WOMEN'S SUMMIT

AGENDA

(1) Project Overview

(2) Getting Started

(3) Tips for Success

PROJECT OVERVIEW

PROJECT OVERVIEW

Our capstone projects are designed to help aspiring data scientists simulate a real-world experience and showcase their skills.

The projects...

(1) Are purposefully-open ended

(2) Require working in teams

(3) Have a heavy emphasis on impact and relevance

Sample projects

correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT

The NLP News Sentiment Factor Trading Strategy for a Portfolio of S&P 500 Stocks

Anastasia Tatarenko, Daria Yurova , Ningyuan Zhang, Yang Su, Rohini Shimpatwar GitHub: shorturl.at/elMRU

Highlights

- Define and obtain NLP sentiment index using BERT and Vader models
- Create 10 equally-weighted portfolios based on sentiment index and test if the sentiment factor is statistically significant
- Develop the trading strategy that goes long on top 10% stocks with high company sentiment and shorts bottom 10% with the low sentiment on the previous day with daily rebalancing
- Check if the industry sentiment index is significant for a portfolio within each industry

Background

Our research aims to answer three questions: Which NLP model is the best for sentiment factor extraction on financial news? Does the news sentiment factor help predict stock returns? Does this strategy beat other benchmark trading strategies, e.g. buying the market portfolio?

Datasets and Pre-processing

Datasets:

1. 'US Financial News Articles' from Kaggle
2. S&P 500 companies and industries from Wikipedia
3. S&P 500 daily stock prices from Yahoo Finance

Data Preprocessing:

1. Remove noise words: removing stopwords, special characters, dates, common names, numbers, etc
2. Labeling: we put sentiment labels on 800 articles as negative(-1), positive(1) and neutral(0) and match all news articles with S&P 500 companies and industries

Data Insights

Top bigrams indicate that the phrases most frequently mentioned are related to financial statements (e.g. "net income", 'GAAP financial'), and CEO compensations (e.g. 'chief executive', 'based compensation'). The top 2 industries mentioned are Consumer and the IT industry.

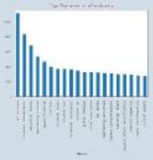


Figure 1 Top Bigrams

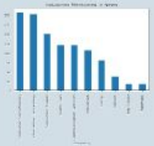


Figure 2 Industries in News

Figure 3 Word Cloud

Sentiment Factor Extraction Using NLP Models

Approaches

1. **Rule-based -- Valence Aware Dictionary and Sentiment Reasoner(Vader) model:** A lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. We directly use Vader to predict sentiment for all news.
2. **Transfer learning -- FinBERT model:** We train a FinBert model based on BertForSequenceClassification(BFSC) model, which is built on BERT(Bidirectional Encoder Representations from Transformers) with an extra linear layer on top. To capture the sentiment in financial news, we perform transfer learning and fine-tune the BFSC model using the 800 labeled news articles and then predict the sentiment for the rest 39,000 news in our news data set.

Model Pipeline for BFSC Model

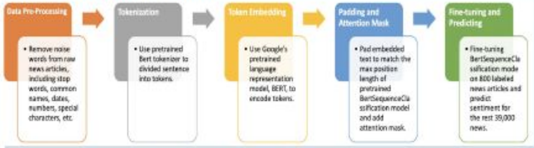


Figure 4 Model Pipeline for FinBERT

Result and Discussion

1. **Vader:** Vader achieves ~0.56 accuracy on labeled news. As shown in figure 5, Vader recognizes positive news well but tends to label neutral/negative news as positive.

Figure 5 Confusion Matrix of Vader

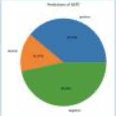


Figure 6 FinBERT Sentiment Predictions

2. **FinBERT:** Our BFSC model achieves ~0.65 valid accuracy and ~0.81 accuracy on labeled news. According to the predictions, negative and neutral news takes up ~60%, which is closed to real-world situations. Overall, FinBert performs better than Vader on financial news texts. However, 800 samples are not sufficient for good transfer learning. Higher accuracy could be achieved with more labeled data and by trying other top layer architectures to better capture data distribution.

Factor Model Based Trading Strategy

Industry and Company Sentiment Index

Using results from our sentiment models, we construct a time series of daily values for industry and company sentiment indices. The company sentiment index is the average sentiment for each company in the industry each day. The company sentiment index is the average of the sentiment on all the news about this company each day.

Factor Model and Trading Strategy

We then test a simple trading strategy by sorting at the end of each day stocks based on their company-specific sentiment index on the previous day, splitting the stocks into 10 portfolios based on the sentiment and forming an equally-weighted long-short portfolio by buying the portfolio with the highest sentiment and short-selling the one with the lowest sentiment on the previous day.

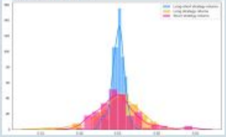


Figure 7 Distribution of daily returns for L, S and L-S portfolios

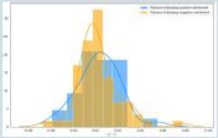


Figure 8 Distributions of returns of the Materials sector L-S strategy following days of positive and negative Industry sentiment index (similar to the other sectors)

Conclusions

1. We use the FinBERT model for sentiment factor extraction and achieve an accuracy of 0.65 on unlabeled news. With the obtained accuracy, the sentiment factor can contain noise and affect the performance of our trading strategy
2. Based on the Chi-squared test we cannot conclude that alphas for each portfolio are jointly significant. Therefore, our sentiment signal does not predict the returns, and investors correctly update their beliefs about the prices based on the information from the news
3. However, we find discrepancies between distributions of returns following positive and negative industry sentiment. Hence, improving the accuracy of sentiment prediction or constructing a more sophisticated sentiment factor strategy may improve the results

References

1. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova https://arxiv.org/abs/1810.04805
2. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models Dogu Tan Araci https://arxiv.org/pdf/1908.10063.pdf
3. Text-Based Industry Momentum, Gerard Hoberg and Gordon Phillip
4. BERT model FinBERT implementation relies on Hugging Face's transformers library. https://huggingface.co/transformers/index.html

Sample projects

correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT

HOW CAN A FINTECH COMPANY IN LATIN AMERICA IMPROVE TIME-TO-MARKET FOR LOAN APPLICATIONS?

