Visual Exploration of Data
The Missing Element in CER Development

17-19 May 2022

Prepared for the 2022 ICEAA Workshop
All data shown in this presentation is notional

Benjamin Kwok
SSC/AC FMCR

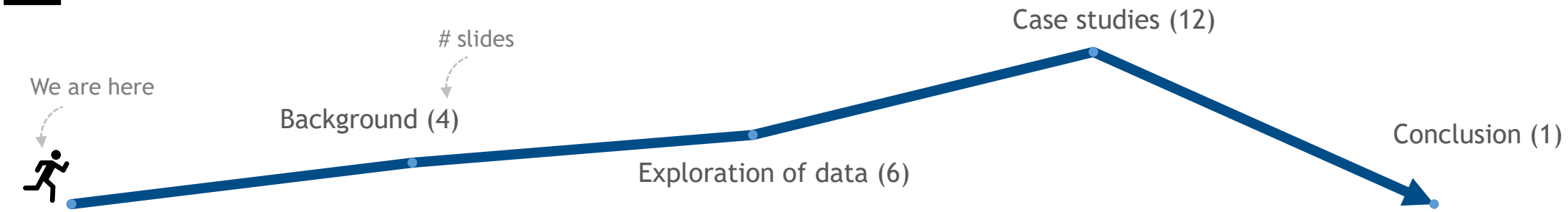
Abstract

Within cost estimating training literature, it is common for the discussion around Cost Estimating relationship (CER) development to focus primarily on its statistical parameters (e.g. correlation, equation form, etc.). An underemphasized component of CER development is the need to first visualize and explore data. This presentation will show how integrating these processes into CER development leads to faster and better results.

Presented at the 2022 ICEAA Professional Development & Training Workshop: www.iceaaonline.com/pit2022

Introduction

Agenda



Presentation goals

- Share visualization-centric lessons learned relating to CER development through the prism of the Unmanned Space Vehicle Cost Model (USCM).
- Inspire the community to incorporate more visualization into analytical processes and try new things.

Presented at the 2022 ICEAA Professional Development & Training Workshop: www.iceaaonline.com/pit2022

Unmanned Space Vehicle Cost Model (USCM)

USCM components

Robust **data collection** and **normalization** process

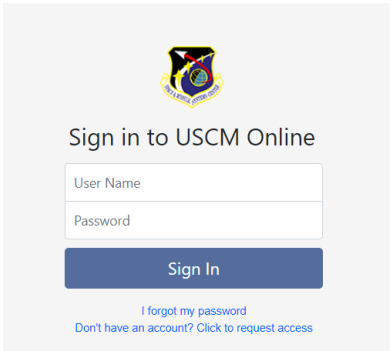
Satellite Database

CERs

Standard | Custom

Independent Research

Access



About

The **Unmanned Space Vehicle Cost Model (USCM)** is a Space Systems Command (SSC) product used to enable the estimation of unmanned, earth-orbiting satellites. The first publication was in November 1969 and has evolved significantly since then.

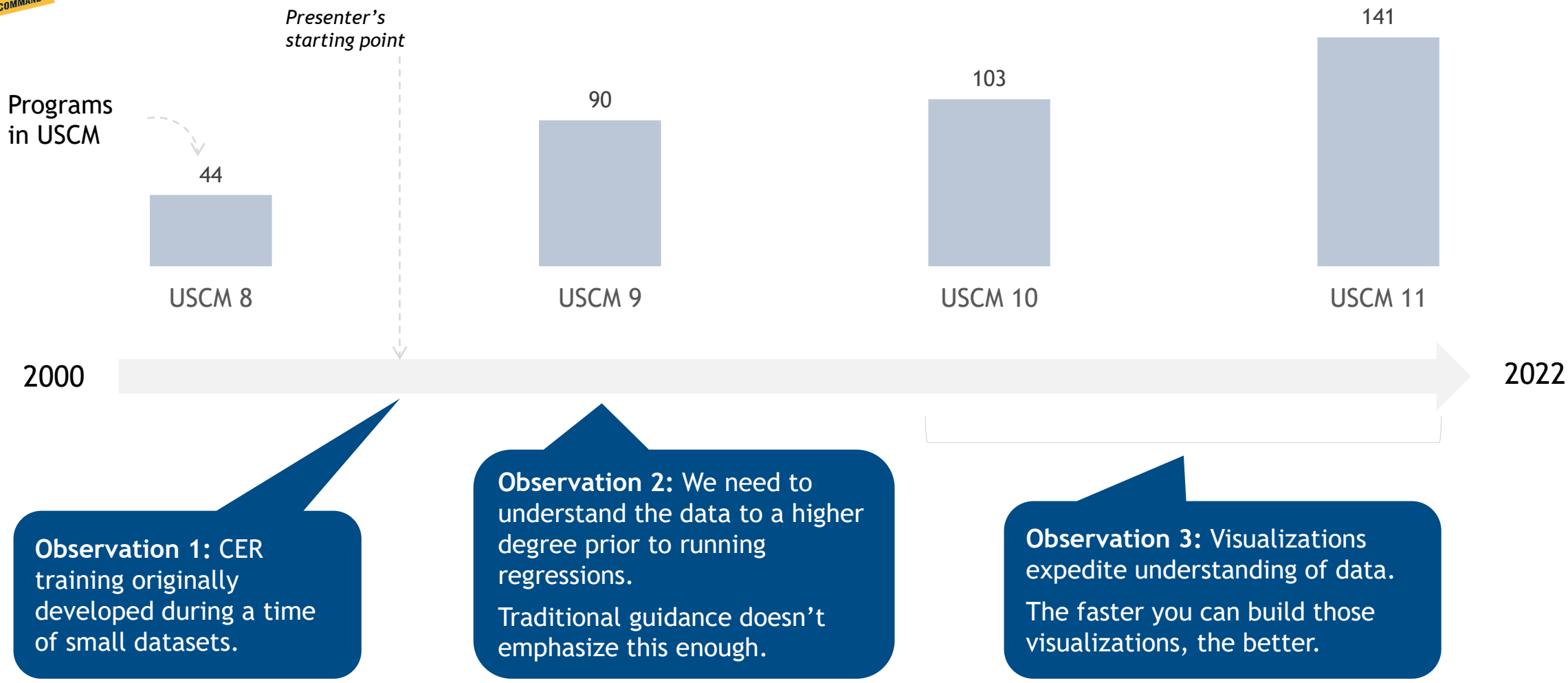
Key Facts

- SSC satellite cost estimating product
- Suite of products
- End-of-program costs
- Standard CERs developed at given points in time / Database updated quarterly

