UCSD

DSC 96

Spring 2019

Class 3

Data Is Messy

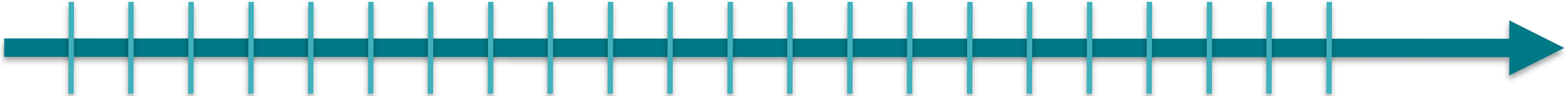
Data is Messy

Colin Jemmott
and
Giorgio Quer
DSC 96

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

Data definition

- Data is too coarse:
    - You needs months, but you only have years
- Data is too granular:
    - You have daily “number of steps”, but you need monthly steps for your statistical analysis



Data collection problems

- We have a great dataset:
  - Physical activity for 1 year from 10M people in US with an activity tracker!
  - We want to describe the physical activity of US citizens !
  - Can we?

Data collection problems

- We have a great dataset:
  - Physical activity for 1 year from 10M people in US who bought an activity tracker!
  - We want to describe the physical activity of US citizens !
  - Can we?
- Ok, let’s collect the data properly:
  - 1000 people randomly selected (any age or physical status or income) in San Diego county
  - 3 months of data (May, June, July)
  - Are we ok now?

Data collection problems

- Sample is not random
    - You have the number of steps, but the population is composed of very active people

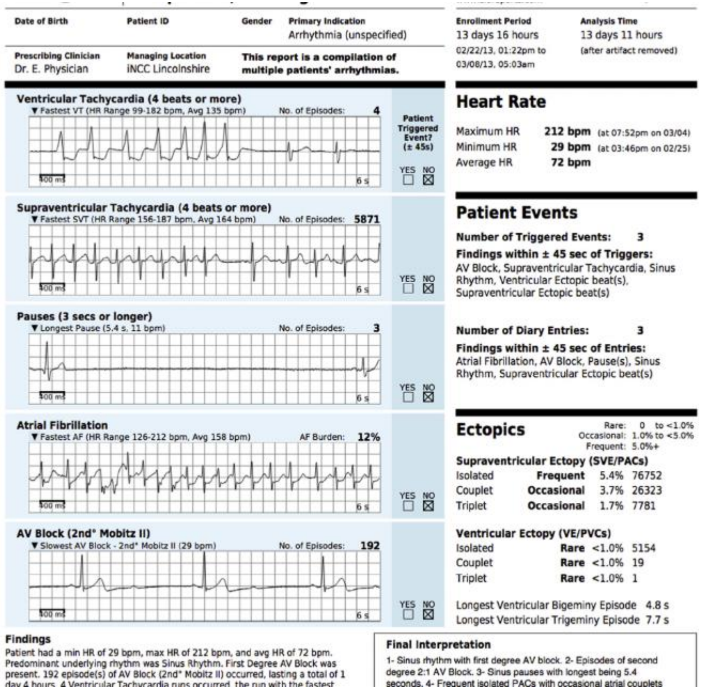
- Seasonal variation
    - You have number of steps from a good population, but only in summer time

- Results are p-hacked
    - The data collection stopped once a significant result

Other data types

Data doesn’t always come in in nicely formatted packages.

- CSV, escaping, and the lack of standards
- Data are in a PDF - what now?
- Images and sound recordings as data



Date of Birth	Patient ID	Gender	Primary Indication
			Arrhythmia (unspecified)

Prescribing Clinician	Managing Location	
Dr. E. Physician	iNCC Lincolnshire	**This report is a compilation of multiple patients' arrhythmias.**

Enrollment Period	Analysis Time
13 days 16 hours 02/22/13, 01:22pm to 03/08/13, 05:03am	13 days 11 hours (after artifact removed)

Ventricular Tachycardia (4 beats or more)
▼ Fastest VT (HR Range 99-182 bpm, Avg 135 bpm) No. of Episodes: **4**
Patient Triggered Event? (± 45s) YES □ NO ☒

