TRUECUE

truecue.com

In partnership with

alteryx IoD

Women + Data Hackathon 2021

How to approach a data project

Skills Session

15th November 2021 6-7pm GMT

Agenda

- Introduction & Technical Checks *18:00-18:05 (~5mins)*
- Data-led, Question-led & Beyond! *18:05-18:10 (~7mins)*
- Data-led Exploration *18:10-18:20 (~10mins)*
- Question-led Exploration *18:20-18:40 (~15mins)*
- Q&As & Close *18:40-18:55 (up to 15mins)*

Introducing ourselves, the two speakers

Erica Reuter, PhD
Manager Solution Engineers, US Public Sector
Global Security Tiger Team - Americas Leader

alteryx

Team Facilitator & Speaker!

&

Sophie Decelle
Consulting Manager
Data Science Lead

TRUECUE

Hackathon Organising Team

TRUECUE

The session will contain Alteryx demos

Did you get Alteryx and materials for this session?

Alteryx Designer

Alteryx Workflow
(packaged)

sustainability.yxzp

Datasets

1. sample data on sustainability,
2. sustainability indicators,
3. list of COP26 signatories

TRUECUE

Two seemingly different ways to approach a data and analytics project…

Data-led

The exploration of one or more datasets and their connections with no priori (this is not a formal definition!)

Typical questions:

- How much data is there?
- What's the format? The structure?
- If multiple datasets, can they be related?
- How much data is missing or looks incomplete?
- What can I infer from it?

Aim: Size, Qualify, Connect

vs

Question-led

A data investigation and analysis aiming to answer a specific question

Typical elements:

- What's my hypothesis to test?
- Is there relevant data to answer my question?
- Is there enough data?
- What kind of analysis do I need to perform to prove or disprove my hypothesis?

Aim: An answer (it can be inconclusive!)

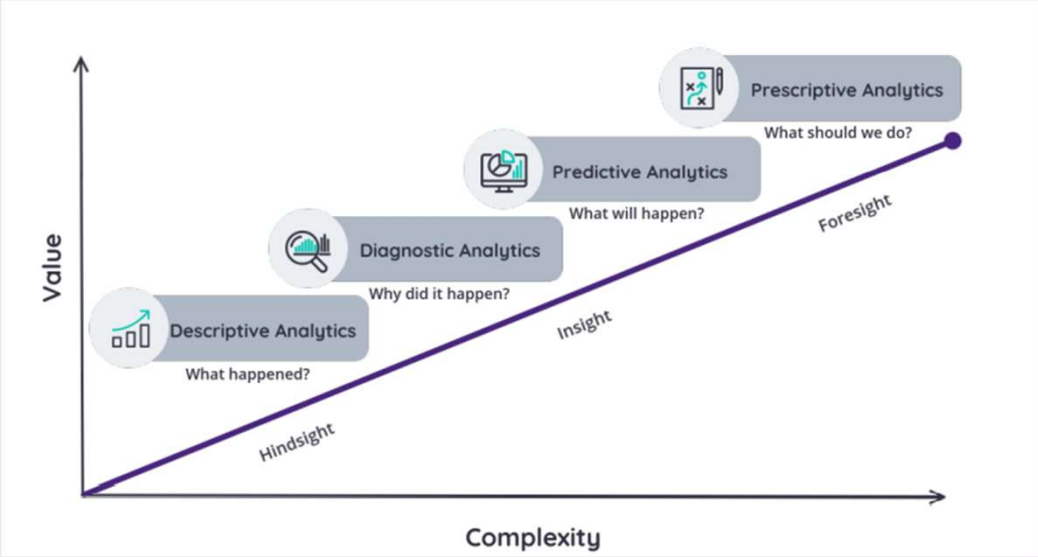
… are often combined!

