GETTING STARTED WITH DATA SCIENCE

YOUR VERY OWN WORKBOOK

#GETGOOD

CODING WITH MAX

FIRST THINGS FIRST

>> WHAT IS A **DATA SCIENTIST**?

>> A data scientist is someone who takes **simple, raw, and unprocessed** data and turns it into an **information gold mine** by transforming, organizing and analyzing the data.

+ DATA SCIENTIST =

>> WHAT DOES A **DATA SCIENTIST** DO?

A data scientist can do just about anything - whether it's analyzing social media data, weather data, stock market data, viewership data, cat data, research data, etc..

All a data scientist needs is... you guessed it! **Data**.

Data is becoming a more and more powerful tool to make money, make a difference, make us smarter, or make the world just a little bit better.

All that data just needs to find a home - and its home is in the arms of a data scientist. The data scientist twists and turns the data until it is coherent and able to tell us something interesting or useful!

DATA SCIENCE PROBLEMS:

1) THE PINEAPPLE DILEMMA

Jim loves pineapples. He loves them so much, he wants to eat them all year round AND he doesn't want to pay more in one season for pineapples than another.

So... Jim wants to find out if his favorite fruit (the pineapple) changes price based on the rainfall, the amount of sun, or because of some other reason. That way - he can predict when prices are going to be high and stock up on lots and lots of pineapples.

2) THE DUMPLING ENTHUSIAST

Emma is a dumpling enthusiast. She wants to figure out which restaurants sell the best dumplings in her country. She also wants to see comparisons between pricing, location, availability of wifi, and reviews.

Problem is, there are over 4,000 dumpling restaurants in her country, so she can't exactly pull up 4,000 tabs and examine each restaurant manually.

3) THE MONEY CURIOSITY (true story - read more here!)

Max wants to see if he can discover a link between Twitter data and company stock prices on the stock exchange. He wants to see if Twitter users react to scandals and controversial news before the stock prices are affected.

For example, did Twitter talk about the Yahoo hack of 2016 before Yahoo's stock price took a hit? (Spoiler: it did!)

How do we solve any of these problems?

DATA SCIENCE MAGIC, OF COURSE

SO, I GUESS ONLY ONE QUESTION REMAINS...

ARE YOU READY TO BECOME A DATA SCIENTIST?

STEP 1:

GRAPHS ARE YOUR NEW BEST FRIENDS

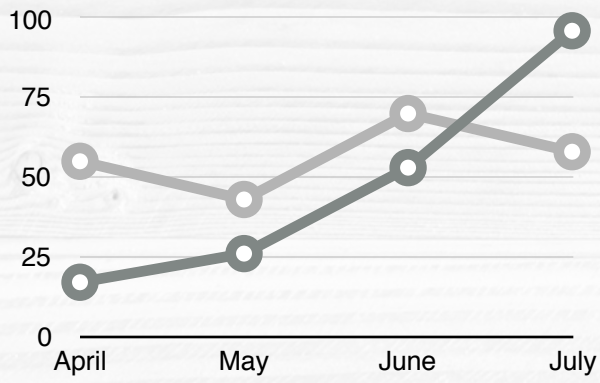
You can't be a good data scientist and have no clue how to analyze a graph. It just doesn't work.

It's like if a mechanic didn't know how to work with a wrench, or a mega pop-star didn't know how a microphone worked. It's just strange.

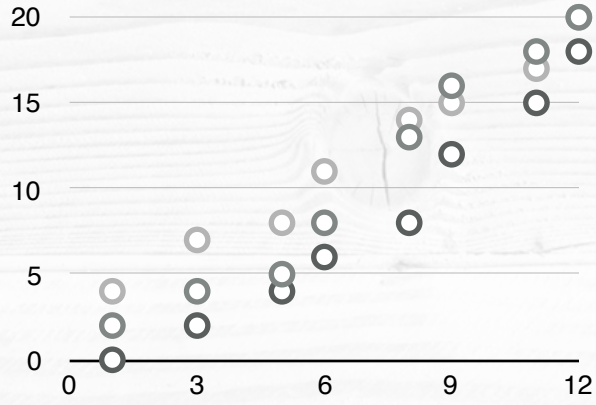
You need to know graphs if you're going to be a stellar data scientist. You need to know them backwards, forwards, up and down.

Luckily for you - this is a cheatsheet, which means I condensed all the info perfectly for you here!