Supraventricular Tachycardia (4 beats or more)
▼ Fastest SVT (HR Range 156-187 bpm, Avg 164 bpm) No. of Episodes: **5871**
YES □ NO ☒

Pauses (3 secs or longer)
▼ Longest Pause (5.4 s, 11 bpm) No. of Episodes: **3**
YES □ NO ☒

Atrial Fibrillation
▼ Fastest AF (HR Range 126-212 bpm, Avg 158 bpm) AF Burden: **12%**
YES □ NO ☒

AV Block (2nd° Mobitz II)
▼ Slowest AV Block - 2nd° Mobitz II (29 bpm) No. of Episodes: **192**
YES □ NO ☒

Heart Rate

Maximum HR	**212 bpm**	(at 07:52pm on 03/04)
Minimum HR	**29 bpm**	(at 03:46pm on 02/25)
Average HR	**72 bpm**	

Patient Events

Number of Triggered Events: 3

Findings within ± 45 sec of Triggers:
AV Block, Supraventricular Tachycardia, Sinus Rhythm, Ventricular Ectopic beat(s), Supraventricular Ectopic beat(s)

Number of Diary Entries: 3

Findings within ± 45 sec of Entries:
Atrial Fibrillation, AV Block, Pause(s), Sinus Rhythm, Supraventricular Ectopic beat(s)

Ectopics

Rare:	0 to <1.0%
Occasional:	1.0% to <5.0%
Frequent:	5.0%+

Supraventricular Ectopy (SVE/PACs)

Isolated	**Frequent**	5.4%	76752
Couplet	**Occasional**	3.7%	26323
Triplet	**Occasional**	1.7%	7781

Ventricular Ectopy (VE/PVCs)

Isolated	**Rare**	<1.0%	5154
Couplet	**Rare**	<1.0%	19
Triplet	**Rare**	<1.0%	1

Longest Ventricular Bigeminy Episode	4.8 s
Longest Ventricular Trigeminy Episode	7.7 s

Findings
Patient had a min HR of 29 bpm, max HR of 212 bpm, and avg HR of 72 bpm. Predominant underlying rhythm was Sinus Rhythm. First Degree AV Block was present. 192 episode(s) of AV Block (2nd° Mobitz II) occurred, lasting a total of 1 day 4 hours. 4 Ventricular Tachycardia runs occurred, the run with the fastest

