

DATA ANALYSIS

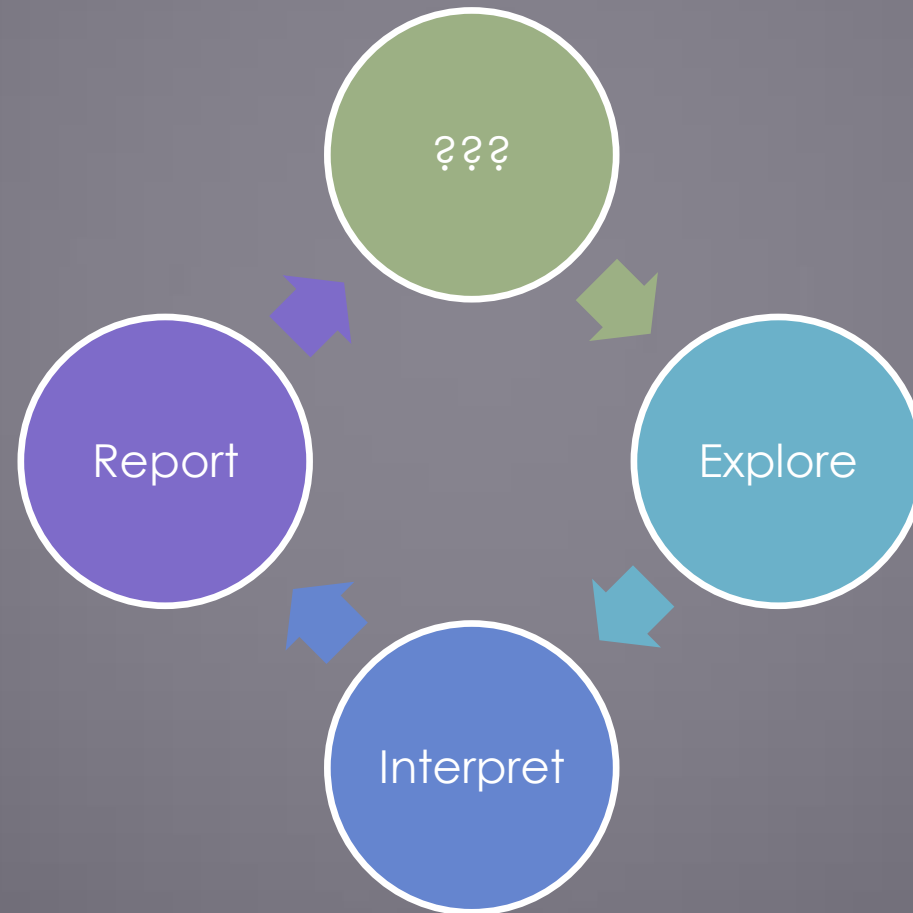
Bryn Mawr Digital Scholarship
Graduate Community of Learning
March 27, 2018



RAWGraphs

WHAT IS DATA ANALYSIS?

THE FOUR PHASES OF DATA ANALYSIS



PHASE 1: REFINE YOUR QUESTION

???

- Are air pollution levels higher on the east coast than on the west coast?
- Are hourly ozone levels on average higher in New York City than they are in Los Angeles?

