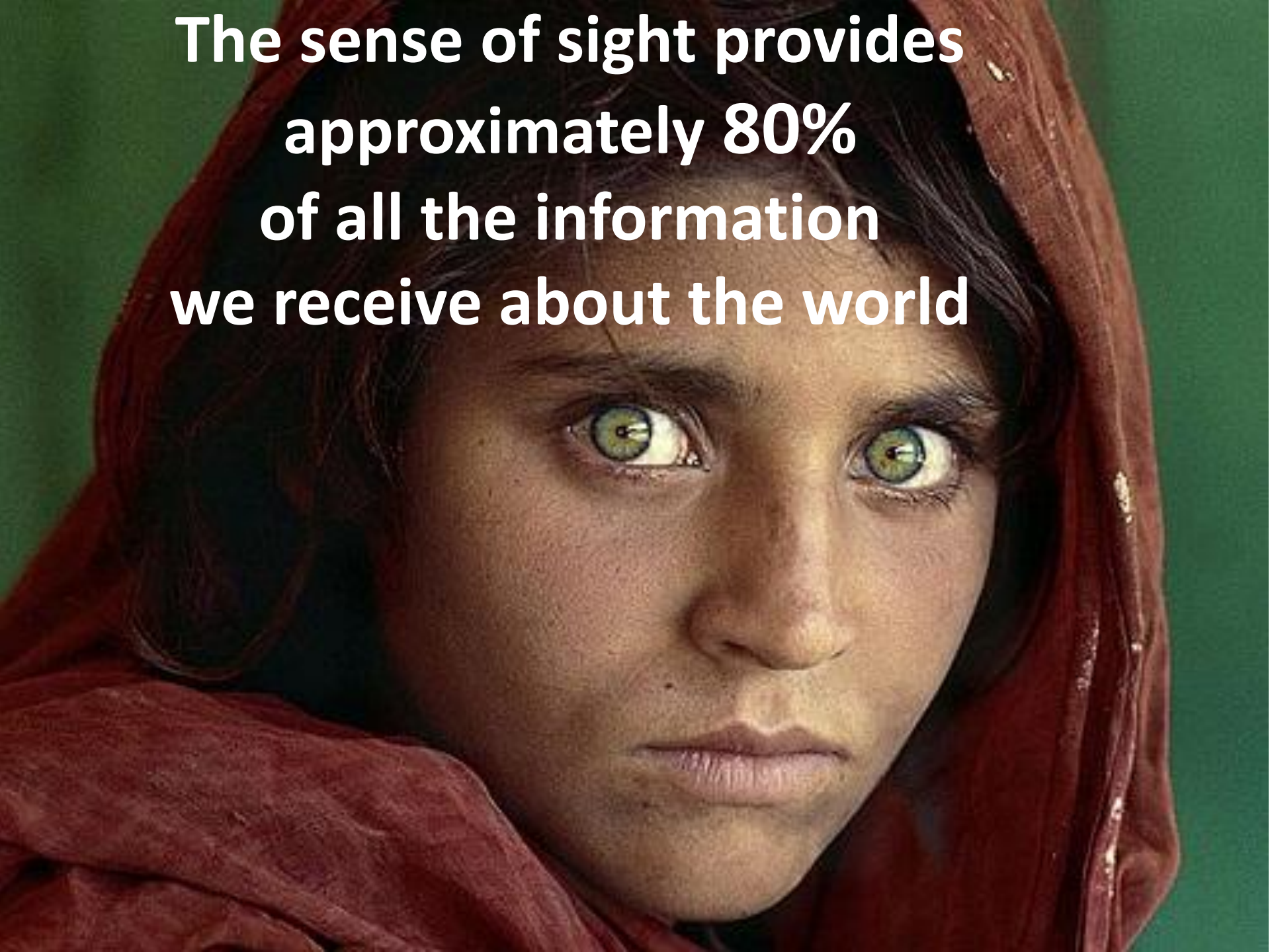

visualization


The sense of sight provides
approximately 80%
of all the information
we receive about the world

"The great fun of information visualization is that it gives you answers to questions you didn't know you had"

Saaien Tist

"The great use of data visualization is that it gives you **clues to questions** you didn't know you had"

changed so profoundly that it takes a while to give it all a place in your head; let alone a blog. Until I remembered this morning why I started this blog in the first place: to help me order my thoughts in the first place. So it might have sped things up instead, actually...