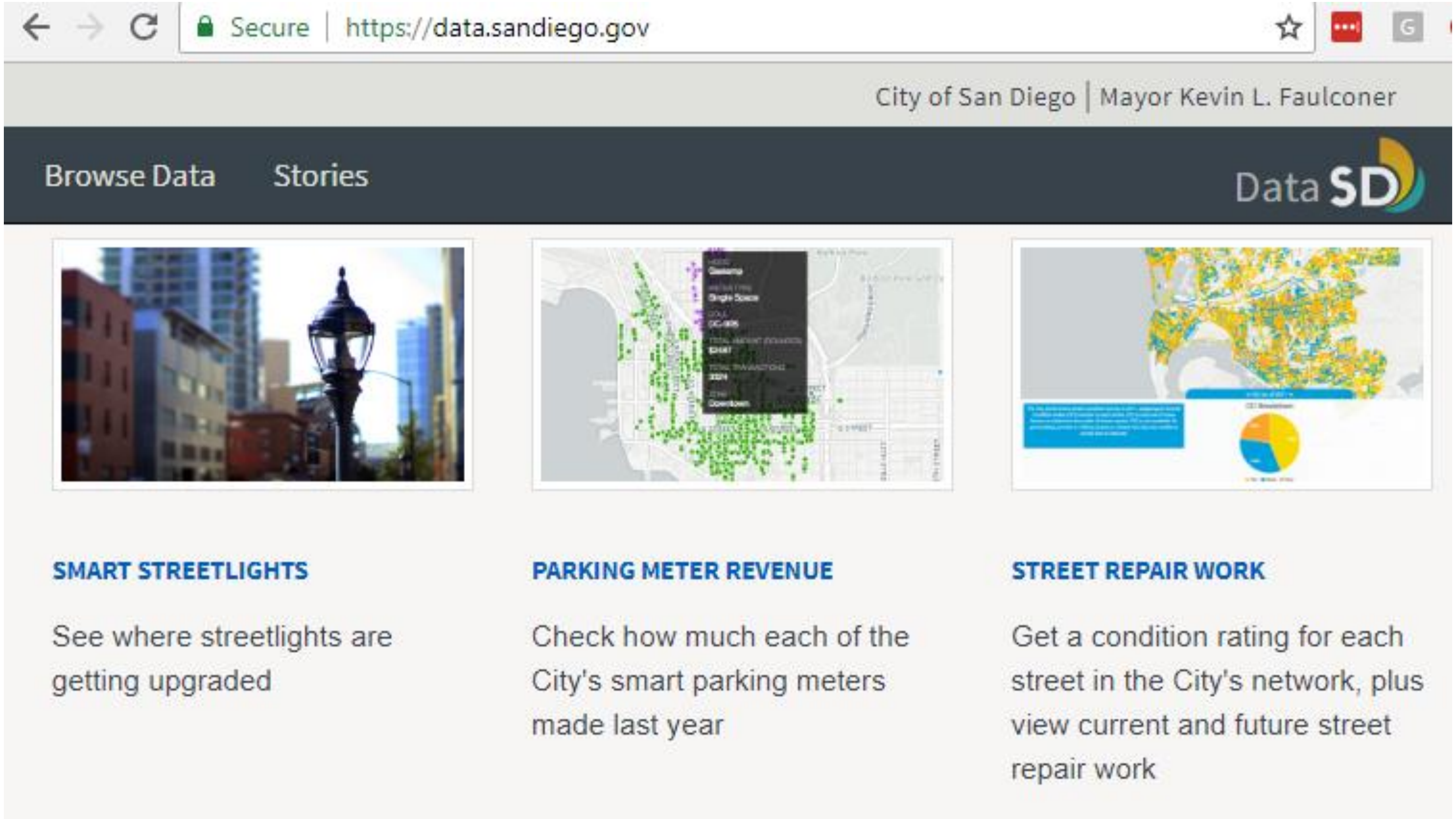
Final Interpretation
1- Sinus rhythm with first degree AV block. 2- Episodes of second degree 2:1 AV Block. 3- Sinus pauses with longest being 5.4 seconds. 4- Frequent isolated PACs with occasional atrial couplets

from: Barrett et al, "Comparison of 24-hour Holter Monitoring with 14-dayNovel Adhesive Patch Electrocardiographic Monitoring"

Vehicle Stop Data

DSC 96

Data Source



Secure
https://data.sandiego.gov
City of San Diego | Mayor Kevin L. Faulconer
Browse Data
Stories
Data SD

