Data 100

Lecture 6: Unboxing the Data

Congratulations!

Box of Data

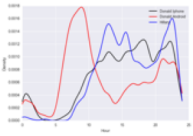
You have **collected** or **been given** a box of data

What do you do next?

Begin predictive modeling and hypothesis testing?

Data Acquisition

Preparing Data for Analysis



Exploratory Data Analysis

Topics For Today

- Understanding the Data
  1. File Format & Organization
  2. Granularity
  3. Quality
- Common Data Anomalies
  - … and how to fix them

Where did the box come from?

Box of Data

Data Deluge:

- Collected from an instrument
- Federal Survey
- Scraped from the Web
- Administrative data set

Three Examples

- SCIENTISTS collected: CO_2 Measurements from the observatory at Mauna Loa
- GOVERNMENT collected: Drug Abuse Warning Network survey
- INFORMAL collection: Housing sale prices in the Bay Area right before the 2009 economic downturn https://www.sfgate.com/homesales/c/a/2008/04/27/REHS.tbl

Three Examples

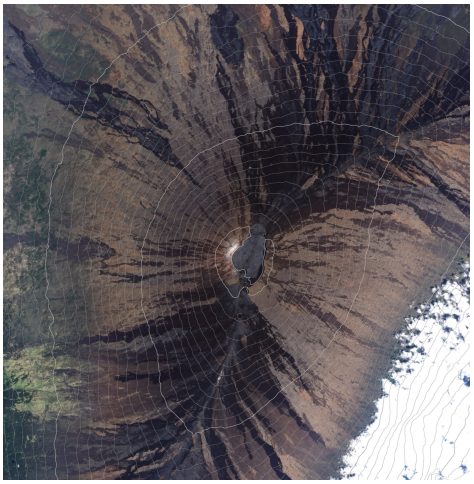
- SCIENTIFIC: tend to be clean and well documented
- GOVERNMENT: tend to be clean, well documented, and categorical
- INFORMAL: tend to need cleaning and not well documented

CO_2 levels at Mauna Loa Observatory

Data Collected by Scientific Instruments

Mauna Loa Volcano

Largest Volcano in world
4 km above sea level
Summit 17 km above base
On the Island of Hawaii



Data and photos available from Scripps Institute and NOAA

Sampling Frame – Mauna Loa Observatory

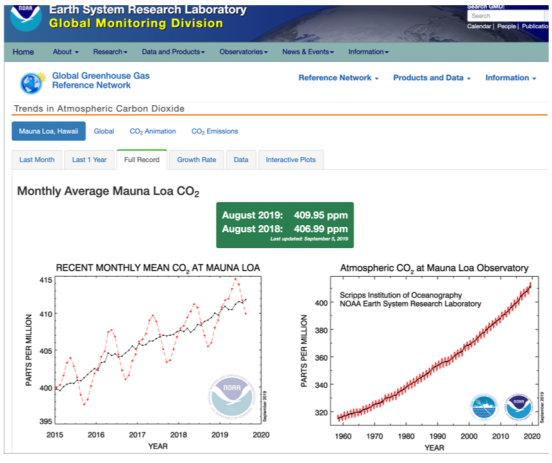
Far from any continent: air sampled is representative of the central Pacific.

High altitude: above the inversion layer where local effects may be present.

Measurements of atmospheric CO_2 since 1958 – longest continuous record

Acquiring the box of data

- Clean
- Well documented
- Simple structure
- Broadly shared
- Reproducibility is key to trusting findings



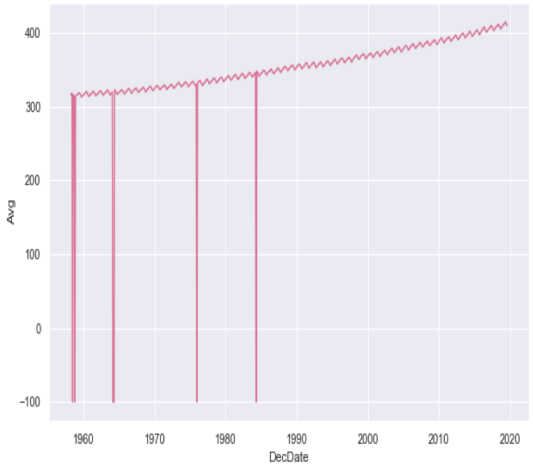
https://www.esrl.noaa.gov/gmd/ccgg/trends/

Start modeling the change in CO2 over time!