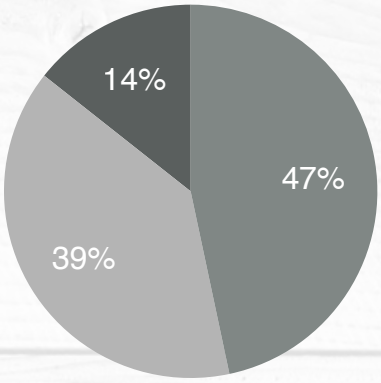
MUST-KNOW GRAPHS



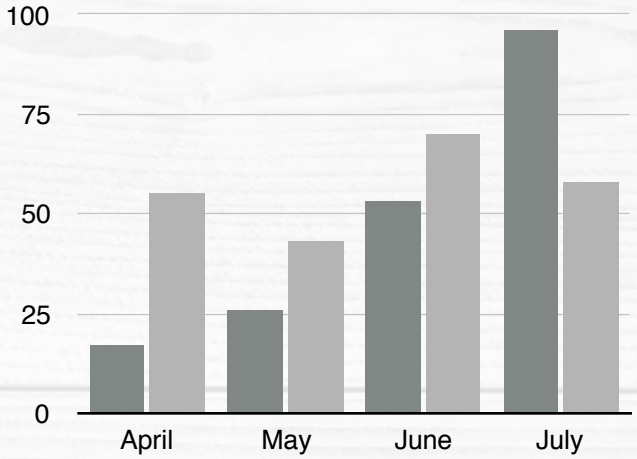
1. **line graph** - used for showing patterns, trends and changes over time



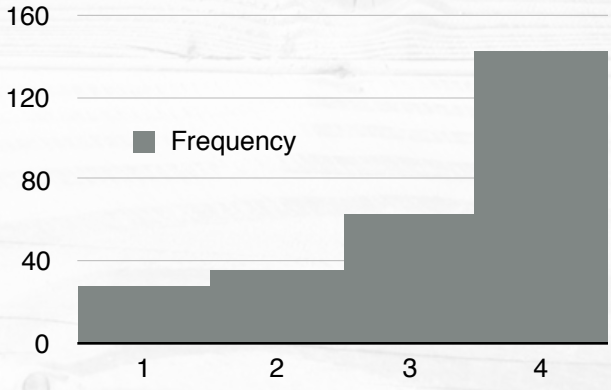
2. **scatter plot** - used for looking at correlations and relationships



3. **pie chart -** used for showing proportional data and percentage data using relative sizes



4. **bar graph** - used for comparing differences or showing changes over time



5. **histogram** - used for showing frequency and the overall distribution of data

6. **box and whiskers plot** - used for showing data distribution, quartiles, averages and maximum and minimum values

STEP 2:

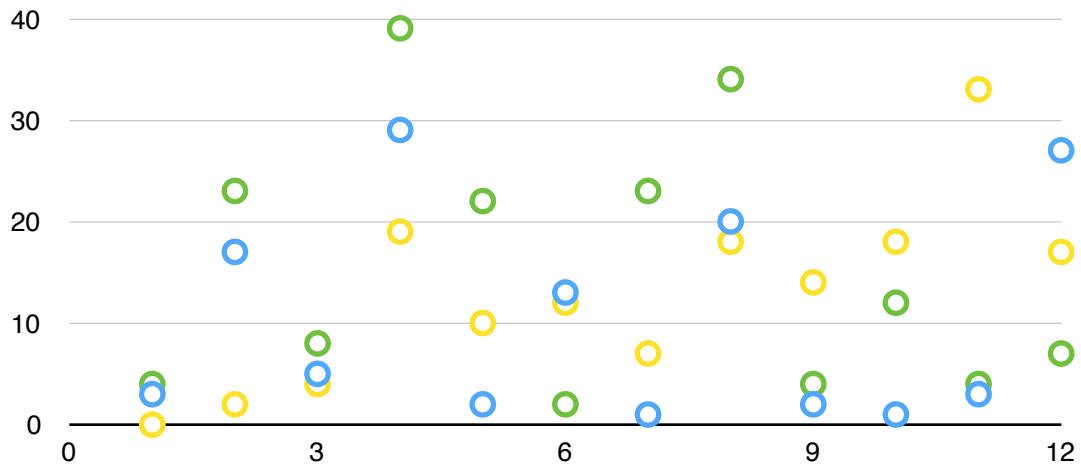
GET COMFY WITH STATISTICS

(YES, I'M SORRY)

Data science is all about interpreting and reading data. You won't be able to get far without some basic knowledge of statistics and how to analyze the information you're being given.

Not only that, proper statistics knowledge lets you distinguish noise from real meaningful data.

*What is **Noise**?*



Noisy data is just having randomness in your data. It often looks like a big mess and can be quite useless.