Andrés Murillo, Renan Añez, Ricardo Ángel Granados, Roger Terán, Fernando Aguirre

IMPACT

Provide underserved Latin Americans with capital quickly, seamlessly, and responsibly

30% reduction in loan application process, improving customer experience

20 FTE reduction due to less need for manual data input in a new application form

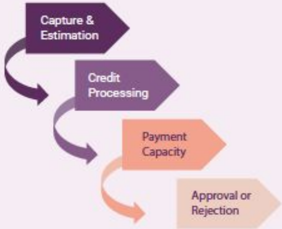
11% improvement in accuracy means more capital to those who need it while minimizing default risk

BACKGROUND

Time-to-market is one of the main reasons clients choose one financial institution over another. The faster you can give them an answer, the better. But you better be right.

DATA & PROCESS

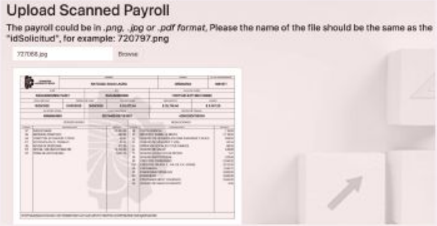
We used 5k+ loan registries that calculate the clients' payment capacity to train an OCR to automate this process which is currently slow, manual, and error prone.
We also used 700k+ historical loans and their outcomes with almost 100 variables to train a risk assessment model to improve the company's current model.

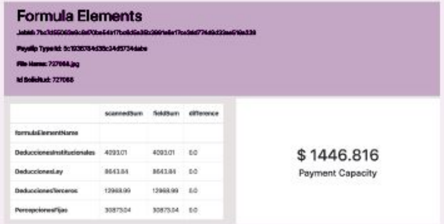


MACHINE LEARNING

For the OCR we used Amazon Textract in combination with Deep Learning to automate the information extraction for the payment capacity from government payroll slips.
For the credit assessment, we experimented with different methods including loss-based, tree-based, and probability-based classifiers. A random forest proved best.

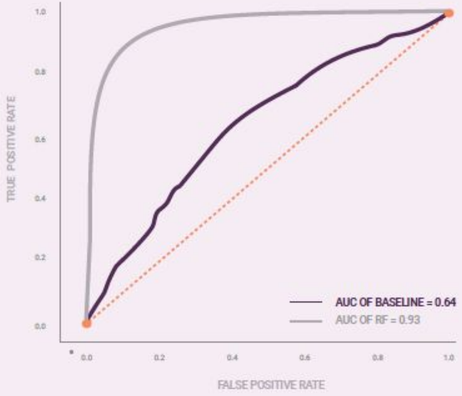
The OCR extracts both income and deductions from a PDF file, calculating a client's payment capacity with near perfect accuracy.





We tackled two challenging problems that delay a loan application process: automating the **payment capacity calculation** and enhancing the **credit risk assessment.**

The payment capacity from the OCR was a key input to improve the risk assessment. By training a random forest with additional features, we improved the accuracy by 11% and AUC by 40% over a baseline model.



