



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

INTRODUCTION TO DATA SCIENCE

Sankara Nayaki K

sankaranayakik@wilp.bits-pilani.ac.in

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

References:

- Introducing Data Science by Cielen, Meysman and Ali
- Storytelling with Data by Cole Nussbaumer Knaflic; Wiley
- Introduction to Data Mining by Tan, Steinbach and Vipin Kumar
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas

TABLE OF CONTENTS

1 COURSE HANDOUT

2 STORY TELLING WITH DATA

COURSE HANDOUT

- M1 Introduction to Data Science
- M2 Data Analytics
- M3 Data Science Process
- M4 Data Science Teams
- M5 Data & Data Models
- M6 Data Wrangling & Feature Engineering
- M7 Data Visualization
- M8 Storytelling with Data
- M9 Ethics for Data Science

TABLE OF CONTENTS

1 COURSE HANDOUT

2 STORY TELLING WITH DATA

- A good story grabs your attention and takes you on a journey, evoking an emotional response. In the middle of it, you find yourself not wanting to turn away or put it down. After finishing it—a day, a week, or even a month later—you could easily describe it to a friend.

- Story is a time-tested structure.
- Story has a clear beginning, middle, and end.
- This can be setup, conflict, and resolution.
- Stories start with balance. Then some event happens, that upsets the balance. The resulting struggle, conflict, and suspense are critical components of the story.

THE BEGINNING

- Introduce the plot, build the context for your audience.
 - The setting, main character, unresolved state of affairs, and desired outcome.
- 1 The setting: When and where does the story take place?
 - 2 The main character: Who is driving the action? (framed in terms of your audience)
 - 3 The imbalance: Why is it necessary, what has changed?
 - 4 The balance: What do you want to see happen?
 - 5 The solution: How will you bring about the changes?

- “what could be”
- The goal is to convince the audience of the need for action.
- Identify what motivates the audience.
- Consider framing your story and the need for action in terms of this.
- Use data to strengthen the story and integrate it to make sense.

- ❶ Further develop the situation or problem by covering relevant background.
- ❷ Incorporate external context or comparison points.
- ❸ Give examples that illustrate the issue.
- ❹ Include data that demonstrates the problem.
- ❺ Articulate what will happen if no action is taken or no change is made.
- ❻ Discuss potential options for addressing the problem.
- ❼ Illustrate the benefits of your recommended solution.
- ❽ Make it clear to your audience why they are in a unique position to make a decision or drive action.

THE END



- Call to action.
- The audience is totally clear on what you want them to do with the new understanding or knowledge that you've imparted to them.

THE NARRATIVE FLOW

- Your story must have an order to it.
- ① Are they a busy audience who will appreciate if you lead with what you want from them?
- ② Are they a new audience, with whom you need to establish credibility?
- ③ Do they care about your process or just want the answer?
- ④ Is it a collaborative process through which you need their input?
- ⑤ Are you asking them to make a decision or take an action?
- ⑥ How can you best convince them to act in the way you want them to?

IMPORTANCE OF CONTEXT

TYPES of ANALYSIS



IMPORTANCE OF CONTEXT



STORYTELLING



THREE
MINUTE
STORY



Knowing exactly what
you want to communicate
reduces reliance on
slides and data

**BIG
IDEA***

*from Nancy Duarte
(Resonate)



A SINGLE SENTENCE that...

the "So what?"
boiled down
further

- ① articulates your point of view
- ② conveys what is at stake
- ③ is one complete sentence

STORY BOARDing

UPFRONT PLANNING to CREATE STRUCTURE



STICKY NOTES help to ...

avoid
attachment
to work done
on computer

force
concise
articulation

easily
rearrange
the flow

①
BRAINSTORM

②
EDIT

③
GET FEEDBACK

STORYBOARDING



Storytelling with Data by Cole Nussbaumer Knaflic

STORYBOARDING



Storytelling with Data by Cole Nussbaumer Knaflic

CHOOSING AN EFFECTIVE VISUAL

**SIMPLE
TEXT**

91%

*Just because you
have numbers doesn't
mean you need a graph!*

TABLE

*What is the
main point
I want to make?*

*OFTEN THERE
ARE MORE
EFFICIENT WAYS*

	A	B	C
CATEGORY 1	15%	22%	47%
CATEGORY 2	40%	36%	20%
CATEGORY 3	35%	11%	39%
CATEGORY 4	30%	27%	58%

*Avoid using tables in
live presentations because people
stop listening & start reading*