By understanding all the main statistical terms and what they mean, you can begin to pull apart the noise from the data.

All the essential terms you should get familiar with + their definitions (*because I'm so nice*) are on the next page.

MUST-KNOW STATISTICAL TERMS

1. **Mean** - (aka the average) calculated by adding up all the numbers and dividing it by the amount of numbers
2. **Median** - middle value of a distribution of numbers
3. **Mode** - most common value in a set
4. **Minimum** - the smallest number in a set
5. **Maximum** - the largest number in a set
6. **Standard Deviation** - tells you how much the data differs from the average of the sample
7. **Variance** - average of the squared differences from the mean, also known as standard deviation squared
8. **Sample size** - number of elements included in your sample
9. **Statistical Significance** - the probability that the relationship between two variables is not random
10. **Percentiles** - percentage of the total cases that are equal or lower than the value
11. **Standard Error of the Mean** - no matter how many people you have in your sample, you will still always have a limited sample size. Because of this, you can never be 100% certain about your mean. The true mean lies within this specific range around the mean

STEP 3:
LEARN A PROGRAMMING LANGUAGE

The ideal programming language for data science should be:

❑ SUPPORTIVE

Is the community helpful? Do you have support? Do you have to start from zero every time you start programming? Or are there pre-written pieces of code to help you out?

❑ EASY TO USE

Is the language easy to use? Is it user-friendly and easy to read? Do you have to spend a lot of time debugging or looking for small errors? Is it easy for others to review your work and for you to review others' work?

❑ FAST TO PROTOTYPE

How long does it take to get something up and running? Are you spending more time coding than actually analyzing the data? Can you write up what you need to analyze the data quickly?

LEARN A LANGUAGE

TOP 6 LANGUAGES FOR **DATA SCIENCE**

#1 PYTHON

Python is one of the most popular languages for data science because it is fast to code in, which makes it efficient and easy to learn

#2 JAVASCRIPT

Javascript has a very big user base, and is very beginner-friendly. It is really great for prototyping!

#3 R

R is used plenty in data science and is particularly advantageous for visualization and breaking down statistical variables without needing much code.

#4 MATLAB

Matlab is similar to R in that it's a math-based programming language. It's used for mathematical computations and data visualization.

#5 SAS

SAS is kind of like a commercialized version of R, and has a wide range of statistical functions and offers tech support.

#6 C++

C++ is a very fast programming language, and is used for big scale implementations. It is, however, complex to use so it's not suggested for beginners.

STEP 4:

GET REALLY GOOD WITH PRACTICE

#GETGOOD
#GETREALLYGOOD

It's time to start your first data science project! Data science is all about playing around, testing and developing solutions. We're really just a bunch of kids trying to figure out the coolest information out there.

YOUR FIRST DATA SCIENCE PROJECT

STAGE 1: THE IDEA

What problem are you trying to solve? What do you want to do? What is your ultimate legacy? (Fill it out below!)

STAGE 2: THE DATA

What data do you need for it? What specific variables are you looking at? What are you analyzing? What information do you need?

Where are you getting the data from? What are your sources? Is it a website? Is it a database?

How are you going to get your data? Is there an available API? Will you do web-scraping? Or will you download a dataset from somewhere (ex. Kaggle)?

STAGE 3: FORMATTING & PROCESSING

What format do you need your data to be in for analysis? *(Tip: Most of the time, the data in its raw format is a huge mess.)*

STAGE 4: ANALYSIS

Now that you have your data formatted and your crazy good statistical knowledge: IT'S ANALYSIS TIME!

What are the key elements you want to look at? What questions do you want answered? What indicators will give you those answers?

Do you want to visualize your data? Will data visualization help you make sense of the data, see patterns or recognize a specific trend?

THAT'S IT. You did it! Your first data science project. *I'm so proud of you*, young grasshopper. Seems like just yesterday when you started on this journey.

STEP 5:
SECRET SAUCE IS...
CONFIDENCE

The ultimate step for getting started is building your confidence in your abilities. And that is why it's very important to do so many of your own projects.

Here's your final checklist to figure out if you're on your way to being an amazing data scientist!

- ☐ I know my **graphs**.
- ☐ I know my **statistics**.
- ☐ I know a good **programming language**.
- ☐ I have done my **practice**.
- ☐ I believe **I'm a good data scientist**.

Whew, we did it! You're ready to be a data scientist.

CODING WITH MAX

This is me - **Max**. I look really serious here but I'm actually a goofball.

If you're looking to get into Data Science or Python - check out my courses. I promise they're every bit as good as watermelon on a sunny day or a cuddly dog on the couch on Friday night.

CHECK OUT COURSES