Exploratory Data Analysis

Building a richer set of questions and insights

IAT 355

SFU

SCHOOL OF INTERACTIVE ARTS + TECHNOLOGY

Traditional analysis

- Deductive
- Questions are already known
- Factors of importance are already set
- Looking for the descriptive or statistically significant answer

Or …..

- Needle in a haystack of intermediate results

SFU

AN OLD FOLKTALE

We need to examine and analyze data from multiple perspectives

SFU

Visual Analytics

is

a scientific approach to combine…

our **visual intelligence**

and **analytics techniques**

with **interactive visualization**

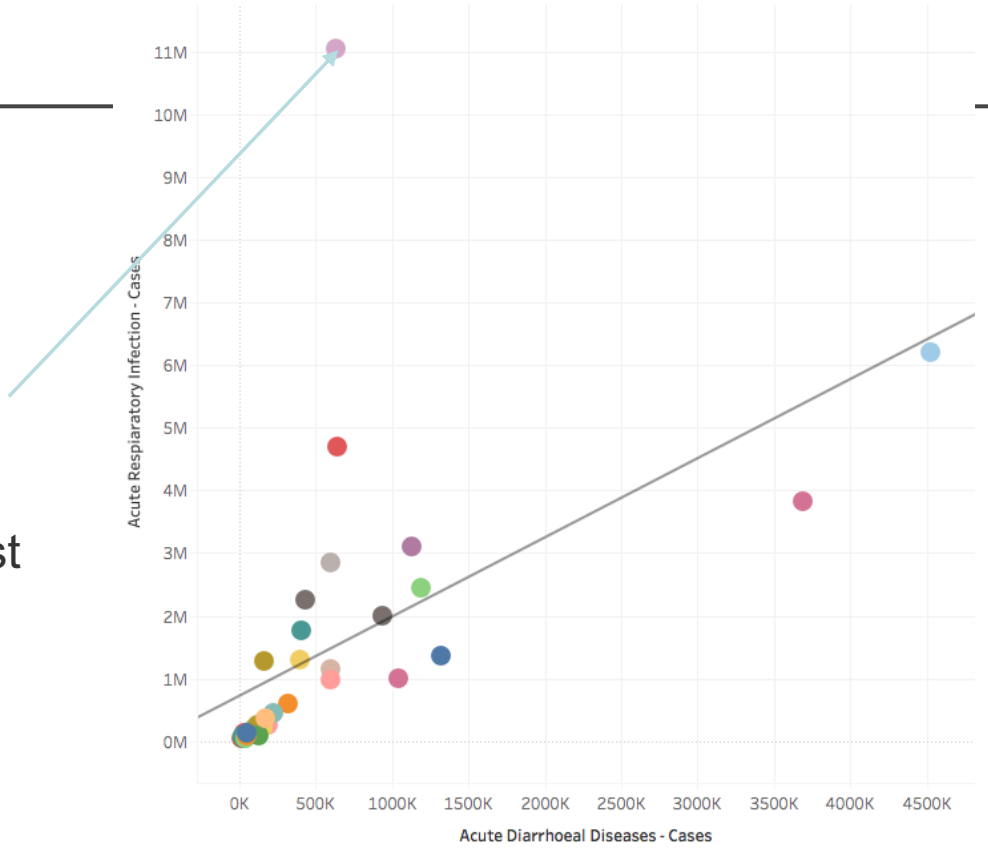
to get **relevant information** out of data

SFU

Visualization helps us derive new questions for analysis

why?

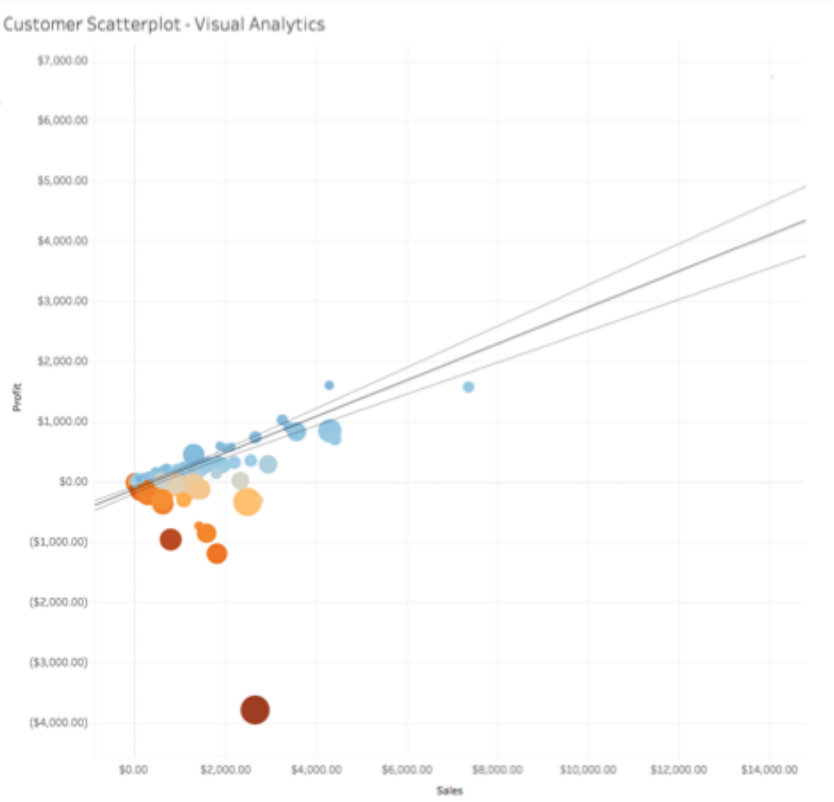
beyond just what?



SFU

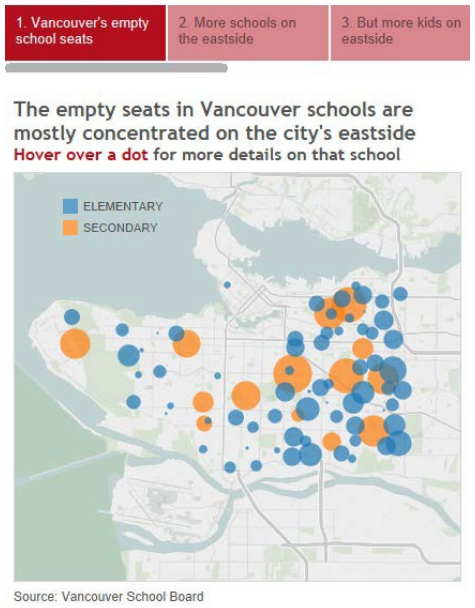
Interaction

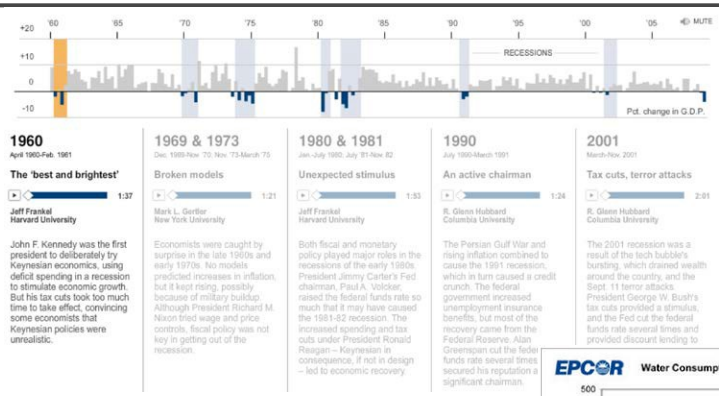
- Interactively exploring data increases cognitive engagement with problem solving
- *Visual analysis* helps us link insights into understanding using structured interactive exploration

