integral part of our humanity. Even in our digital age, stories continue to appeal to us just as much as they did to our ancient ancestors. Stories play a vibrant role in our daily lives—from the entertainment we consume to the experiences we share with others to what we conjure up in our dreams.

Modern-day storytelling is often associated with the popular [TED conference series](https://www.ted.com/talks) and its slogan of “Ideas Worth Spreading.” Analysis of the most popular 500 TED Talk presentations found that [stories made up at least 65%](http://www.forbes.com/sites/carminegallo/2014/02/28/how-sheryl-sandbergs-last-minute-addition-to-her-ted-talk-sparked-a-movement/#1e705d8c6ea1) of their content. Throughout time, storytelling has proven to be a powerful delivery mechanism for sharing insights and ideas in a way that is memorable, persuasive, and engaging.

For some people, crafting a story around the data may seem like an unnecessary, time-consuming effort. They may feel the insights or facts should be sufficient to stand on their own as long as they’re reported in a clear manner. They may believe the revealed insights alone should influence the right decisions and drive their audience to act. Unfortunately, this point of view is based on the flawed assumption that business decisions are based solely on logic and reason.

In fact, neuroscientists have confirmed decisions are often based on emotion, not logic. USC professor Antonio Damasio found patients, who had brain damage in an area that helped to process emotions (prefrontal cortex), [struggled to make basic decisions when choosing between alternatives](https://www.youtube.com/watch?v=1wup_K2WN0I,%20last%20accessed%20on%2021/09/2015). Deciding on where to eat or when to schedule an appointment turned into lengthy cost-benefit debates for these individuals. Interestingly, these patients’ decision-making skills were significantly impaired by the lack of emotional judgment. Emotion actually plays an essential role in helping our brains to navigate the alternatives and arrive at a timely decision.

When you package up your insights as a data story, you build a bridge for your data to the influential, emotional side of the brain. When neuroscientists observed the effects detailed information had on an audience, brain scans revealed it only activated two areas of the brain associated with language processing: Broca’s area and Wernicke’s area. However, when someone is absorbed in a story, they discovered it [stimulated more areas of the brain](http://www.nytimes.com/2012/03/18/opinion/sunday/the-neuroscience-of-your-brain-on-fiction.html). [People hear statistics, but they feel stories](https://twitter.com/intent/tweet?url=http%3A%2F%2Fwww.forbes.com%2Fsites%2Fbrentdykes%2F2016%2F03%2F31%2Fdata-storytelling-the-essential-data-science-skill-everyone-needs%2F&text=People%20hear%20statistics%2C%20but%20they%20feel%20stories). This subtle but important difference pays dividends for data storytellers in a few key ways:

* **Memorability:** A study by Stanford professor Chip Heath (Made to Stick author) found [63% could remember stories](http://mannerofspeaking.org/2009/10/13/making-it-stick-tell-stories/), but only 5% could remember a single statistic. While 2.5 statistics were used on average in the exercise and only 10% of the participants incorporated a story, the stories are what caught people’s attention.
* **Persuasiveness:** In another study, researchers tested two variations of a brochure for the [Save the Children](http://www.savethechildren.org/site/c.8rKLIXMGIpI4E/b.6115947/k.B143/Official_USA_Site.htm) charity organization. The story-based version [outperformed the infographic version by $2.38 to $1.14](http://www.nationalpost.com/story.html?id=8cf455d9-8634-4cb1-972f-03dd9621350b) in terms of per participant donations. Various statistics on the plight of African children were far less persuasive than the story of Rokia, a seven-year-old from Mali, Africa.
* **Engagement:** Researchers also discovered people enter into a trance-like state, where they [drop their intellectual guard and are less critical and skeptical](http://www.fastcocreate.com/1680581/why-storytelling-is-the-ultimate-weapon). Rather than nitpicking over the details, the audience wants to see where the story leads them. As mathematician John Allen Paulos observed, “In listening to stories we tend to suspend disbelief in order to be entertained, whereas in evaluating statistics we generally have an opposite inclination to suspend belief in order not to be beguiled.”

In a previous article, I shared an [account of Ignaz Semmelweis](http://www.forbes.com/sites/brentdykes/2016/02/09/a-history-lesson-on-the-dangers-of-letting-data-speak-for-itself/#57113f667ced), a mid-nineteenth century obstetrician, who discovered hand washing could save countless lives but failed to communicate his findings effectively to a skeptical medical community. His data was ignored, his life-saving ideas were rejected, and he was sadly discredited by his colleagues.

Many bold, incredible insights will suffer a similar fate if they are not successfully molded into data stories. Uncovering key insights is one skill and communicating them is another—both are equally critical to deriving value from the data your business is now amassing. Data storytelling represents an exciting, new field of expertise where art and science truly converge. My hope is more data storytellers—from across an organization—will emerge to ensure the survival and adoption of more transformative insights.

Follow Brent Dykes on Twitter:

### Prework: Python

Python and data science go together like peanut butter and jelly. Python is the most popular programming language for data science tasks and comes with a robust ecosystem full of built-in tools and libraries. This subunit is packed with resources that will set you up for success as you kick off your data science journey. We've partnered with DataCamp to bring you resources that go over the foundations of the Python language.

A picture containing clipart

Description automatically generatedIf you completed our Data Science Prep course, the DataCamp interface may remind you of NextTech. However, the DataCamp resources feature video lessons from experts prior to the hands-on instructions.

You need to create a DataCamp account to start the course. Springboard School of Data has partnered with DataCamp to provide you with free access to interactive Python programming courses while you’re actively enrolled in the data science career track. Please reach out to [support@springboard.com](https://www.springboard.com/workshops/data-science-career-track-f2c/support@springboard.com.) to receive access.

Jupyter Notebook  
Jupyter Notebook is an application that lets you create and share documents that contain live  
code, equations, visualizations, and markdown text. It allows you to prototype code rapidly and  
combine it with useful documentation. Throughout your course, you’ll receive Jupyter Notebooks  
with assignment instructions, and you’ll upload completed notebooks for grading. In this section,  
you’ll install Jupyter Notebook.  
The only thing you need to do to complete this section is to make sure that Jupyter is  
installed and running on your machine. But since any development environment is unfamiliar  
at first, you’re welcome to check out any of the following optional resources to get more  
comfortable with the environment:  
- This video gives you a few awesome tips and tricks to become more adept with Jupyter.  
- This video is a neat and concise introduction to Jupyter’s many features (Optional).  
- The official Jupyter documentation provides a more in-depth account of Jupyter’s  
features (OptionaJupyter Project Documentation — Jupyter Documentation 4.1.1 alpha  
documentationl).  
Psst! If you’re struggling to set the default Jupyter kernel to Python 3 due to having multiple  
versions of Python on your machine, these extra steps should help you smooth t

When you reach the statistics unit in this course, you'll start working with a book called The Art of Statistics by Sir David Spiegelhalter. Spiegelhalter is a Cambridge professor knighted for his contributions to the fields of statistics, risk, and probability. His research and professional work directly contributed to efforts to combat sports doping, reduce child mortality in heart surgeries, and helped bring one of England’s most notorious killers to justice. As you can see, someone who understands statistics can make a real difference in the world. Statistics are more than distributions, the applications of this nuanced field have an impact and better lives.

Sir David is a powerful orator and a world-renown teacher. Watch him in action in this **optional** lecture. We recommend you start at 3:56 and watch until at least the hour mark.