TRUECUE

A data-led approach helps discovery whilst a question-led approach develops your analyses

- Discovering and describing the data is the first step to any analytics project and any data-driven strategy

- Questions shape your analysis of the data to focus on finding an answer…but not all questions are equal



A useful framework is the analytics maturity model

Note: There are “advanced” techniques which have “no a-priori” for example unsupervised learning…

TRUECUE

Let’s explore our datasets for today

1. Sustainability & country indicators – sample data

Similar to the hackathon main dataset (prepared)

2. Sustainability Indicators

Raw data, very large dataset, containing in particular education indicators (unprepared)

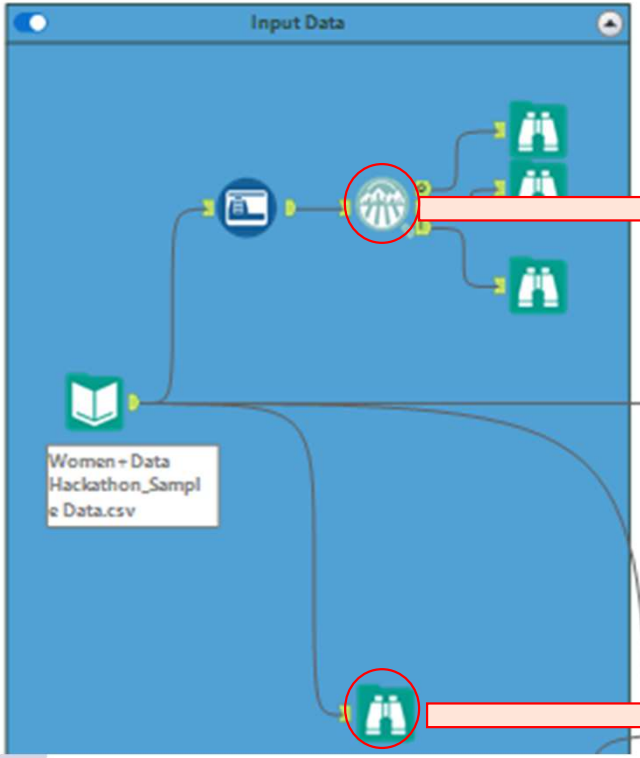
3. List of COP26 signatories

Additional data (prepared)

DEMO - Alteryx

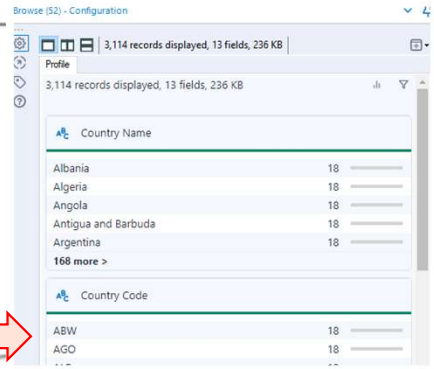
TRUECUE

Using Alteryx designer or a data preparation tool how can I bring the datasets together…



22 of 22 Fields | Cell Viewer | 13 records displayed, 147 KB | Search | Data | Metadata

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	% of Population using the Internet	Numeric	0	99.652849	25	29.386654	4.110469	2448	33.117667
2	Access to electricity (% of population) - EG.ELC.A...	Numeric	1.27018	100	98.695902	30.226875	2.344252	1660	79.480966
3	Gross national expenditure (% of GDP) - NE.DAB...	Numeric	40.110847	264.782066	102.842706	18.805056	7.19332	2891	104.615966
4	Inflation, consumer prices (annual %) - FP.CPI.TO...	Numeric	-18.10863	359.936614	3.466427	11.160364	5.908799	2930	5.504176
5	Life Expectancy (Years)	Numeric	40.369	84.934146	72.563415	9.205975	0.545922	2927	70.156799
6	Population	Numeric	69671	1392730000	8363404	141354616.937355	0.224791	3107	38902822.540714
7	Proportion of seats held by women in national p...	Numeric	0	63.75	16.666667	11.165726	6.13359	887	18.250062
8	Renewable electricity output (% of total electricit...	Numeric	0	100	17.987102	33.591342	16.570328	2132	32.337535
9	Renewable energy consumption (% of total final...	Numeric	0	98.342903	22.6563	29.838381	0	3016	32.609397
10	Women Business and the Law Index Score (scale...	Numeric	23.75	100	72.5	18.111931	2.312139	106	69.981303
11	Year	Numeric	2001	2018	2009.5	5.188961	0	18	2009.5
12	Country Code	String	[Null]	[Null]	[Null]	[Null]	0	173	[Null]
13	Country Name	String	[Null]	[Null]	[Null]	[Null]	0	173	[Null]

