Fundamentals of storytelling

DATA COMMUNICATION CONCEPTS

Hadrien Lacroix
Curriculum Manager

datacamp

What is data storytelling?

> *Data storytelling is the practice of building a narrative around a set of data and its accompanying visualizations to help convey the meaning of that data in a powerful and compelling fashion*

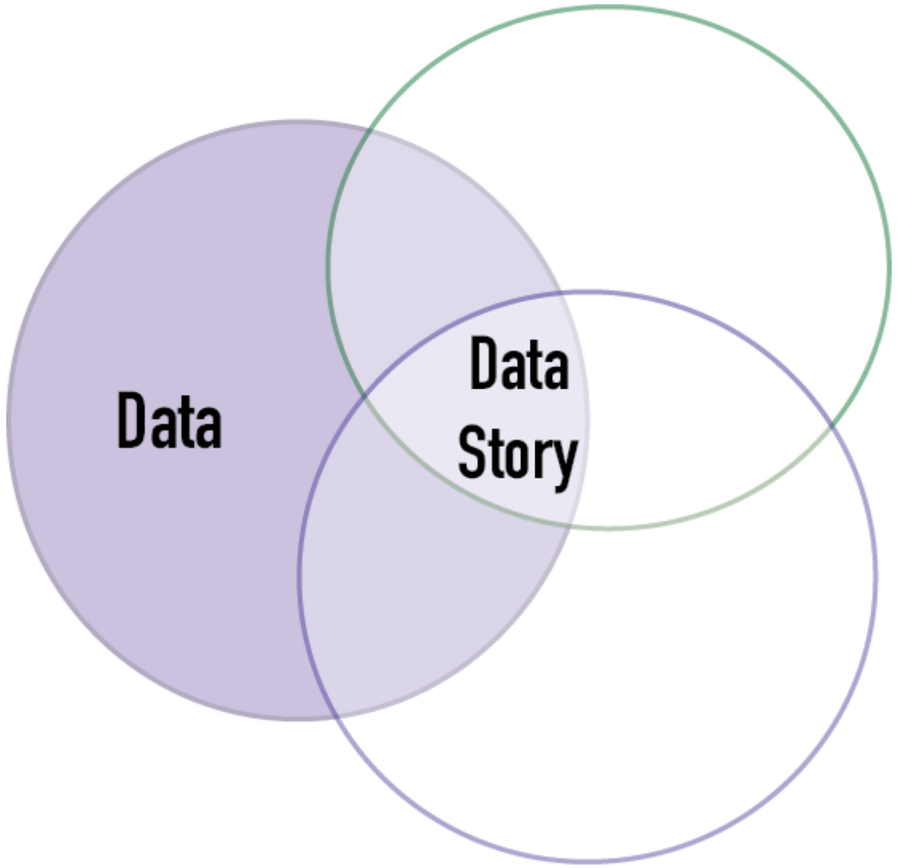
- **3-minutes story:**
  - What would you say in 3 minutes?
- **Big idea:**
  - Unique point of view
  - One sentence

Results →

1. **Insightful**
2. **Explanatory**
3. **Concise**

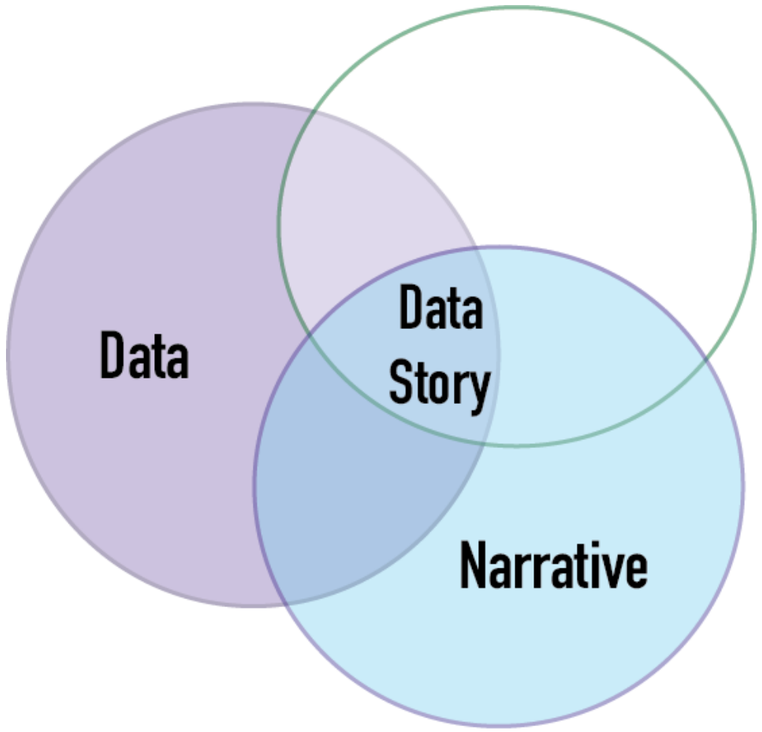
[1] https://tdwi.org/portals/what-is-data-storytelling-definition.aspx

Data



- **Results** (e.g predictions) and **findings** (e.g. data analysis)
- **Relevant** (methods or results and implications)
- **Accurate** and reliable
- **Actionable** insights

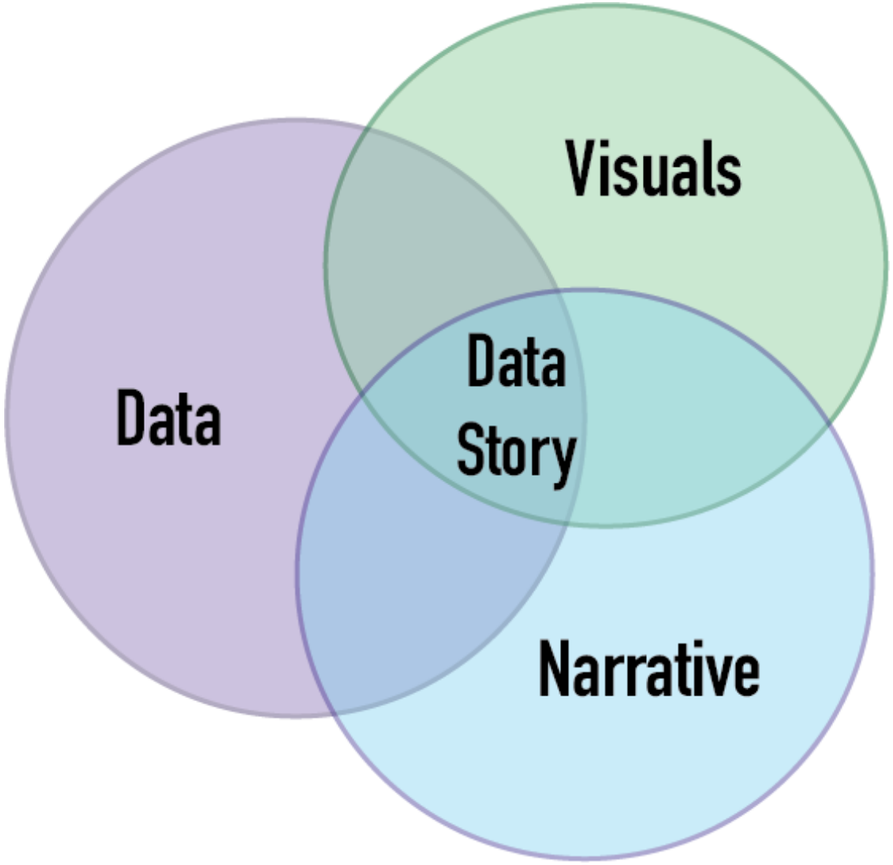
Narrative



- Main point:
  - **Avoid disconnected** facts
  - **Central** insight
- Explanatory context:
  - Understand **background** and audience
  - **Clarify** facts to that audience
- Linear sequence
- **Compelling** and **easy** to understand
- Prioritize **essential** points
- Drive **change**

> *A description of connected events that organizes information to engage the audience and make them care for the results or information shared*