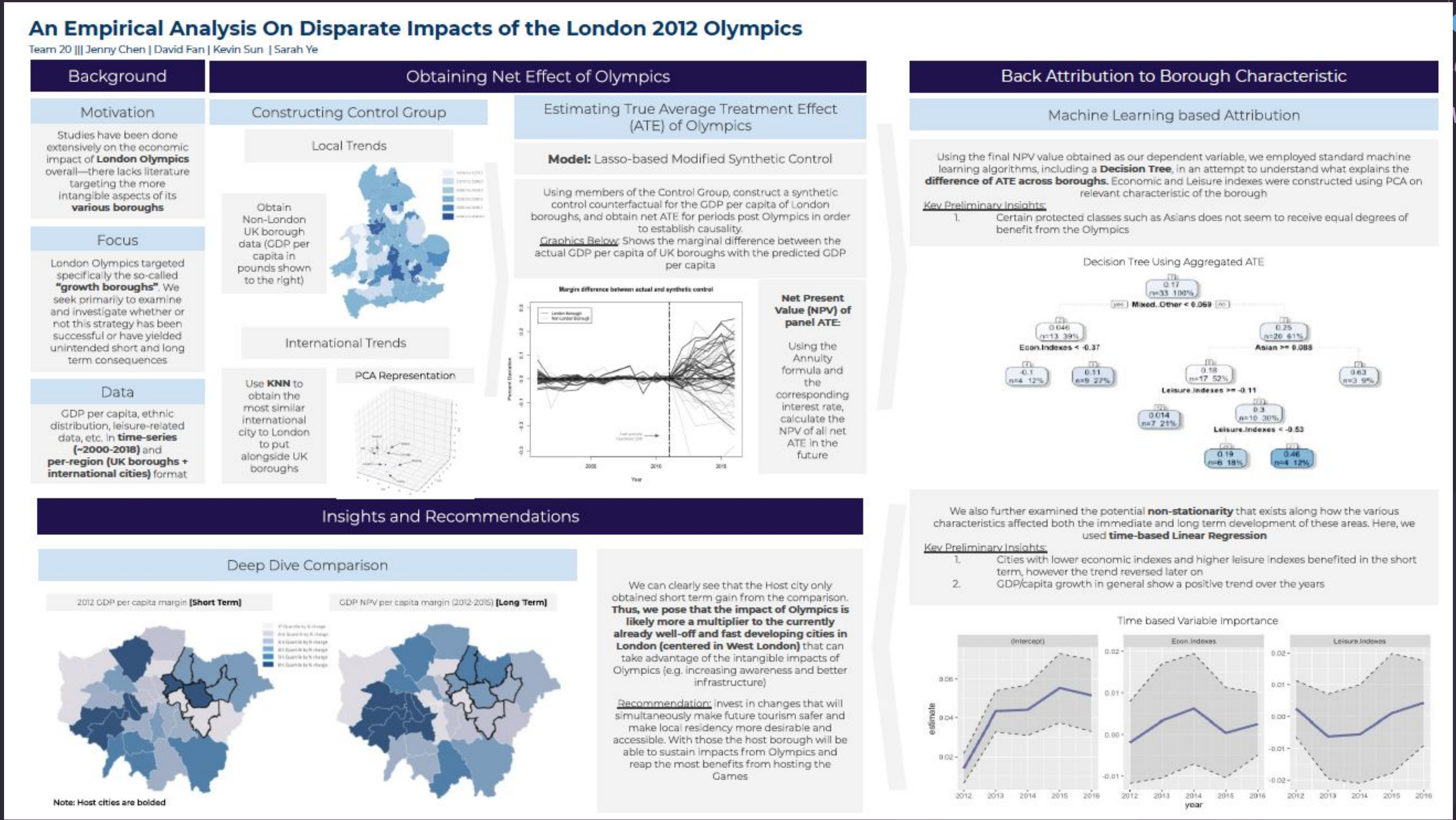
FUTURE WORK

During our data exploration, we found that both the amount of capital and term of the loan are highly correlated with the clients' probability of default. This could be an opportunity to determine the optimal amount to lend given a risk profile.

Sample projects

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT



An Empirical Analysis On Disparate Impacts of the London 2012 Olympics

Team 20 ||| Jenny Chen | David Fan | Kevin Sun | Sarah Ye

Background

Motivation

Studies have been done extensively on the economic impact of **London Olympics** overall—there lacks literature targeting the more intangible aspects of its **various boroughs**

Focus

London Olympics targeted specifically the so-called **"growth boroughs"**. We seek primarily to examine and investigate whether or not this strategy has been successful or have yielded unintended short and long term consequences

Data

GDP per capita, ethnic distribution, leisure-related data, etc. in **time-series (~2000-2018)** and **per-region (UK boroughs + international cities)** format

Obtaining Net Effect of Olympics

Constructing Control Group

Local Trends

Obtain Non-London UK borough data (GDP per capita in pounds shown to the right)

International Trends

Use **KNN** to obtain the most similar international city to London to put alongside UK boroughs

PCA Representation

Estimating True Average Treatment Effect (ATE) of Olympics

Model: Lasso-based Modified Synthetic Control

Using members of the Control Group, construct a synthetic control counterfactual for the GDP per capita of London boroughs, and obtain net ATE for periods post Olympics in order to establish causality.

Graphics Below: Shows the marginal difference between the actual GDP per capita of UK boroughs with the predicted GDP per capita

Margin difference between actual and synthetic control

Net Present Value (NPV) of panel ATE:

Using the Annuity formula and the corresponding interest rate, calculate the NPV of all net ATE in the future

Back Attribution to Borough Characteristic

Machine Learning based Attribution

Using the final NPV value obtained as our dependent variable, we employed standard machine learning algorithms, including a **Decision Tree**, in an attempt to understand what explains the **difference of ATE across boroughs.** Economic and Leisure indexes were constructed using PCA on relevant characteristic of the borough

Key Preliminary Insights:

1. Certain protected classes such as Asians does not seem to receive equal degrees of benefit from the Olympics

Decision Tree Using Aggregated ATE

We also further examined the potential **non-stationarity** that exists along how the various characteristics affected both the immediate and long term development of these areas. Here, we used **time-based Linear Regression**

Key Preliminary Insights:

1. Cities with lower economic indexes and higher leisure indexes benefited in the short term, however the trend reversed later on
2. GDP/capita growth in general show a positive trend over the years

Time based Variable Importance

Insights and Recommendations

Deep Dive Comparison

2012 GDP per capita margin **[Short Term]**

GDP NPV per capita margin (2012-2015) **[Long Term]**

Note: Host cities are bolded

We can clearly see that the Host city only obtained short term gain from the comparison. **Thus, we pose that the impact of Olympics is likely more a multiplier to the currently already well-off and fast developing cities in London (centered in West London)** that can take advantage of the intangible impacts of Olympics (e.g. increasing awareness and better infrastructure)

Recommendation: Invest in changes that will simultaneously make future tourism safer and make local residency more desirable and accessible. With those the host borough will be able to sustain impacts from Olympics and reap the most benefits from hosting the Games

PROJECT TIMELINE

Date	Description
Sept 18	Introductions to projects and team formation
Sept 19	Project Proposals due
Sept 23	Project selections / mentor pairings announced
Sept 27	Project Outlines due
Oct 4	Report Drafts due (Complete through 2B)
Oct 11	Final Reports due
Oct 14	Final Presentations due
Oct 16	Final Presentation day

correlation.one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

FINAL PRESENTATION DAY

Start	End	Session
10:00 AM	12:00 PM	Project Symposium
12:00 PM	12:30 PM	Break
12:30 PM	1:00 PM	Keynote -- Sheena Iyengar
1:00 PM	2:00 PM	Top Project Showcase
2:00 PM	2:30 PM	Keynote -- Matthew Granade
2:30 PM	3:00 PM	Awards & Closing Remarks
3:00 PM	4:00 PM	Networking Breakouts

* Each project team will have 10 minutes to present

GETTING STARTED

PROJECT GUIDELINES

correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT

Guidelines Document Linked Here

(also shared on #projects)

PROJECT PROPOSALS

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

(1) Provided Prompts

Correlation One and select partners have provided a few points to inspire project ideas.