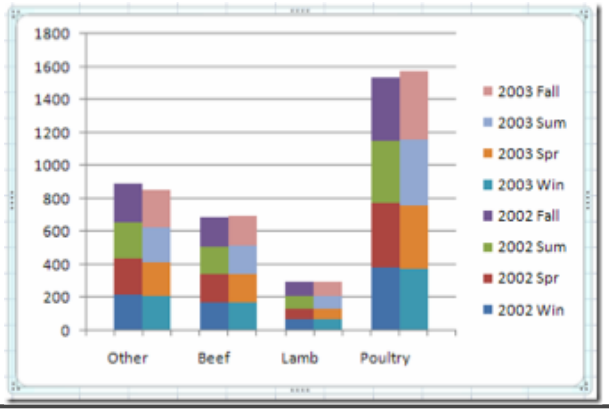


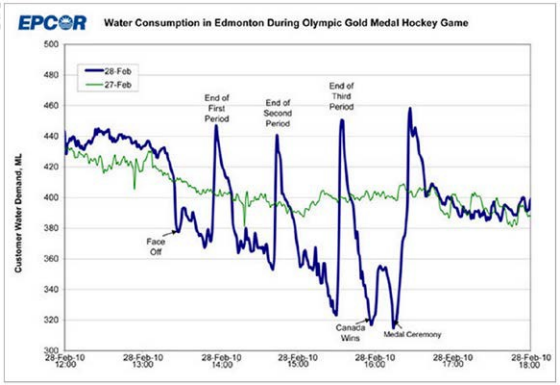
SFU

Different visual forms support different visual discovery









SFU

There's just so much data

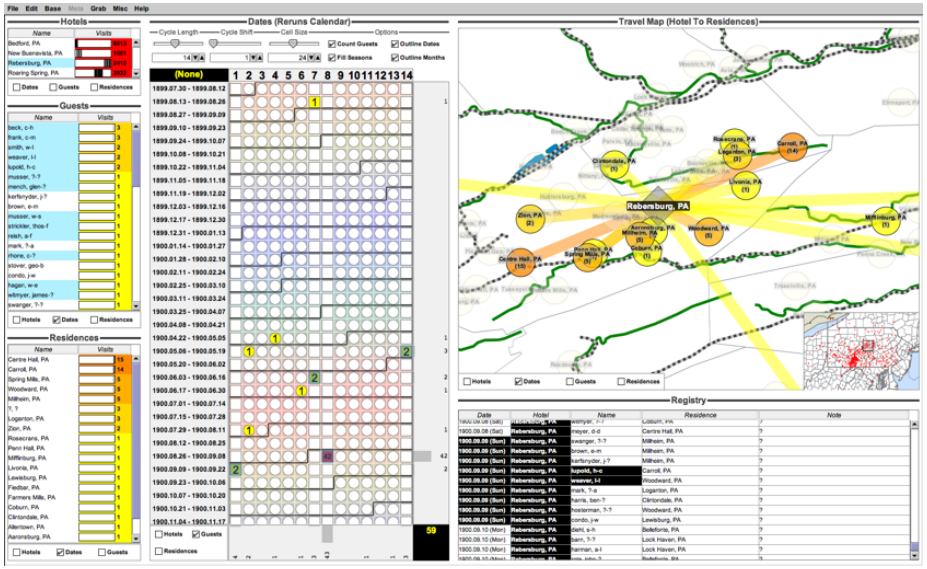
- Visualizations work well with 3-5 dimensions
- Reduce complexity?

- Reduce data
  - Descriptive statistics
  - Aggregation

- Promote visual thinking?
- Facet the data/ dimensions into coordinated views
  - Choose view sets to support primary questions!

SFU

Visual analysis is the process of pursuing and linking these insights through **analytical reasoning** facilitated by **interactive visual** interfaces

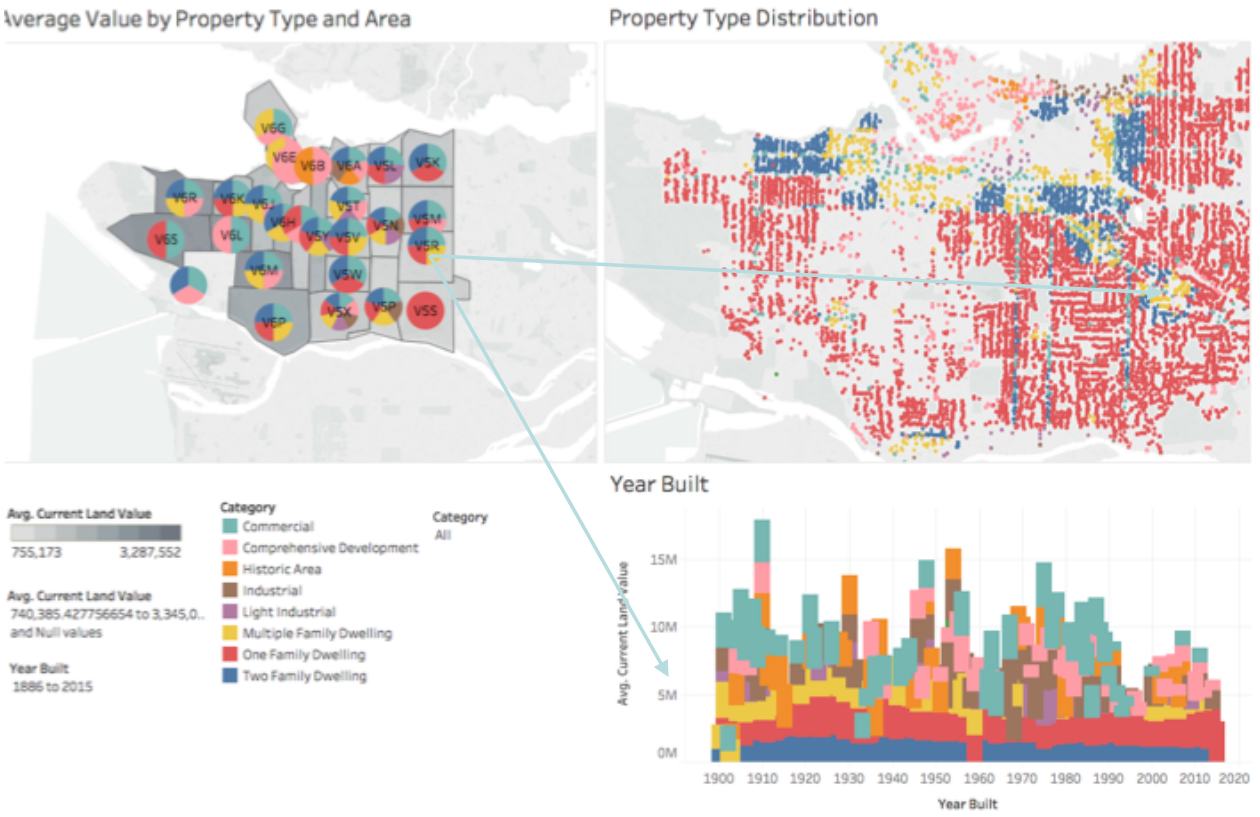


http://www.mdpi.com/1660-4601/14/9/1056

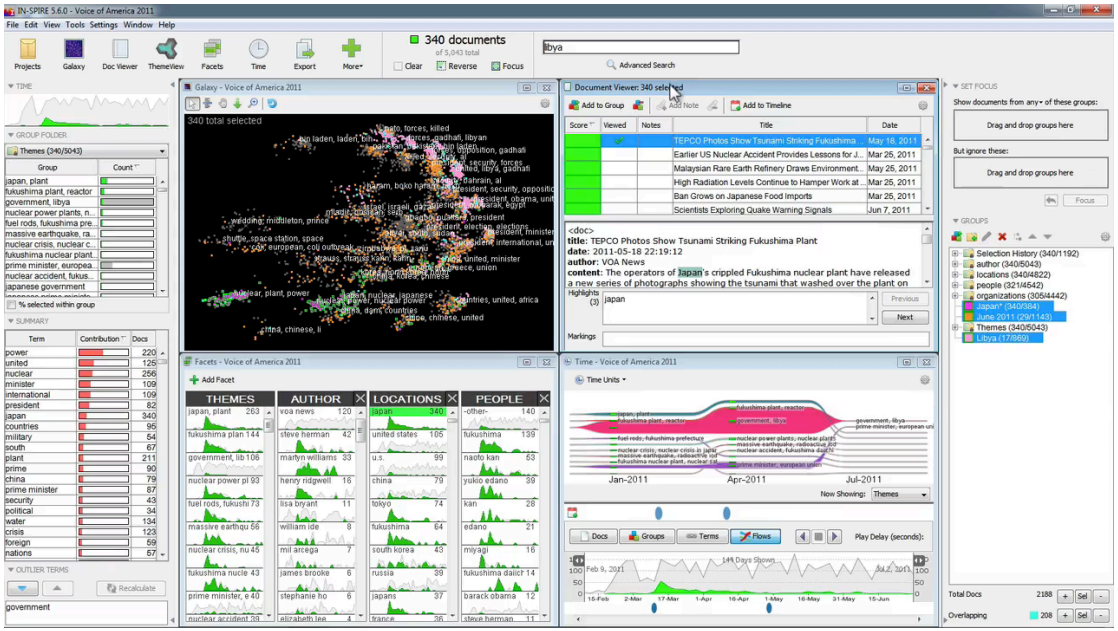
SFU

Dashboards

Highlighting and derived views linked by interaction



Tools for analysis



[IN-SPIRE™ Visual Document Analysis]
PNNL

SFU

Choices for designing the visualization

- What research questions do you want to explore?
- What type of visual encoding and visualization suits the data and the questions?