Visuals



- Graphs should be:
  - simple
  - engaging
  - **not misleading**

Focus on impact

Instead of

- *Use a non-relational database to make efficient nested queries.*
- *Number of rooms shows correlation of 0.7 with a house price.*

Focus on

- *Changing the storage approach will save a lot of time.*
- *The more rooms in the house, the higher the price.*

Humility

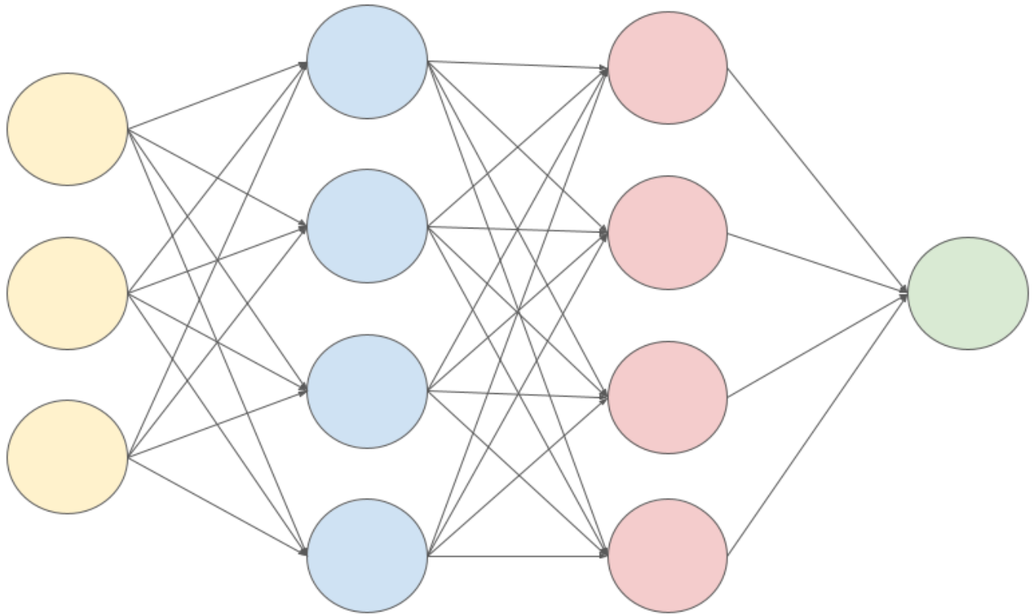
- Be receptive
- Proactively ensure understanding
- Explain differently

ADEPT

- **A**nalogy
- **D**iagram
- **E**xample
- **P**lain English
- **T**echnical definition

Analogies

Instead of



Use

[1] Alpha, "Liam is an expert on the shape sorter", Creative Commons

Building narrative

- **Change over time**: Chocolate lower in summer and higher in winter.
- **Correlation**: Chocolate rating vs. price
- **Comparison**: Two age groups vs. chocolate consumption
- **Clustering**: Groups with different coffee and chocolate consumption

Narrative structure

Background

- What **motived** the analysis?
- What **changed**?
- Who is the **focus** of the analysis?

Explain with a line plot that the company usually has a churn rate of 5%, but last year that rate suddenly increased to 15%.

Our background: Total profit decreased

[1] Dykes, Brent. Effective Data Storytelling. Wiley.

Narrative structure

- **What contributed to the problem?**
- Only relevant information

Using boxplots, show that the percentage of churn customers with more than one dependent in their household has increased, affecting the total rate.

Our insight: Chips 20% increase. Sweets 30% decrease.

Narrative structure

- Add **supporting evidence**
- Help better explain the cause of problem

Add further evidence by showing that a higher percentage of customers with more than one dependent in their household with DSL service churn.

More insights: Most popular chocolate 50% decreased.

Narrative structure

- Central insight
- **What would happen if there is no change**

Show a barplot that reveals that monthly charges are the most important predictor of customer churn.

Our climax: Loss $10M next year.

Narrative structure

- Potential solutions
- Course of action
- Proactive

Recommend to implement promotional prices to churn-intending customers and show that this will result in 10% more earnings with a barplot.

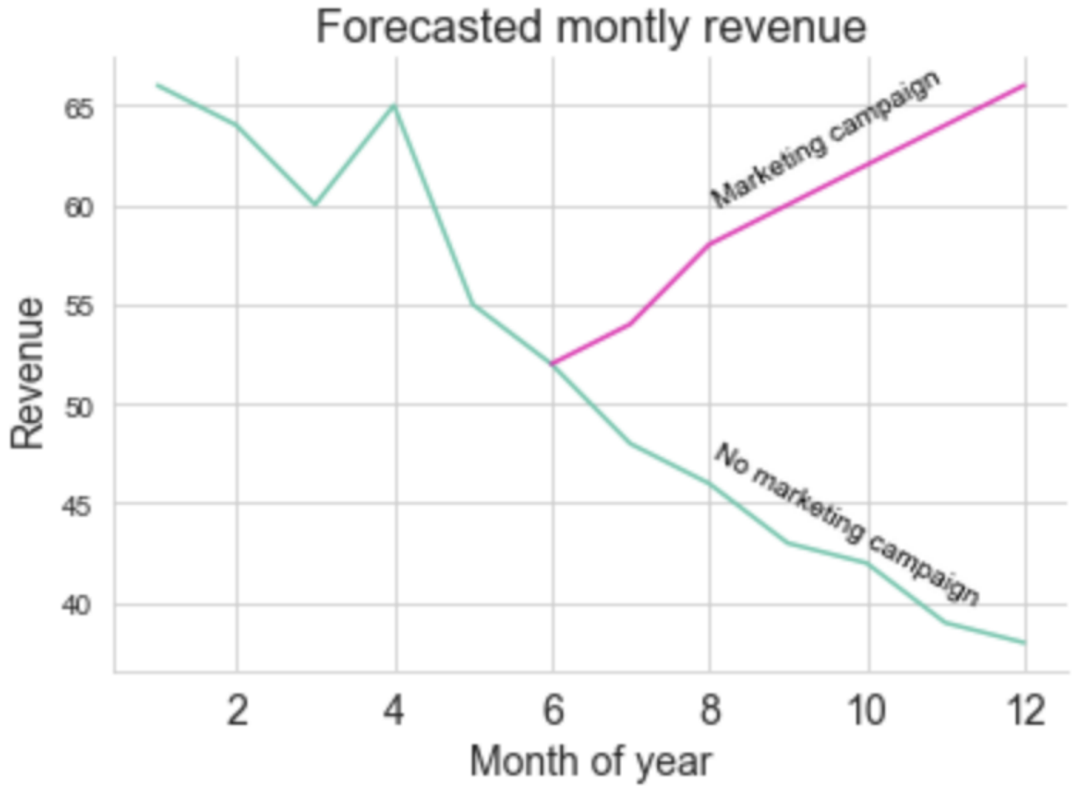
Our next steps: Rebrand chocolate.

Story

- Background:
  - Increase in defaulting percentage over last 5 years.
  - Predicting which customers had a high probability of default.
- Insight: People with more unemployment periods tends to default more
- Insight: People with lower income tend to default more
- Climax: Possible to predict which people is more likely to default with an accuracy of 95%
- Next steps: Run a trial on a control population