Why "Saaien tist"? Because it's pronounced as 'scientist', and means 'boring bloke' in Flemish

"statistics is about proving what you expect, while visualization is about **discovering what you didn't expect**"

As do many others, I see a strong connection between statistics and data visualization. Taking a bit of a shortcut here, you could say that **statistics is about proving what you expect, while visualization is about discovering what you didn't expect** and refining those expectations.

View my complete profile

Data exploration via visualization

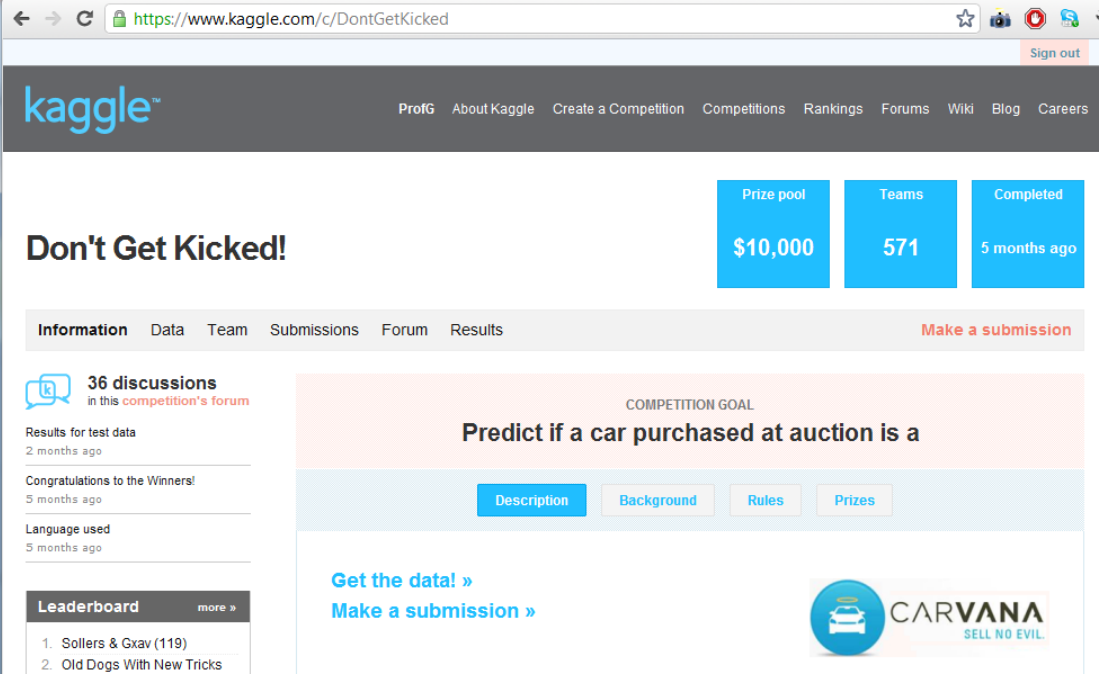
- Examine suspected patterns
- Discover unexpected patterns and exceptions

This is part of

Business Intelligence

Easy to manipulate
Interactive
Intuitive
Fast

V. 4.0



Data Files

File Name	Available Formats
Carvana_Data_Dictionary	.txt (2.75 kb)
test	.csv (9.16 mb)
	.zip (1.90 mb)
training	.csv (13.82 mb)
	.zip (2.84 mb)
example_entry	.csv (449.33 kb)

You only need to download one format of each file. Each has the same contents but use different packaging methods.