Project Prompts (also on #projects)

(2) Topic of Your Choice

We encourage you to brainstorm interesting project ideas that align with your passions, interests, and prior research.

FIRST STEP: PROPOSAL

correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT

Tomorrow, each team will submit <u>two</u> project proposals which answers the following questions:

(1) What question do you want to investigate?

(2) Why is this question interesting and/or relevant?

(3) Which datasets will help you answer this question?

(4) Which analysis techniques and technologies do you plan to utilize?

TIPS FOR SUCCESS

TIPS FOR SUCCESS

1. Start with good data
2. Pick a relevant, interesting question to ask
3. Tell a compelling story with the data
4. Leverage the strengths of your team

START WITH GOOD DATA

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

1. Data.gov
2. Healthdata.gov
3. Data.worldbank.org
4. WHO Open Data Repository
5. EU Open Data Portal
6. Kaggle.com
7. Data.world
8. AWS OpenData Registry
9. FiveThirtyEight
10. 100 more public data sources

HOW TO ASK THE RIGHT QUESTION?

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

A good question is...

- **Specific:** Can you visualize a possible answer to your question? The more clearly you can see it, the more specific the question is.

HOW TO ASK THE RIGHT QUESTION?

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

A good question is...

- **Measurable:** Is the answer something you can quantify? It’s hard to make decisions based off things that aren’t measured well with data.

HOW TO ASK THE RIGHT QUESTION?

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

A good question is...

- **Actionable:** If you had the answer to your question, could you do something useful with it? How relevant is the question to important stakeholders?

HOW TO ASK THE RIGHT QUESTION?

correlation.one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

A good question is...

- ❖ **Realistic:** Can you get an answer to your question with the data you have? If not, can you get the data that would get you an answer?

HOW TO ASK THE RIGHT QUESTION?

A good question is...

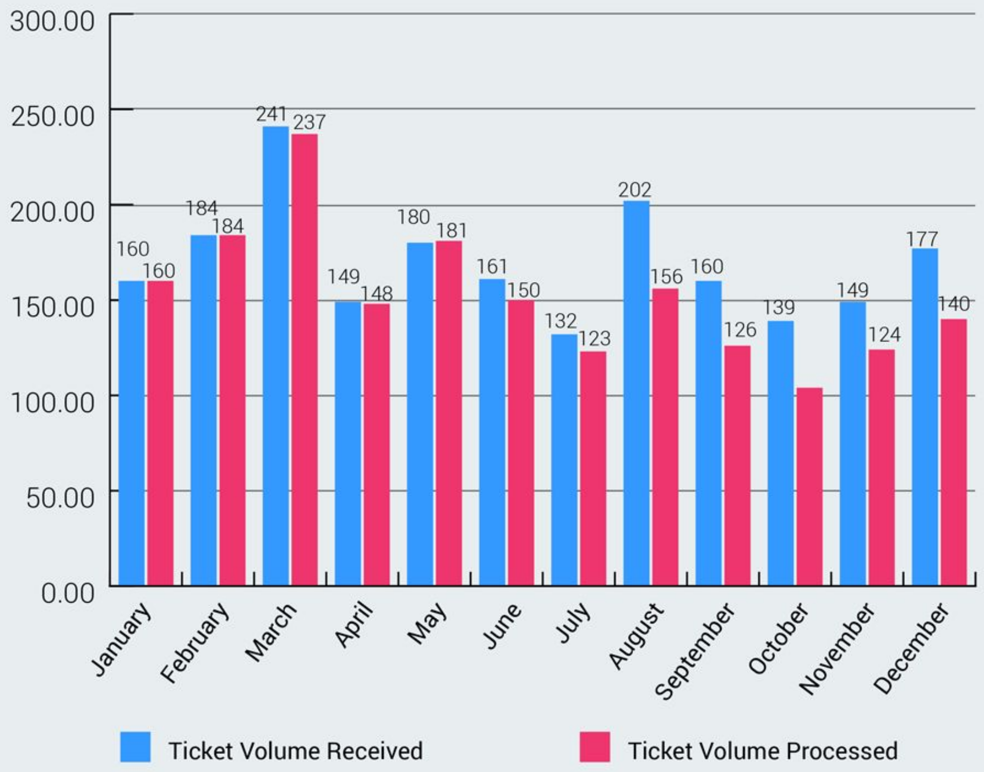
- **Timely:** Can you get an answer in a reasonable time frame, or at least as before you need it?

HOW TO TELL A STORY WITH DATA

1. Know Your Audience:

What are their interests and goals? How much background knowledge do they already have? Do they want the details, or just the high-level summary?