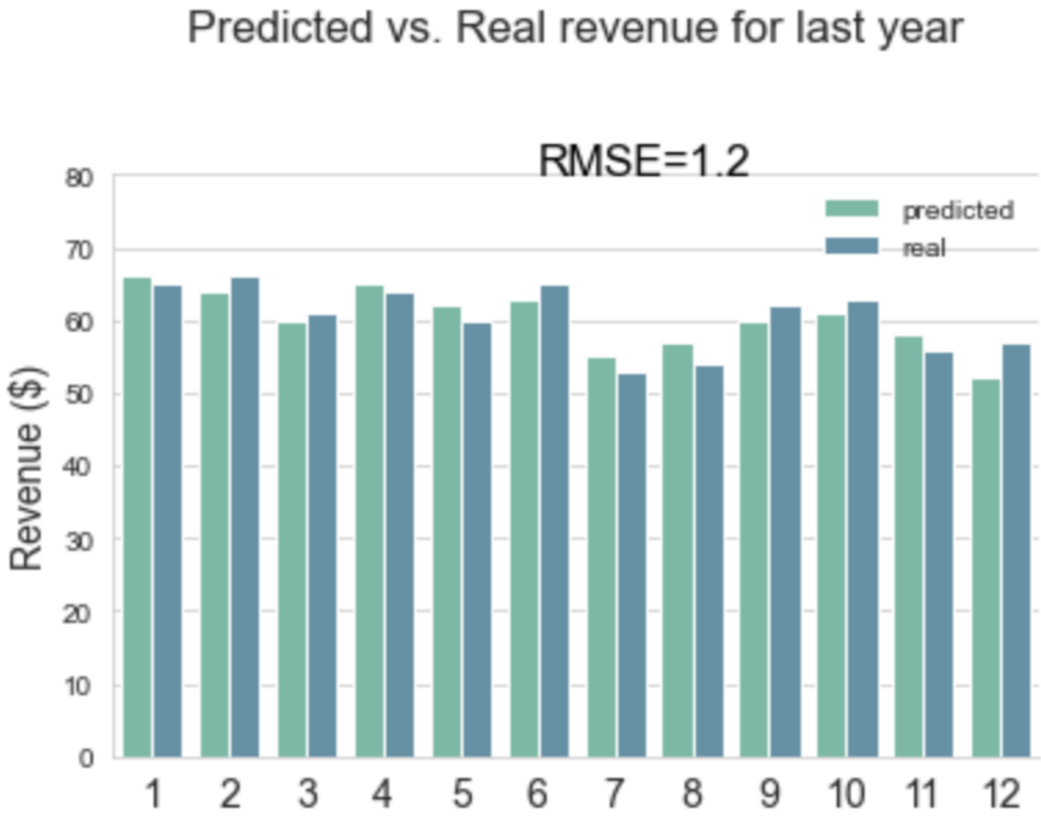
Executive team

- **Interest:** Inform their decisions based on findings



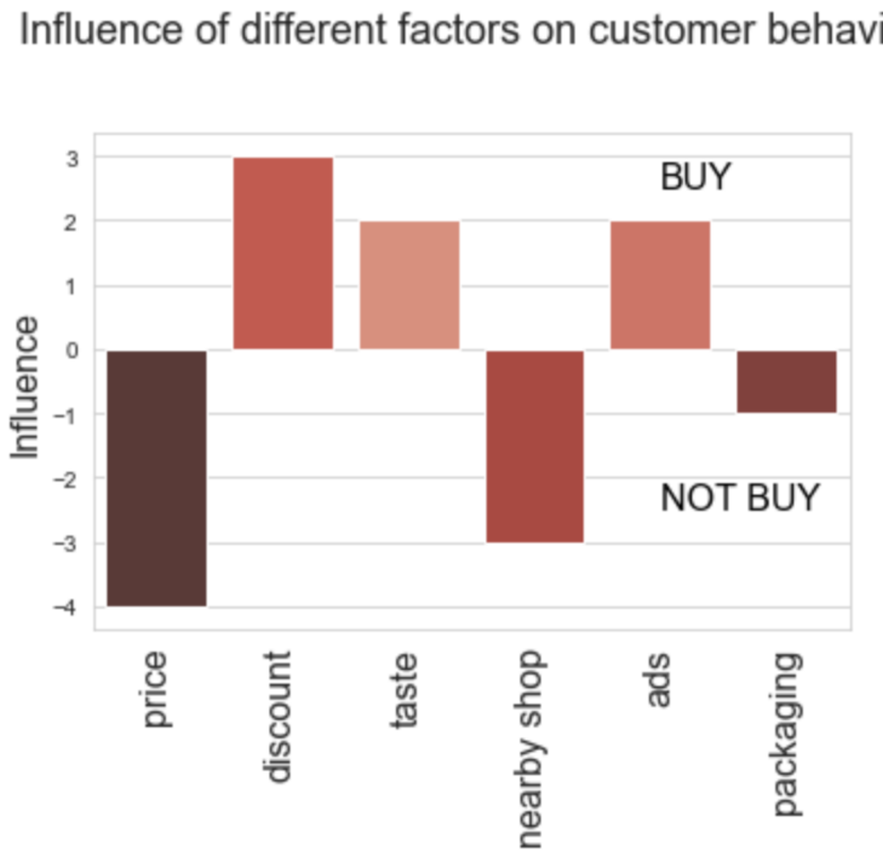
Data nerd

- **Interest:**
  - Replicate project
  - Continue project



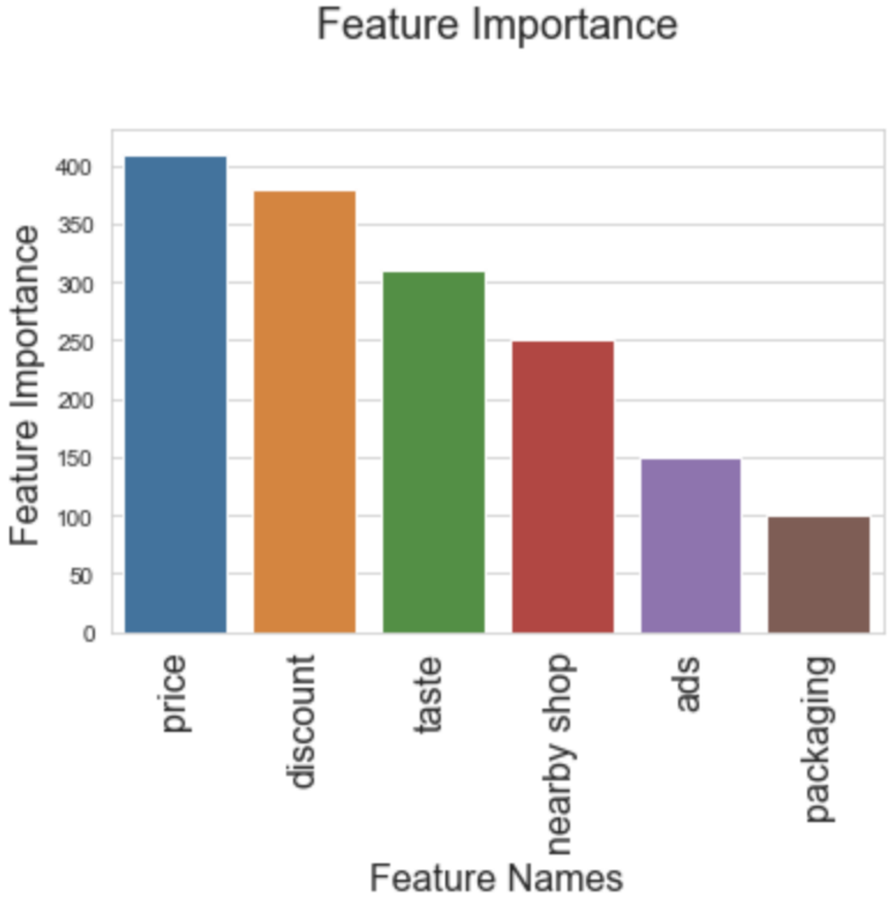
Executive team

- **Interest:** Inform their decisions based on findings



Data nerd

- **Interest:**
  - Replicate project
  - Continue project



Project manager

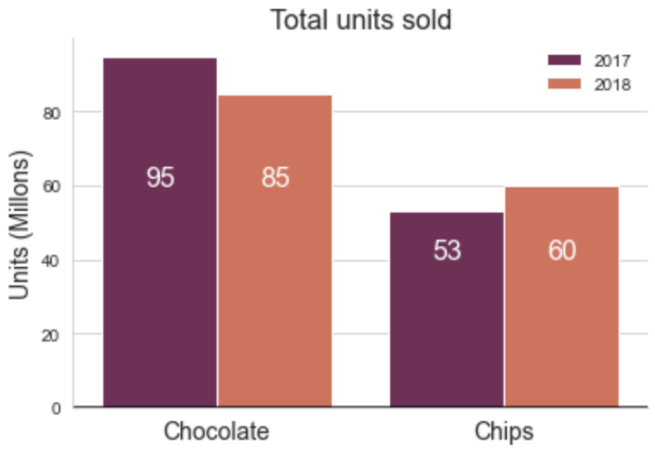
- **Interest:** Project aligns with company objectives

