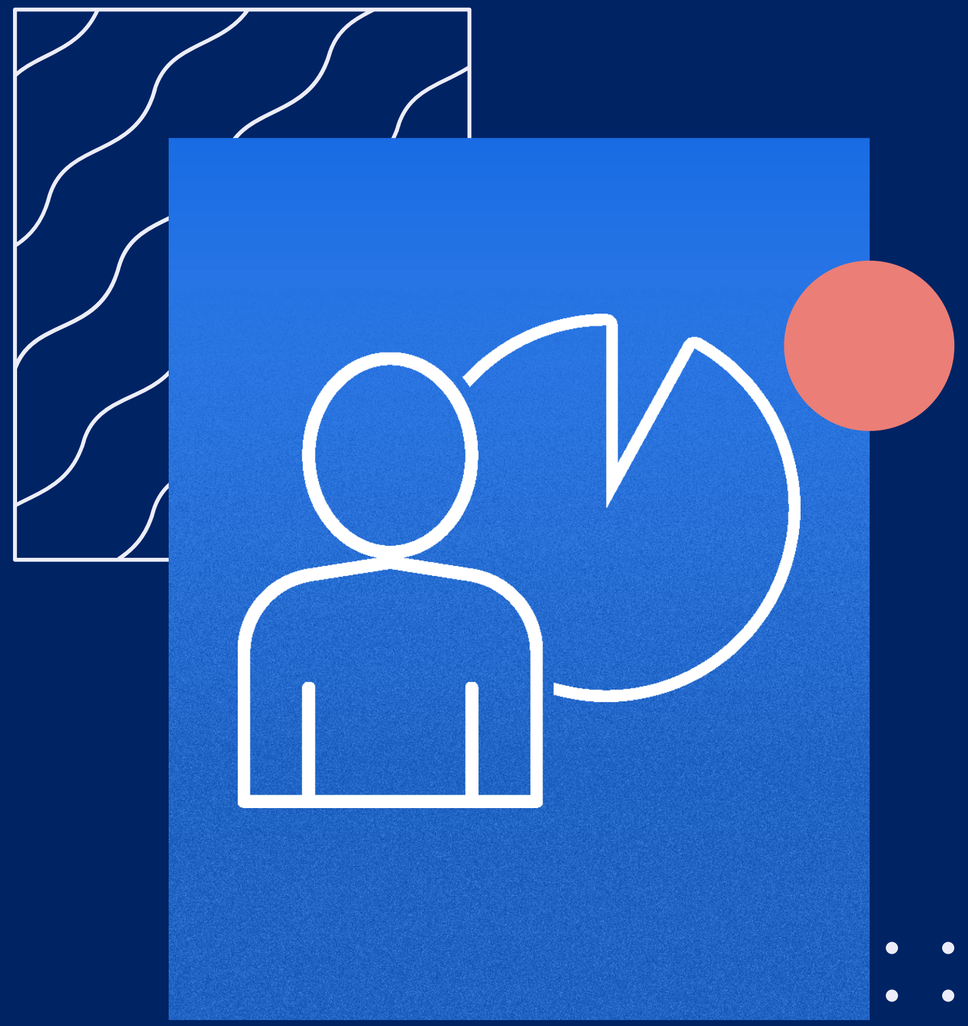
Telling your Data Story

A GUIDE TO A SUCCESSFUL DATA SCIENCE PROJECT



Presented by Jessica Uwoghiren for AfroTech Girls x IHS Towers
International Day of Women and Girls in Science Conference
THEME: Thriving as a Woman in STEM
DATE: Thursday, 11th February, 2021

About Speaker

- Data Analyst and Machine Learning enthusiast
- Over 4 years experience working as an Engineer in Energy and Manufacturing sectors. Ex. General Electric, Seplat Petroleum
- Discovered my passion for Data Science & Analytics in 2020
- Regular contributor to Towards Data Science Online Publication
- Run free online community, DataTech Space, for aspiring data analysts and scientists
- Dream Goal: Get into Google & retire to full-time consulting
- Hobbies: Football, Reading and Binge-watching TV series

Outline

What we will learn today

- Data and Data Sources
- What we do with Data
- Case Study: A Full Cycle Data Science Project
- The Data Science Process
- Project Outcome
- Conclusion

In one word, what does DATA mean to you?