TICKET TREND

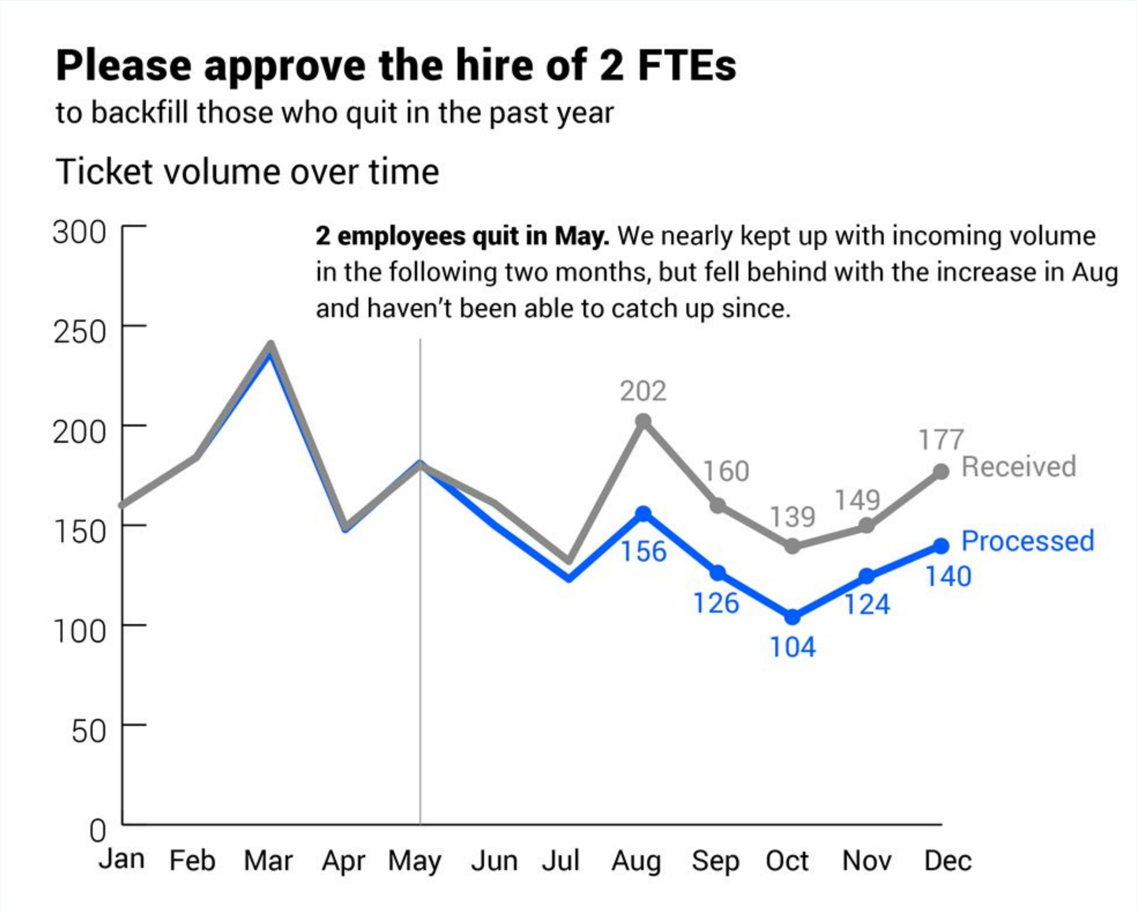


Source: Storytelling With Data by Cole Nussbaumer Knaflic

correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT



Source: Storytelling With Data by Cole Nussbaumer Knaflic

