Data is (still) Messy

Giorgio Quer (adapted from Colin Jemmott)
DSC 96

Much of this is adapted from the outstanding "Quartz Bad Data Guide" https://github.com/Quartz/bad-data-guide

Last week: identifying messy data

- Are the data types correct?
- String type fields are have consistent values?
- No missing values that we don’t understand?
- All values look in a reasonable range?

The data was perfect, right? HA!

How do we deal with the messiness we found?

Last week: identifying messy data

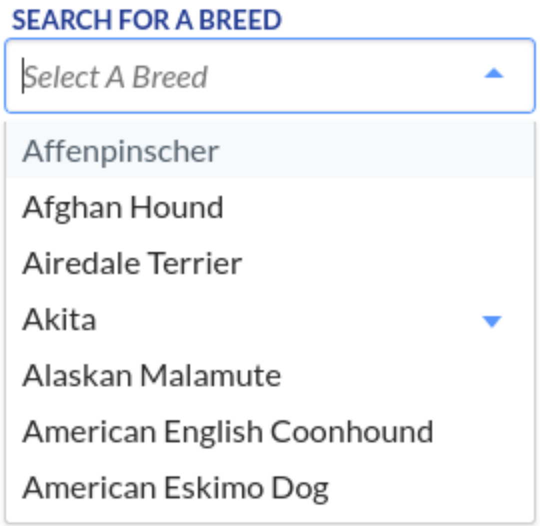
- Are the data types correct?
  - Mostly. Did a little convenience conversion
- String type fields are have consistent values?
  - Case Type, Sex, Ethnicity
  - Solutions: Re-map values (calculated field), filter values, etc...
- All values look in a reasonable range?
  - Age
  - Solutions: filter, smooth,...
- No missing values that we don’t understand?
  - Age, Time, Search, Arrested,....
  - Solutions: filter, imputation, create a new binary variable

Human entered data

The dog licensing website for Cook County, Illinois gave a text field to type your dog breed into. As a result this database contained at least 250 spellings of Chihuahua!

How can this be fixed?

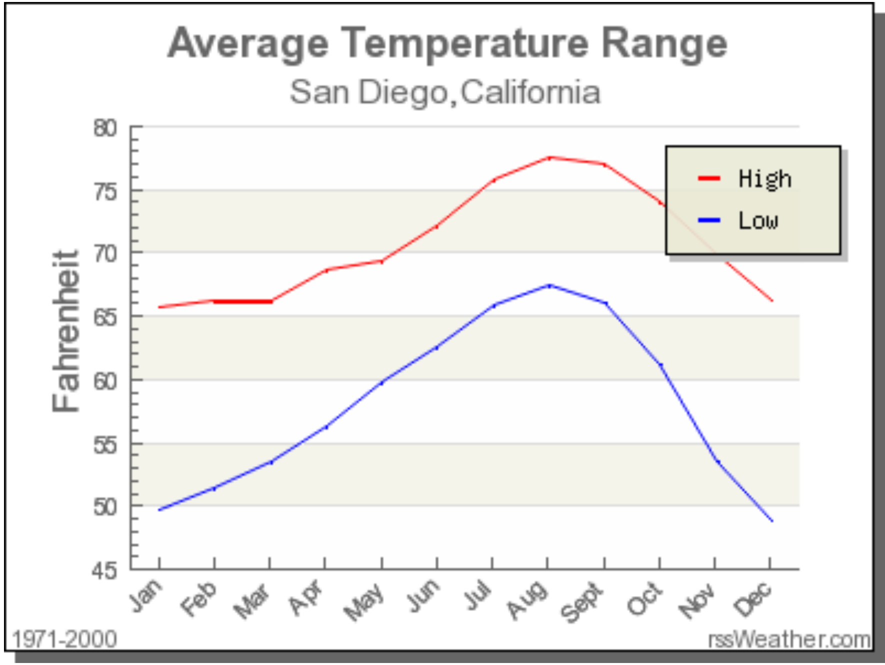
One solution: limit choices



Non-Stationary Data

The average low temperature in San Diego is 57 F (14 C). If it is July do you need to bring a sweater?

Sheldon graduated from UCSD CSE in 2010 and got an entry level job paying $60,000. After working his way up, he is now earning $68,000. That is more money, right?



Outliers and “Incorrect” Values

- Consistently “nonsense” values
  - Is it a product of the data ingestion process? Time field has year 1899? Is it an inferred “default” value?
  - Solution: Change the value to the correct one!
- Abnormal artifacts from the data collection process
  - E.g. unreasonable spikes in recorded ages at round numbers (25, 35, 45)
  - Solution: Try “smoothing” (e.g. binning the ages)
- Unreasonable outliers
  - Data points with unrealistic and highly unreasonable values. E.g. age=200
  - Solution: filter it? Maybe it points to bugs in the data collection? Maybe it’s **real** and you should investigate!