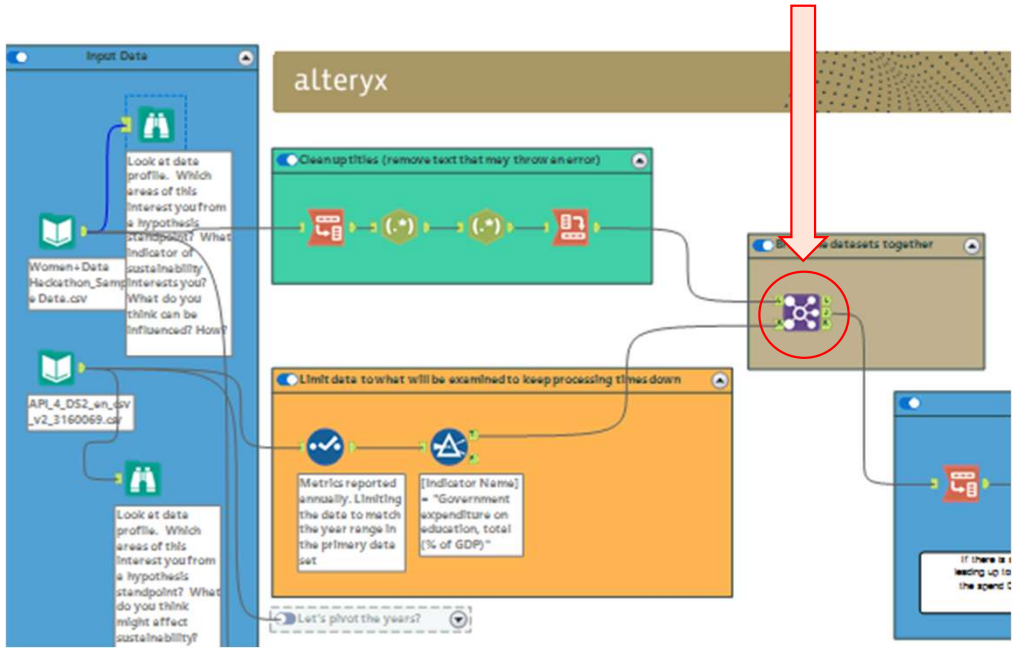


- How much data is there?
- What's the format? The structure?
- How much data is missing or looks incomplete?
- How are the values in each field distributed?

DEMO - Alteryx

TRUECUE

Using Alteryx designer or a data preparation tool how can I bring datasets 1 & 2 together…