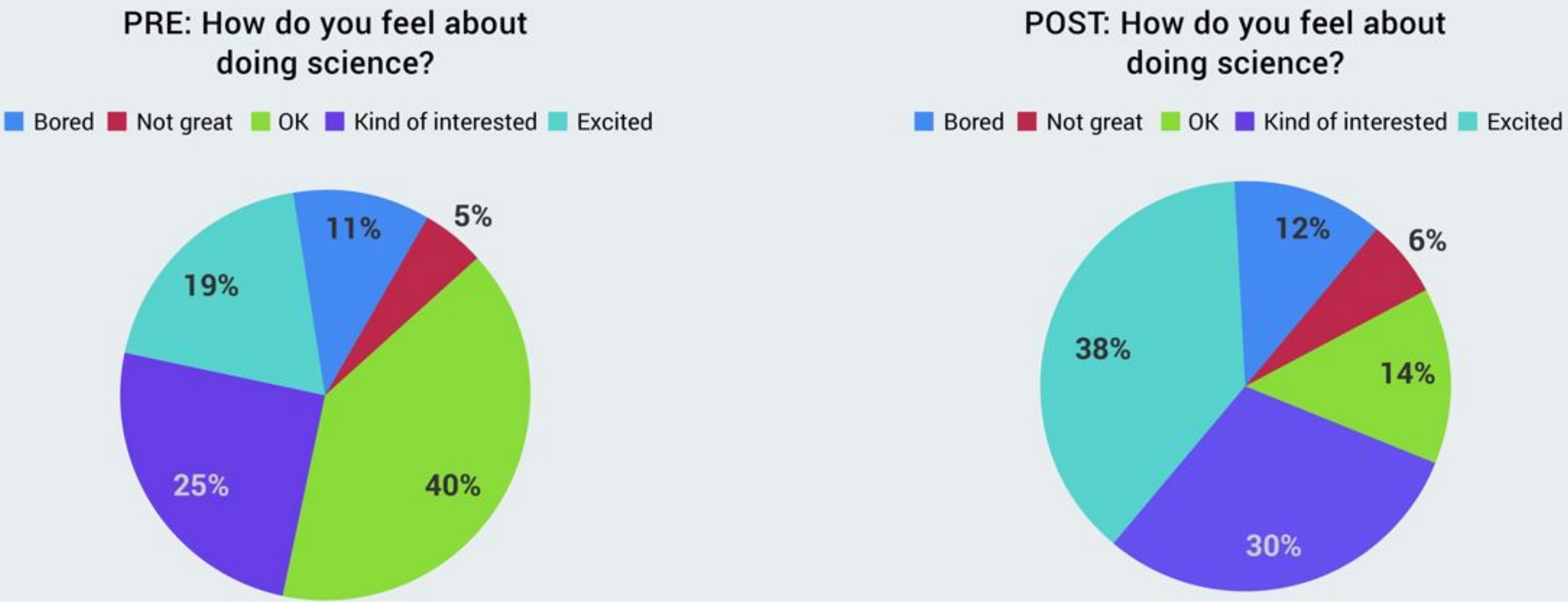
HOW TO TELL A STORY WITH DATA

correlation.one DATA SCIENCE 4 ALL WOMEN'S SUMMIT

2. Tell a Compelling Story

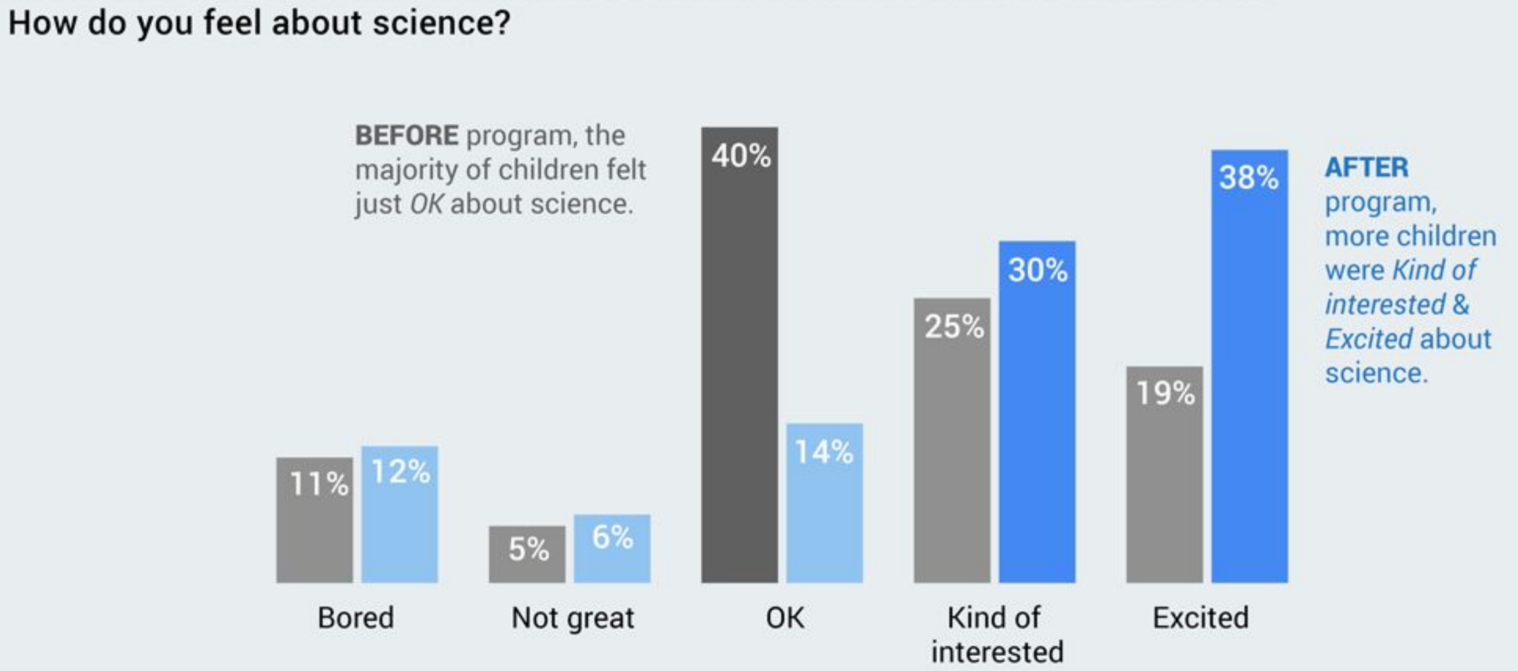
People remember stories, not data. Take them on your journey.