PHASE 2: EXPLORE THE DATA

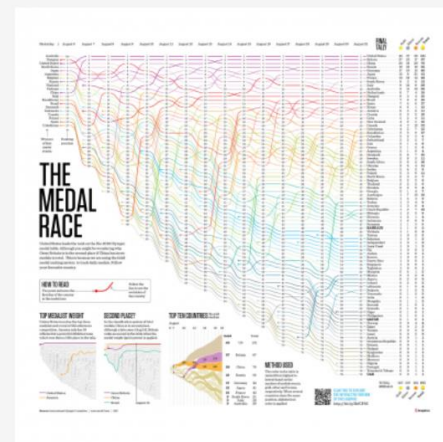
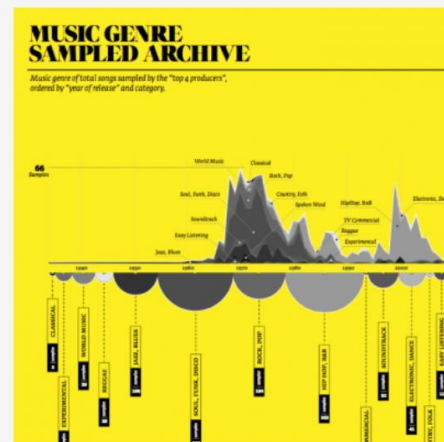
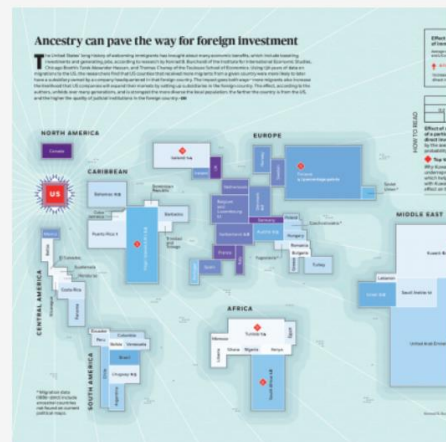
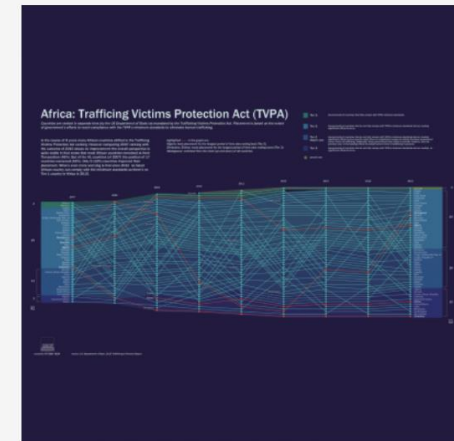
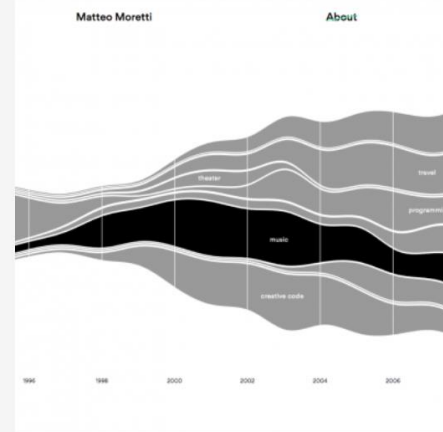
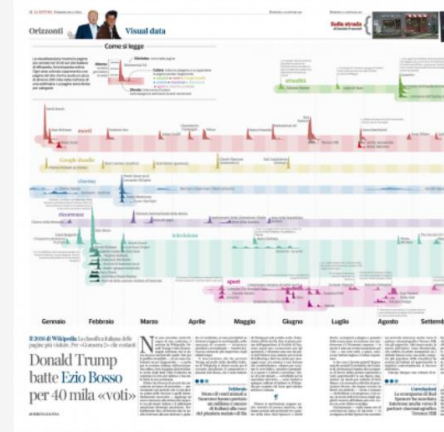
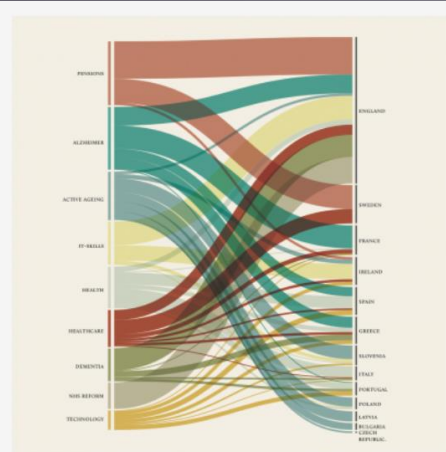
Goals of Exploratory Data Analysis:

- Summarize and understand the structure of a dataset
- Identify any assumptions you have about the data
- Formulate some hypotheses based on your initial observations of the data



Explore

DATA VISUALIZATION



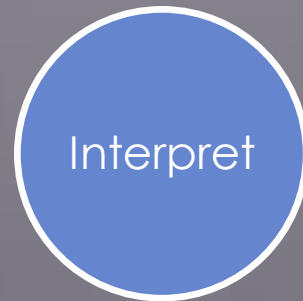
DATA VISUALIZATION

- **EXPLORE**: model your data in different ways
- **INTERPRET**: provide graphical evidence for trends
- **REPORT**: support arguments and convey information