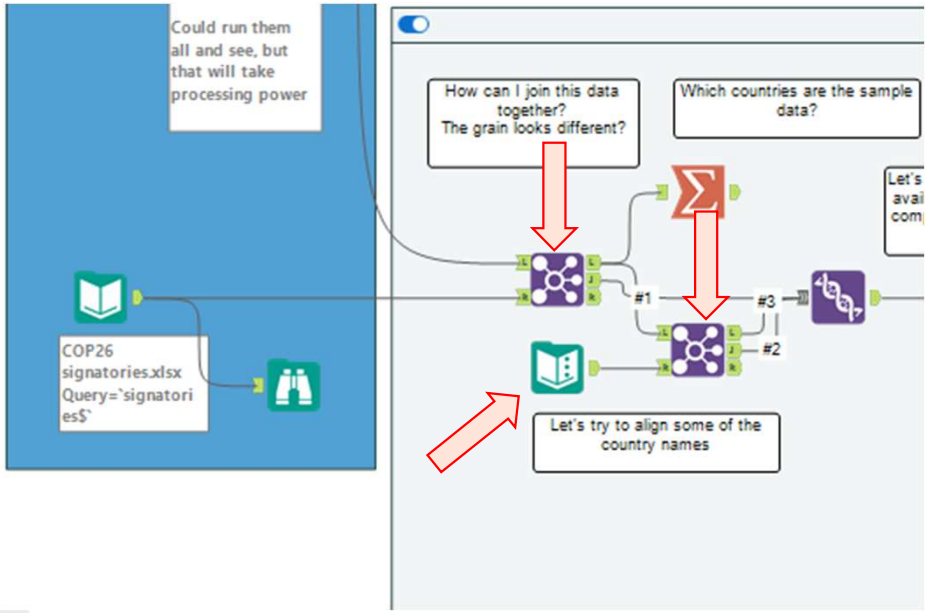


- How to bring the data together?
- Which “fields” or “columns” to use from either side?

DEMO - Alteryx

TRUECUE

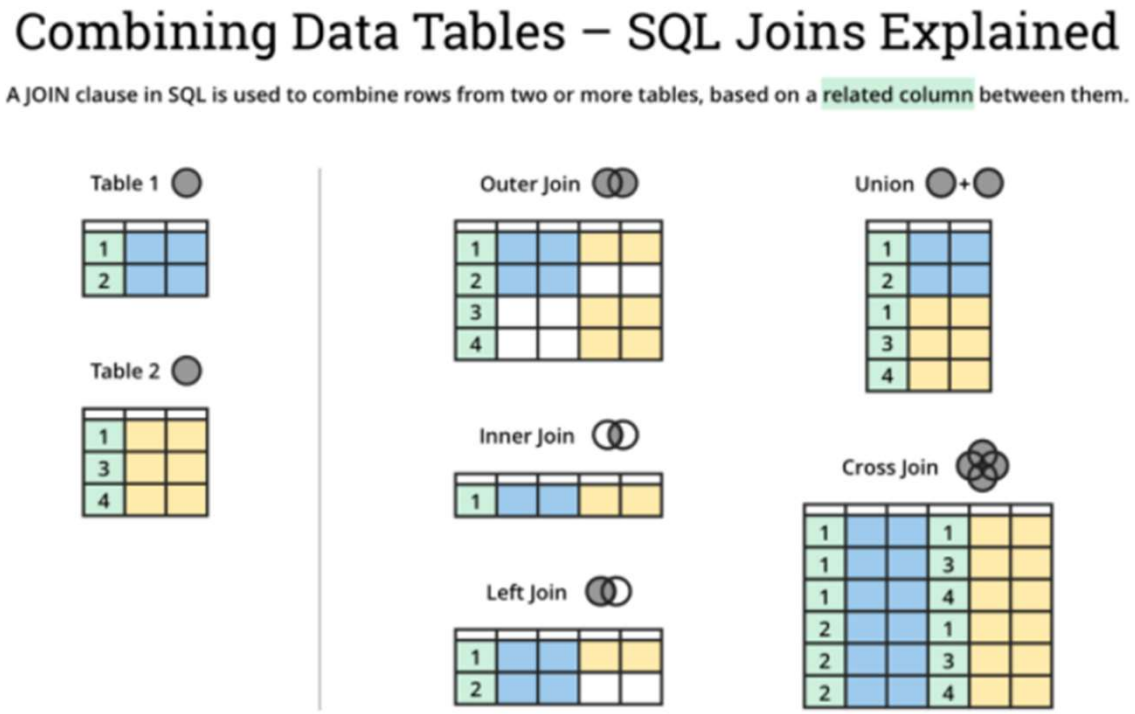
Using Alteryx designer or a data preparation tool how can I bring datasets 1 & 3 together…



Similarly…

- How to bring the data together?
- Which “fields” or “columns” to use from either side?
- When records don’t match – what does it mean? Do I try to map those not matching?

A quick note about joining datasets…



- How do the datasets relate to each other?
- Which fields or columns are in common throughout the different datasets?

Source: https://dataschool.com/how-to-teach-people-sql/sql-join-types-explained-visually/
Author: Tim Miller

TRUECUE

How the datasets are joined depends on the question you want to answer, here are some examples...