Customers, Other department

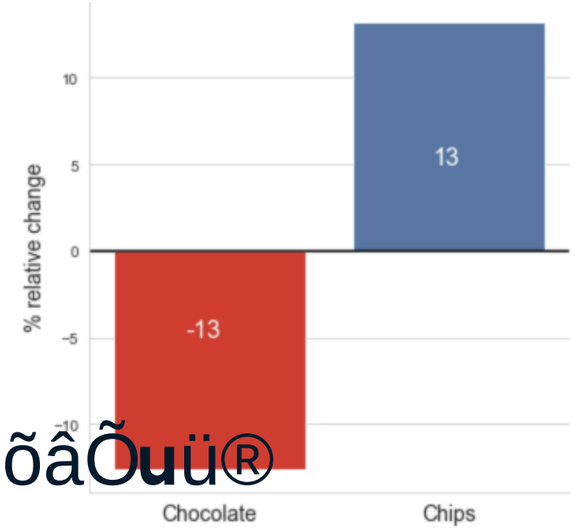
- **Interests:**
  - To understand the general results and impact of the project

Variations of data

Absolute



Relative

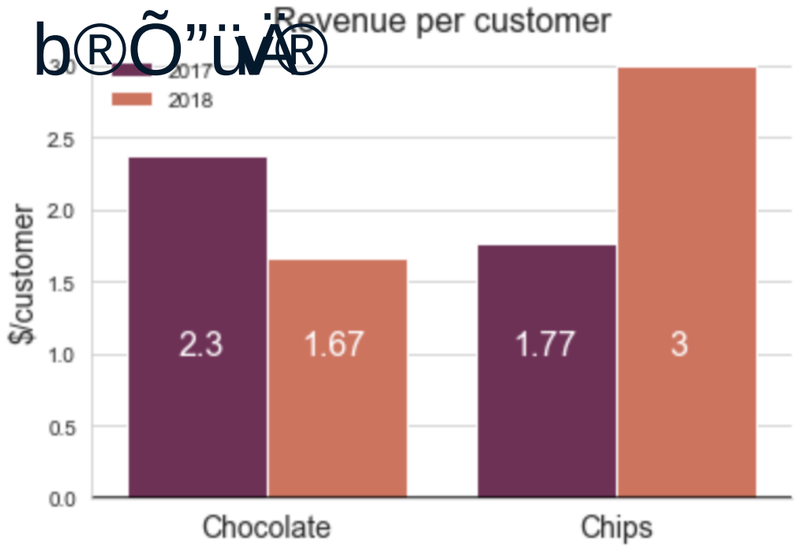


¢õâÕuü®

Ratio

- Quotient of two variables
  - Revenue per customer (**total product revenue/number customers**)
- Normalize values = **better comparisons**

b®Õ”üÄ®



Aggregates

- Representative value:
  - Totals / counts
  - Mean
  - Median
- Mean can be misleading (outlier)
- Distribution of the data
- Example:
  - 2019 US **average** salary: **$51**,916.27
  - 2019 US **median** salary: **$34**,248.45

p-value

What is p-value?

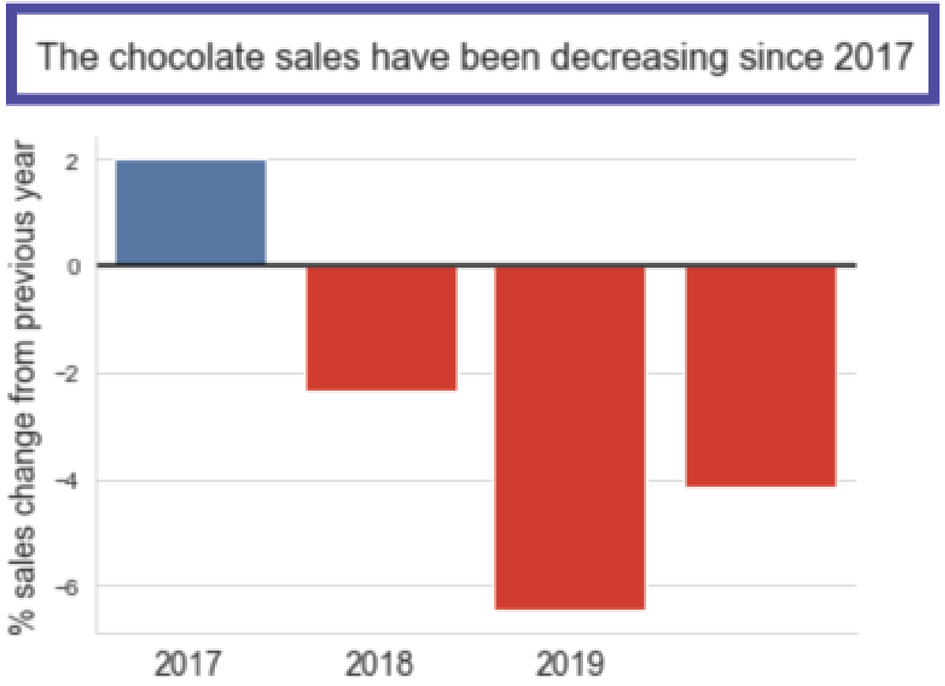
- Convention:
  - Value less than 0.05: statistical significance
  - Values close to 0.05: weak indicator

What is it not?

- Not proof of evidence
- Consider alternatives or complementary metrics

McCandless method

1. **Introduce visualization by name**
   - Graph headline
   - Clear and obvious
   - y vs x technique
2. **Anticipate audience's questions**
   - Focus on story not on decoding graph
3. State insights
4. **Help the audience relate**
   - Importance
   - Action items



[1] https://artscience.blog/home/the-mccandless-method-of-data-presentation

Oral communication

Advantages

- Relationship with the audience
- Immediate feedback
- Non-verbal cues

Disadvantages

- No permanent record of communication
- Not suitable for long messages

Written communication

Advantages

- Permanent record of communication
- Shared easily with a large audience
- Less emotional reaction to message
- Suitable to share code with colleagues

Disadvantages

- Hard to see if the message was understood
- No immediate feedback

Chapter 3

How to structure a written report?

- Types of reports
- Reproducibility
- Write precise and clear reports

Written reports

- Explain data analysis project
  - Sentiment analysis on product reviews
- Communicate findings
  - 30% negative ratings for delayed shipping
  - Predict ratings with 90% accuracy
- Standards
- Give recommendations to drive change

Analytical

- Analysis (recommendations)
- Varies (short or long)
- Strict structure
- Data-driven decisions

Report structure

- Introduction
- Body
- Conclusions

Summary report

Elements

- Key findings and recommendations
- Visuals

Format

- Short (< 5 pages)
- Summary of final report
- Link to main document

Final report

Elements

- Data analysis
- Findings and results
- Visuals

Format

- Long

Report structure

- **Introduction**
  - Purpose
    - Analysis of the product reviews gathered from website
    - Rating prediction based on review
  - Contextual information
    - Increase in negative reviews
  - Question of analysis
    - Factor affecting bad user experience

Report structure

- Introduction
- **Body**
  - Data
    - Description and tables
  - Methods
    - NLP and Random Forest
  - Analysis
    - Visuals
      - Graphs with most common words
  - Results
    - Description and visuals
      - 30% negative ratings associated with words "delayed" and "shipping".

Report structure

- Introduction
- Body
- **Conclusions**
  - Restate question
  - Summarize important results
  - Add recommendations

Example

The purpose of this report is to describe the results obtained from a model that predicts and identifies customers that will likely churn.

The data, gathered from the website, contains categorical data, such as gender, and internet service. It was converted to either 0 or 1 columns.

The dataset was split into train (70%) and test (30%) set. A K-Nearest Neighbors model was trained and model performance was evaluated.

As you can see in the graph, this reports analyzes the importance of different features such as monthly charges, contract type and phone service.

The model has an accuracy of 92% in predicting customer churn. The internet service type and discounts correlates with customer churning.

In summary, discounts on premium phone services should be implemented in order to retain customers.