Use Cases

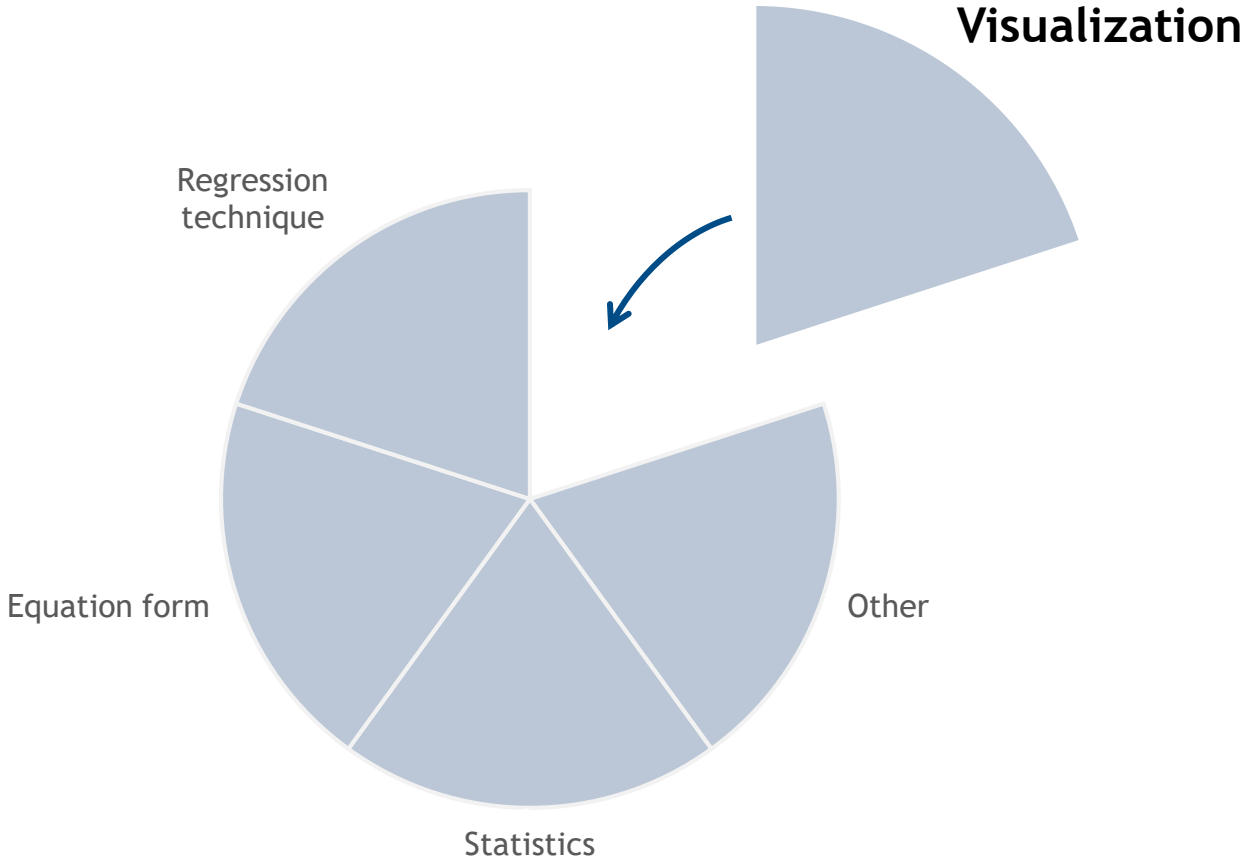
- Source selection
- Setting budgets
- Crosschecks
- Education
- Glean insights regarding SMC satellite costs

USCM CER development observations *through the years*



Perspectives through the lens of the Unmanned Space Vehicle Cost Model (USCM)

Perspectives on CER training material



Elements of CER training

Focus areas of traditional CER training are good but not complete.

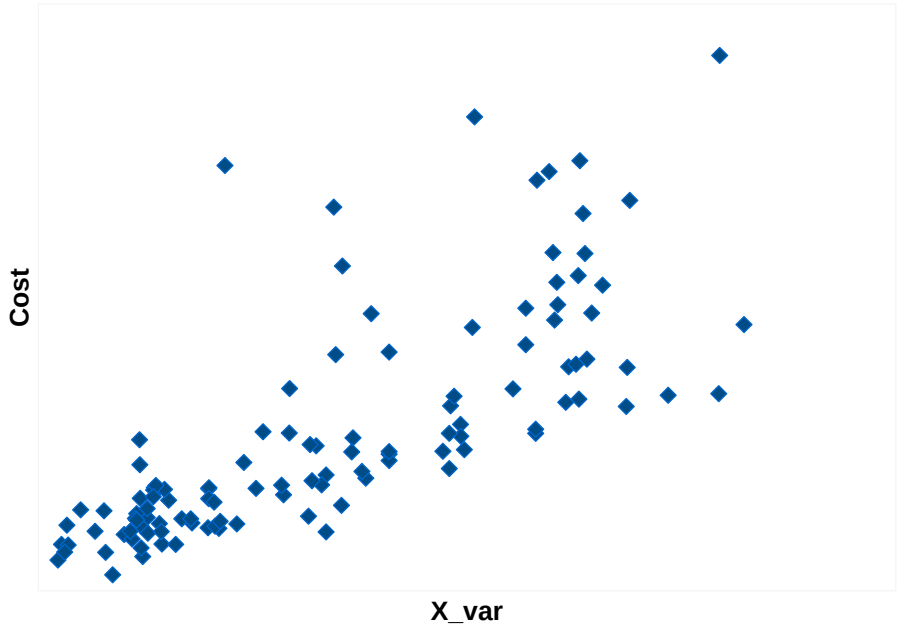
Why wasn't visualization emphasized as much as other topics?