CHOOSING AN EFFECTIVE VISUAL

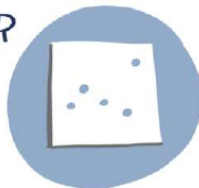
HEAT MAP

	A	B	C
CATEGORY 1	15%	22%	40%
CATEGORY 2	40%	36%	20%
CATEGORY 3	35%	17%	39%
CATEGORY 4	30%	29%	50%

EYES CAN EASILY PICK OUT BIG DIFFERENCES IN COLOR INTENSITY, but smaller ones don't stand out

Can work well when beginning to explore data and deciding where to dig further

SCATTER PLOT



Good for encoding data simultaneously on two axes to identify what relationships exist

CHOOSING AN EFFECTIVE VISUAL

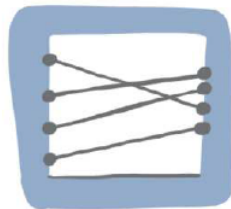
LINE



Rule: The lines that connect the dots have to make sense! Most effective with continuous data, often time

SLOPE GRAPH

↑
A FANCY WORD FOR A LINE GRAPH WITH ONLY 2 POINTS



Useful to focus on change between two points in time or difference between groups

CHOOSING AN EFFECTIVE VISUAL

BAR CHARTS

Great for categorical data

Easy for our eyes - comparing heights to a consistent baseline

Rule:
Must have a zero baseline.
No exceptions!

VERTICAL

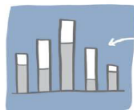


HORIZONTAL

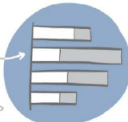


Good when category names are long

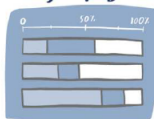
STACKED



OFTEN MISUSED...
EASIER TO COMPARE TOTAL & FIRST SERIES,
BUT SEGMENTS UP
THE STACK DON'T LINE UP

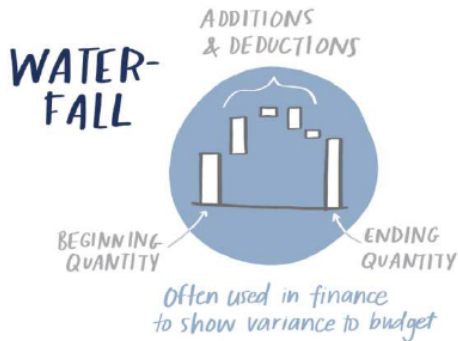


100% STACKED

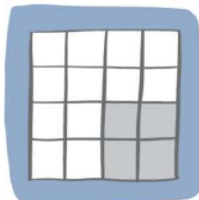


TWO BASELINES
FOR COMPARISON

CHOOSING AN EFFECTIVE VISUAL



SQUARE AREA
(AKA WAFFLE CHART)



THE GRID IS IMPORTANT BECAUSE WE TEND TO OVERESTIMATE AREAS

Good for showing numbers of very different magnitudes, or as an alternative to a pie chart

IMPROVING THE TABLE

New client tier share

Tier	# of Accounts	% Accounts	Revenue (\$M)	% Revenue
A	77	7.08%	\$4.68	25%
A+	19	1.75%	\$3.93	21%
B	338	31.07%	\$5.98	32%
C	425	39.06%	\$2.81	15%
D	24	2.21%	\$0.37	2%






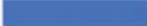






IMPROVING THE TABLE: OPTION

New client tier share

TIER	ACCOUNTS		REVENUE	
	#	% OF TOT	\$M	% OF TOT
A+	19	2%	\$3.9	21%
A	77	7%	\$4.7	25%
B	338	31%	\$6.0	32%
C	425	39%	\$2.8	15%
D	24	2%	\$0.4	2%
All other	205	19%	\$0.9	5%
TOTAL	1,088	100%	\$18.7	100%