Survey Results



Source: Storytelling With Data by Cole Nussbaumer Knaflic

Pilot program was a success



Source: Storytelling With Data by Cole Nussbaumer Knaflic

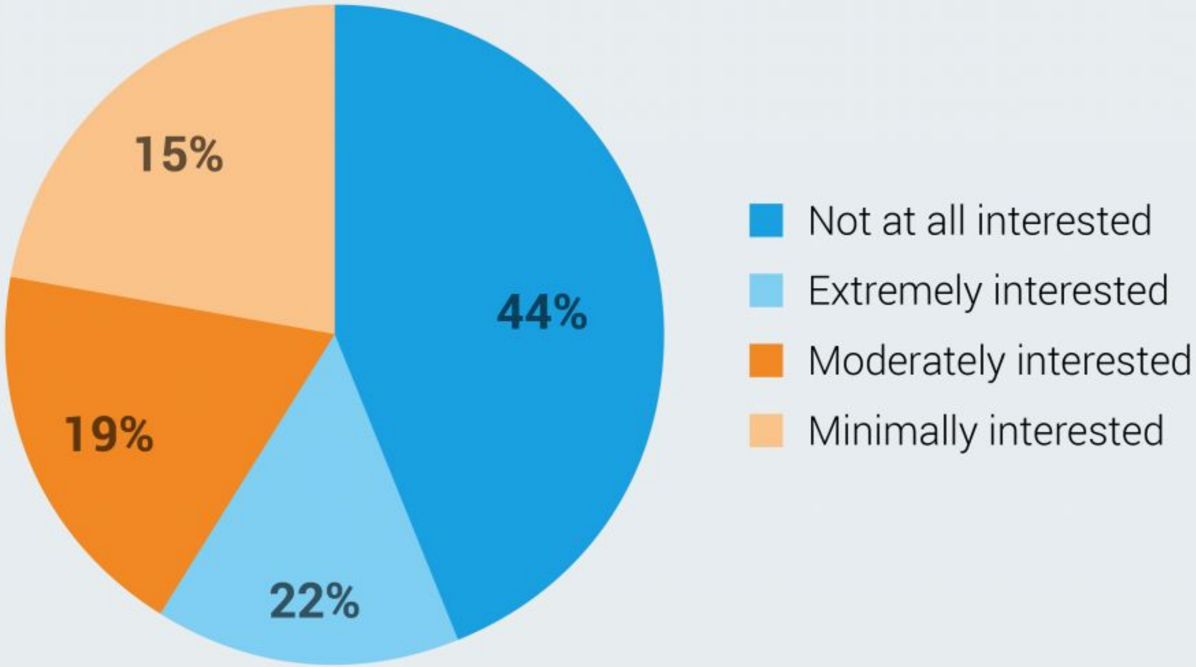
HOW TO TELL A STORY WITH DATA

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

3. Be clear and concise:

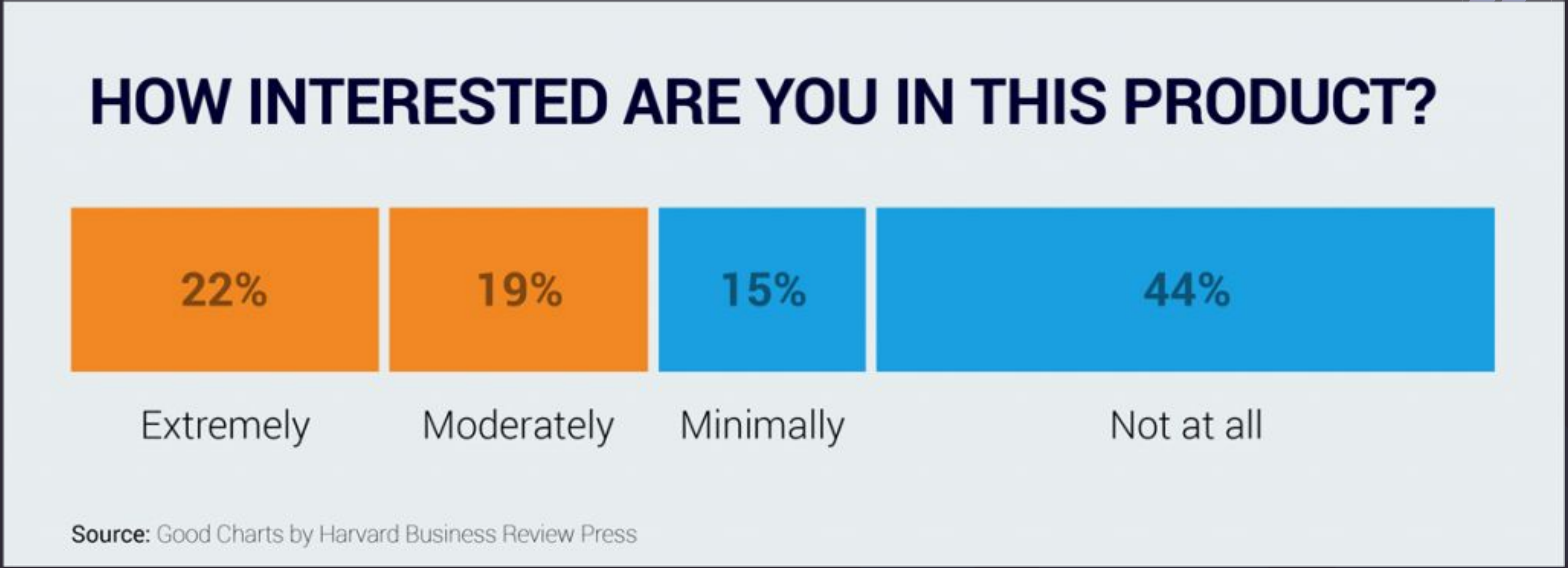
Make sure your graphic supports the story you are telling, and remove everything that is not part of your story.

HOW INTERESTED ARE YOU IN THIS PRODUCT?