Let's look at two example datasets

1 **Cat Owners**

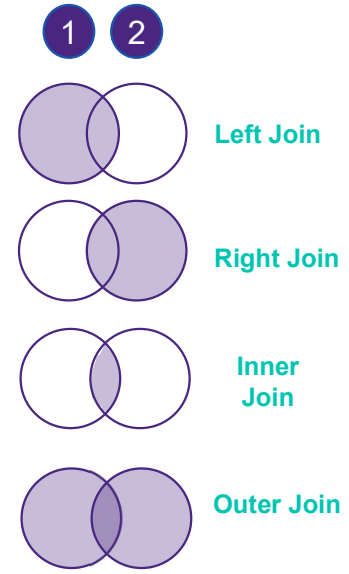
	#
Ahsley	2
Mike	3
Sarah	1
Aaron	3

2 **Dog Owners**

	#
Sarah	3
Aaron	1
Helen	5
Rebecca	6

Questions:

- How many dogs and cats, the Cat Owners have?
- How many dogs and cats, the Dog Owners have?
- How many cats and dogs, owners of both pets have?
- How many cats and dogs, do all pet Owners have?



Answering question 1, requires a left join, which results in the following dataset

How many dogs and cats, the Cat Owners have?

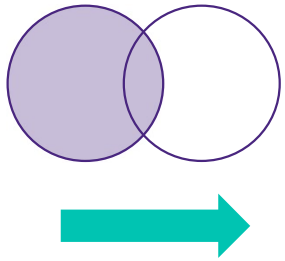
1

Cat Owners	
	#
Ahsley	2
Mike	3
Sarah	1
Aaron	3

2

Dog Owners	
	#
Sarah	3
Aaron	1
Helen	5
Rebecca	6

Left Join



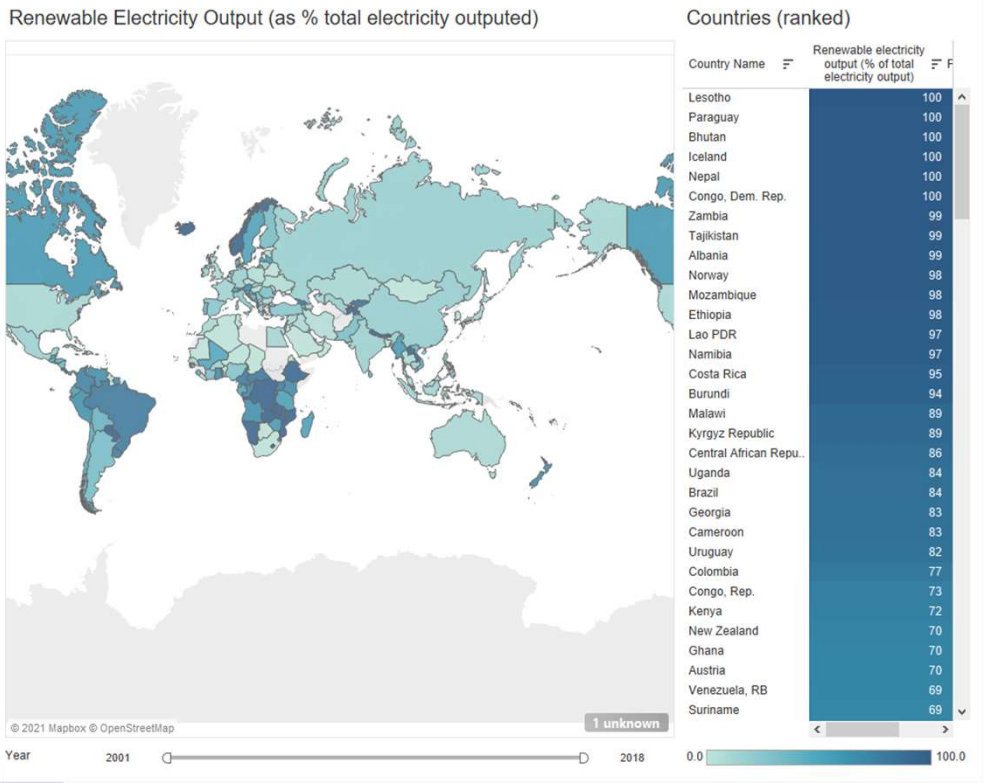
	Cat Owners	Dog Owners
Ahsley	2	
Mike	3	
Sarah	1	3
Aaron	3	1

Can you answer the other questions?