- What approach and tool is most appropriate to your purpose?
  - explore and analyze your data ? Tell a story? Communicate/ collaborate with others?
- Who is your audience?

SFU

The sensemaking loop of visual analysis

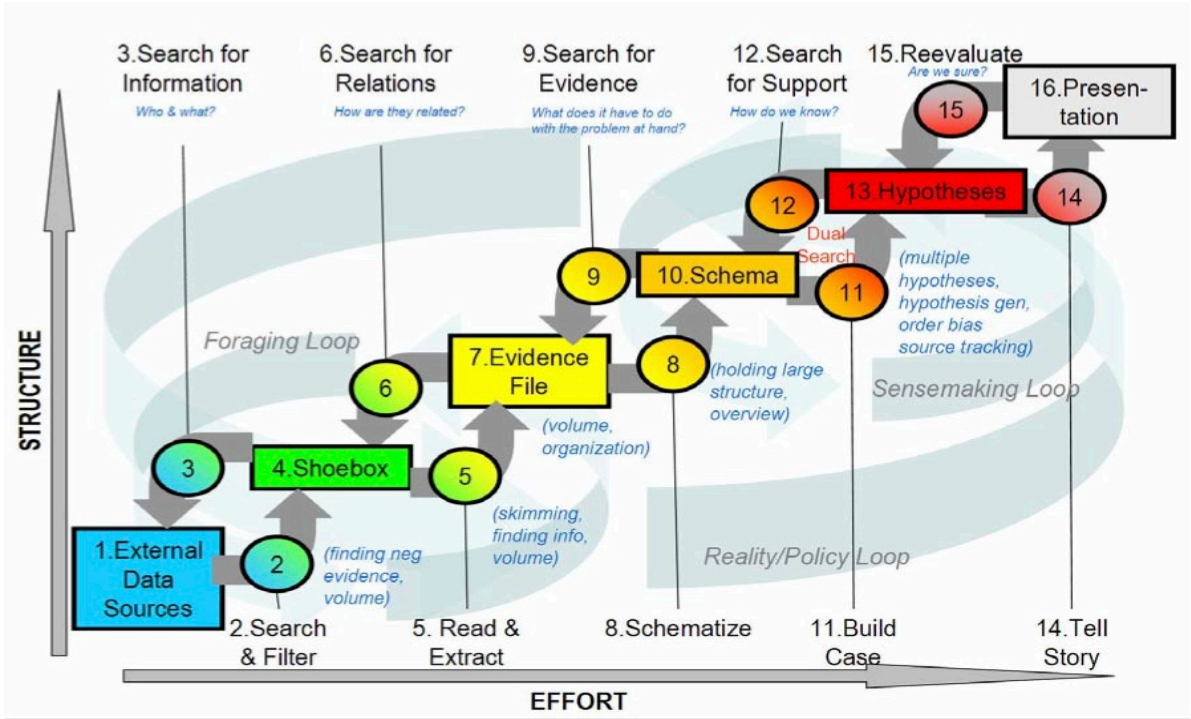


Figure 2. Notional model of sensemaking loop for intelligence analysis derived from CTA.

SFU

Supporting analysis through visual thinking

- Analytics tasks: search for information, forming hypotheses, asking questions, and evaluating and organizing evidence
- Visual task: search, compare, find relations, see similarities, find trends, see distribution
- Interactions: search, select, navigate, change view, filter, brush, get details, collapse dimensions, stitch together views.

SFU

Four Stages of Decision-making

1. Intelligence
   - Discovering, identifying, and understanding the problems occurring in the organization
2. Design
   - Identifying and exploring solutions to the problem
3. Choice
   - Choosing among solution alternatives
4. Implementation
   - Making chosen alternative work and continuing to monitor how well solution is working

SFU

Choices for supporting visual analysis

- What data do you need to analyze? What questions does your data suggest to you?
- What research questions do you want to explore?
- How do your individual answers relate to these larger questions?

- What data do you need?
  - (probably not your whole data set)?
  - More data!

SFU

Analytical reasoning

All students who work part-time take longer to finish their program

Deductive

- Applies and tests theories against specific situations

Why are some students taking longer to finish their programs?

Inductive

- Forms theories and hypotheses to explain the data

How long will students take to finish their programs? What other factors contribute?

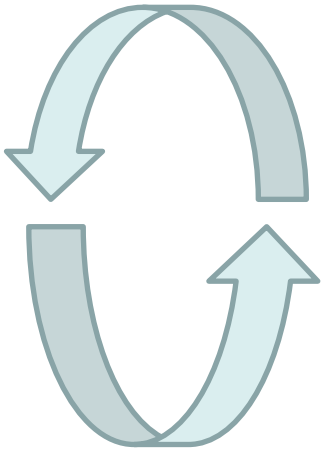
Abductive

- Works from incomplete data to predict likeliest outcomes

SFU

Basic strategies

- Bottom up: Given a data set, explore it to discover the questions and hypotheses it might present/suggest
  - What can you find out?
  - Multiple views and slices of the data
- Top-down: given research question(s), what data do you need to explore what is important?
  - Multiple data sets
  - Context!



Search for information

- Search for patterns
- Important factors and measures (what? When? Who?)
- Slice and visualize - subsets
- It's not about aesthetic perfection: it's about different views to explore data aspects
- Explore outliers

What are the critical observations in student program completion data?

SFU

Search for relations

- Search for patterns across dimensions
- Important relations: trends, correlations, sequences, interactions
- Combine multiple views : brush and link, filter,
- It's not about aesthetic perfection: it's about combining views to get different perspectives
- Start forming hypothesis about phenomenon you observe
- Question everything; Ask "why" often

Do students with a part-time job finish more quickly in certain programs?

SFU

Search for evidence

- Develop a schema of your analysis
  - re-representation or organized marshalling of the information so that it can be used more easily to draw conclusions.

Properties of student
Properties of program

- Build initial hypothesis(es) from your observations
  - the tentative representation of those conclusions

Students with part-time jobs finish faster in programs with multiple schedule options for gateway courses

- Combine views to explore support (or not!) for the hypothesis

SFU

Exercise 1

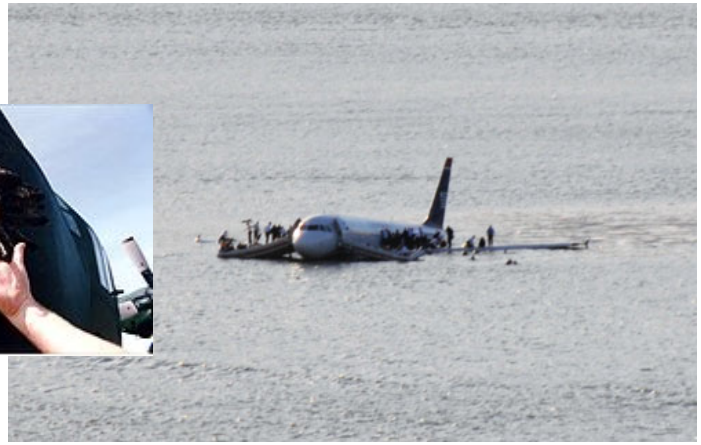
Wildlife Damage to Aircraft
(courtesy of the FAA)

SFU

VA Challenge – bird strike

Dangerous

Costly



US Airways 1549 – Airbus A320
LaGuardia-Charlotte (15 Jan 2009)

- *Sources: http://helicopterems.blogspot.ca/2013/10/bash-bam-boom-living-through-bird-strike.html / http://en.wikipedia.org/wiki/US_Airways_Flight_1549#mediaviewer/File:Plane_crash_into_Hudson_River_(crop).jpg*