- ❑ It's too obvious and unnecessary.
- ❑ The world began with small data sets and the idea of **exploring data** probably didn't make sense.

Presented at the 2022 ICEAA Professional Development & Training Workshop: www.iceaaonline.com/pit2022

An example of the problem

Notional Data Set (n = 115)



CER	Adj R^2 (MUPE)	SPE
Linear	.41	60%
Log-linear	.55	52%
USCM11	.81	39%

2 CERs generated with the real data set produce mediocre results. By comparison, the equivalent USCM11 CER is much better.

Grab data → Run diff. eq. forms → Success?

Exploratory Data Analysis (EDA)

An approach or philosophy for data analysis, not a structured set of techniques

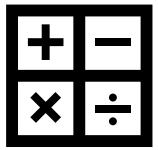
Heavy reliance on statistical graphs to…

- maximize insight into a data set
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- …

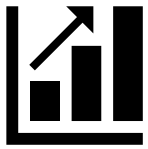
The main role of EDA is to open-mindedly explore.

The power of human perception + statistics

Our brains are excellent at finding patterns… especially if those patterns are presented visually



Statistics are really good at precise quantification

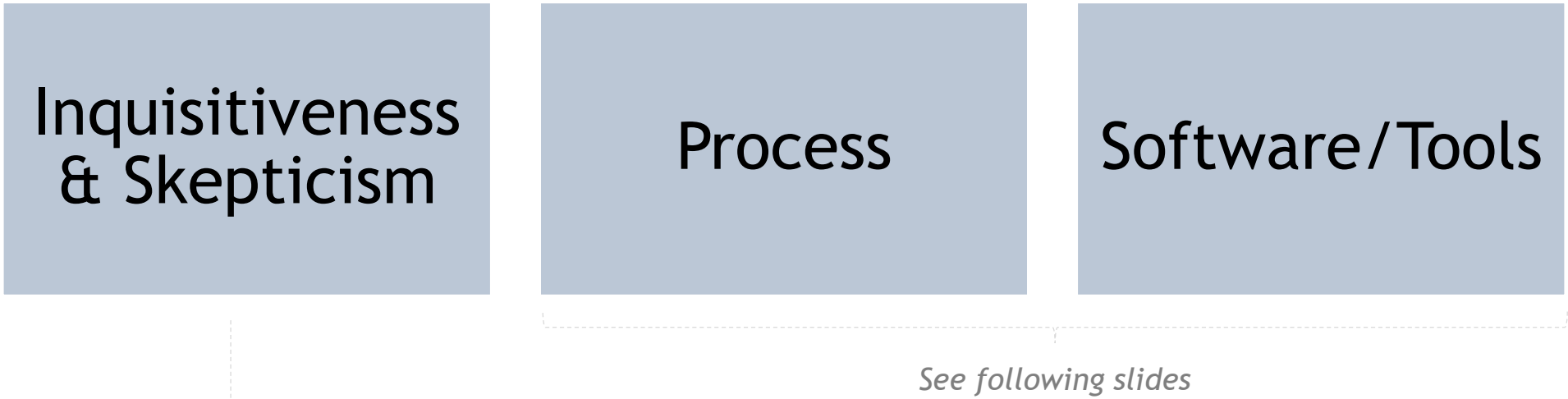


Using data visualization to combine these strengths can provide powerful insights about our data

- Bar charts + average lines
- Box plots
- Regression lines + confidence intervals

Requirements

Main components that enable you to visually explore your data.



Guidance

- Research and learn.
- Utilize subject matter experts.
- Never have 100% trust in the data or the stats.

Process: getting started with open minded exploration

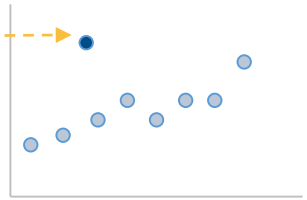
1) Start with the big picture objectives

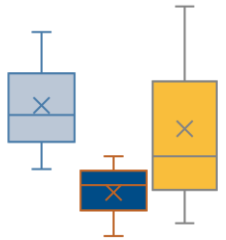
- Data content
- Goodness or quality of data
- Relationships, patterns…
- Group identification