ChatGPT

1. Project Title: The title of your project should be clear and concise.

2. Project Overview: Provide a brief overview of your project, including the problem or question being addressed, the data used, and the methodology applied.

3. Installation: Include instructions on how to install any necessary software or packages required to reproduce your analysis.

4. Data: Describe the data used in your project, including how it was obtained, the format, and any cleaning or preprocessing steps that were performed.

5. Methodology: Describe the methodology or techniques used to analyze the data, including any statistical or machine learning models applied.

6. Results: Provide a summary of the results of your analysis, including any visualizations or tables that you created.

7. Future Work: Discuss any limitations of your project and suggest possible directions for future work or improvement.

8. Acknowledgments: Acknowledge any sources of support or assistance you received in completing your project.

9. References: Include a list of references or sources used in your project.

Report structure

- Business context
- 1-3-25
  - 1 page of abstract
  - 3 max pages of executive summary
  - 25 max pages of detail

Summary report structure

- Introduction
  - Purpose
  - Contextual information
  - Question of analysis
- Body
  - Data
  - Results: Key findings
- Conclusions
  - Restate question
  - Central insight
  - Add recommendations

The report pursues the feasibility of automatizing a loan approval algorithm.

The project was motivated by the increasing percentage of defaulting customers over the last 5 years.

Is it possible to automatize the loan approval to customers with a low likelihood of default?

The data contains the loan details about borrowers from Loanme Bank. It was gathered over the last 10 years.

As you can see in the boxplot, the median age of defaulting customers is lower than the non-defaulting ones.

It is possible to predict which people is more likely to default with an accuracy of 95%.

If the automatization is implemented, the percentage of customers will decrease by 25% next year.

ChatGPT

1. Introduction: Start with a brief introduction that describes the problem or question that your project addresses. Explain the context and the motivation behind the project.

2. Data: Provide an overview of the data that you used for your project, including how you collected or obtained it, the size and format of the data, and any data cleaning or preprocessing steps you performed.

4. Results: Present the results of your analysis, including any visualizations or tables that you created. Describe the insights that you gained from the data and how they relate to the original problem or question.

5. Conclusion: Summarize your findings and draw conclusions from your analysis. Explain how your project contributes to the broader field of data science and its potential real-world applications.

6. Code: Provide a link to the code for your project, preferably in a GitHub repository or other code-hosting platform. Include any relevant data files or instructions for reproducing your analysis.

7. Future Work: Discuss any limitations of your project, and suggest possible directions for future work or improvement.

Reproducibility

- Data project
  - Run analysis again - **same results**

Replicability

- Data project
  - Different environment

Virtues

- Prevents duplication effort
- Build upon preexisting work
- Focus on new challenges
- Peer review
- Tool agnostic

Best practices

1. Keep track of how results were produced
   - Well document scripts
     - Comments in code
   - List packages and environment used
   - Version control

Best practices

1. Keep track of how results were produced
2. Avoid manual data manipulation
    - Data versioning
    - Store raw data and intermediate steps
    - Adapt and resolve problems

Best practices

1. Keep track of how results were produced
2. Avoid manual data manipulation
3. Control randomness
    - Random seeds for ML pipelines
    - Controls confounding variables

Best practices

1. Keep track of how results were produced
2. Avoid manual data manipulation
3. Document randomness
4. Interpretability
    - The degree to which a human can understand the cause of a decision or predict model results
    - Story with compelling **narrative**
    - Link with reproducibility

[1] Molnar C. Interpretable Machine Learning. 2019.

Best practices

1. Keep track of how results were produced
2. Avoid manual data manipulation
3. Document randomness
4. Interpretability
5. Cite bibliography correctly
    - **APA style**:
        - In text citations (author, date)
    - **Business context**
        - Less strict
        - Simpler (hyperlink)
        - ==> information available and retrievable

Reference Tools

- Reference management tools
    - Easier to keep track
    - Change between styles
    - Search for reference online
    - Options:
        - EndNote
        - Mendeley
        - RefWorks
- Business context
    - Less strict
    - Simpler (hyperlink)

Repo Structure by ChatGPT

- R/: Folder for your R scripts, including the script(s) used to generate the report.
- data/: Folder for any data files used in your project.
- docs/: Folder for your report file(s), in this case a .Rmd file and a rendered .pdf file.
- renv/: Folder for renv package management. This folder will be created and populated by renv.
- .gitignore: File specifying which files/folders should be ignored by Git when committing changes.
- LICENSE: File specifying the license for your project.
- README.md: File with information about your project, including how to install dependencies and generate the report.

Write precise and clear reports

DATA COMMUNICATION CONCEPTS

Hadrien Lacroix
Curriculum Manager

datacamp

Empty phrases

- Contain no information
  - It is interesting to note that
  - The fact that
  - It should be pointed out that
  - It is well known that
  - It is obvious that
- Distracting
- ==> should be removed

Bad

Another important point is the fact that negative ratings were associated with the words "delayed" and "shipping"

Good

Negative ratings are associated with the words "delayed" and "shipping"

[1] Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

Concrete nouns

- Write concrete nouns
- Avoid "this", "that", "it"
  - Adds cognitive load
  - Distracts them from insights
- **Active voice**: emphasis on the author

Bad

This shows an accuracy of 80% when predicting customer churn.

Good

The model shows an accuracy of 80% when predicting customer churn.

[1] Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

Redundant adjectives and adverbs

- Phrases that say the same thing twice
  - Introduce a new
  - Done previously
- Eliminate redundant adjective and adverbs

[1] Nolan D, Stoudt S. Communicating with Data. OUP Oxford. 2021.

Run-on sentences

- Two or more independent clauses connected incorrectly
  - There is a correlation between delayed shipping and customer rating, the shipping delay is the cause for negative review.
- Correction
  - Make two sentences
  - Use dependent clause by introducing words like because or so

Case study: report on credit risk

DATA COMMUNICATION CONCEPTS

Hadrien Lacroix
Curriculum Manager

Credit risk

- Credit risk: probability of defaulting
- Loanme bank wants to predict if a customer is likely to default
- Raw data available
- Data Exploration Analysis
- Model training and evaluation

Audience

- Non-technical stakeholders
- Bank decision-makers

Story

- Background:
  - Increase in defaulting percentage over last 5 years.
  - Predicting which customers had a high probability of default.
- Insight: People with more unemployment periods tends to default more
- Insight: People with lower income tend to default more
- Climax: Possible to predict which people is more likely to default with an accuracy of 95%
- Next steps: Run a trial on a control population

Tech or non-tech

- Translate technical results

The right data

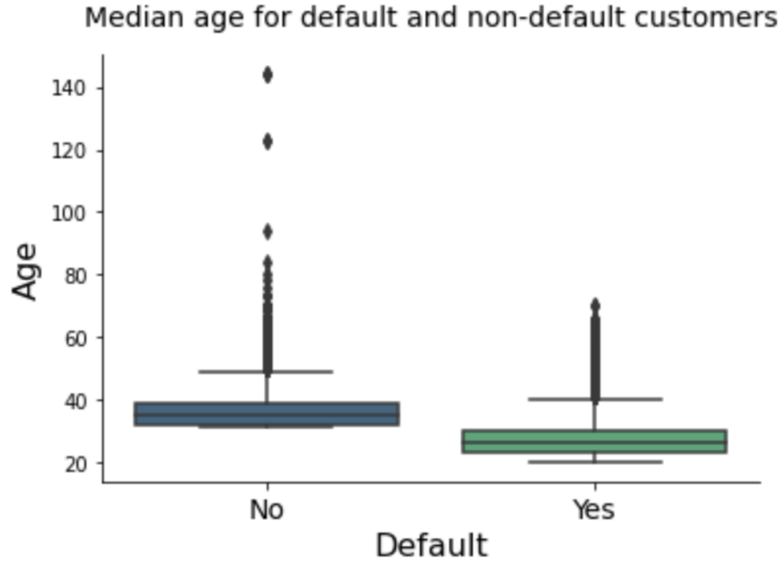
- Audience persona
  - **Role:** Financing Department Director
  - **Interest:** Decision on implementing an automated loan rejection system
  - **Appropriate data:**
    - Relationship between unemployment or income and loan default
    - Percentage customer defaulting over the next months