SMART STREETLIGHTS

See where streetlights are getting upgraded

PARKING METER REVENUE

Check how much each of the City's smart parking meters made last year

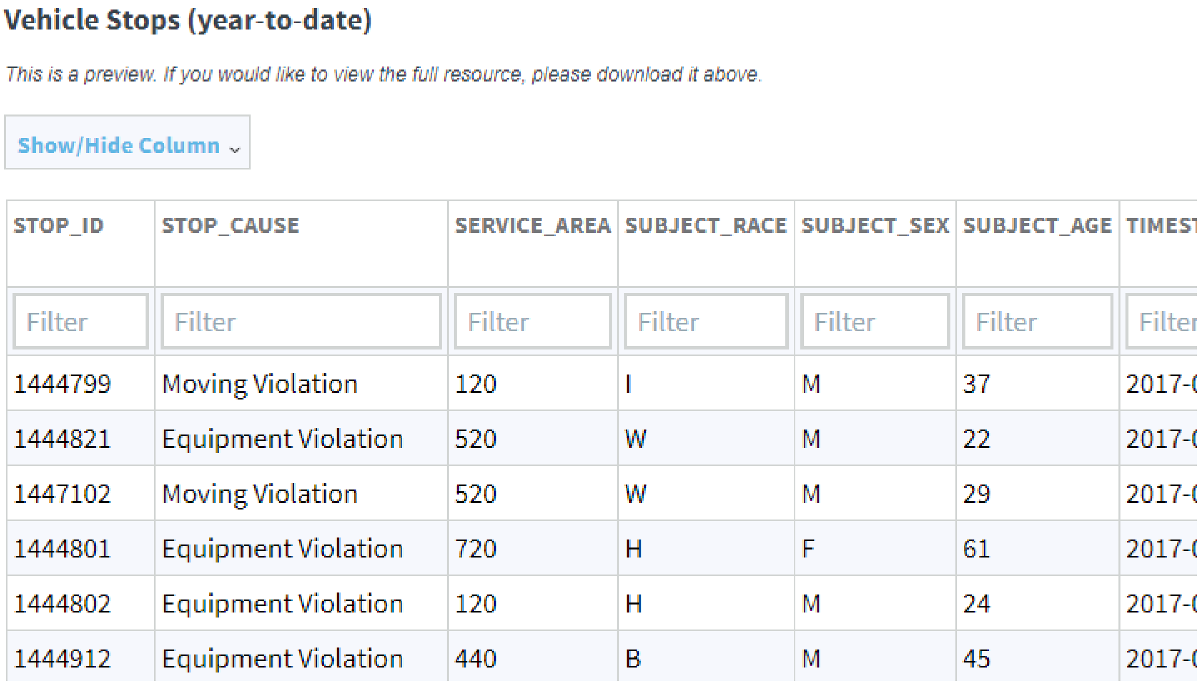
STREET REPAIR WORK

Get a condition rating for each street in the City's network, plus view current and future street repair work

Why Police Data?

Police Vehicle Stops

Vehicle stops made by the San Diego Police Department. Vehicle Stops files contain all vehicle stops for a given year.



Vehicle Stops (year-to-date)

This is a preview. If you would like to view the full resource, please download it above.

STOP_ID	STOP_CAUSE	SERVICE_AREA	SUBJECT_RACE	SUBJECT_SEX	SUBJECT_AGE	TIMEST
Filter	Filter	Filter	Filter	Filter	Filter	Filter
1444799	Moving Violation	120	I	M	37	2017-
1444821	Equipment Violation	520	W	M	22	2017-
1447102	Moving Violation	520	W	M	29	2017-
1444801	Equipment Violation	720	H	F	61	2017-
1444802	Equipment Violation	120	H	M	24	2017-
1444912	Equipment Violation	440	B	M	45	2017-

SDPD Vehicle Stop Data

1. Plot count of stops by age. Notice any issues? What should we do?
2. Make some time series plots! For example, stops by hour of day, day of week, month, etc. might be interesting.
3. Explore the “stop cause” variable. Notice any issues? What should we do?

Finally, explore and answer questions. When you find bad data, bring it up to the class.

Other info on the vehicle_stop dataset

- Where is it?
  - https://github.com/gquer/dsc-96_winter19/tree/master/02_data_messy/data

- Where do we start?
  - https://github.com/gquer/dsc-96_winter19/blob/master/02_data_messy/README.md

Data is (still) Messy

Giorgio Quer and Colin Jemmott
DSC 96

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

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?

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!

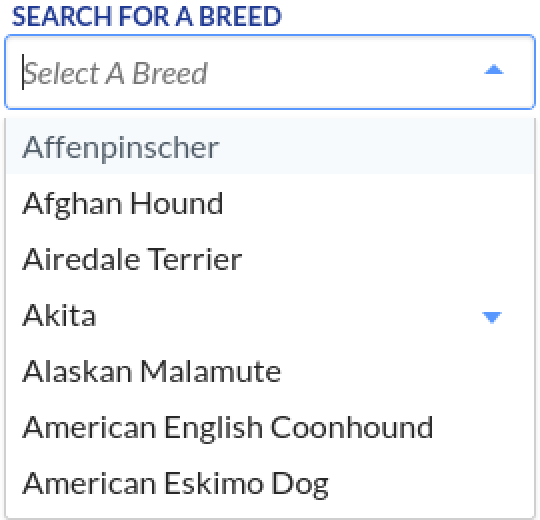
How can this be fixed?

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!