2) Come up with questions

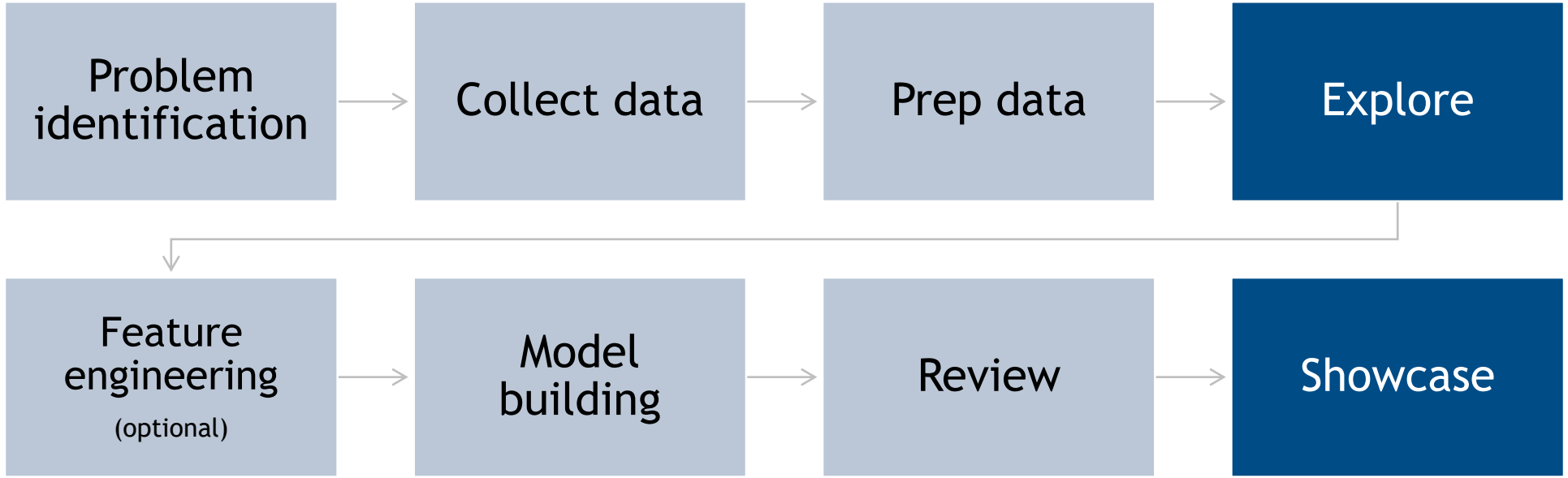
- What kind of data is in my data set? (e.g. numerical? Categorical?)
- Does the data need tidying?
- Are there holes in the data?
- How are the values distributed?
- Who are the outliers?
- What is the relationship between variables?
- Which data points are the most/least ___?
- What undefined groups belong in this data set?
- How do the groups within the data set behave?
- Is there a bias with respect to another variable?

3) Plot the data





A revised CER development process



Data exploration can often uncover new variables.

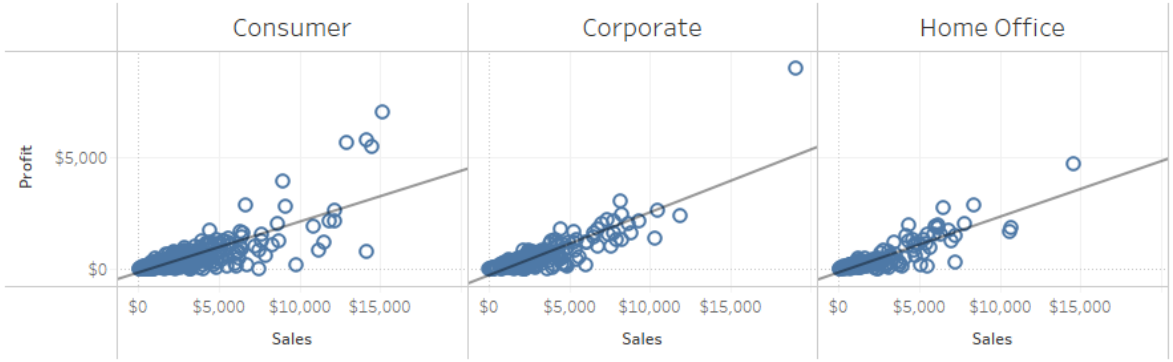
Visuals matter to the users as well.

Software tools for exploration

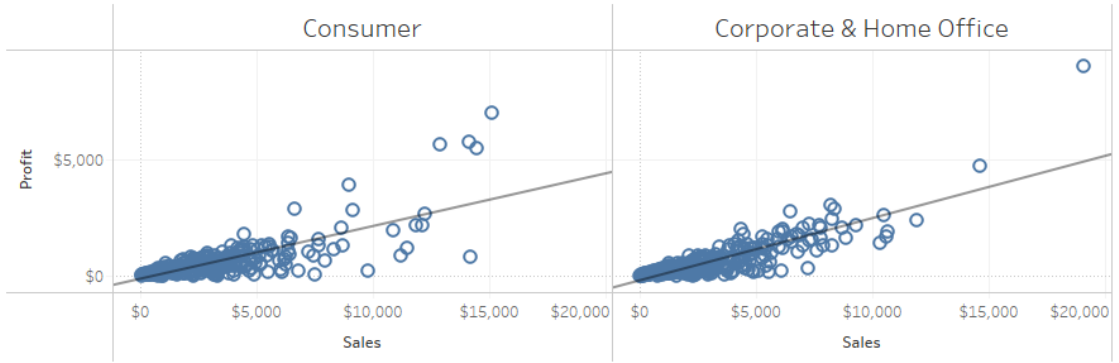
Guidance

- Opinions (of diff. tools) are a mixed bag because it really depends on use case.
- Make sure the software can generate the charts and features you need.
- Use something that will keep pace with the questions you want to ask.