Source: Good Charts by Harvard Business Review Press

correlation one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

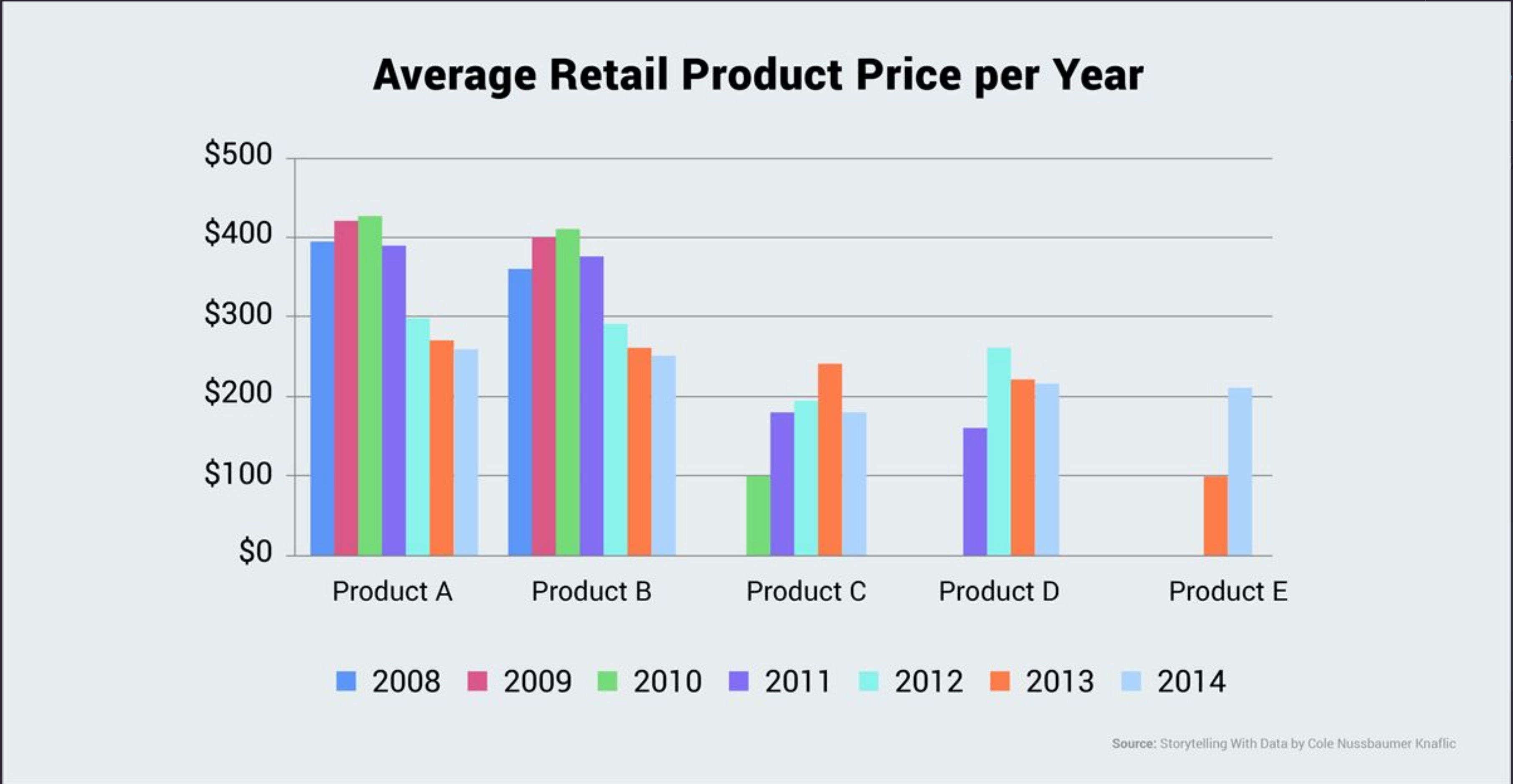

HOW INTERESTED ARE YOU IN THIS PRODUCT?
22%
19%
15%
44%
Extremely
Moderately
Minimally
Not at all
Source: Good Charts by Harvard Business Review Press

HOW TO TELL A STORY WITH DATA

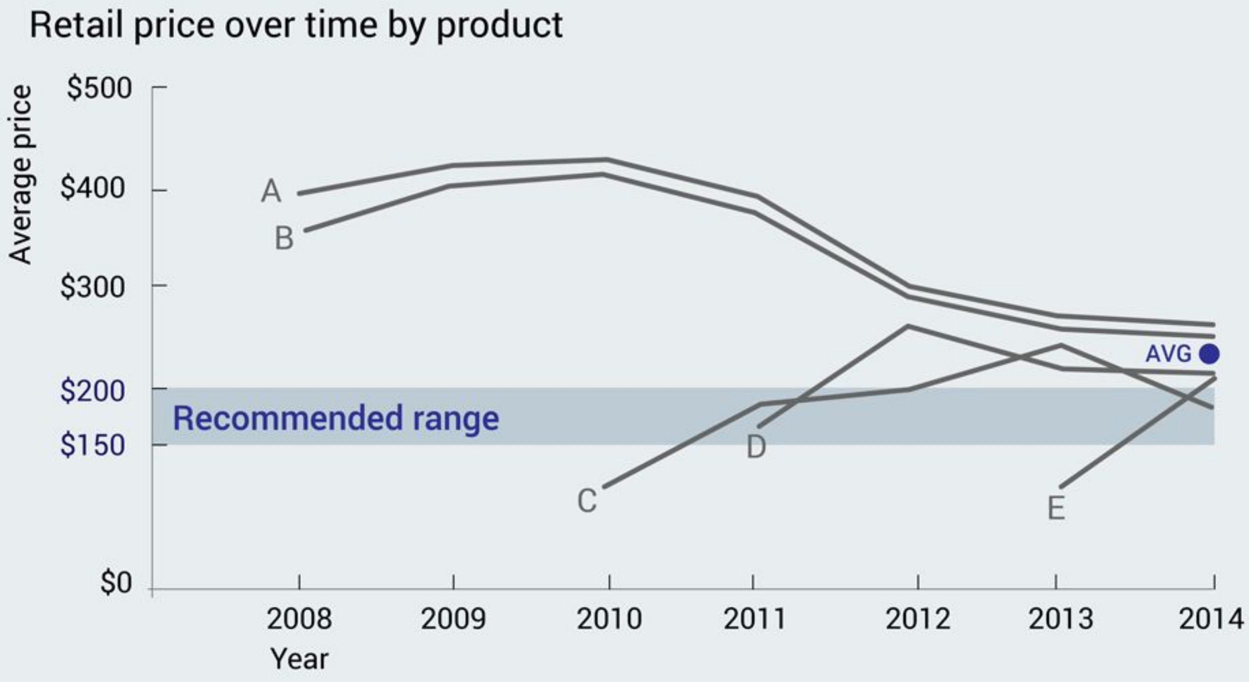
correlation.one DATA SCIENCE 4 ALL WOMEN'S SUMMIT

4. Provide context:

Compare metrics over time or to industry benchmarks. Numbers are meaningless without context.


Average Retail Product Price per Year
$500
$400
$300
$200
$100
$0
Product A
Product B
Product C
Product D
Product E
2008
2009
2010
2011
2012
2013
2014
Source: Storytelling With Data by Cole Nussbaumer Knaflic

To be competitive, we recommend introducing our product below the $223 average price point in the **$150-$200 range**



Source: Storytelling With Data by Cole Nussbaumer Knaflic

WORKING WITH YOUR TEAM

correlation·one
DATA SCIENCE 4 ALL
WOMEN'S SUMMIT

Our Goal	Help you develop great relationships with other young data scientists and experienced mentors
Group Lead	Main point of contact from each group that will coordinate your teams schedule and communicate with C1 Team and mentors
Communications	*Group Slack channel (strongly recommended)*
Mentors and TAs	Each group will have an assigned TA and mentor(s) who will provide project support

correlation one

DATA SCIENCE 4 ALL

WOMEN'S SUMMIT

QUESTIONS?

correlation one

DATA
SCIENCE
4 ALL

WOMEN'S SUMMIT