Statistics

- Median age and income
- Percentage of change

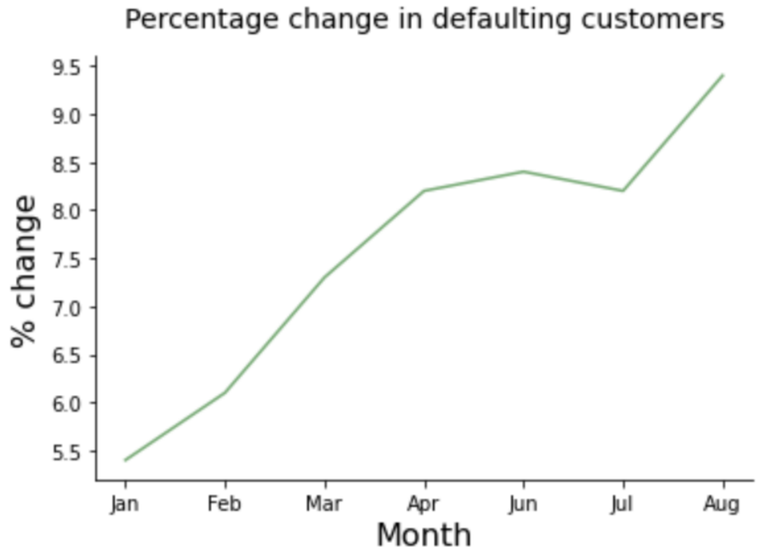
Visuals

- Boxplot with age vs. default condition



Visuals

- Boxplot with age vs. default condition
- Lineplot with % change defaulting customers



Correct format

- Who? **Financial Department director**
- Why? **Important decisions ahead**
- Content: **Key findings and recommendations**
- Channel: **Send the results before the meeting**

Report

- Written report
- Summary report or final report?

Report

- Summary report
- Informational report vs. analytical report?

Report

- Summary report
- Analytical report

Let's practice!

DATA COMMUNICATION CONCEPTS

Planning an oral presentation

DATA COMMUNICATION CONCEPTS

Hadrien Lacroix
Curriculum Manager

Message

What is the central message? After one week: 90% forgotten ==> What do we want to stick?

- **Opening statement**
  - Capture audience's attention
    - Negative ratings scare customers away from our website
- **Central message**
  - One sentence
    - Delayed shipping is the main cause of negative reviews and immediate actions are needed to revert the situation.
- **Closing statement**
  - Sums up presentation and strengthens central message
    - There is a decrease in sales. Negative reviews have been increasing. Delayed shipping is causing negative ratings. Actions are needed to revert situation.

Structure

- **Introduction**
  - Provide background information
  - Catch audience attention
  - Glimpse of presentation content
- **Methods, analysis and model outputs**
  (only for technical audience)
- **Conclusions and takeaways**
  - Refers back to the introduction
  - Contains call-to-action statement or/and next steps

Outline

- Graphs and visuals
- Sections (five or less smaller parts)
  1. Reason for analysis
  2. Exploratory analysis
  3. Sentiment analysis
  4. Conclusions
  5. Follow-up actions

Structure

- **Introduction**
  - Provide background information
  - Catch audience attention
  - Glimpse of presentation content
- **Methods, analysis and model outputs** (only for technical audience)
- **Conclusions and takeaways**
  - Refers back to the introduction
  - Contains call-to-action statement or/and next steps

Example

- Increase in private sector salaries over the last 5 years
- Analysis of software development salaries from different companies
- Forecast salaries for the software sector for the next 5 years
- Employees will be leaving due to higher salaries on other companies
- Company should offer competitive salary ranges to retain talented staff

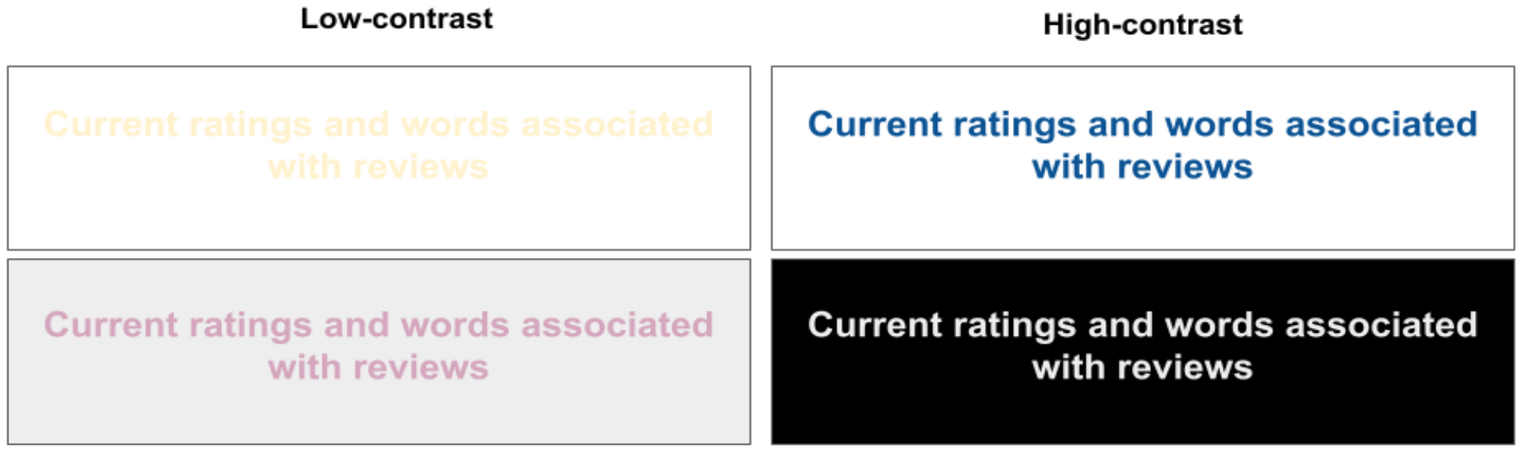
From planning to building

- Slides
  - Support story
  - Refined slides
  - Slide count or timing = bad metric
  - **One message per slide**

[1] https://www.slidecow.com/blog/how-many-slides

Color

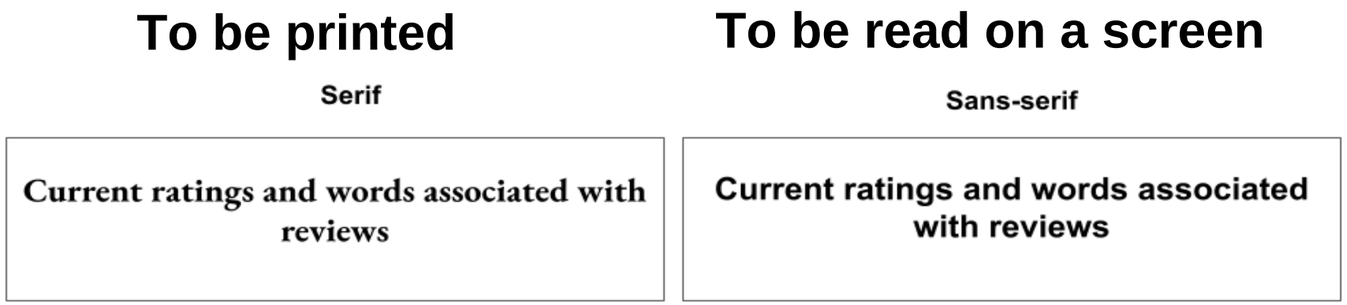
- No more than **3 colors**
- Good **contrast** between words and background
- **Inclusive**
  - Color deficiency
  - Example: green and red