"Small multiples" chart example



Merging charts with a click



Data shown is notional

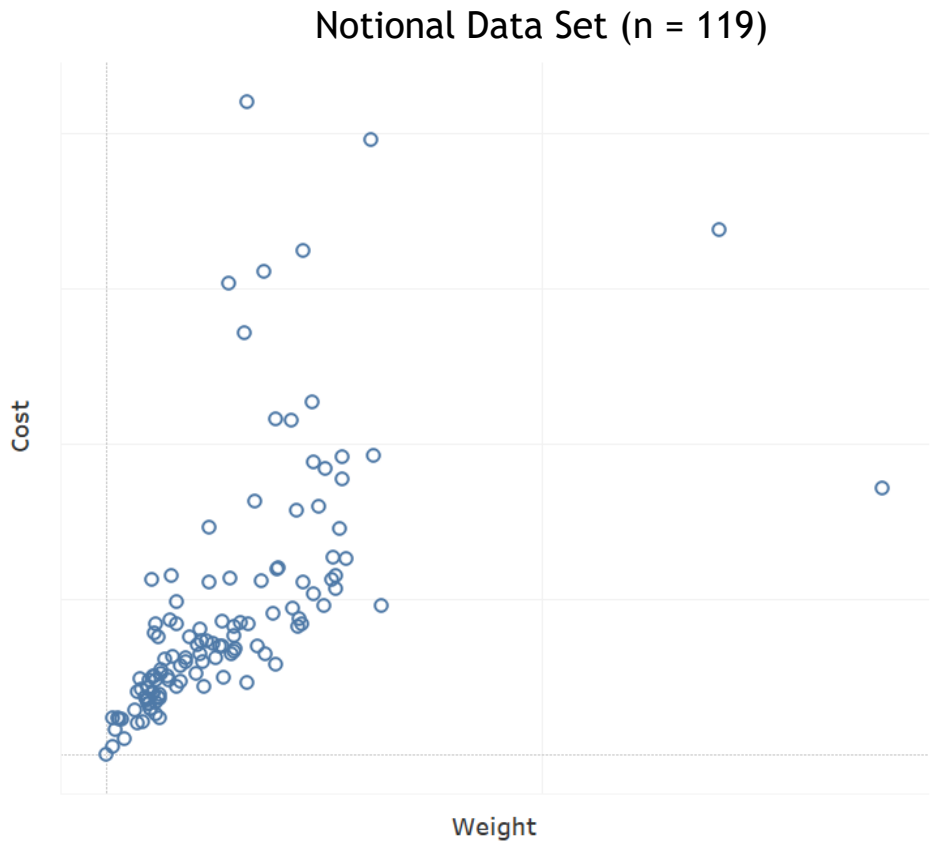
Case study 1: the common CER

Scenario

- You need to develop a CER for a satellite Work Breakdown Structure (WBS) element.
- The data set is new to you.
- You first explore the data.

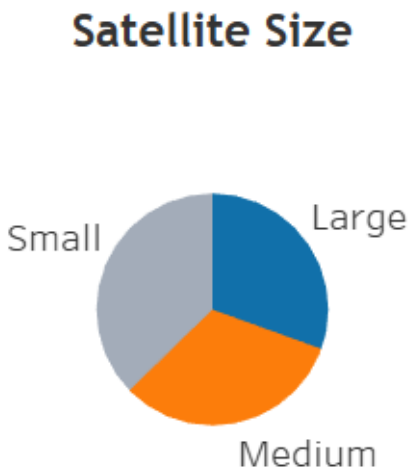
Key points

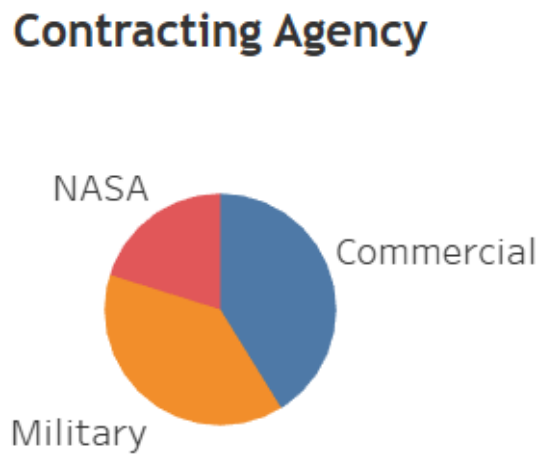
- Exploring data isn't a structured, linear process.
- It's often messy and full of rabbit holes.



Group identification -> what's in the data?