- The challenge of this competition is to predict if the car purchased at the Auction is a good / bad buy.
- All the variables in the data set are defined in the file Carvana_Data_Dictionary.txt
- The data contains missing values
- The dependent variable (IsBadBuy) is binary (C2)
- There are 32 Independent variables (C3-C34)

One of the biggest challenges of an auto dealership purchasing a used car at an auto auction is the risk that the vehicle might have serious issues that prevent it from being sold to customers. **The auto community calls these unfortunate purchases "kicks".**

Training.csv (~73K rows, 34 columns)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	RefId	IsBadBuy	PurchDate	Auction	VehYear	VehicleAg	Make	Model	Trim	SubModel	Color	Transmiss	WheelTyp	WheelTyp	VehOdo	Natior
2	1	0	12/7/2009	ADESA	2006	3	MAZDA	MAZDA3	i	4D SEDAN	RED	AUTO	1	Alloy	89046	OTHEF
3	2	0	12/7/2009	ADESA	2004	5	DODGE	1500 RAM	ST	QUAD CAE	WHITE	AUTO	1	Alloy	93593	AMER
4	3	0	12/7/2009	ADESA	2005	4	DODGE	STRATUS \	SXT	4D SEDAN	MAROON	AUTO	2	Covers	73807	AMER
5	4	0	12/7/2009	ADESA	2004	5	DODGE	NEON	SXT	4D SEDAN	SILVER	AUTO	1	Alloy	65617	AMER
6	5	0	12/7/2009	ADESA	2005	4	FORD	FOCUS	ZX3	2D COUPE	SILVER	MANUAL	2	Covers	69367	AMER
7	6	0	12/7/2009	ADESA	2004	5	MITSUBISI	GALANT 4	ES	4D SEDAN	WHITE	AUTO	2	Covers	81054	OTHEF
8	7	0	12/7/2009	ADESA	2004	5	KIA	SPECTRA	EX	4D SEDAN	BLACK	AUTO	2	Covers	65328	OTHEF
9	8	0	12/7/2009	ADESA	2005	4	FORD	TAURUS	SE	4D SEDAN	WHITE	AUTO	2	Covers	65805	AMER
10	9	0	12/7/2009	ADESA	2007	2	KIA	SPECTRA	EX	4D SEDAN	BLACK	AUTO	2	Covers	49921	OTHEF
11	10	0	12/7/2009	ADESA	2007	2	FORD	FIVE HUNI	SEL	4D SEDAN	RED	AUTO	1	Alloy	84872	AMER
12	11	0	12/14/2009	ADESA	2005	4	GMC	1500 SIERF	SLE	REG CAB 4	SILVER	AUTO	1	Alloy	80080	AMER
13	12	0	12/14/2009	ADESA	2001	8	FORD	F150 PICK	XL	REG CAB 4	WHITE	MANUAL	1	Alloy	75419	AMER
14	13	1	12/14/2009	ADESA	2005	4	DODGE	CARAVAN	SE	MINIVAN	RED	AUTO	1	Alloy	79315	AMER
15	14	0	12/14/2009	ADESA	2005	4	NISSAN	ALTIMA	Bas	4D SEDAN	WHITE	AUTO	2	Covers	71254	TOP LI
16	15	0	12/14/2009	ADESA	2006	3	DODGE	CARAVAN	SXT	MINIVAN	GOLD	AUTO	1	Alloy	74722	AMER

Open data

Three options:

Open > File

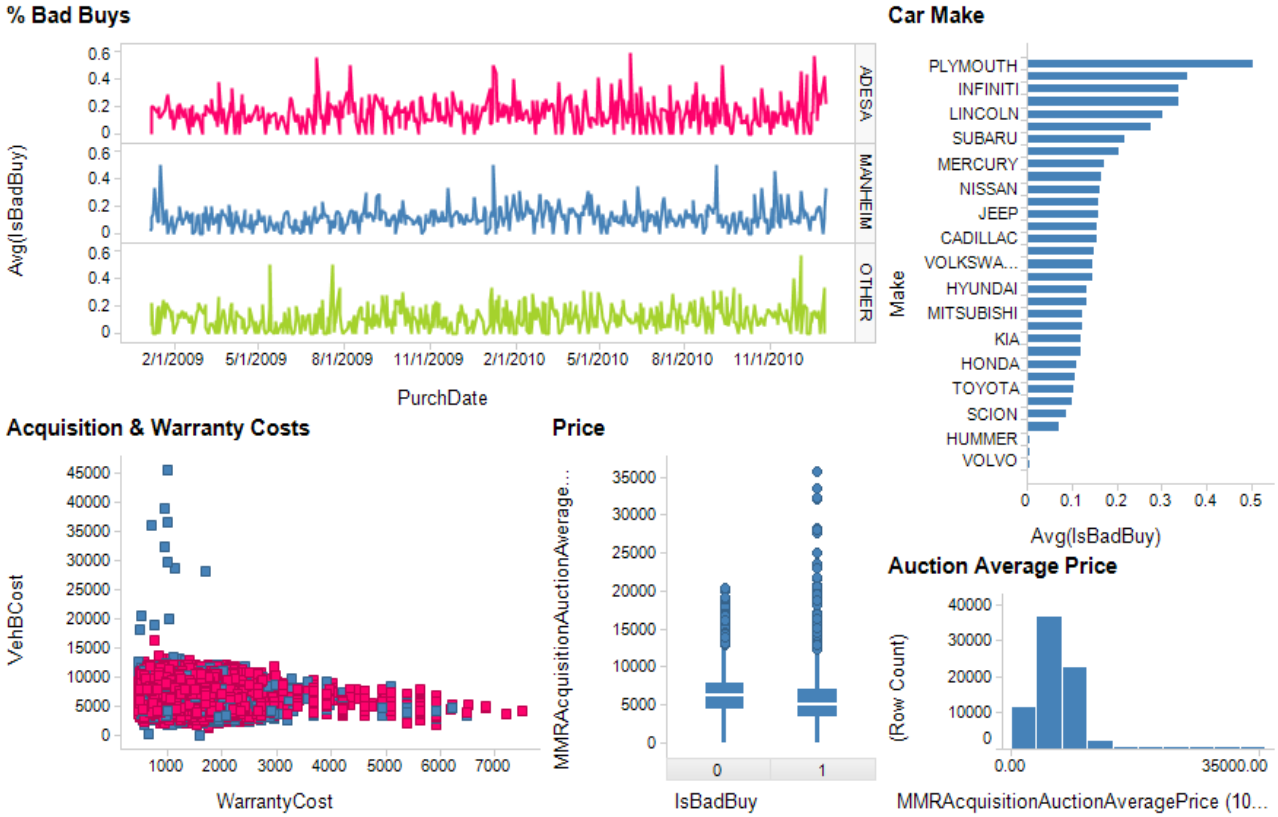
Copy from spreadsheet; Paste into Spotfire