What is Data and its science?

It's simpler than we think

Data is a collection of facts, such as numbers, words, measurements, observations or descriptions of things.

Data Science is a process of collecting, cleaning, analyzing, interpreting large amounts of data and building models that best fit the data for future use.

59 Zetabytes

That's the amount of data available in the world.*

200,000,000,000,000 MP3 Songs (of 5-minute length)

59,000,000,000,000 4K Videos (of 3 minute length)

59,000,000,000 Laptops (of about 1 TB capacity)

*As at December, 2020
Sources: statista.com & howtogeek.com

What do we use data for?

Our everyday decisions are based on data

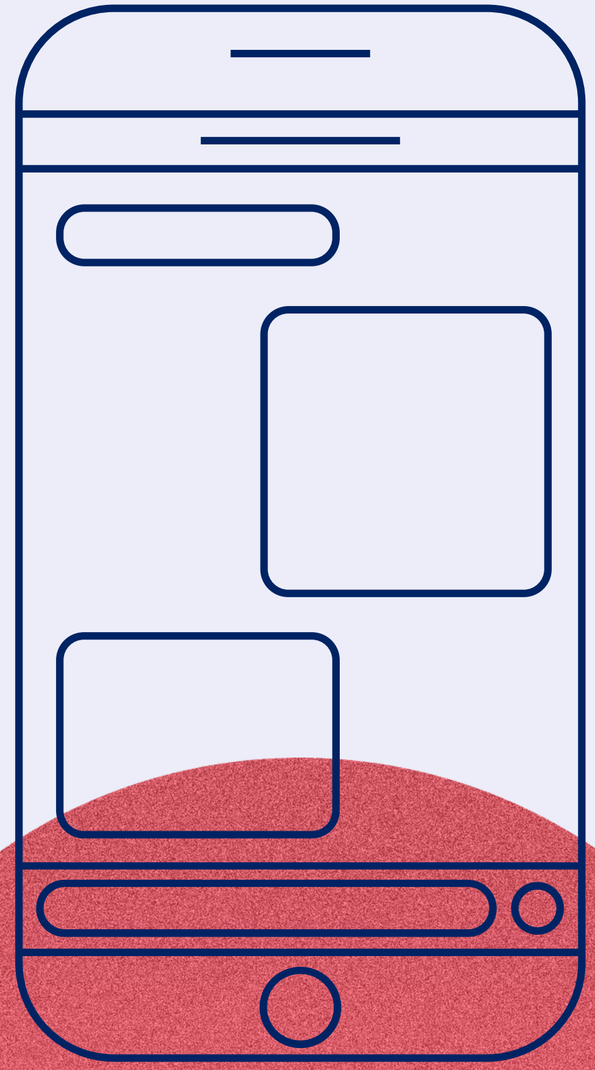
- Business Decisions
- Solve Problems
- Share Information
- Marketing Strategy
- Identification Purposes

Where can I get this data?

Data is everywhere.

- Online Sources - Google, Kaggle, DataHub
- Specific Websites - Glassdoor, Amazon
- Government Portals - NCDC (COVID-19 data),

  Nigerian National Bureau of Statistics
- Surveys - Questionnaire, Polls
- Social Media - Twitter, Instagram



Case Study

Now, an actual Data Science project

Title: Analyzing Twitter Users' Reflections on the Year 2020 using NLP

A Sentiment analysis project to get insights on how these users felt about the year based on over 50,000+ tweets

Stats: Over 1000 positive reactions on Medium, LinkedIn, Twitter etc.
Why Twitter? Over 500 million tweets a day
Programming Language: Python
Technique: Natural Language Processing (A subset of Machine Learning)
Outcome:

- Most common words used to describe the year 2020
- Time of the day when Twitter users are more active
- The proportion of positive, negative, and neutral tweets
- The country with the most tweets

Data Science Process Pt. 1

Trust the process, you'll get to the finish line

Problem Definition

What is my end goal? What problem am I trying to solve?