Make visualizations



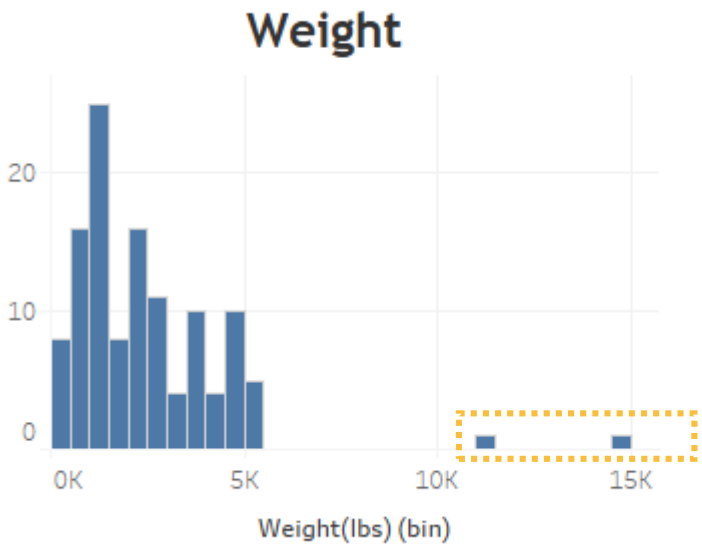


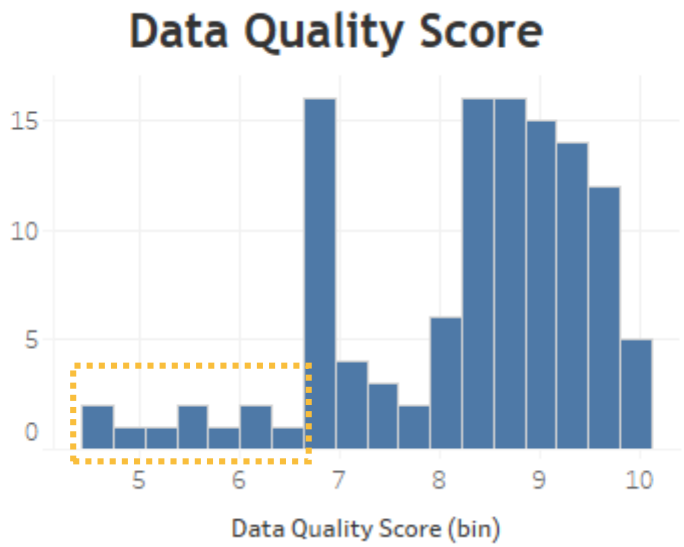
Ask more questions

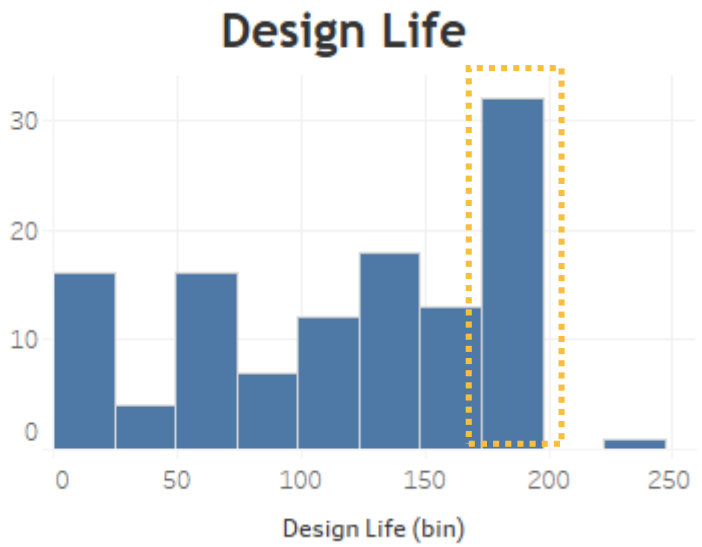
- How does a given group behave?
- Do I have the right data represented?
- Will group X bias the results higher/lower? (Bring in your expertise)

Goodness of data -> who are the outliers?

And are the outliers obvious?







Probable outliers

Sometimes you can just see them.

Potential outliers

In USCM, we derive a quality score to expedite outlier analysis.

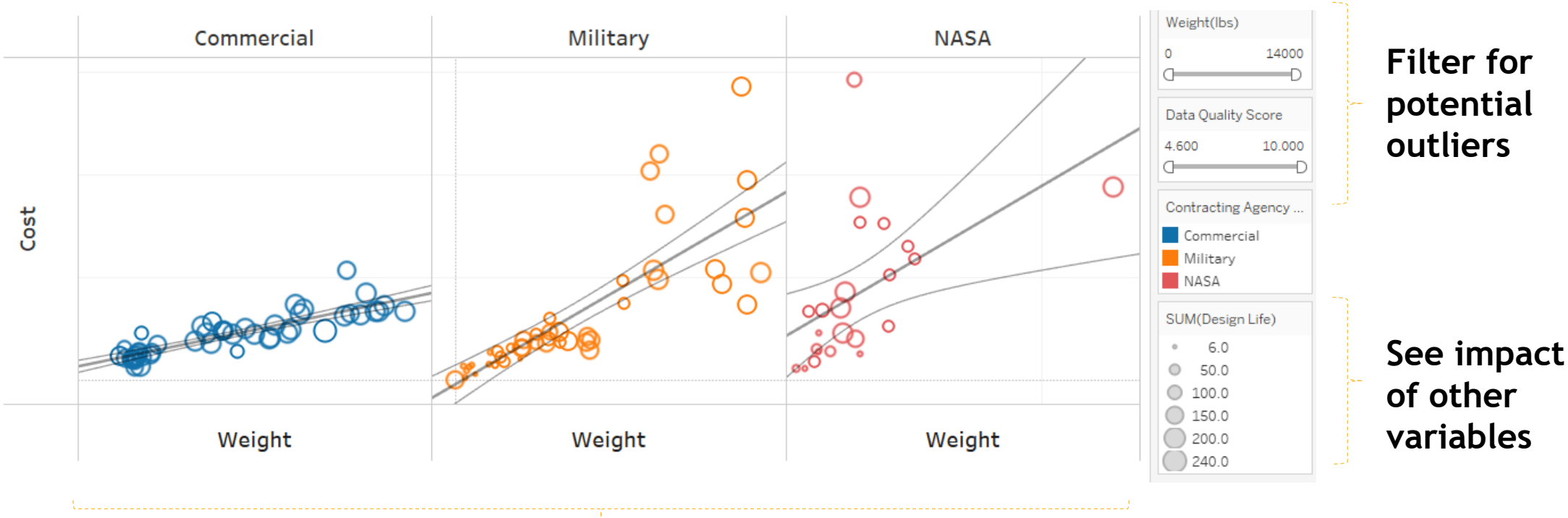
A new path to explore

Sometimes when you look for outliers you find something else.

Data shown is notional

Putting it together

View relationships by pre-identified groups



Filter for potential outliers

See impact of other variables

Small multiples chart

Data set used is notional.

Case Study 2: Automating your visuals

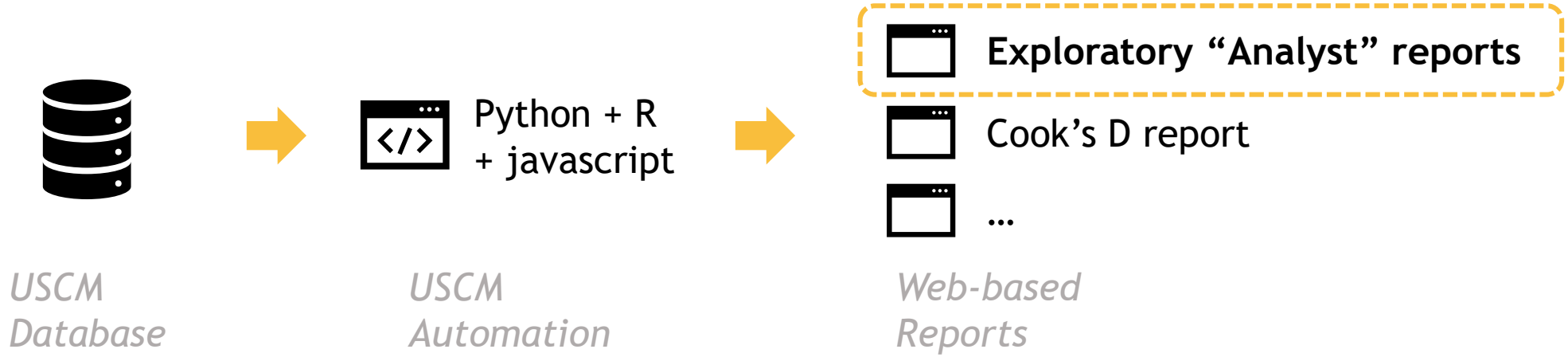
Scenario

- You are tasked with needing to update a significant number of existing CERs.
- Use automation to enable your exploration.