Drag file into Spotfire

Basic charts

Discover patterns and exceptions

Bar chart
Histogram
Boxplot
Line chart
Scatter plot



Ben Shneiderman's Mantra

Overview first,
zoom and filter,
then details-on-demand

Interaction

Change variables

Compare

Sort

Aggregate

Add variables

Re-scale

Zoom, pan

Filter

Aggregate

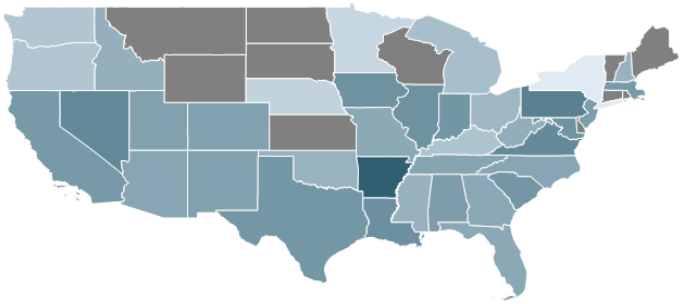
Re-visualize

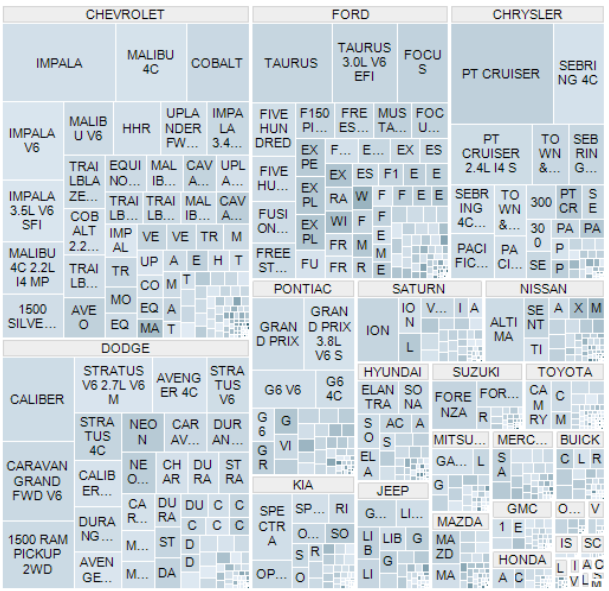
Access details on demand

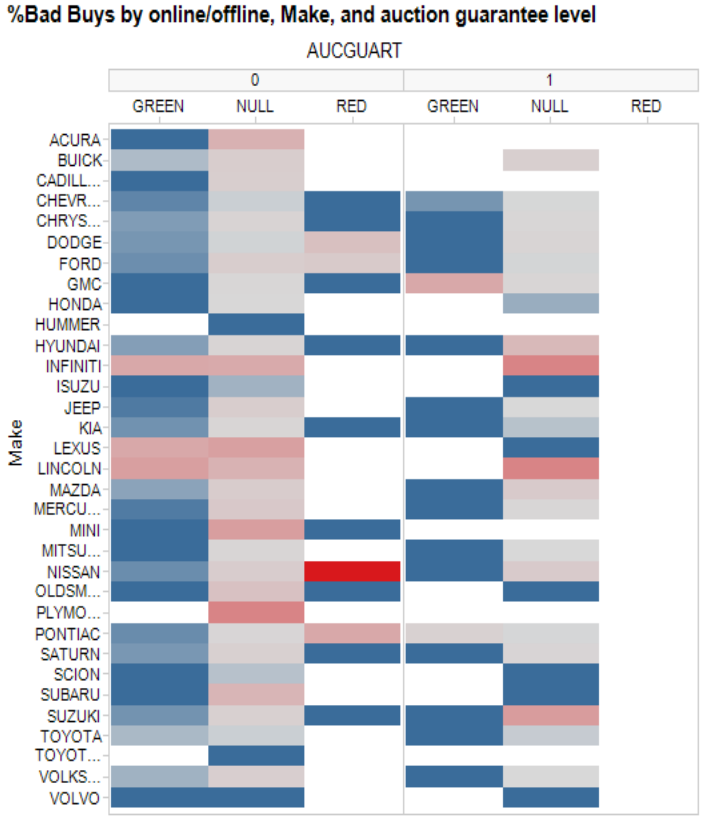
Annotate

Specialized charts

For special data structures







Visualization for a data mining task

Supervised learning

- Focus on relationship between output and inputs
- Numerical vs. categorical output

Unsupervised learning

- Relationships between all variables

See more in textbook

Online sharing