SFU

The Data

- Dataset: StrikeReport.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	OPERATOR	ATYPE	NUM_ENGS	INCIDENT_D	TIME_OF_DA	STATE	HEIGHT	SPEED	DISTANCE	PHASE_OF_F	DAMAGE	Damage text	SPECIES	BIRDS_STRUCK	SIZE	REMARKS	
2	MILITARY	B-707	4	6/25/92	Day	TX	30	140		Approach	N	No damage	Unknown bird - m	2 to 10	Medium	REMARKS - CROWS/RAVENS?; AIRC	
3	MILITARY	KC-135E		10/1/92	Night	TN	200	160		Climb	N	No damage	Unknown bird - m	1	Medium	REMARKS - ; AIRCRAFT - KC - 135 -	
4	MILITARY	T-38A		10/1/92	Dusk	TX	0	100	0	Landing Roll	N	No damage	Ring-necked pheas	1	Medium	REMARKS - RUNWAY 17R.; AIRCRAF	
5	MILITARY	A-10A		9/6/91	Day	SC	130	140		Climb	N	No damage	Red-winged blackb	1	Small	REMARKS - ; AIRCRAFT - A - 10 - ; I	
6	MILITARY	T-38A		9/10/91	Dusk	AZ	10	160	0	Landing Roll	N	No damage	Unknown bird - m	1	Medium	REMARKS - ; AIRCRAFT - T - 38 - ; I	
7	MILITARY	KC-135R		5/14/93	Day	CA		200		Approach	N	No damage	Mourning dove	1	Small	REMARKS - FOUND BY GROUND CRE	
8	MILITARY	KC-135		5/18/93	Day	CA	112	150		Approach	N	No damage	Eastern meadowla	1	Small	REMARKS - ; AIRCRAFT - KC - 135 - ;	
9	MILITARY	C-130H		5/18/93	Dusk	WI	0	100	0	Take-off run	N	No damage	Unknown bird or b	1	Medium	REMARKS - ; AIRCRAFT - C - 130 - H	
10	MILITARY	B-52H		5/22/90	Night	MI	1000	150		Approach	N	No damage	Unknown bird - m	2 to 10	Medium	REMARKS - ; AIRCRAFT - B - 52 - H;	
11	MILITARY	RF-4C		5/22/90	Day	ID	2000	300		Approach	N	No damage	Unknown bird - m	1	Medium	REMARKS - BOISE AIR TERMINAL; AI	
12	MILITARY	C-130H		5/30/95	Night	WI	200	140		Approach	N	No damage	Unknown bird - m	1	Medium	REMARKS - ; AIRCRAFT - C - 130 - H	
13	MILITARY	T-38A		5/29/96	Dawn	TX	0	160	0	Take-off run	N	No damage	Unknown bird or b	1	Medium	REMARKS - ; AIRCRAFT - T - 38 - ; I	
14	MILITARY	T-38A		5/29/96	Day	TX	0	130	0	Take-off run	N	No damage	Unknown bird or b	1	Medium	REMARKS - ; AIRCRAFT - T - 38 - ; I	
15	MILITARY	C-130	4	6/2/96	Dawn	HI	0	100	0	Take-off run	N	No damage	Cardinals, bunting	1		REMARKS - ; AIRCRAFT - C - 130 - ;	
16	MILITARY	T-1A		6/3/96	Day	TX	0	75	0	Landing Roll	N	No damage	Unknown bird or b	1	Medium	REMARKS - ; AIRCRAFT - T - 1 - ; IM	
17	MILITARY	T-37B		12/17/92	Dusk	TX	20	90		Approach	N	No damage	Unknown bird - m	2 to 10	Medium	REMARKS - ; AIRCRAFT - T - 37 - B;	
18	MILITARY	T-37B		12/17/92	Dusk	TX	1700	200		Climb	N	No damage	Unknown bird - m	1	Medium	REMARKS - ; AIRCRAFT - T - 37 - B;	

SFU

Exercise -- Investigation

Goal: To use visual analysis to investigate your data
To analyze and report possible findings

1. In which state do most incidents occur where there is damage?
2. What wildlife causes the most damage?
3. Do more birds result in more damage?
4. At what points during the flights do incidents occur? Anything special to note?
5. During what month do the most incidents occur?
6. In NY state, what are the top 5 wildlife offenders?

- Explore and choose the most effective visualizations. What works best?

SFU

Discussion

Investigation and Hypothesis. For each, what did you find as the most effective visualization?

1. In which state do most incidents occur where this is damage?
    1. NULL, TX, CA, NY
2. What wildlife causes the most damage?
    1. Unknown, Gulls, Deer
3. At what points during the flights do incidents occur?
    1. Approach
4. During what month do the most incidents occur?
    1. August
5. In NY state, what are the top 5 wildlife offenders?
    1. Unknowns, Gulls

What else?

- Are there other questions you might find interesting?
- What hypotheses did you come up with?
- What couldn’t you answer?