CASE STUDY

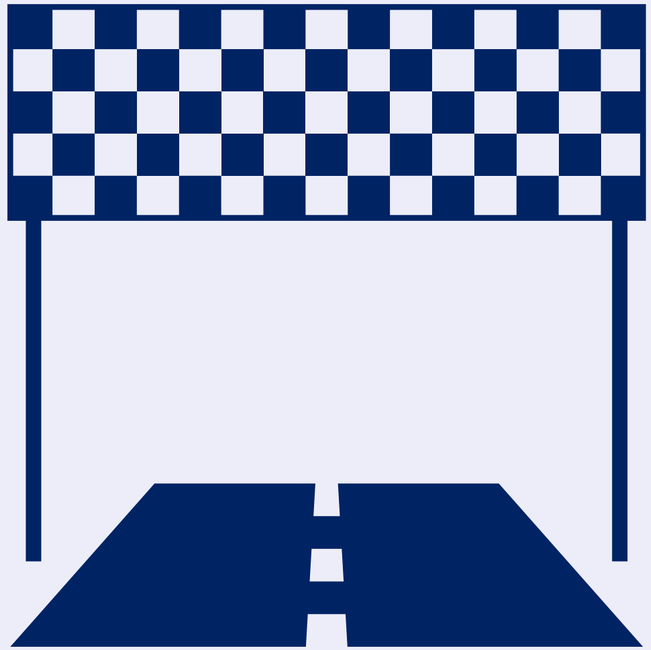
- How did people feel about the year?
- What time were these people tweeting?
- Where were the users tweeting from?

Data Gathering

Where is my data? Do I need multiple sources?

Data Cleaning

Is my data ready for analysis? Are there missing values?



Data Science Process Pt. 1

Trust the process, you'll get to the finish line

Problem Definition

What is my end goal? What problem am I trying to solve?

CASE STUDY

- How did people feel about the year?
- What time were these people tweeting?
- Where were the users tweeting from?

Data Gathering

Where is my data? Do I need multiple sources?

CASE STUDY

- Source: Twitter
- Use the right "search words" to fetch data via Twitter's API

Data Cleaning

Is my data ready for analysis? Are there missing values?

Data Science Process Pt. 1

Trust the process, you'll get to the finish line

Problem Definition

What is my end goal? What problem am I trying to solve?

CASE STUDY

- How did people feel about the year?
- What time were these people tweeting?
- Where were the users tweeting from?

Data Gathering

Where is my data? Do I need multiple sources?

CASE STUDY

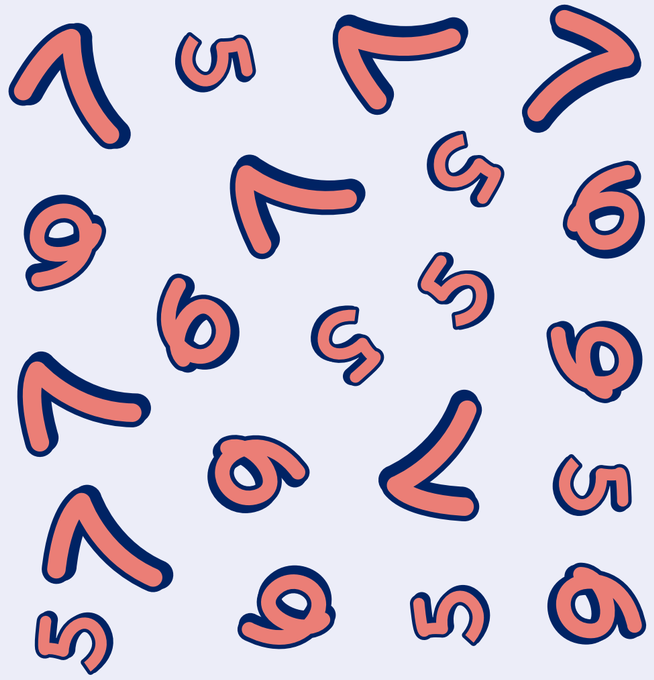
- Source: Twitter
- Used the right "search words" to fetch data via Twitter's API

Data Cleaning

Is my data ready for analysis? Are there missing values?

CASE STUDY

- Used Python methods such as pd.drop_duplicates() to clean the data set; Remove punctuation and stop words



Data Science Process Pt. 2

Trust the process, you'll get to the finish line

Data Mining & Exploration

What does my data look like? Are there patterns in my data?

<u>CASE STUDY</u>

- Used a word cloud to explore patterns in my data
- Image was generated with my code in Python
- Other examples are Box plots, Statistical measures

Machine Learning

What Algorithm is best suited for my analysis?