Fonts

- Serif vs sans-serif
- Context
- Support
- Size

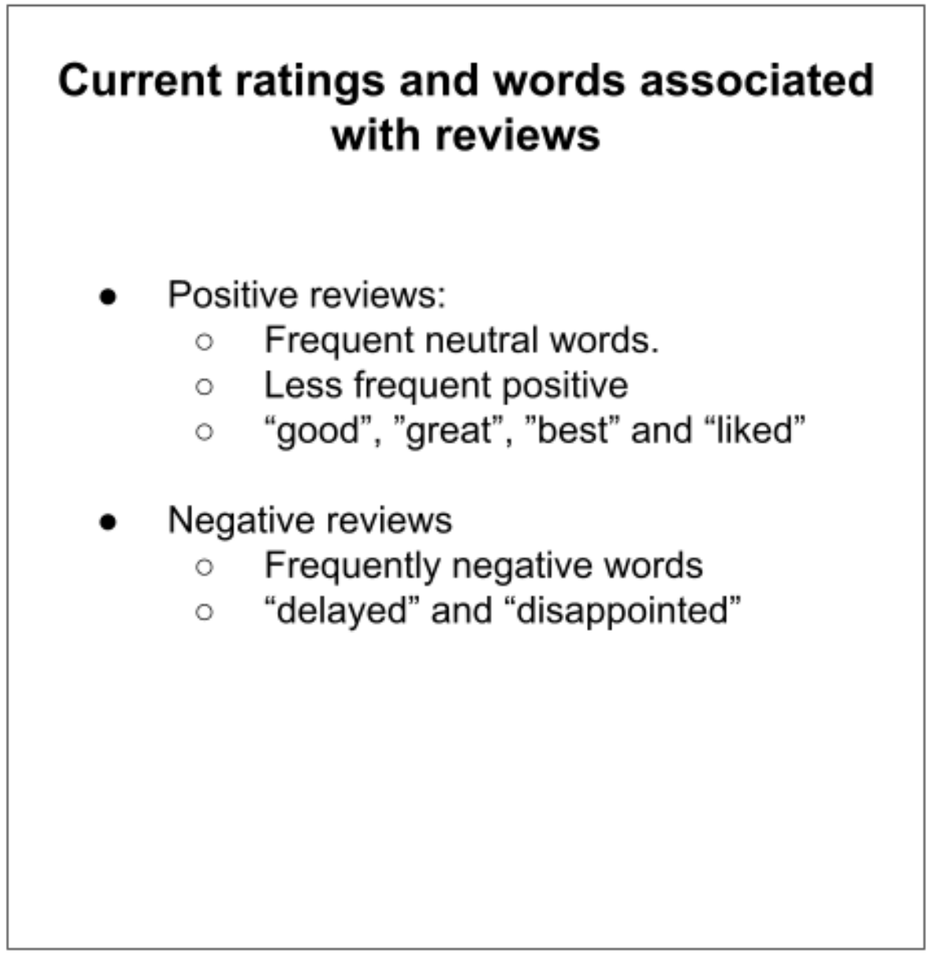
- Several fonts
- Spacing of letters and lines
- **Bold**, italic and sizes



For **positive reviews**, some of the words that appear frequently do not have a particular connotation and can be interpreted as **neutral**.

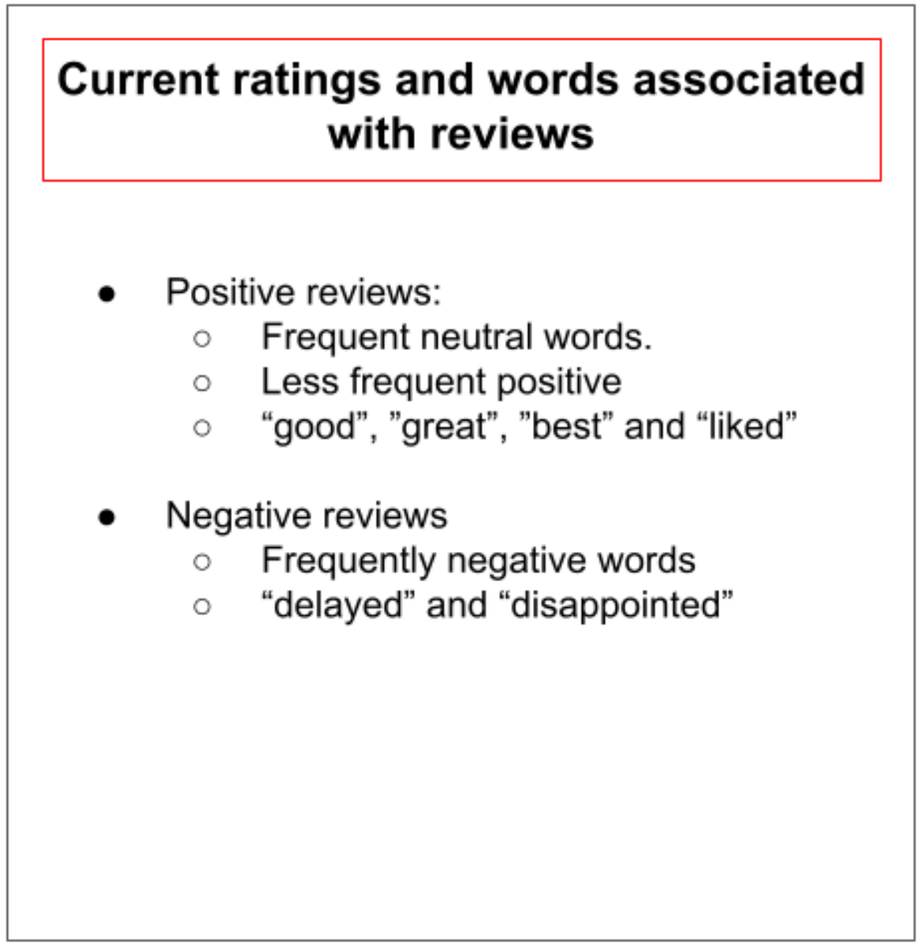
Text slide

- **Main points**
  - Don't dual purpose the slide deck



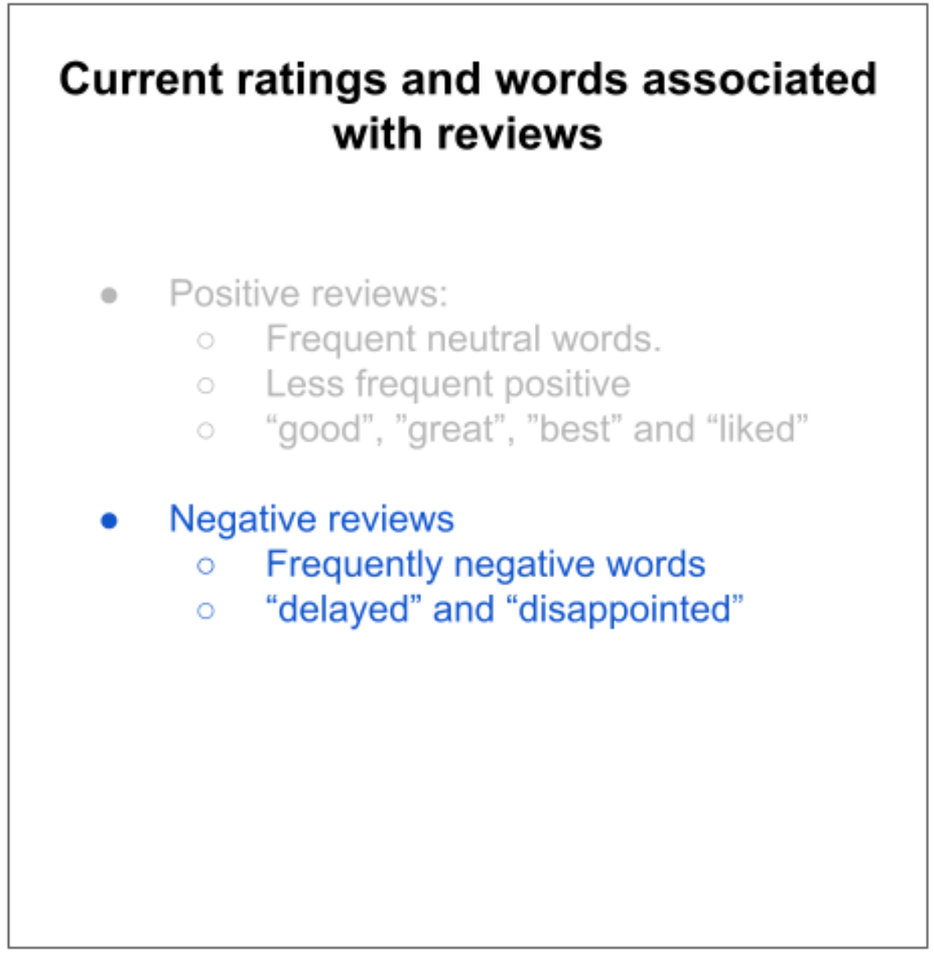
Text slide

- Less text
- **Headline**
  - Highlight main point
  - Specific and concise
  - Big size