DEMO - Visualisation

TRUECUE

For data exploration, visualisation tools are extremely useful but have less out-of-the-box rigor



- Exploring the data visually, does it lead to more questions?
- Are there visual patterns to check?

TRUECUE

Now we know our data better, what questions could we answer?

1. Sustainability & country indicators – sample data
2. Sustainability Indicators
3. List of COP26 signatories

Is there a historical relationship between spend on education and some of the sustainability indicators?

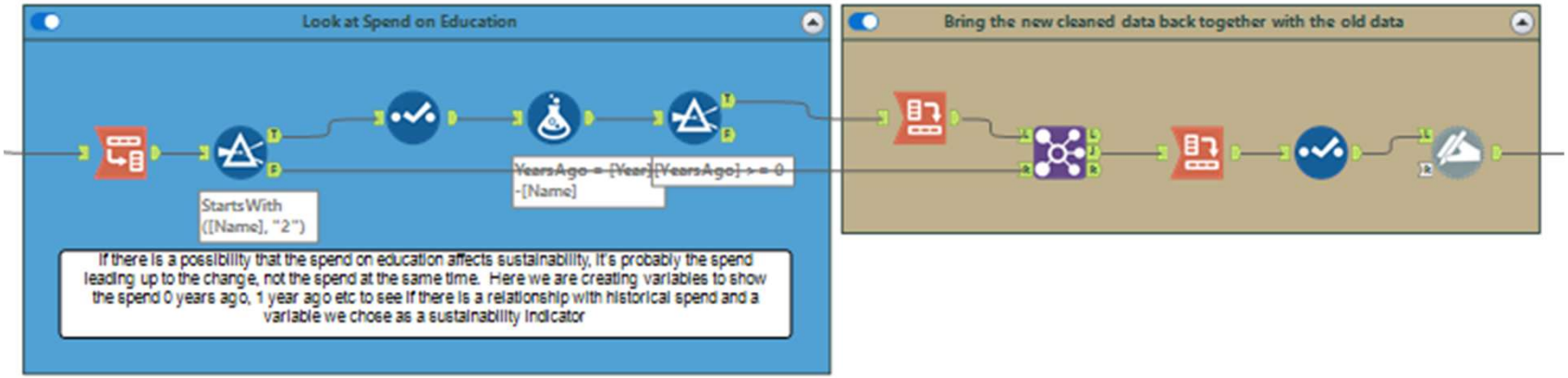
Which countries are more likely to have signed COP26?

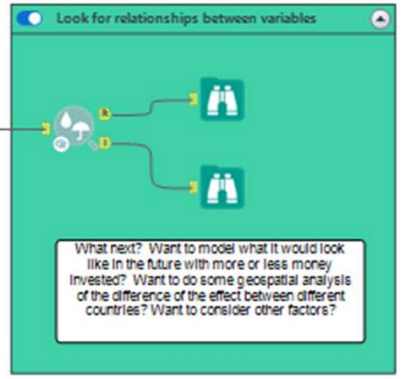
DEMO - Alteryx

TRUECUE

How to prepare my data and which methodology to explore the relationship between education spend and sustainability?

- How to **shape** the data to be able to test the relationship between spend on education and sustainability metrics?





- How to **test for significance** of relationships between spend on education and sustainability metrics?

What next?

