
DATA ANALYTICS: THE VALUE OF DATA

HELLO!

- ▶ ebernstein.GA@gmail.com
- ▶ generalassemb.ly
- ▶ facebook.com/gnrlassembly
- ▶ Cell 617.606.1078



Course Goals

- Input, transform, and format data for more productive use.
- Analyze data to improve decision making.
- Organize and package data for external consumption.
- Learn how to easily manipulate datasets using pivot tables and other functions.
- Find out how to leverage the onboard data analysis tools to make effective conclusions on data.
- Master the art of shortcuts to streamline the process of analysis in Excel.



Today's Objectives

- ▶ Describe the value of data and how it can lead to informed decisions that have positive impact on many facets of an organization
- ▶ Identify the steps and goals of the analytics workflow
- ▶ Explore Excel both as software (basic layout, navigation, keyboard shortcuts, worksheet organization) and as a data analysis platform (basic math formulas, visualization)
- ▶ Practice basic summary tactics used to familiarize yourself with a dataset



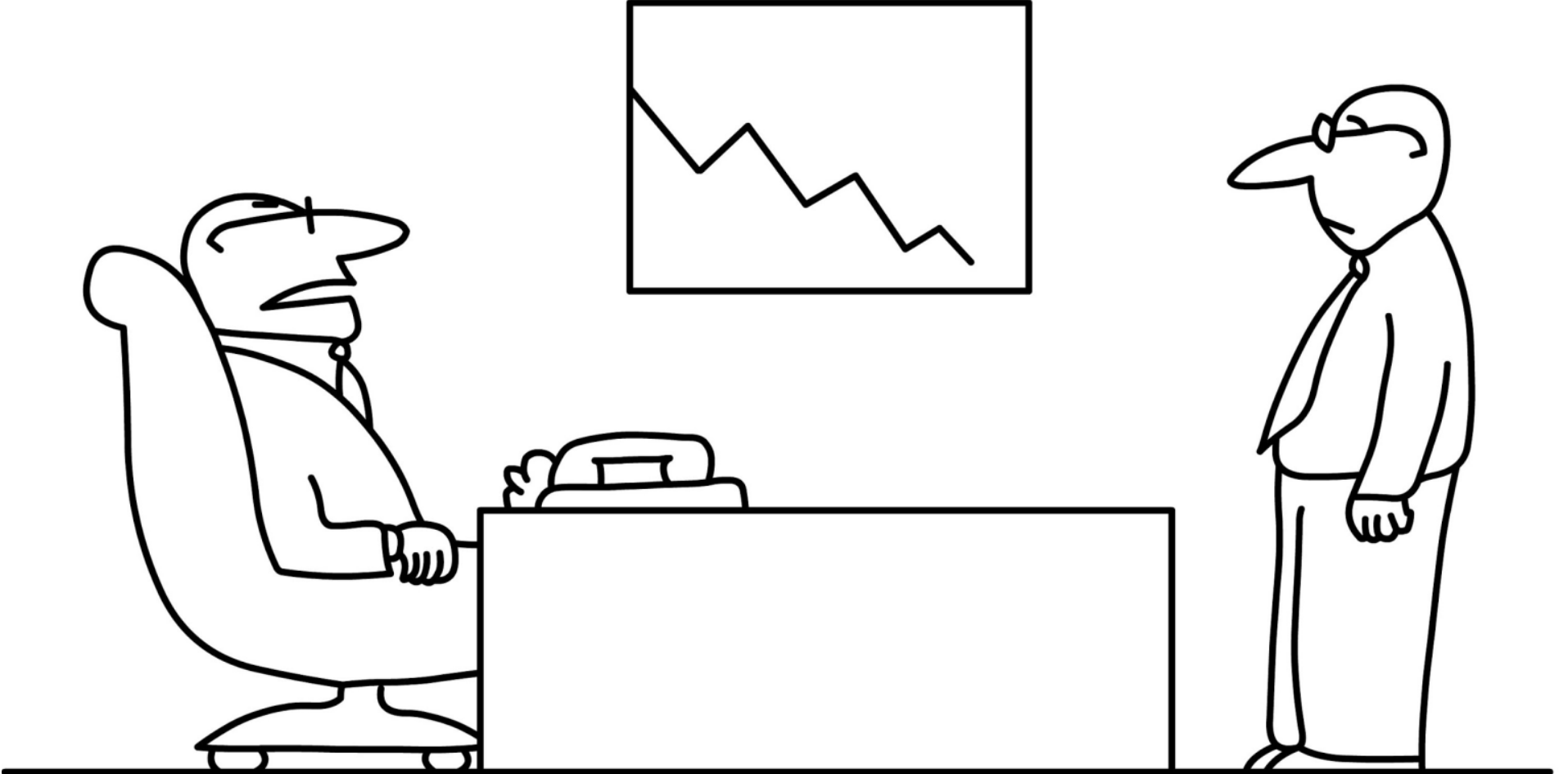
What is Data Analytics?