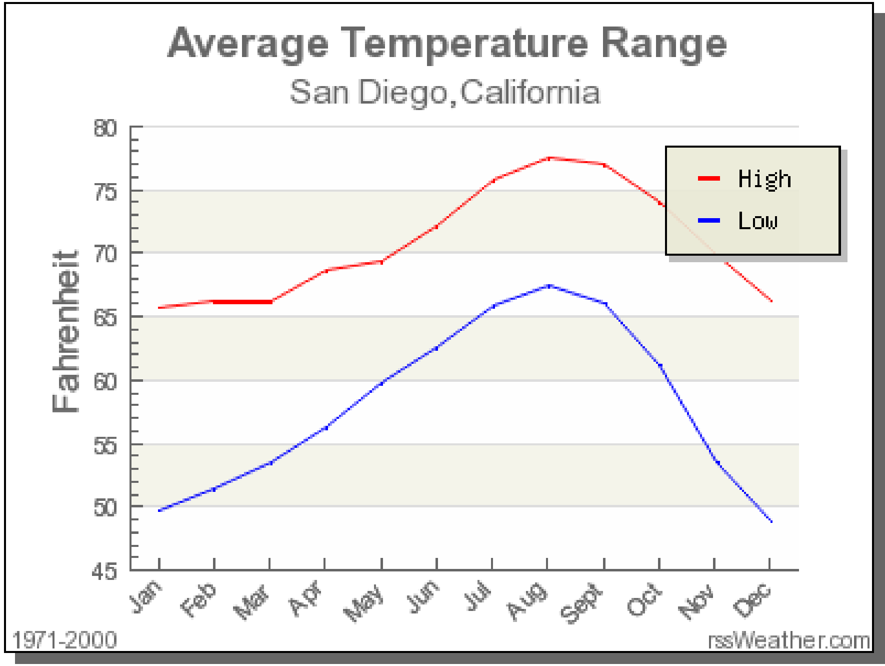
How can this be fixed?

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?