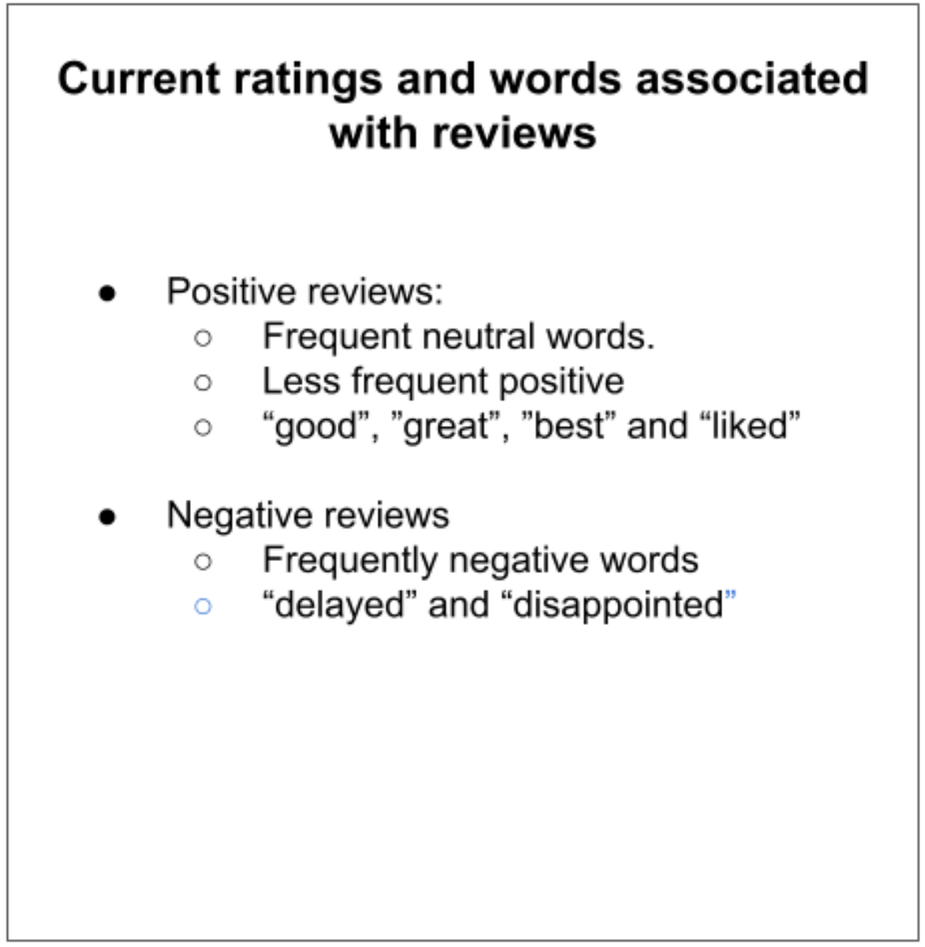
Text slide

- Less text
- Headline
- **Layering approach**
  - Breaks complex slide into smaller points
  - Present each point on its own



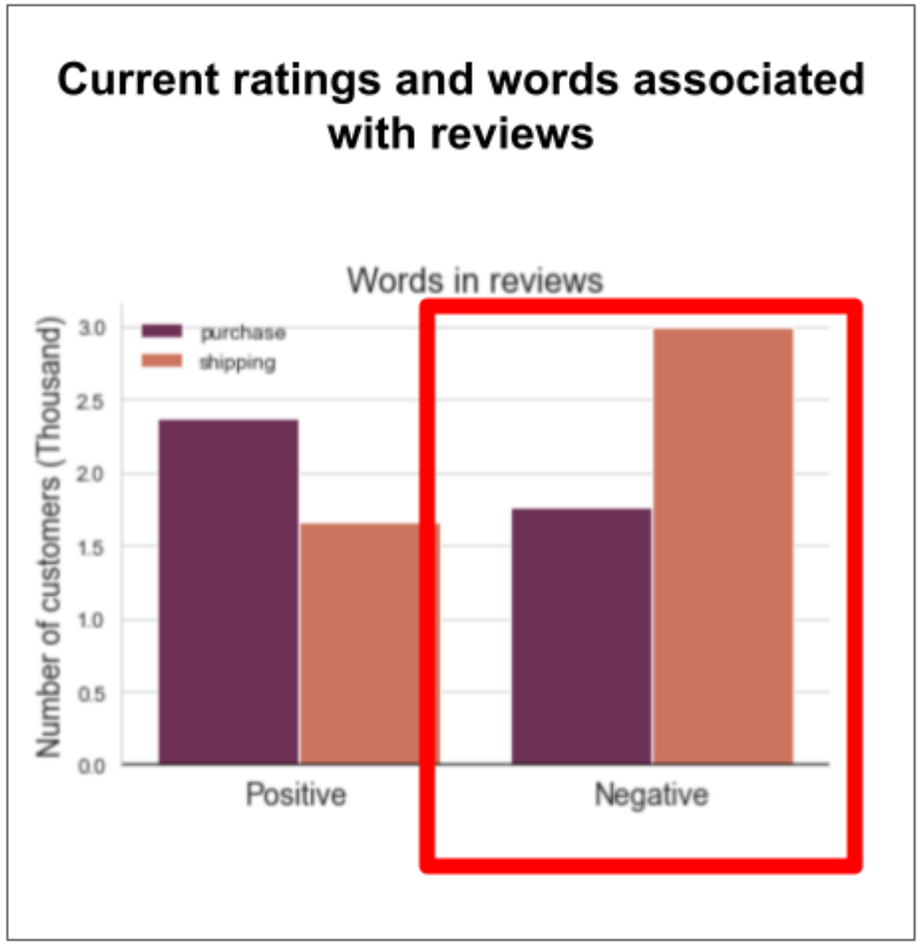
Text slide

- Less text
- Headline
- **Layering approach**
  - Breaks complex slide into smaller points
  - Present each point on its own
  - Displayed together



Visualization slide

- Replace many sentences
- **Use layering and highlighting**
- Headline (If needed)
- **One or two full-size graphs**
  - One message per slide
  - No overcrowding



Delivering the presentation

DATA COMMUNICATION CONCEPTS

Hadrien Lacroix
Curriculum Manager

datacamp

Practice

- Write script
- Don't memorize
- Become familiar with content
- Anticipate follow-up questions
- **Rehearsal**
  - Stand up
  - Use the slides
  - Speak out loud
  - Detect distracting patterns (um, so, like, basically, actually)
  - Find linking statements
  - Answer to Q&A
- **Be aware of emotions**
  - Confidence vs. unsure
- Short attention span
- Talk **to** audience (not _at_ them)
- Develop a relationship
- Timing (around 20 minutes)
- Pace
- **Open up for questions**
  - During or at the end of the presentation

Audience involvement

- **Engage** and **involve** audience
- **Strong introduction**
  - *Good morning! My name is Hadrien, and I'm here today to present how negative ratings are impacting the company profits.*
- **State key assumptions**
- **Ask questions**
  - Know answer
  - Hook for next slide
- **Reiterate** to main idea

Guide audience

- Sequence of information
- Keep audience's attention
- Do not leave all findings to the end

Body language

- What matters is the message...
- ...but the speaker is at the center of the presentation
- Emphasis by natural gesture and movements
- Attracts attention
- **Supports message**

Voice tonality

- Use different voice tonalities
  - **Speed**
    - **Fast**: urgency, excitement, and emotion
    - **Slow**: importance, and new ideas introduction
  - **Volume**
    - Live: speak loud
    - Online: check mic
  - **Intonation**