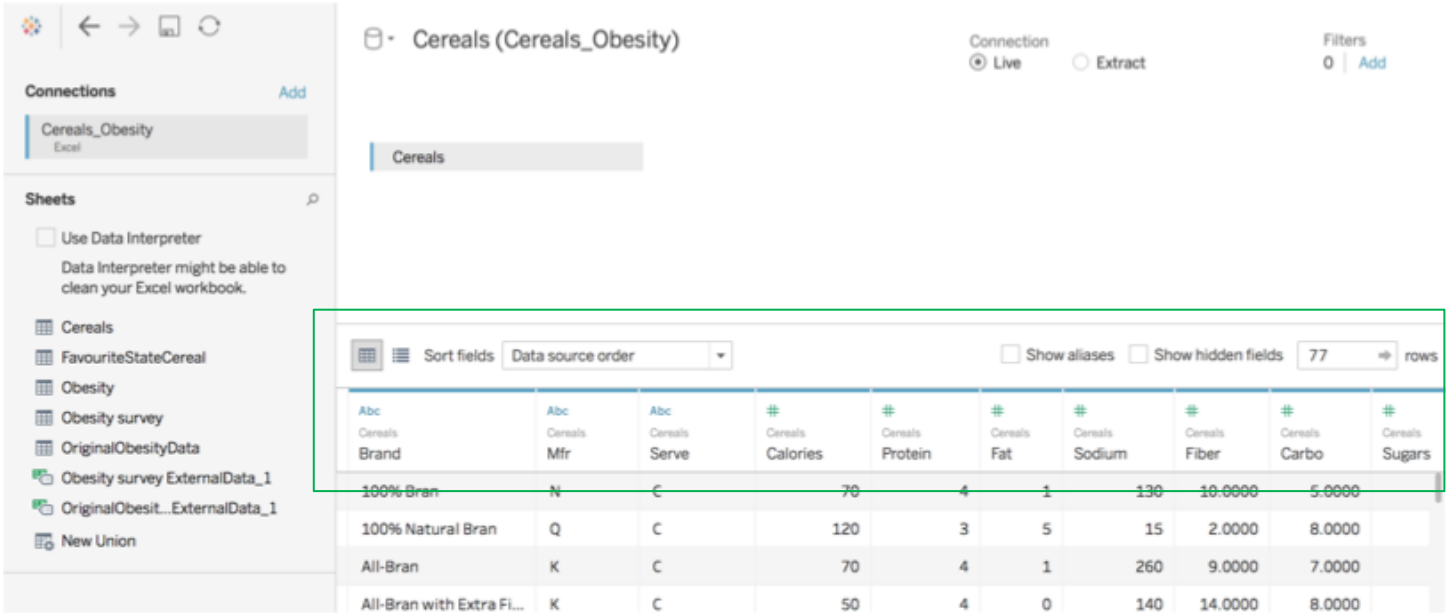
SFU

example

- Dataset: Popular breakfast cereals

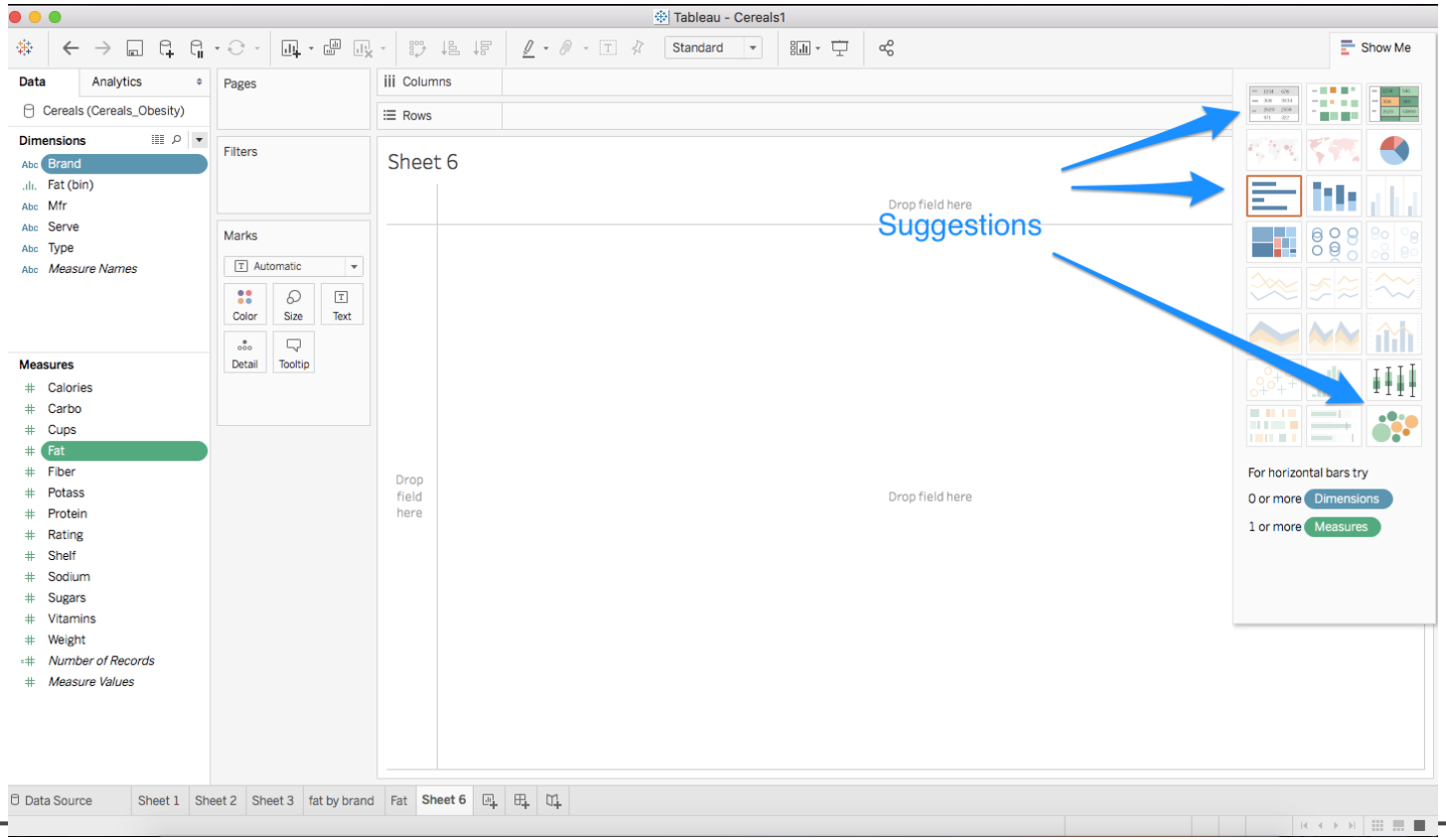
- What kinds of cereal are the most fattening?
- What makes a cereal high in calories?
- Are high-calorie cereals more popular?

- Cereal data



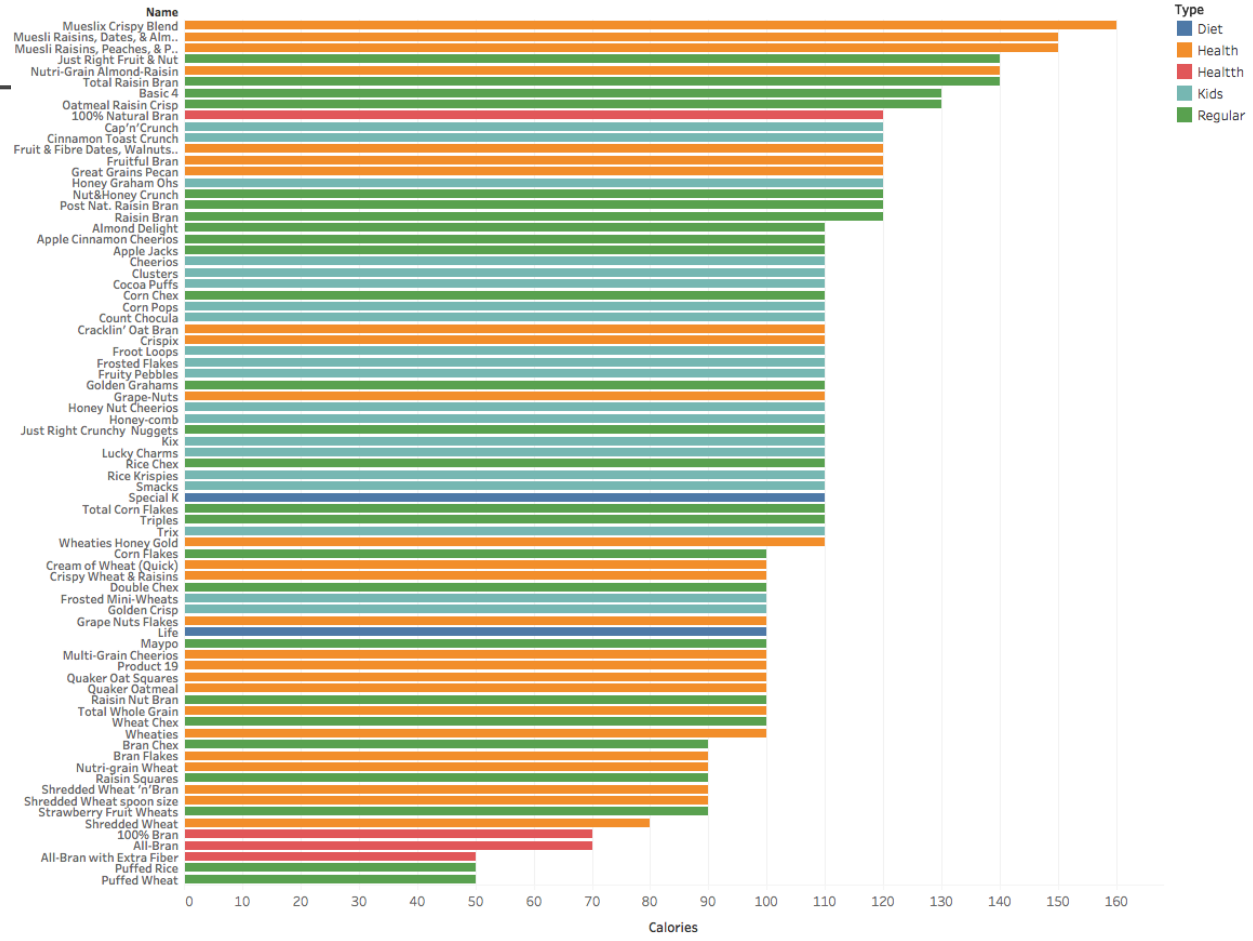
SFU

What kinds of cereal are the most fattening?