Model Deployment

Where else can my ML model be applied? Can I scale it?

Data Science Process Pt. 2

Trust the process, you'll get to the finish line

Data Mining & Exploration

What does my data look like? Are there patterns in my data?

CASE STUDY

- Used a word cloud to explore patterns in my data
- Image was generated with my code in Python
- Other examples are Box plots, Statistical measures

Machine Learning

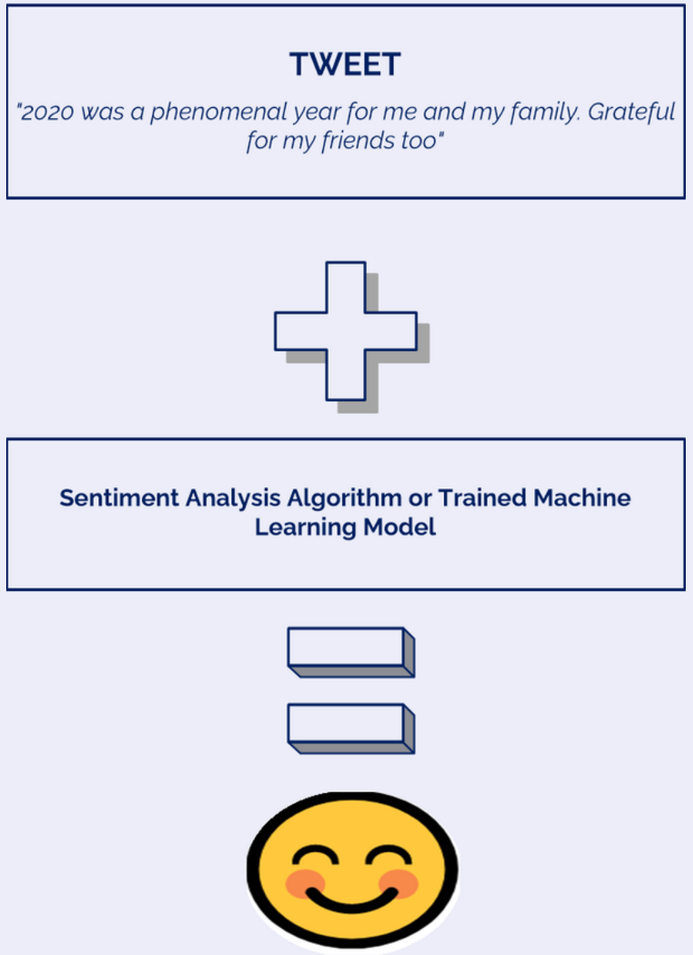
What Algorithm is best suited for my analysis?

CASE STUDY

- Algorihtms: NLTK (Natural Language Toolkit) & TextBlob
- Typically, you analyse different algorithms to find the best

Model Deployment

Where else can my ML model be applied? Can I scale it?



Data Science Process Pt. 2

Trust the process, you'll get to the finish line

Data Mining & Exploration

What does my data look like? Are there patterns in my data?

CASE STUDY

- Used a word cloud to explore patterns in my data
- Image was generated with my code in Python
- Other examples are Box plots, Statistical measures

Machine Learning

What Algorithm is best suited for my analysis?

CASE STUDY

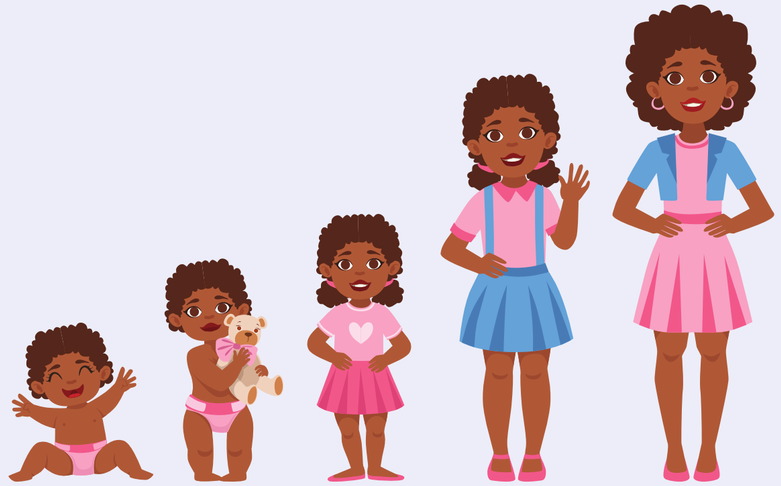
- NLP Algorihtms: NLTK (Natural Language Toolkit) & TextBlob
- Typically, you analyse different algorithms to find the best

Model Deployment

Where else can my ML model be applied? Can I scale it?