PHASE 3: INTERPRETATION

Testing and Explanation

- Discipline-specific methodologies
- Matching expectations to the results of exploratory data analysis



PHASE 4: REPORT RESULTS

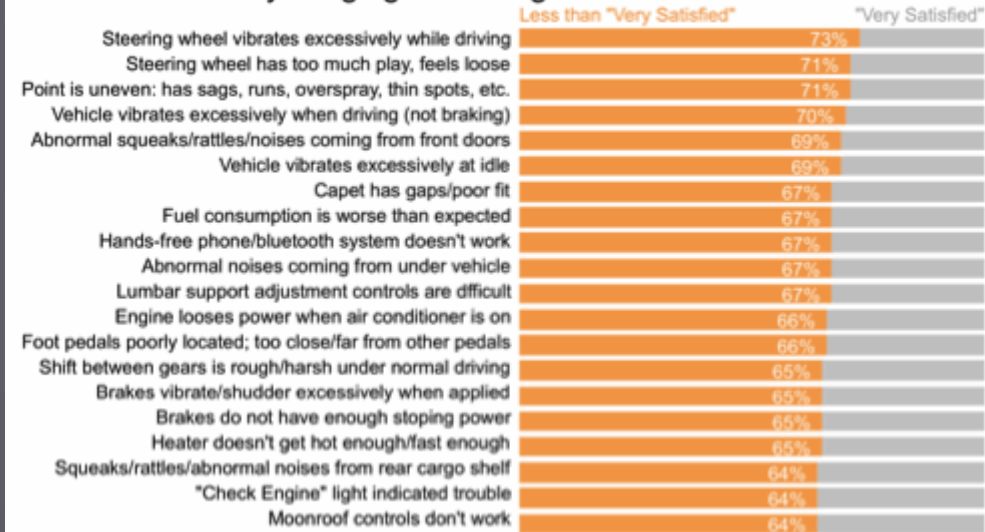
Communicating Your Results

- Clean versions of exploratory visualizations
- Storytelling with data



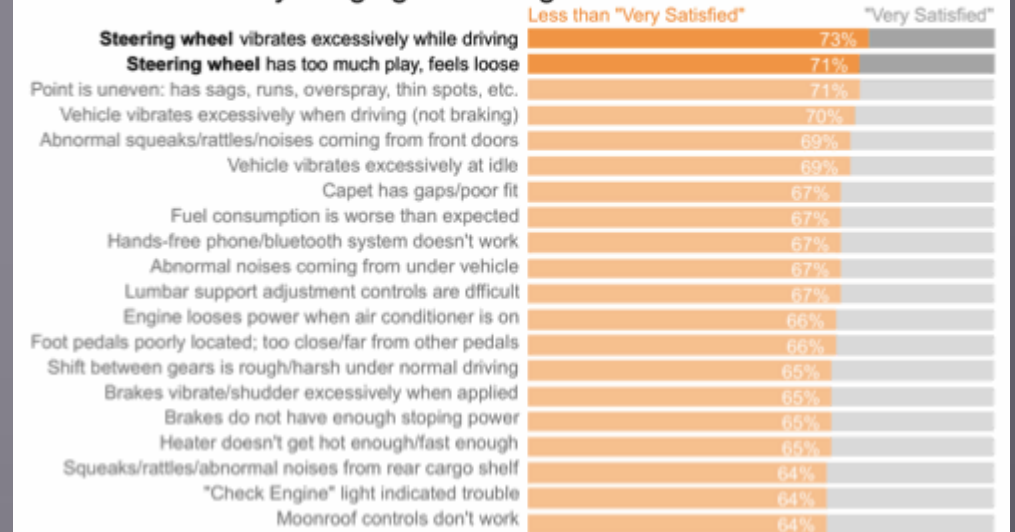
STORYTELLING WITH DATA

Satisfaction loss by things gone wrong

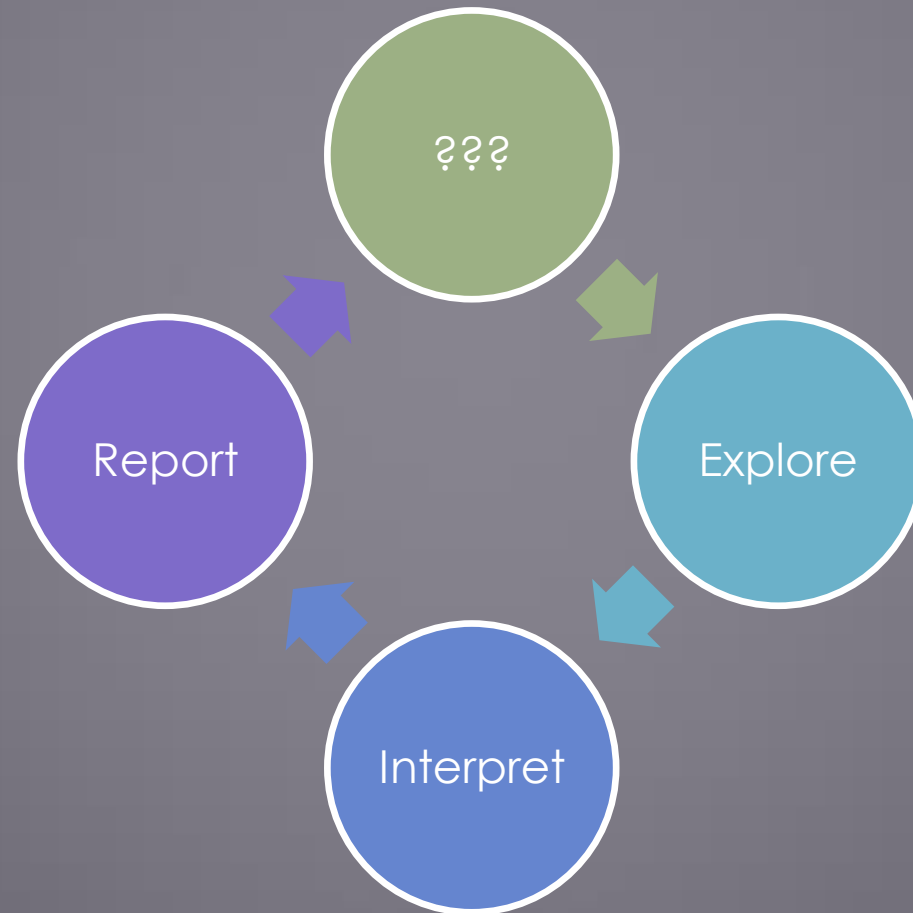


The **top 2 issues** with the highest percentage of non-complete satisfaction are **both steering wheel related**.

Satisfaction loss by things gone wrong



THE FOUR PHASES OF DATA ANALYSIS



THE FOUR PHASES OF DATA ANALYSIS



DATA!

CLEAN DATA

Checklist:

1. Each variable in the dataset is placed in its own column
2. Each observation is placed in its own row
3. Each value is placed in its own cell
4. Missing values are all treated the same way
5. Duplicate observations are removed
6. Data is stored as an appropriate type

The image shows three versions of a data table with annotations. The first table has vertical double-headed arrows between columns, labeled 'variables'. The second table has horizontal double-headed arrows between rows, labeled 'observations'. The third table has circles around individual cells, labeled 'values'.

country	year	cases	population
Afghanistan	1999	175	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	213766	1280425583

variables

country	year	cases	population
Afghanistan	1999	175	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	172006362
Brazil	2000	80488	174004898
China	1999	212258	1272015272
China	2000	213766	1280425583

observations

country	year	cases	population
Afghanistan	99	75	987071
Afghanistan	00	666	095360
Brazil	99	31737	172006362
Brazil	00	80488	174004898
China	99	212258	1272015272
China	00	213766	1280425583

values

CLEAN DATA

Clean data is...

- VALID** - not empty; correct data type
- ACCURATE** - conforming to true values
- COMPLETE** - missing values treated properly
- CONSISTENT** - items across dataset match
- UNIFORM** - same units of measurement or text formats