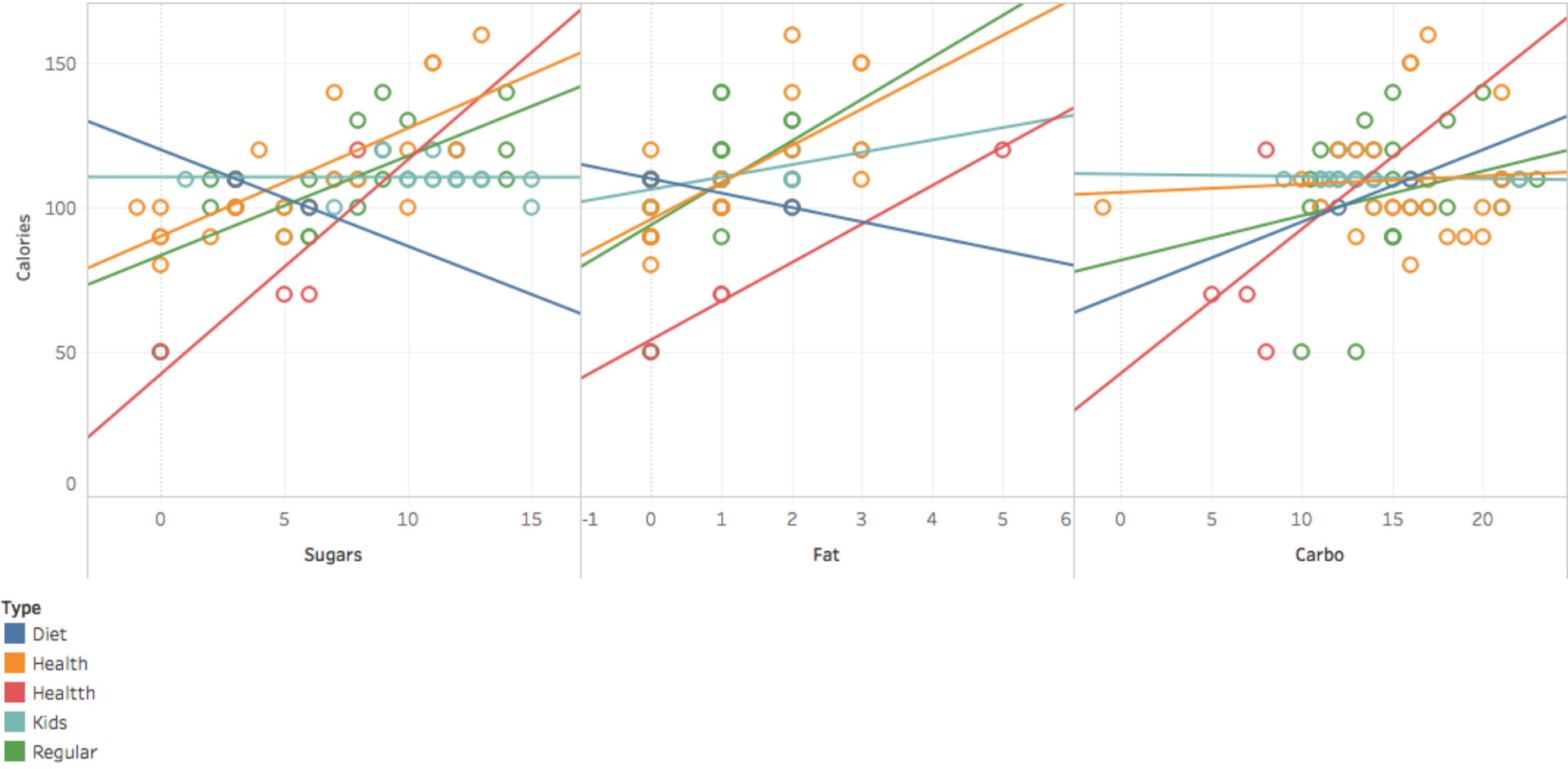
SFU

What kinds of cereal are the most fattening?



SFU

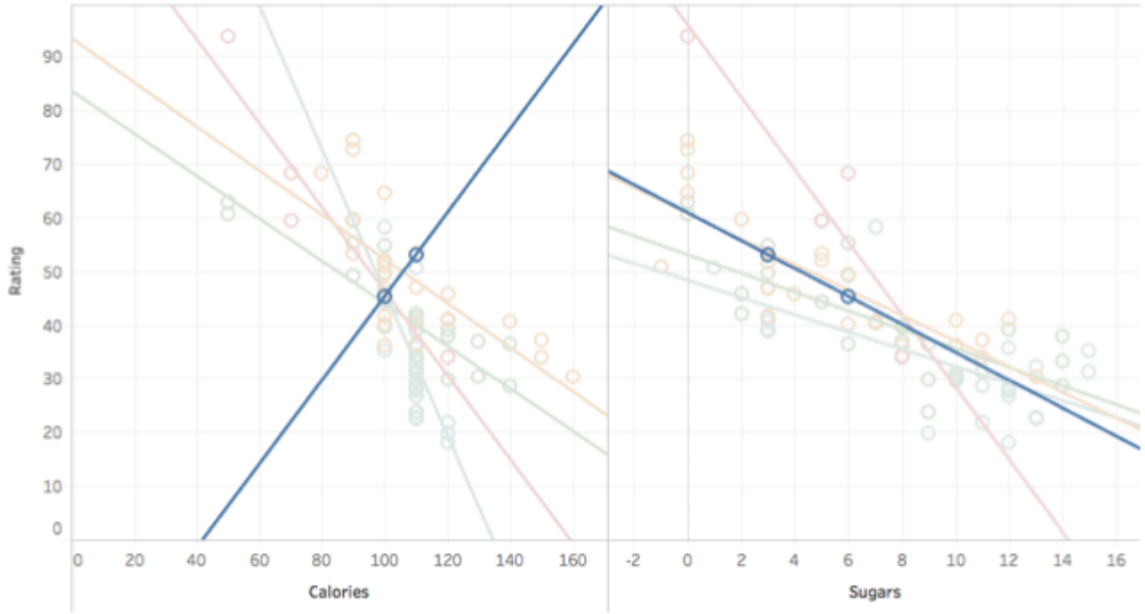
What makes a cereal high in calories?



SFU

Are high-calorie cereals more popular?

- any relation between cereal properties and rating ?
- Does this differ by type?



SFU

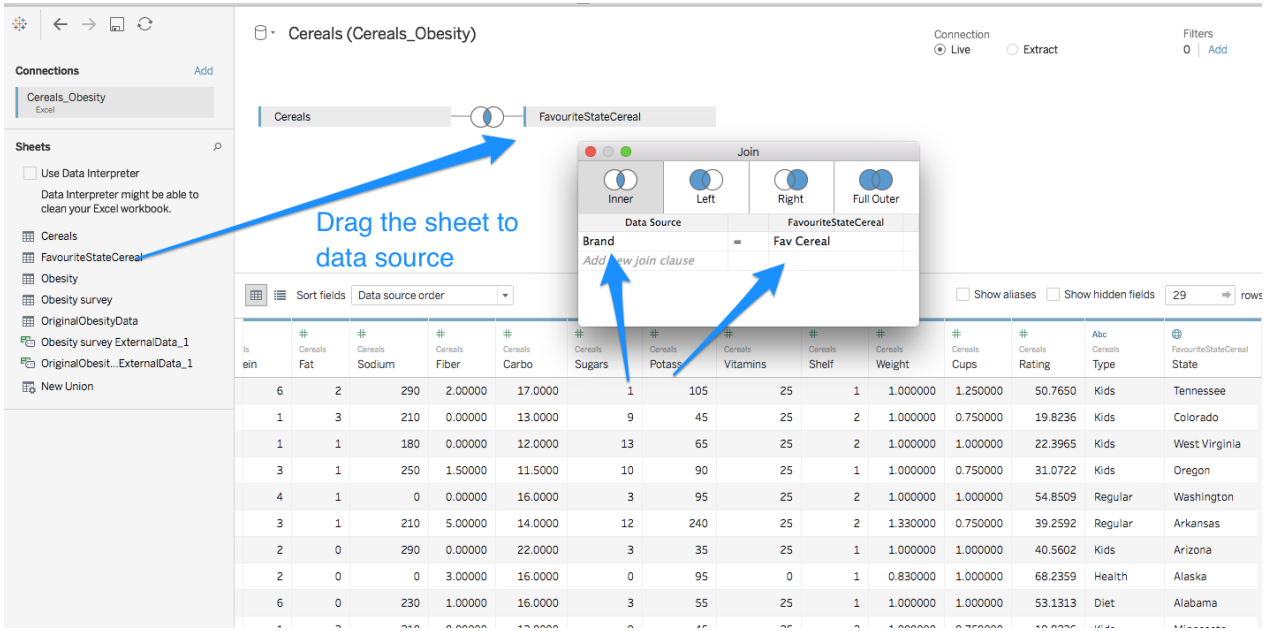
What other questions might we have?

- Where are brands popular?
- How do breakfast cereals relate to health?

These questions could not be answered by the Cereal data !