- ▶ Learn to make sense of data; tell a story; defend your proposal
- ▶ We can store data points, but learning from them is an entirely different skill.



Value of Data

- ▶ What is the value of Data in the Modern World?



©2008 www.timoelliott.com

"It would appear, Hopkins, that your gut feel was only indigestion"



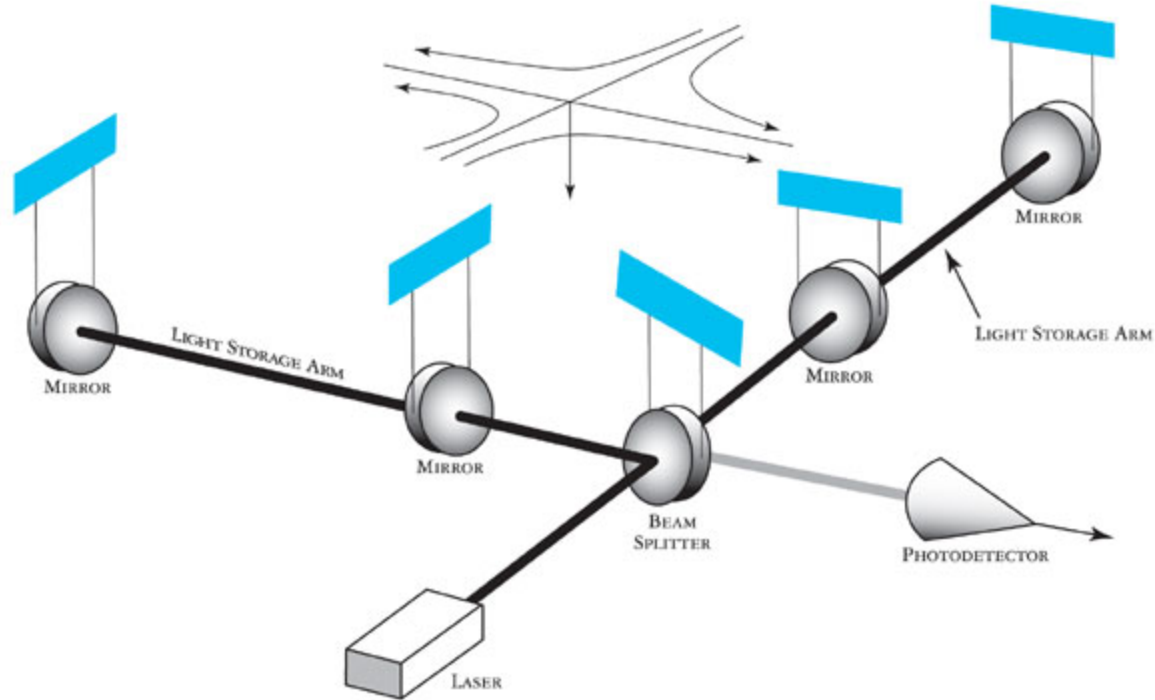
LIGO Collaboration



Gravitational Waves Detected 100 Years After Einstein's Prediction

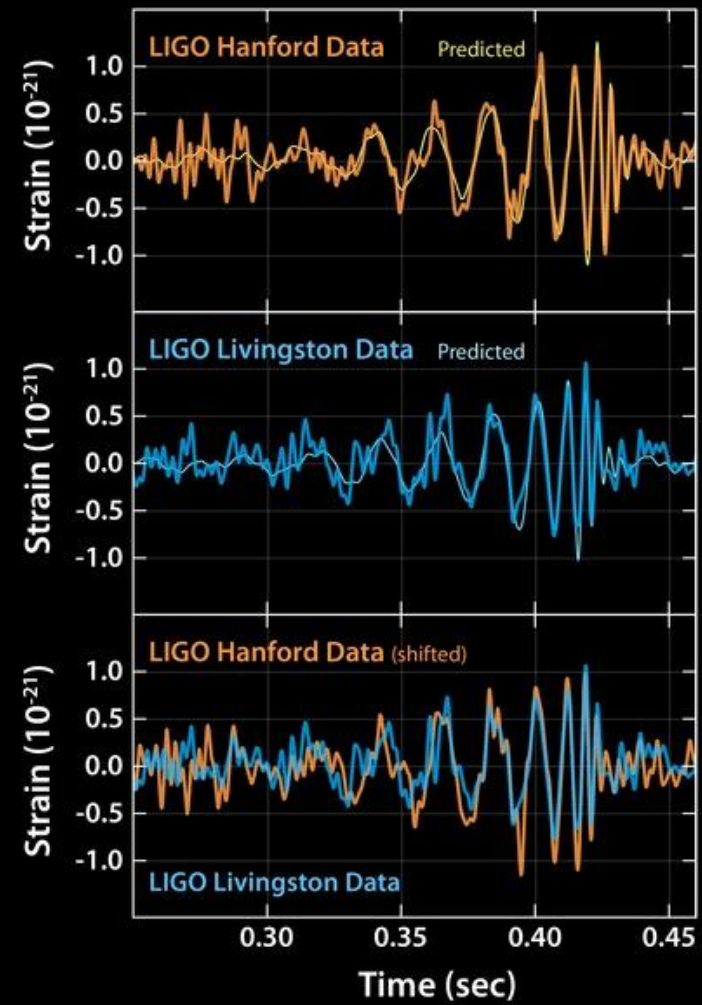


LIGO -- Laser Interferometer Gravitational Wave Observatory





LIGO data



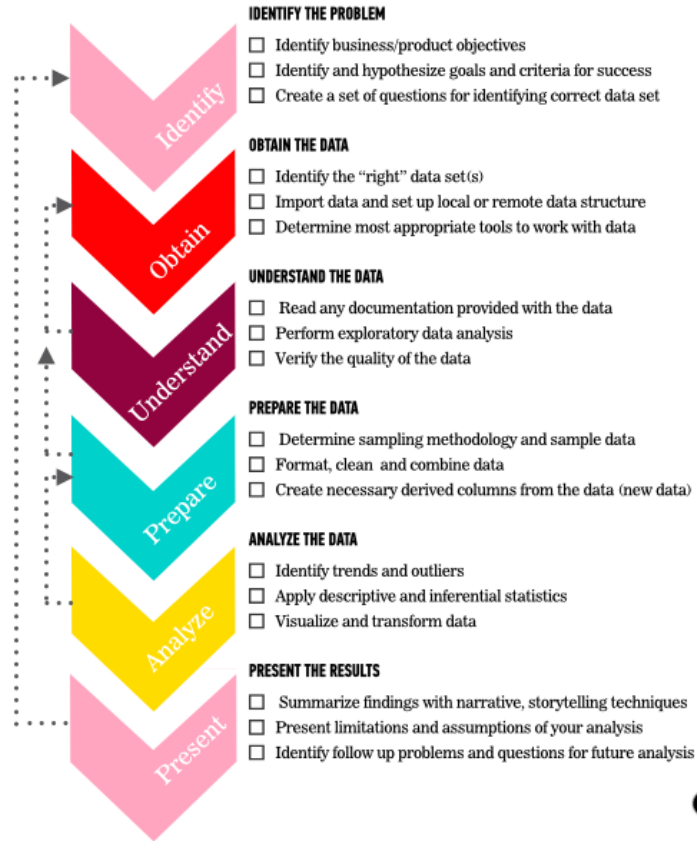


Class Exercise

- ▶ With the 1 or 2 people sitting next to you, brainstorm the steps you would include in an analytical workflow. Think about the process from start to finish.

Data Analytics Workflow

ANALYTICS WORKFLOW





Why is having a workflow framework important?

- ▶ Guidance for you
- ▶ Computers need help!
- ▶ Prevent over-analysis
- ▶ Defined feedback loops
- ▶ Planning ahead => saving time
- ▶ Defining goals, business context prevents irrelevant decision making



When should you reference the workflow?

- ▶ From the beginning to the end of any analytical project
- ▶ NOTE: It won't give you specific details about each step but it guides you through higher-level tasks

Here's the situation...

- ▶ Congratulations! You've just started a new role as a product analyst at Kickstarter, a site for crowdsourcing fundraising. Your manager would like to gain a better understanding of the site's projects, and what factors contribute to a project's success.
- ▶ Let's start by accessing the data and navigating around it in Excel
- ▶ If you haven't already, [download](#) the Kickstarter data collected about projects and backers.