Missing data

vehicle_stops_2016_datasd

stop_id	stop_cause	service_area	subject_race	subject_sex	subject_age	timestamp	stop_date	stop_time	sd_resident	arrested	searched
1308198	Equipment Violation	530	W	M	28	2016-01-01 00:06:00	2016-01-01	0:06	Y	N	N
1308172	Moving Violation	520	B	M	25	2016-01-01 00:10:00	2016-01-01	0:10	N	N	N
1308171	Moving Violation	110	H	F	31	2016-01-01 00:14:00	2016-01-01	0:14			
1308170	Moving Violation	Unknown	W	F	29	2016-01-01 00:16:00	2016-01-01	0:16	N	N	N
1308197	Moving Violation	230	W	M	52	2016-01-01 00:30:00	2016-01-01	0:30	N	N	N
1308200	Moving Violation	710	H	M	24	2016-01-01 00:30:00	2016-01-01	0:30	Y	N	N
1308174	Moving Violation	Unknown	O	M	20	2016-01-01 00:35:00	2016-01-01	0:35	Y	N	N
1308199	Moving Violation	440	H	M	50	2016-01-01 00:45:00	2016-01-01	0:45	Y	N	N
1308979	Moving Violation	310	H	F	25	2016-01-01 01:03:00	2016-01-01	1:03	Y	N	Y
1308965	Moving Violation	240	W	F	23	2016-01-01 01:10:00	2016-01-01	1:10	Y	N	N
1308175	Moving Violation	120	O	M	54	2016-01-01 01:20:00	2016-01-01	1:20	Y	N	N
1308176	Moving Violation	520	W	F	53	2016-01-01 01:39:00	2016-01-01	1:39	Y	N	N
1308177	Moving Violation	520	W	M	35	2016-01-01 01:57:00	2016-01-01	1:57	N	N	N
1308178	Moving Violation	520	W	M	29	2016-01-01 02:00:00	2016-01-01	2:00	N	Y	N
1308180	Moving Violation	510	B	M	38	2016-01-01 03:24:00	2016-01-01	3:24	Y	N	N
1308182	Moving Violation	310	W	M	24	2016-01-01 06:40:00	2016-01-01	6:40	Y	N	N
1308969	Moving Violation	Unknown	W	F	38	2016-01-01 06:45:00	2016-01-01	6:45	Y	N	N
1308181	Equipment Violation	830	H	M	18	2016-01-01 06:50:00	2016-01-01	6:50			
1308191	Moving Violation	230	W	M	25	2016-01-01 07:52:00	2016-01-01	7:52	N	N	N

Missing data

- Missing by Design (MD)
  - The field being absent is deterministic.
- Missing Completely at Random (MCAR)
  - The missing value isn't associated to the (actual, unreported) value itself, nor the values in any other fields.
  - The participants with completely observed data are in effect a random sample of all the participants
  - The analysis performed on the data is unbiased
  - Example: additional questions in a survey are posed on a random sample of respondents
- Missing at Random (MAR)
  - A missing value may depend on values of other fields, but not its own
  - Example: service workers are less likely to report income.
- Not Missing at Random (NMAR)
  - A missing value depends on the value of the (actual, unreported) variable that's missing.
  - Example: people with high income are less likely to report income.

Missing data

- See example ipython!

- Missing by Design (MD)
  - The field being absent is deterministic.
- Missing Completely at Random (MCAR)
  - The missing value isn't associated to the (actual, unreported) value itself, nor the values in any other fields.
  - The participants with completely observed data are in effect a random sample of all the participants
  - The analysis performed on the data is unbiased
  - Example: additional questions in a survey are posed on a random sample of respondents
- Missing at Random (MAR)
  - A missing value may depend on values of other fields, but not its own
  - Example: service workers are less likely to report income.
- Not Missing at Random (NMAR)
  - A missing value depends on the value of the (actual, unreported) variable that's missing.
  - Example: people with high income are less likely to report income.

Null Values: MD, MCAR, MAR, NMAR?

- Attrition due to natural processes?
- Built into the data collection process (intentional)?
- Random issues in (the mechanics of) the data collection process.
- Non-response or refusal

It’s very tricky to distinguish between these with certainty!

Can you come up with examples from SDPD dataset?

Null Value Imputation (what to do about them)

- Missing by Design
  - Fill them in? Drop them? Recode the variable?
- Missing Completely at Random (MCAR)
  - Dropping them is ok (if there aren’t too many)
- Missing at Random (MAR)
  - Careful! Dropping data will skew your dataset!
  - Replace with mean/mode (perhaps by an associated group)
  - Train a model to replace the missing values
- Not Missing at Random (NMAR)
  - Difficult! Proceed with caution!
  - Train a model to replace the missing values