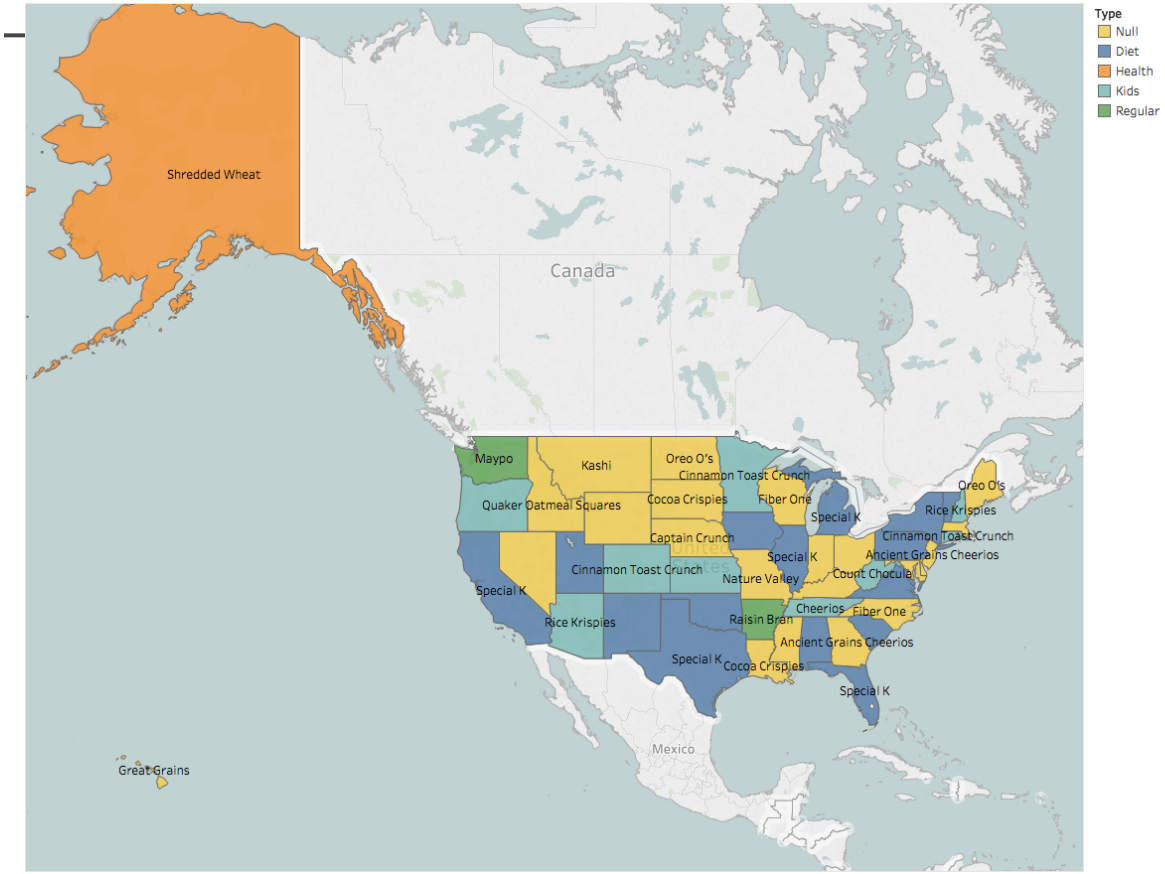
SFU

Add a second dataset



SFU

Where are brands popular?



Note: Some missing data!

Yellow = Missing Type

SFU

How do breakfast cereals relate to health?

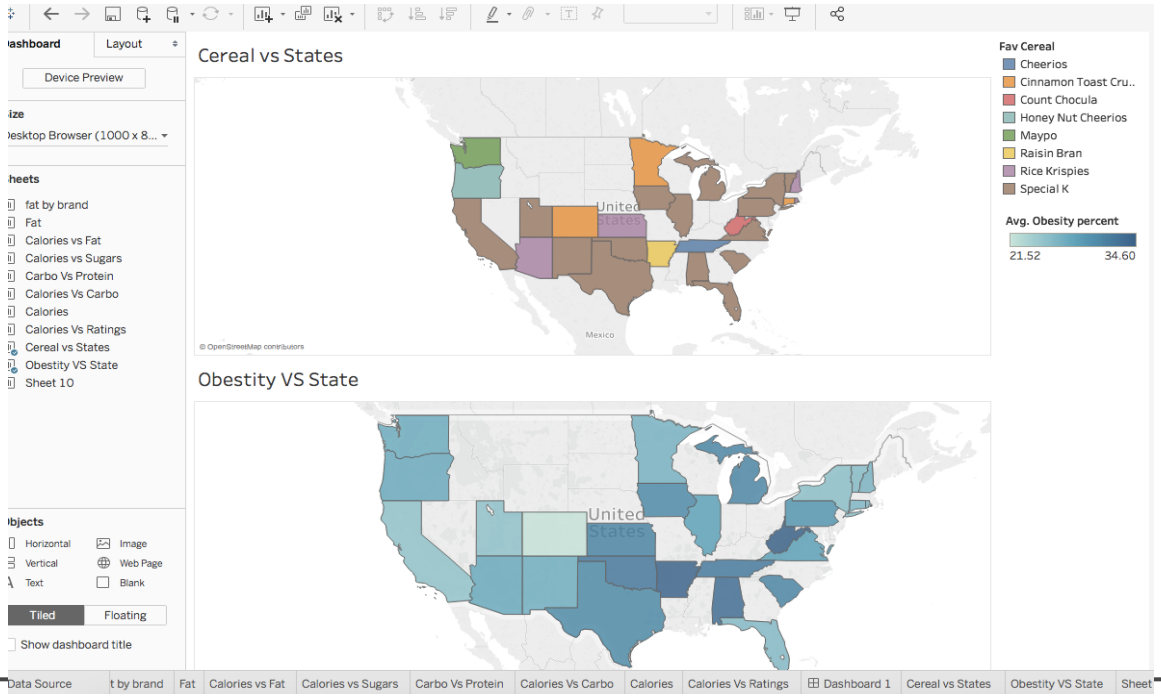
- How do we answer questions like that ?

- What is the obesity level in each state ?
- We can look at obesity rate and high sugar cereals?

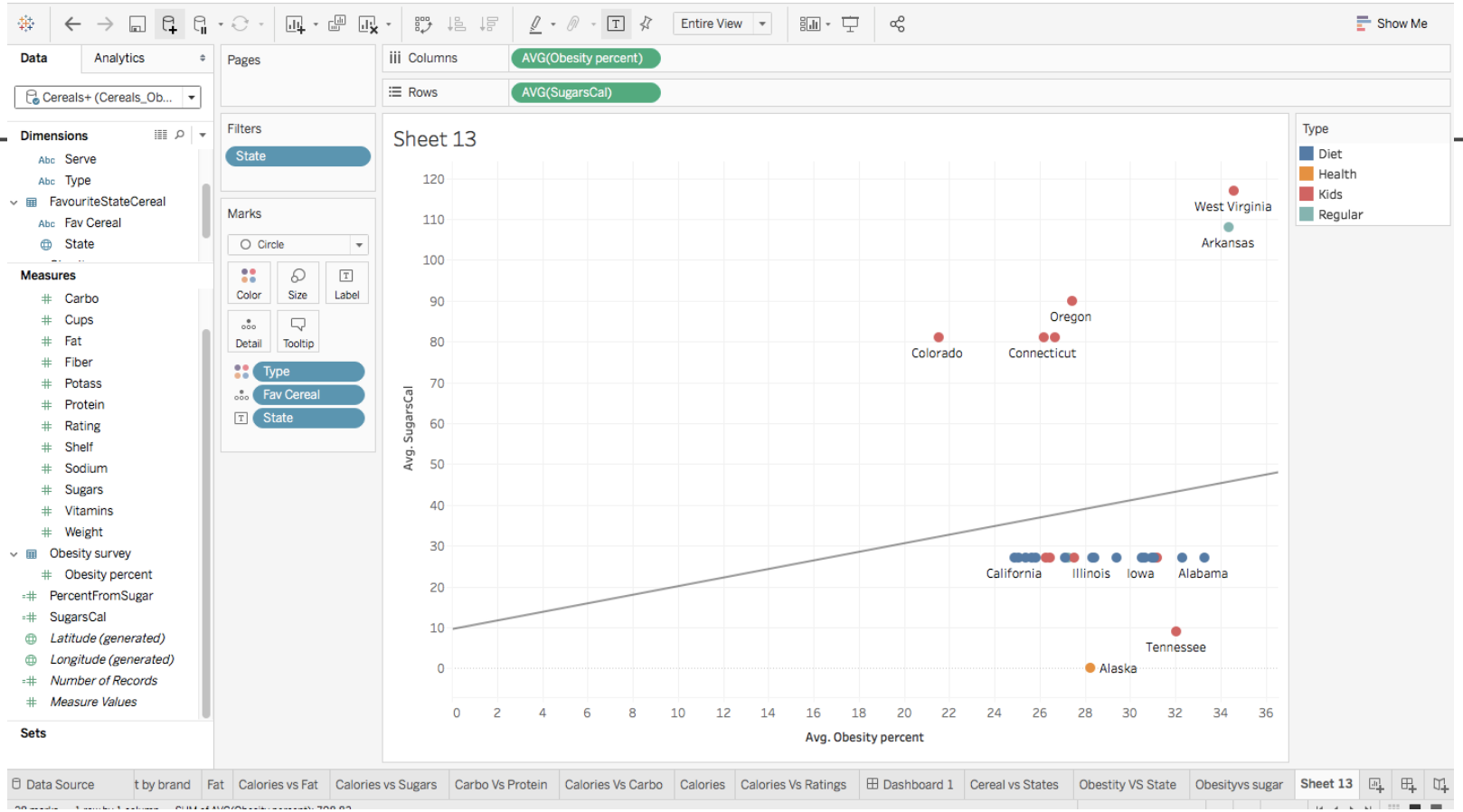
SFU

How do breakfast cereals relate to health?

- More data !