Key points

- Scripting + data pipelines = huge potential
- Build upon your lessons learned – if you've done this before, don't start from scratch

The original motivation of USCM Automation



- Development of USCM11 involved the use of automating the generation of web-based interactive reports.
- Published USCM11 CER reporting displays this capability but we did not start there… it started with **automating exploratory analysis**.

Automated Exploration

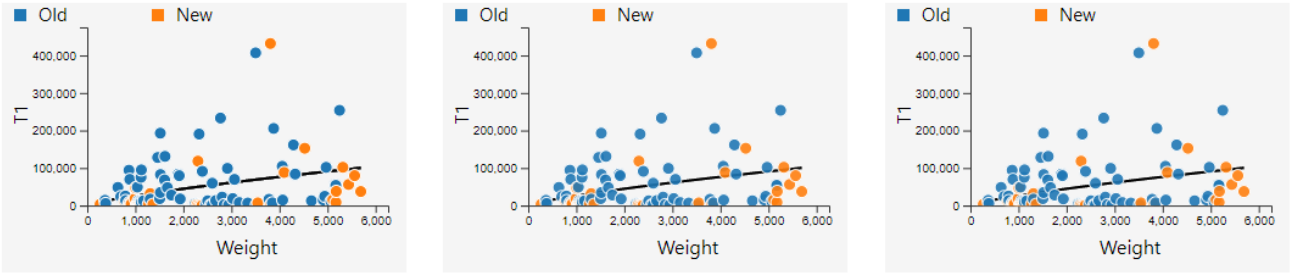
Outliers | New/Old | Agency | Small Sat | New/Modified

Dynamic legend categories

MUPE Regression Results:

Can zoom, move, and delete points. Table can be sorted.

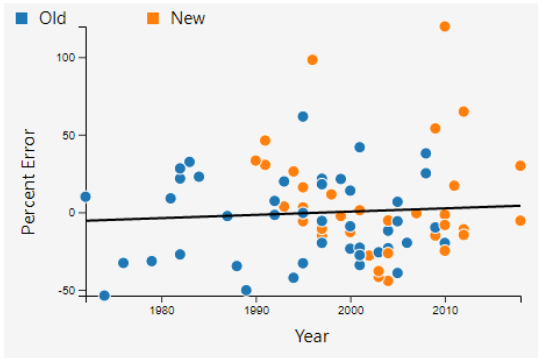
Hover over the data points to see program details.



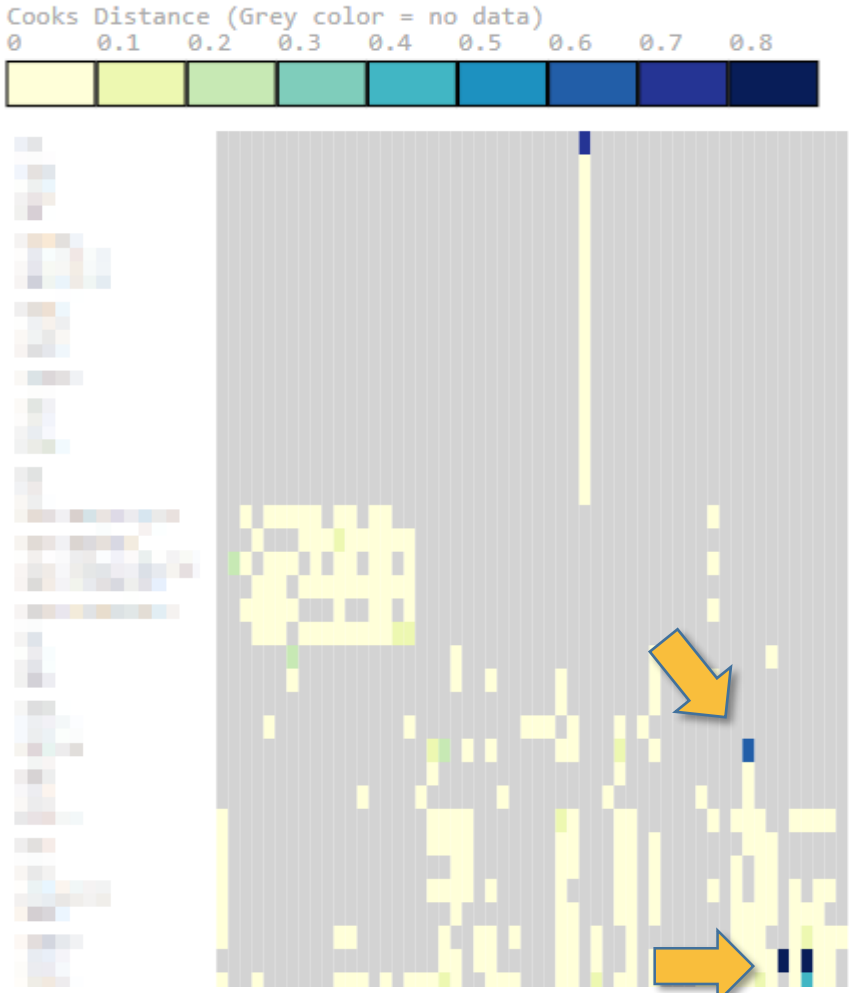
Submit | 1985

Time trend analysis

Space Vehicle T1



Cook's D heat map



Data shown is notional

Useful visual exploration features

- Zoom
- Data point labels on hover over
- Adjustable legend categories
- Re-run regression real time
- Bias plots
- …

Case study 3: Showcasing your CER

Scenario

- You developed a CER for an estimate; it's been peer reviewed but will be new to decision makers.
- You need to provide some insight of the method utilized in an estimate review board.