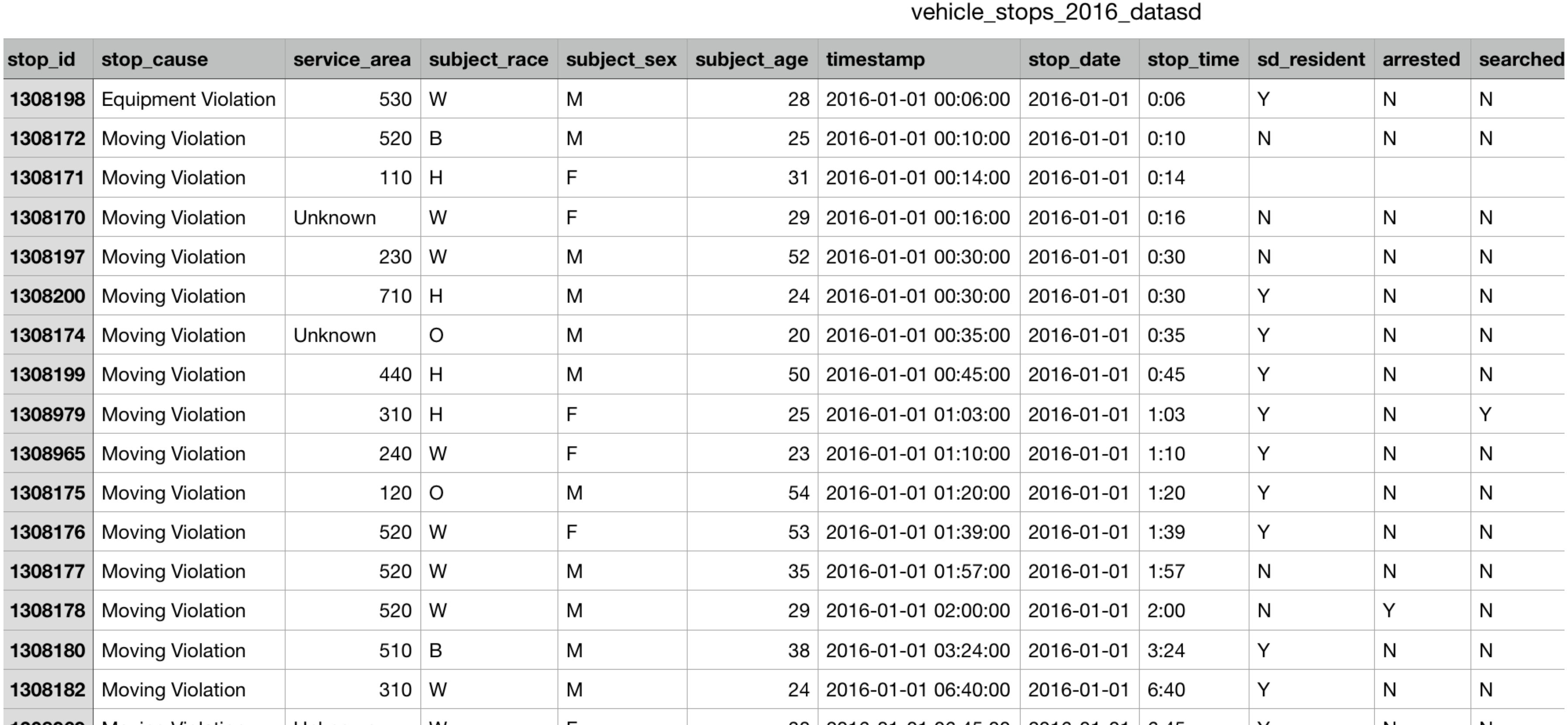


Sheldon graduated from UCSD CSE in 2010 and got an entry level job paying $60,000. After working his way up, he is now earning $68,000. That is more money, right?

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
[illegible]	Moving Violation	Unknown	W	F	[illegible]	2016-01-01 06:45:00	2016-01-01	6:45	Y	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)

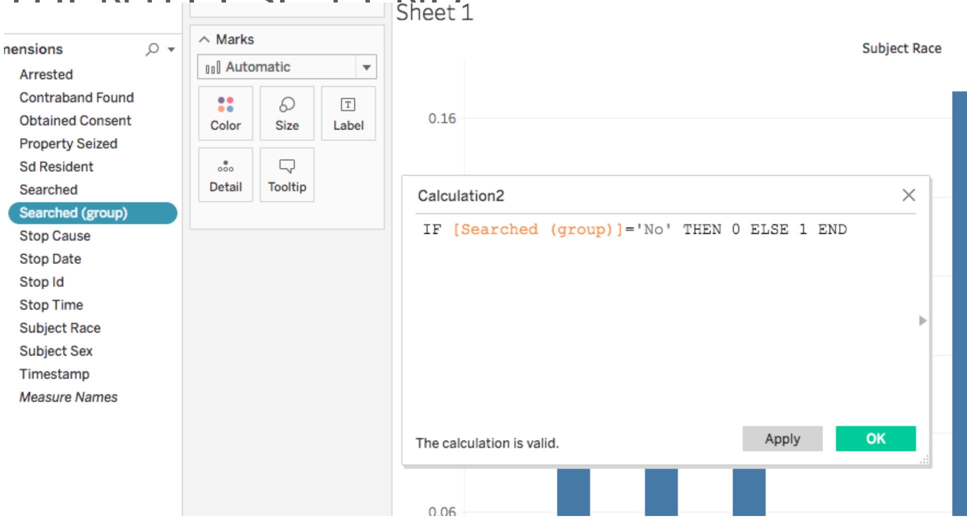
- Missing by Design
    - Fill them in? Drop them? Recode the variable?
- Missing Completely at Random (MCAR)
    - Dropping them is ok (if there aren’t too many)
- Missing at Random (MAR)
    - Careful! Dropping data will skew your dataset!
    - Replace with mean/mode (perhaps by an associated group)
    - 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

SD police stop data

1. age:
   - how are they distributed?
   - What issues you observe? anything strange?
   - Divide by sex and age
2. ethnicity:
   - which races do you see? Can you rename them?
   - which are more represented? should we group them?
   - stop vs searched (or arrested): anything conclusion we can see here?
3. time series plot:
   - plot by quarter, month, day.. any issue you see?
   - Plot by minute?
   - are there any abnormality low/high to discuss?

Searched

- Data is Y y N n Null
- Group them: create - group, N n and Null in the same group
- Change format:
  - Create - Calculated Field:
    - IF [Searched (group)]='No' THEN 0 ELSE 1 END
- Move to Measures
- Now we can calculate the average !!!



```
IF [Searched (group)]='No' THEN 0 ELSE 1 END
```

The calculation is valid.