SFU


Entire View
Show Me
Data
Analytics
Cereals+ (Cereals_Ob...
Dimensions
Serve
Type
FavouriteStateCereal
Fav Cereal
State
Measures
Carbo
Cups
Fat
Fiber
Potass
Protein
Rating
Shelf
Sodium
Sugars
Vitamins
Weight
Obesity survey
Obesity percent
PercentFromSugar
SugarsCal
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Sets
Pages
Filters
State
Marks
Circle
Color
Size
Label
Detail
Tooltip
Type
Fav Cereal
State
Columns
AVG(Obesity percent)
Rows
AVG(SugarsCal)
Sheet 13
Avg. SugarsCal
Avg. Obesity percent
West Virginia
Arkansas
Oregon
Colorado
Connecticut
California
Illinois
Iowa
Alabama
Tennessee
Alaska
Type
Diet
Health
Kids
Regular
Data Source
t by brand
Fat
Calories vs Fat
Calories vs Sugars
Carbo Vs Protein
Calories Vs Carbo
Calories
Calories Vs Ratings
Dashboard 1
Cereal vs States
Obestity VS State
Obesityvs sugar
Sheet 13

SFU

Exercise 2

Wildlife Damage to Aircraft v2
(courtesy of the FAA)

SFU

Expanding analysis

- Are there other questions you might find interesting?
- What hypotheses did you come up with?
- What couldn’t you answer?

Expanding analysis

- What can airports/pilots/Boeing do to reduce effect of bird strikes??

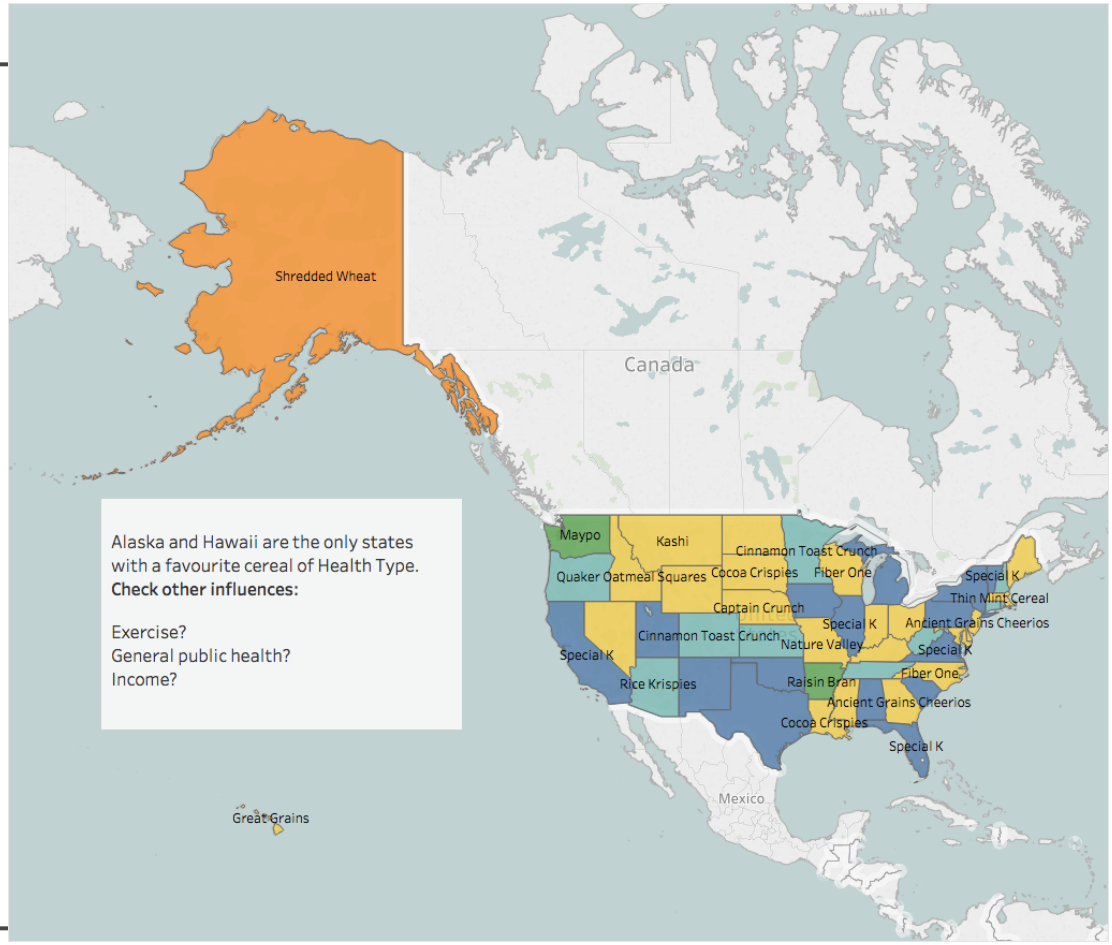
- Bring in additional data
- data set (StrikeReport/StrikeData2): New dimensions
  - Location, time of day, aircraft type, operator ….....

- Expand your hypotheses
- Identify critical factors and relations – find evidence

SFU

Explain your process to yourself and others

- Keep track of your insights!
- Annotate your visualizations
- Keep analysis notes



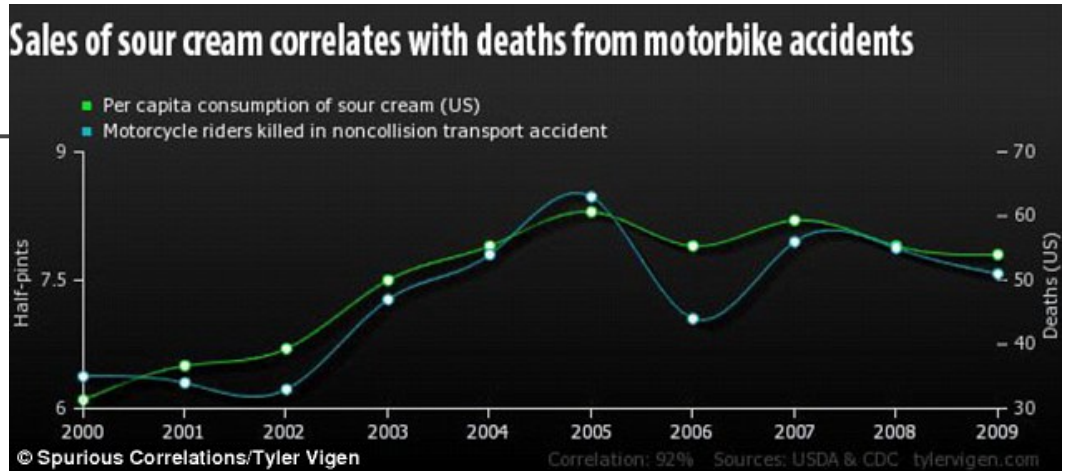
SFU

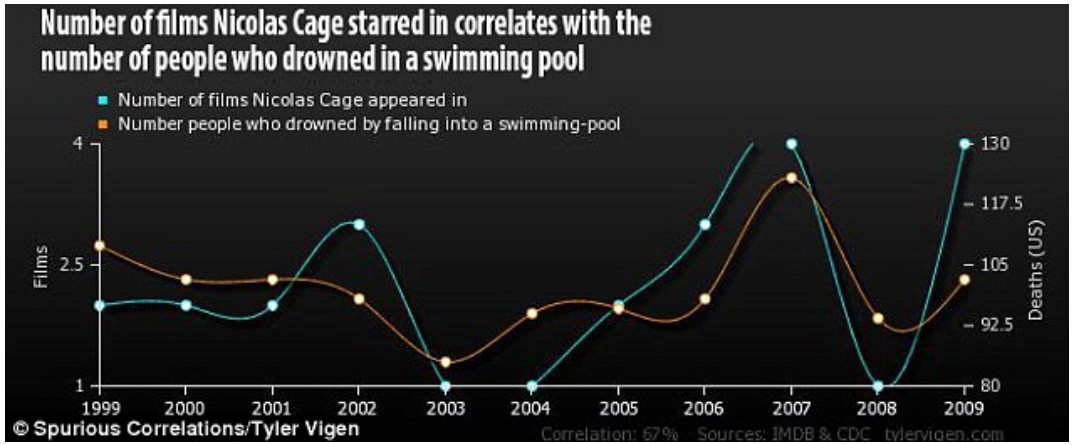
Watch out for …

- Biases
- Decision Making
  - Mere exposure effect
  - Tendency to be positive just because you are familiar with something
  - For instance, “Better the devil you know”
- Selection bias
  - Distortion of evidence arising from the way the data were collected

Watch out for …

- Correlation does not imply causation !!!





SFU

Certainty and Analytic Bias

Would you rather choose:

- A - 100% chance to receive $100

  Repeat this 100 times
  You would gain $10,000

- B - 90% chance to receive $120
  (10% chance of getting nothing)

  Repeat this 100 times
  You would gain $10,800

SFU


SFU