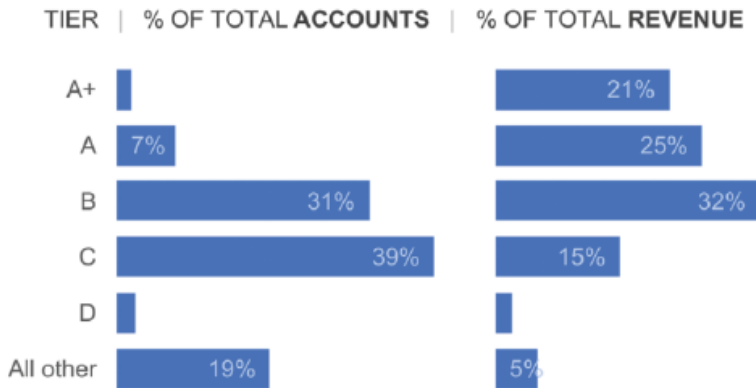
IMPROVING THE TABLE: OPTION

New client tier share

TIER	ACCOUNTS		REVENUE	
	#	% OF TOT	\$M	% OF TOT
A+	19		\$3.9	
A	77		\$4.7	
B	338		\$6.0	
C	425		\$2.8	
D	24		\$0.4	
All other	205		\$0.9	
TOTAL	1,088	100%	\$18.7	100%

IMPROVING THE TABLE: OPTION

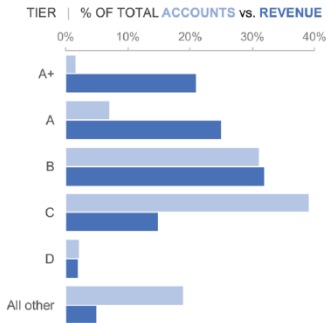
New client tier share



Storytelling with Data by Cole Nussbaumer Knaflic

IMPROVING THE TABLE: OPTION

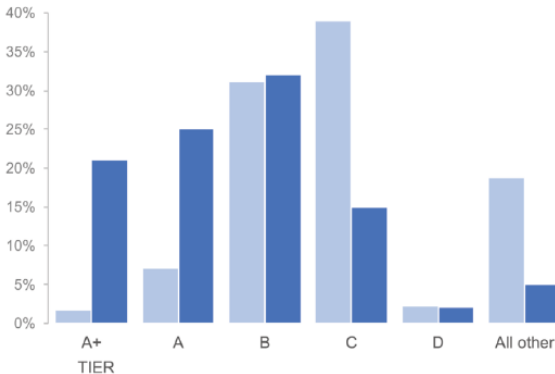
New client tier share



IMPROVING THE TABLE: OPTION

New client tier share

% OF TOTAL ACCOUNTS vs. REVENUE

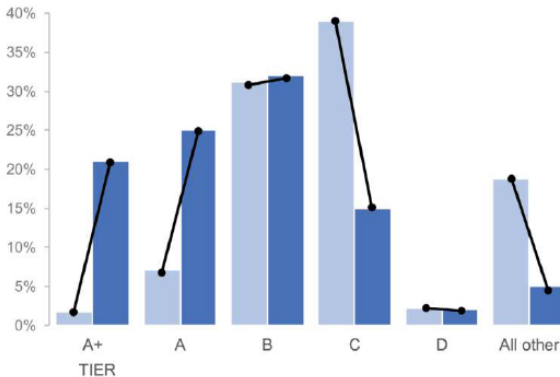


Storytelling with Data by Cole Nussbaumer Knaflic

IMPROVING THE TABLE: OPTION

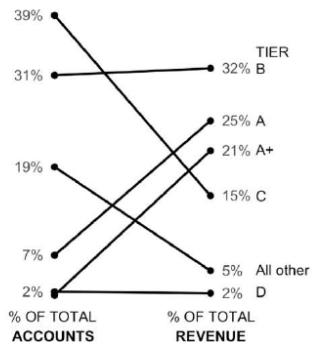
New client tier share

% OF TOTAL ACCOUNTS vs. REVENUE



IMPROVING THE TABLE: OPTION

New client tier share



Meals served over time

Campaign Year	Meals Served
2010	40,139
2011	127,020
2012	168,193
2013	153,115
2014	202,102
2015	232,897
2016	277,912
2017	205,350
2018	233,389
2019	232,797