CASE STUDY

- Think about applying this to 1,000,000 tweets
- Discover what people think about COVID-19 or a new product release



What were my results?

We made it!

50%

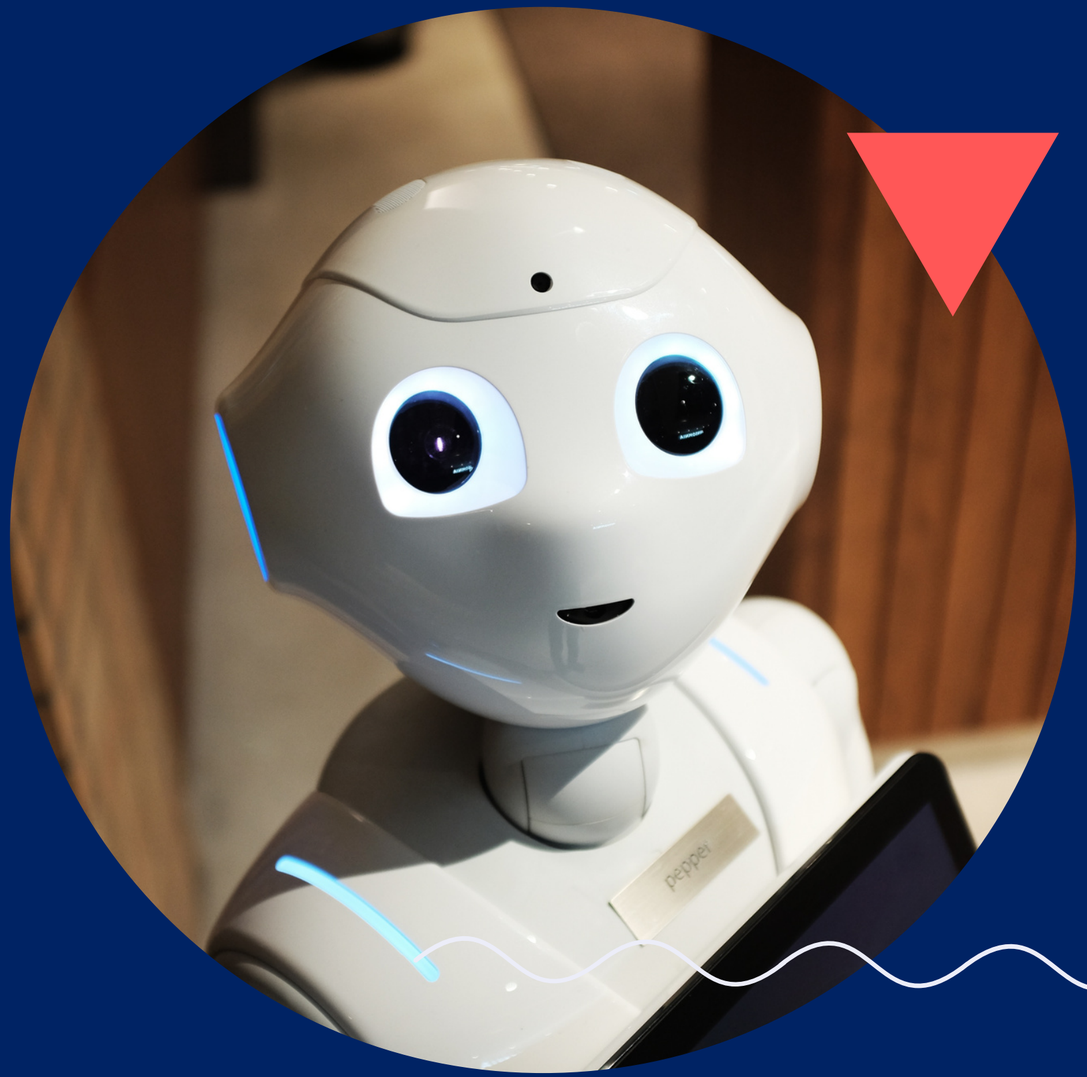
of tweets had positive sentiments

United States

had the most tweets

5PM GMT

was the time with most engagement