Intro to New Material

Data Types

- ▶ Categorical (also Qualitative)
 - Categorical variables represent types of data which may be divided into groups
 - Ex: race, sex, age group, and educational level
- ▶ Numerical (also Quantitative)
 - Values of a quantitative variable can be ordered and measured
 - Ex: age, height, sales, volume
 - Numbers are not always numerical data.
 - Ex: Gender (0=Male, 1=Female)



Conditional Formatting

- ▶ Visualizing data helps us understand the relationship between different data points.
- ▶ **Conditional formatting** is a great way to add quick and easy visual queues to our raw data.



Key Metrics

- ▶ Numerical data can be evaluated using a few summary metrics. Here are four very useful ones with which to start:
 - Count()
 - Min()
 - Max()
 - Average()



Getting to know your data

- ▶ Populate the summary box at the top with formulas to fill them in and think about the following questions:
 - Why did we arrange our dataset this way?
 - What are some observations you had while examining the data?

Data Type	
Count	
Min	
Max	
Average	



Defining Business Problems

- ▶ To do this, we must know: What does Kickstarter do?
 - Kickstarter is a site for crowdsourcing fundraising
 - Let's assume this is Kickstarter's only source of revenue
 - Kickstarter has a strong incentive to ensure projects make it over this threshold, and the further the better



Exploratory Analysis / Surface Analysis

- ▶ Answers to exploratory analysis questions are critical in helping us understand the structure and shape of our data.
- ▶ Example exploratory analysis questions:
 - ▶ What is the highest Goal amount in this dataset?
 - ▶ What project generated the highest Pledged amount?
 - ▶ What funded years are represented here?


Independent Practice - Surface Analysis of Kickstarter data

- ▶ Take clean data - from Kickstarter - use shortcuts and functions to do a surface analysis and describe the data contained in each column, specifically answering these questions and creating these deliverables:
 - ▶ What data types are there?
 - ▶ Create documentation in a new worksheet (like a readme) with information around each unique, important column for each table.
 - ▶ Organize your workbook in structured fashion that is easy to understand
 - ▶ Identify the count, min, max, mean and median for number data-type columns
- ▶ Identify the problem statement / business needs and relevant questions to be answered on a provided prompt.
- ▶ Answer the relevant questions using aggregate Excel functions (feel free to ask other questions you feel are important):
 - ▶ How many projects are contained in the Kickstarter dataset?
 - ▶ What % of all projects were successful?
 - ▶ What was the average amount raised per project?
 - ▶ What was the most common duration for projects?



Today's Objectives

- ▶ Describe the relationship between functions and parameters
- ▶ Use nested functions
- ▶ Find, replace, and change text
- ▶ Remove duplicate records
- ▶ Remove spaces from text
- ▶ Use barcharts



Examples of “Dirty” data

- ▶ **Example:** Let’s look at our Kickstarter data? Take a look at column D or E. If we examine the unique values in these columns we see a few permutations:
 - Film & Video
 - Film & Video
 - Country & Folk
 - Country & Folk



So, why Prepare data?

- ▶ We need clean data in order to perform a legitimate analysis and find information that would otherwise be hidden in the data.
- ▶ Data, especially when taken from the web, has a **ton** of missing values and problems to tackle before any meaningful analysis can take place.

Why is Preparing, Cleaning your data Difficult?

- ▶ There is no simple algorithm that deals with cleaning data; you only get good with practice and flexible rules.
- ▶ We will give you tips; you will begin to create your own processes for cleaning



Types of Dirty Data

- ▶ Duplicate rows
 - Your job is to determine what data is considered duplicate



Types of Dirty Data

- ▶ Case text inconsistencies
 - User input makes this really difficult!
- ▶ Spaces
- ▶ Non-print characters
- ▶ Numbers and number signs
- ▶ Dates, times, and custom formats



Types of Dirty Data

► Irrelevant columns

- Up to you to decide which columns will provide substance to an analysis.
- NEVER delete a column entirely, keep it in an original copy of the data.