Key points

- Goodness of your CER is not the only decision criteria.
- Visuals can help instill confidence quickly.

Considerations for developing CER reports

Medium

e.g. ppt, web, 1 pager Word

Audience background

e.g. education, focus areas, etc.

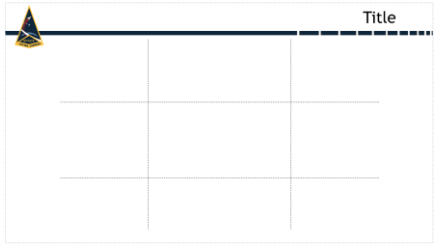
Duration for analyzing

e.g. glance for 10 secs, discuss for 10 mins?

- Size and scope of effort
- Level of detail

Process

1) Layout

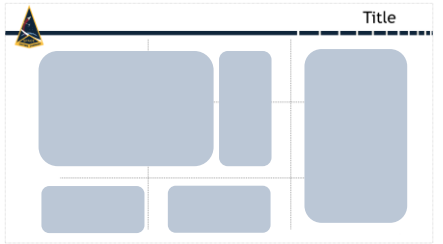


Tip: utilize a grid to evenly space your elements.

2) Determine key elements

- Scatter plot
- Equation
- Dataset visuals
- Documentation
- …

3)Arrangement



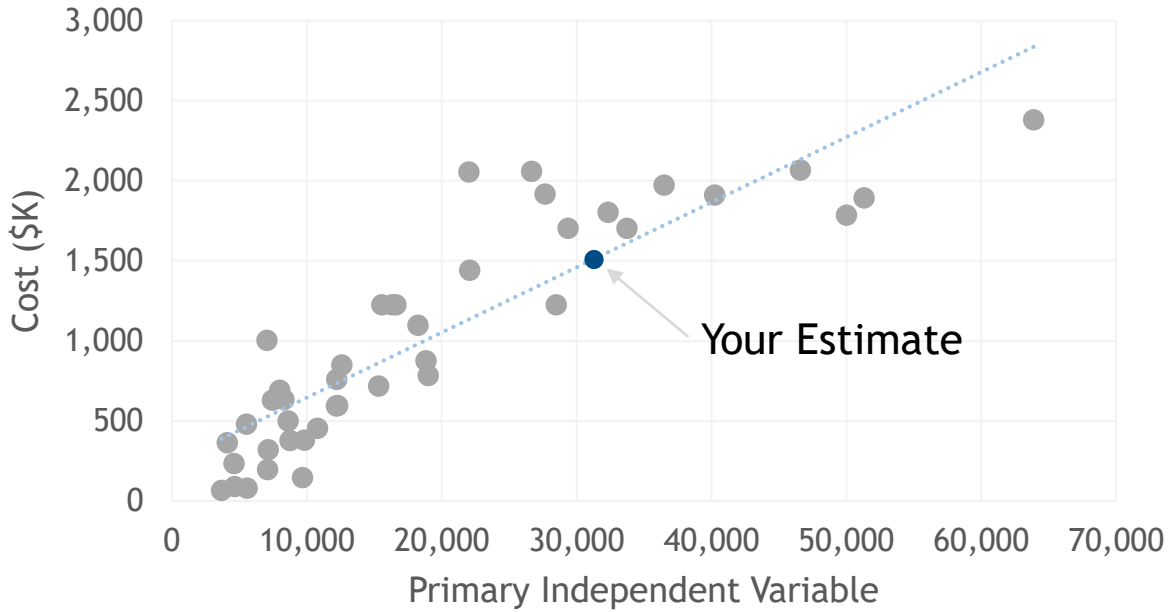
Tip: Use positioning and size to influence priority.

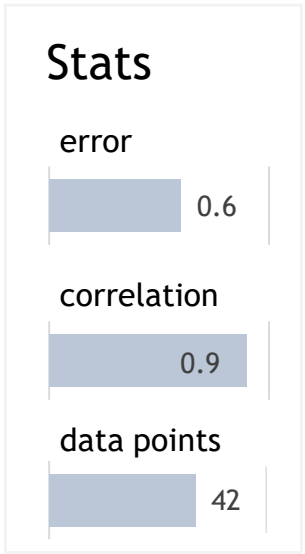
Tip: Group like elements.

4) Polish

- Annotate
- Remove what is unnecessary (entirely or to another section)
- Use color for (selective) emphasis

Notional CER example





CER Parameters

$Y = 22.37 * X$

Where,

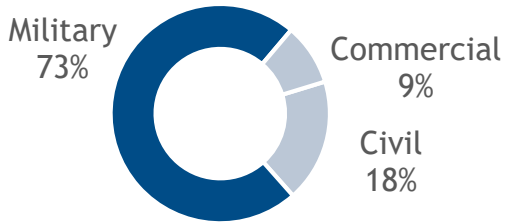
Y = Cost of ___ (FYXX, $K)

X = Variable name

Additional details

- Caveats
- Risks
- …

Data Set Composition



Data set more recent than older

Concluding Thoughts

Summary

- Visualization is a critical element in CER development… don't gloss over it.
- Visualization plays different roles throughout the process.

Resources and References

- USCM public website (www.uscmonline.com)
- USCM11 ICEAA 2021 presentation (https://www.iceaaonline.com/ready/wp-content/uploads/2021/06/MLD07-ppt-Kwok-USCM11-%E2%80%93-an-Evolution-of-Techniques.pdf)
- NIST Exploratory Data Analysis webpage (https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm)
- Feature Engineering Wikipedia (https://www.kdnuggets.com/2018/12/feature-engineering-explained.html)
- ICEAA 2019 Data Visualization Presentation (https://www.iceaaonline.com/ready/wp-content/uploads/2019/06/CV03-Data-Visualization-Kwok.pdf)

Acknowledgements

This work was funded by the Space Systems Command, contract FA8802-21-F-0004. The authors thank Ms. Adriana Contreras and Mr. Raj Palejwala for their support of this effort.