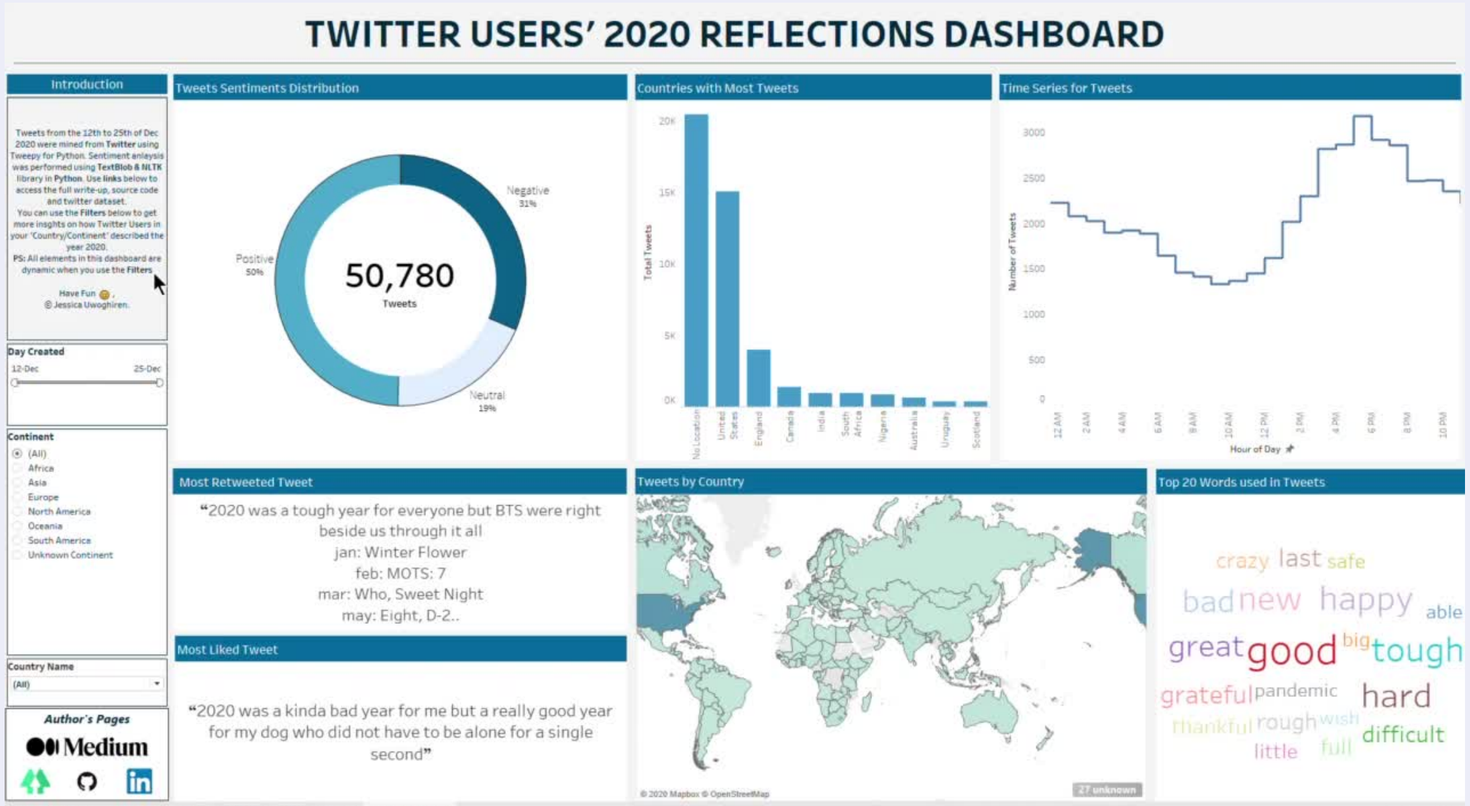
Presenting your results

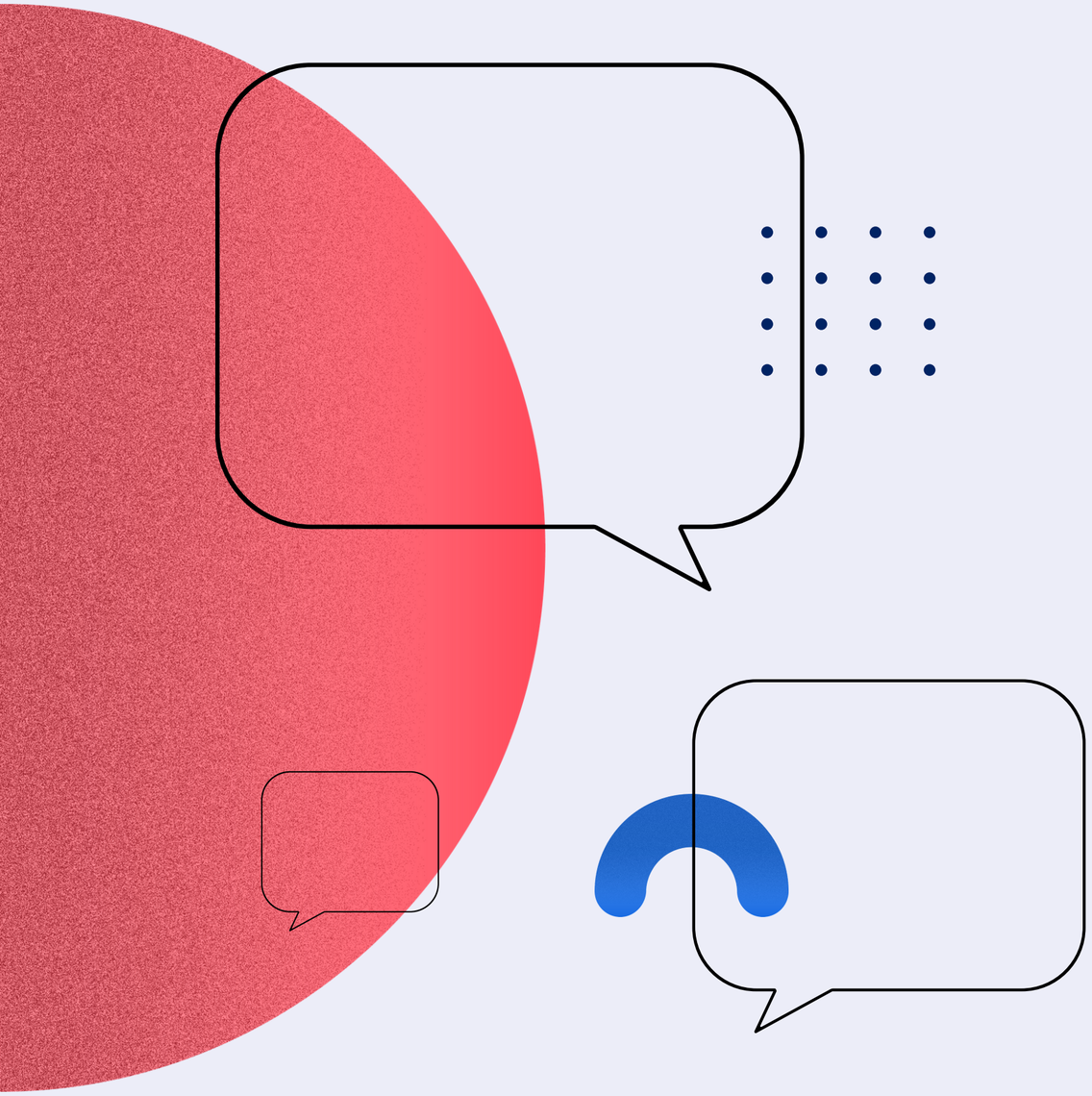
How did I present my findings?

Your mode of presentation depends largely on your audience

Some channels to use:

- Dashboards (Tableau or PowerBI)
- Social Media
- Blog post (Medium or WordPress)
- LinkedIn
- Company website or e-mail





THE DATA SCIENCE PROCESS IS A REPETITIVE ONE

To read more about this project, visit

https://tinyurl.com/twitter-sentiments-2020

To learn more about more projects, visit

www.jess-analytics.com

Connect with me

 @jessica_xls

 @jessica_xls

 Jessica Uwoghiren