Functions vs. Parameters vs. Output

- ▶ How do functions relate to parameters?
- ▶ How do functions and parameters lead to output?
- ▶ A **function** can take **parameters** which are just values you supply to the **function** so that the **function** can do something utilizing those values.
 - COUNT(cell_range)
 - FIND(substring, string)

String Manipulation

► Let's try this together

- Open the AN_lesson2_kickstarterscrape_student.xlsx
- We'll do the following:
 - Use FIND() and REPLACE() to remove 'amp' strings
 - Remove duplicates
 - Separate location into city and state
 - Create TRIM() column
 - Create LEN() column
 - Create conditionals

More String Manipulation

- ▶ Remember, a lot of data is in text form
 - If we can get it standard, great! But it's rare
- ▶ Strings are simply pieces of text that can be manipulated, queried, moved, and edited using additional standard Excel functions
 - CONCATENATE()
 - LEFT()
 - RIGHT()
 - MID()

Text Manipulation

- ▶ Use [AN_lesson_2_student_kickstarter_model_practice.xlsx](#) to practice combining text and selecting parts of strings.
- ▶ Use CONCATENATE or “&” to create formulas that correctly populate the “Full Name with prefix” and “Full Address” columns:
 - combine columns D, E, and F to produce full name in column M
 - An example “Full Name with prefix” is “Ms Jane Smith”
 - combine columns G, H, I, and J to produce full address in column N
 - An example “Full Address” is “5 Gardeners Square, Evesham, Worcestershire WR11 1DZ, UK”
- ▶ Extract the “Segment number” from column B to fill the “Segment” column
 - Hint: You can use conditional logic functions such as IF,AND and OR

Logic Operators

- ▶ A lot of work in Excel involves *comparing data* in different cells.
- ▶ To make this easy, Excel provides “Boolean” or “Logical” operators
- ▶ Return either one of two possibilities
 - True or False
 - Used to determine what to display in the cell or do next in your formula



Boolean Operators

► =, !=, <, <=, >=, >

Logical Operators

- ▶ IF - checks whether a condition is met, returns one value if True and another value if False; often use boolean operators within the IF condition
- ▶ AND - a functional operator you can use to chain comparisons or other boolean operators
- ▶ OR - a functional operator you can use to chain comparisons or other boolean
- ▶ ISBLANK() - checks whether or not a cell is blank

Nested Functions and Operators

- ▶ =IF(OR(AND(A1>300,B1="Blue"),AND(A1<300,C1="Monday")),
"Correct","Wrong")
- ▶ Let's walk through this nice and slow

Nested Functions and Operators

- ▶ =IF(OR(AND(A1>300,B1="Blue"),AND(A1<300,C1="Monday")),
"Correct","Wrong")
- ▶ AND(A1>300,B1="Blue")
 - “If A1 is over 300 and B1 is “Blue”, return ‘Correct’

Nested Functions and Operators

- ▶ =IF(OR(AND(A1>300,B1="Blue"),AND(A1<300,C1="Monday")),
"Correct","Wrong")
- ▶ AND(A1<300,C1="Monday")
 - if A1 is less than 300 and C1 is “Monday”, return ‘Correct’.

Nested Functions and Operators

- ▶ =IF(OR(AND(A1>300,B1="Blue"),AND(A1<300,C1="Monday")),
"Correct","Wrong")
- ▶ So all together:
 - “If A1 is over 300 and B1 is “Blue”, return ‘Correct’ | OR | if A1 is less than 300 and C1 is “Monday”, return ‘Correct’ | Otherwise, it’s ‘Wrong!’”



Why visualizations?

- ▶ For the remainder of unit 1, we'll be looking at visualizations each class
- ▶ They're imperative to analysis because with them, we can uncover patterns not visible in an Excel table



Bar Charts

- ▶ Best for comparing discrete values in a dataset
 - Discrete data is countable, ex: number of successful projects
 - Continuous data can take any value, ex: project backing total
- ▶ Translation: “A bar chart encodes two pieces of information: categories and quantities for each category”



Common Issues with Bar Charts

- ▶ To many categories
- ▶ To many stacks in stacked bar charts
- ▶ Inappropriate axis interval
- ▶ Coloring, 3D

Creating Different Bar Charts

- ▶ Kickstarter is thinking about providing a consulting service to project founders help its customers create more successful crowdfunding campaigns. You have been asked elaborate on your initial exploratory analysis. To do this, you'll need to do some data transformation, preparation, and cleaning, just like you learned today.
- ▶ In your “Summary” tab - of the [AN_lesson2_kickstarterscrape_student.xlsx](#) workbook - fill in the two summary tables: one with two columns - number of projects and successful projects for each different categories; then one with the same two columns but for different countries. Be sure to constructor two bar charts from each of your summary tables. Use the technical requirements below...

CLOSING DISCUSSION