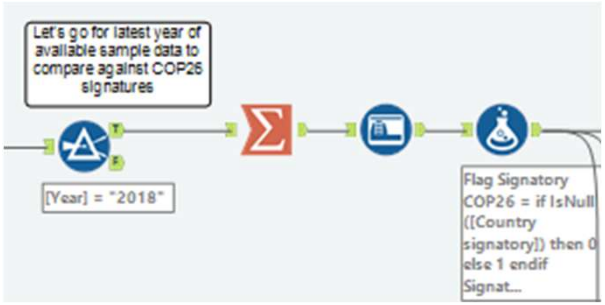
- Want to model what it would look like in the future with more or less money invested?
- Want to do some geospatial analysis of the difference of the effect between different countries?
- Want to consider other factors?

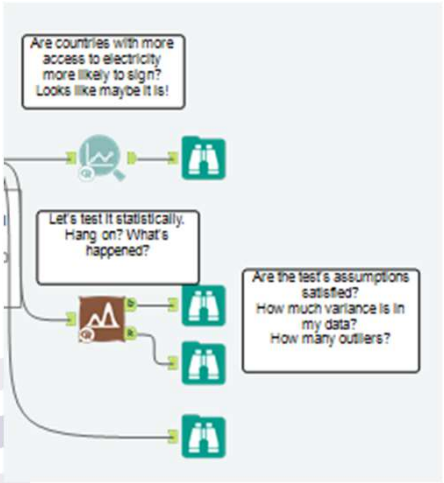
DEMO - Alteryx

TRUECUE

How to prepare my data and which methodology to explore which types of countries are more likely to sign COP26?

- **Which data to use** and how to **shape** the data to be able to identify differences or similarities between COP26 signatories?





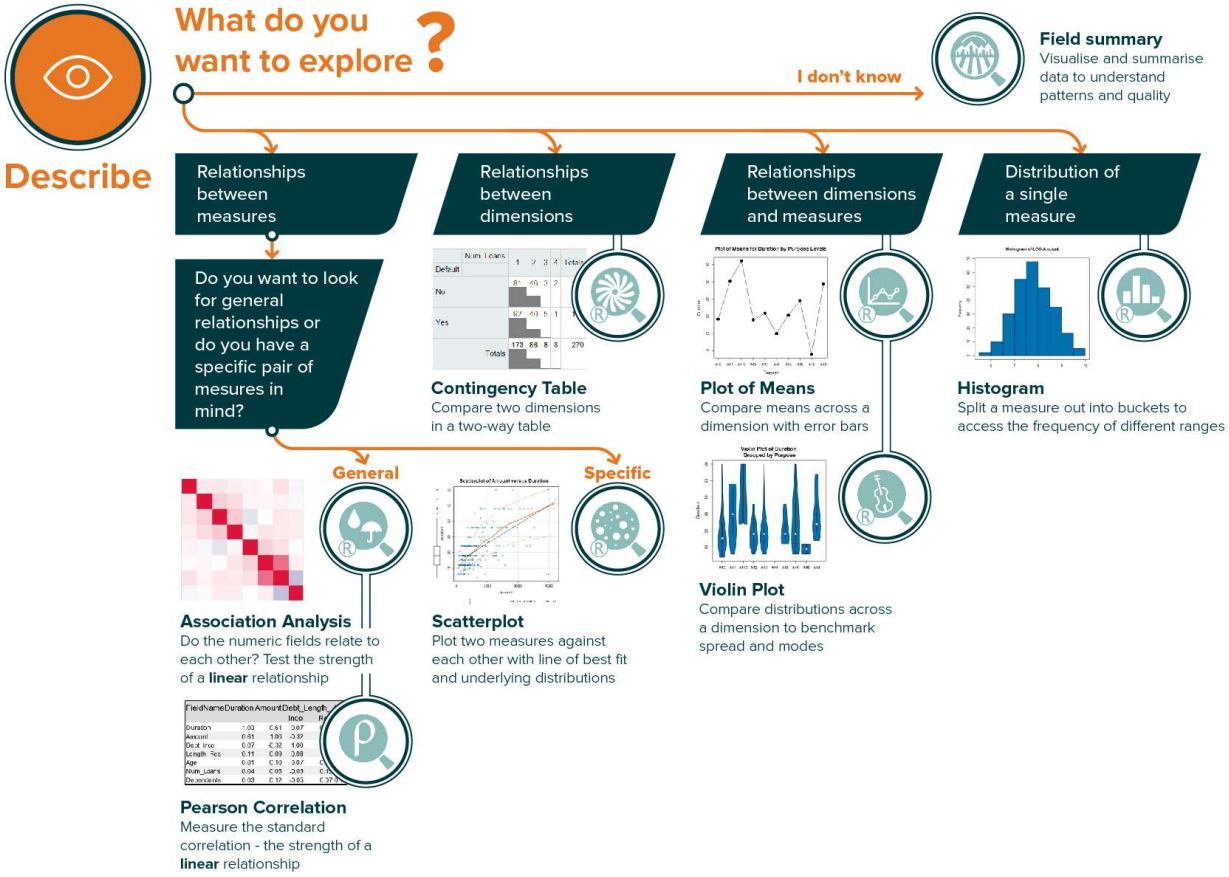
- How to **test** for **significant differences or similarities?**