VISUALIZE: OPTIONS

Meals served over time

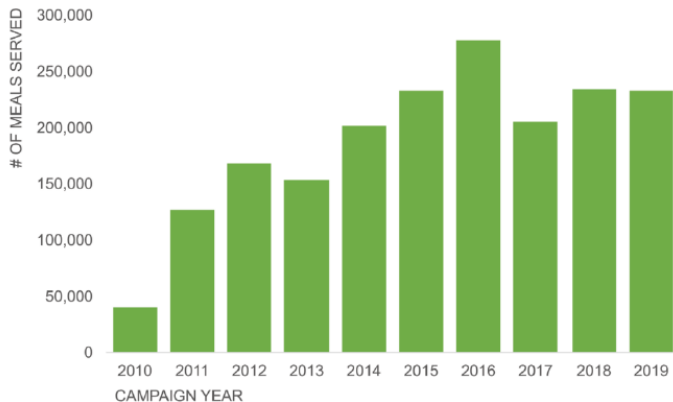
Campaign Year	Meals Served
2010	40,139
2011	127,020
2012	168,193
2013	153,115
2014	202,102
2015	232,897
2016	277,912
2017	205,350
2018	233,389
2019	232,797

Storytelling with Data by Cole Nussbaumer Knaflic

VISUALIZE: OPTIONS



Meals served over time



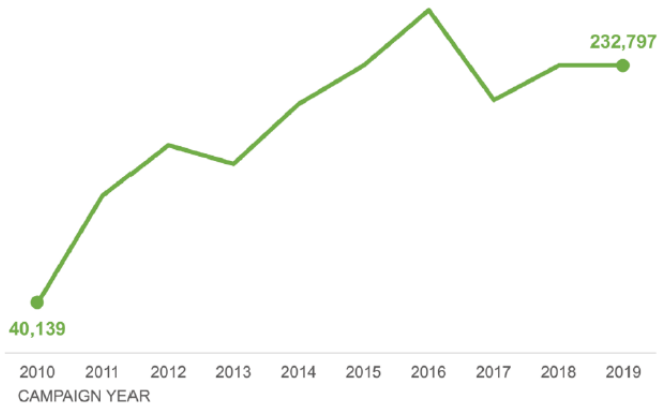
Storytelling with Data by Cole Nussbaumer Knaflic

VISUALIZE: OPTIONS



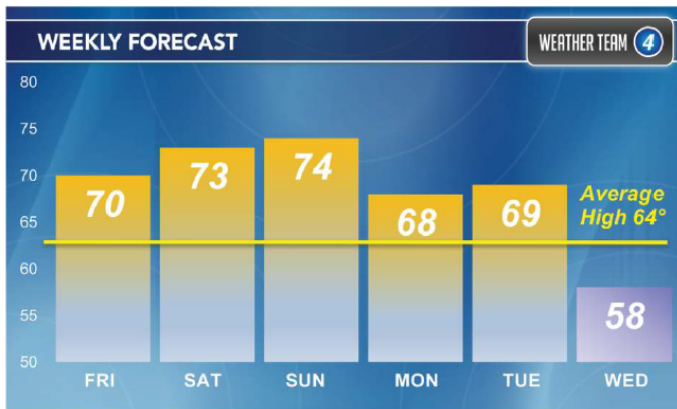
Meals served over time

OF MEALS SERVED



Storytelling with Data by Cole Nussbaumer Knaflic

CORRECT THE VISUALIZATION



Storytelling with Data by Cole Nussbaumer Knaflic

THANK YOU