What next?

- Want to dig deeper as to why some countries are more likely to sign COP26?
- Want to gather additional data to try and prove why some countries are more likely than others to sign?
- Want to model whether a country is likely to sign in the future?
- Want to cluster

TRUECUE

Alteryx descriptive analytics sheet cheat…from Alteryx enthusiasts



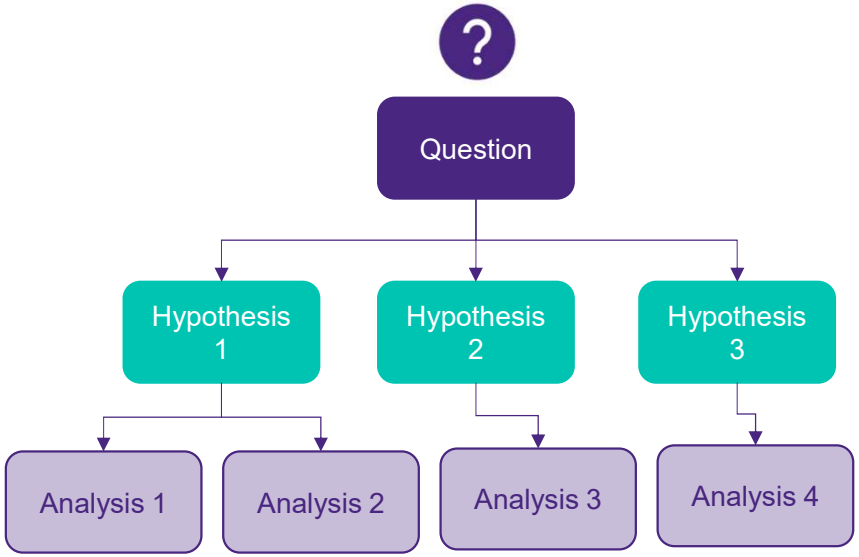
But also, check out the Alteryx Gallery https://gallery.alteryx.com/ or the Alteryx community https://community.alteryx.com/. If you cannot find a tool already made in the designer, likely someone has already built it and shared it!

Source: https://truecue.com/resources/blog/alteryx-predictive-flow-chart/
Blog Author: Bingqian Gao, Chart Creator: Katelyn Weber, Designer: Jakub Szepietowski

Back to your hacking…how to structure your questions on the data to plan your hackathon final output?

Some tips:

1. Decide on which areas to focus on as sustainability is a very broad topic!
2. Brainstorm and start gathering of all your team's questions
3. Group and summarize them! Try to be as Mutually Exclusive and Collectively Exhaustive as possible (MECE)
4. Data feasibility: do you have enough data to tackle the questions?
5. To answer each question, what are the hypotheses you want to test?
6. Which methodology to use to prove or disprove each?
7. What visuals best illustrate your point?



What next?

As part of our next sills session, we will introduce the "Pyramid Principle" to help structure your analyses among other helpful tips!

TRUECUE

In partnership with

alteryx IOD

Any questions ?