YIKES!

What happened?

We didn't clean our data…



Here's the Data: co2_mm_mlo.txt

Start Over with More Care

NOW WHAT?

- How big is it?
- What is the encoding?
- How is it formatted?

Look at it

```
!file data/co2_mm_mlo.txt
```

```
data/co2_mm_mlo.txt: ASCII text
```

```
!wc data/co2_mm_mlo.txt
```

```
     810    5804   51131 data/co2_mm_mlo.txt
```

```
!head -n 10 data/co2_mm_mlo.txt
```

```
# ----------------------------------------------------------------------
# USE OF NOAA ESRL DATA
#
# These data are made freely available to the public and the
# scientific community in the belief that their wide dissemination
# will lead to greater understanding and new scientific insights.
# The availability of these data does not constitute publication
# of the data.  NOAA relies on the ethics and integrity of the user to
# ensure that ESRL receives fair credit for their work.  If the data
# are obtained for potential use in a publication or presentation,
```

These are Unix commands that we run from the Jupyter notebook

There are similar Python commands in the utils library, e.g., line_count() head()

Look At It

What do you see?
Make 4 observations about these data

```
#
# NOTE: In general, the data presented for the last year are subject to change,
# depending on recalibration of the reference gas mixtures used, and other quality
# control procedures. Occasionally, earlier years may also be changed for the same
# reasons.  Usually these changes are minor.
#
# CO2 expressed as a mole fraction in dry air, micromol/mol, abbreviated as ppm
#
#  (-99.99 missing data;  -1 no data for #daily means in month)
#
#            decimal     average   interpolated    trend    #days
#              date                             (season corr)
1958   3    1958.208      315.71      315.71      314.62     -1
1958   4    1958.292      317.45      317.45      315.29     -1
1958   5    1958.375      317.50      317.50      314.71     -1
1958   6    1958.458      -99.99      317.10      314.85     -1
1958   7    1958.542      315.86      315.86      314.98     -1
1958   8    1958.625      314.93      314.93      315.94     -1
1958   9    1958.708      313.20      313.20      315.91     -1
1958  10    1958.792      -99.99      312.66      315.61     -1
```

Observations about the file

- File appears to be plain text
- Column names appear on two lines of file
- Fields line up from one row to the next
- White space between fields
- Seven variables
- -99.99 appears in some rows for the "Average"
- -1 appears in all the first 5 rows for "days"

Read the Data into a Data Frame

```
co2 = pd.read_csv('co2_mm_mlo.txt',
    header = None, skiprows = 72, sep = '\s+',
    names = ['Yr', 'Mo', 'DecDate', 'Avg', 'Int', 'Trend', 'days'])
```

	Yr	Mo	DecDate	Avg	Int	Trend	days
0	1958	3	1958.208	315.71	315.71	314.62	-1
1	1958	4	1958.292	317.45	317.45	315.29	-1
2	1958	5	1958.375	317.50	317.50	314.71	-1
3	1958	6	1958.458	-99.99	317.10	314.85	-1
4	1958	7	1958.542	315.86	315.86	314.98	-1

	Yr	Mo	DecDate	Avg	Int	Trend	days
733	2019	4	2019.292	413.32	413.32	410.49	26
734	2019	5	2019.375	414.66	414.66	411.20	28
735	2019	6	2019.458	413.92	413.92	411.58	27
736	2019	7	2019.542	411.77	411.77	411.43	23
737	2019	8	2019.625	409.95	409.95	411.84	29

Identify the Structure & Granularity

- What is the shape?
- What does a record represent?
- Have the data been aggregated,?
- Do we need to aggregate?

Identify the Structure & Granularity

- What is the shape?
  Rectangular –
  7 columns & 738 rows
- What does a record represent?
  One month of CO_2 measurements
- Have the data been aggregated,?
- Do we need to aggregate?
  Yes, they are aggregated to the month, via an average. We don't need to further aggregate.

Ideas for confirming data quality?

Can you think of some ways for us to check that the data are what we expect?

What about ways to check consistency between the variables?

Yr -- 4 digits, from 1958 to 2019
Mo – 1 to 12
DecDate – Jan 1, 1958 = 1958 + 1/365
Avg -- Average monthly CO2
Int – Interpolated CO2, if Avg is missing
Trend – fitted trend
days -- days in operation

Ideas for confirming data quality

- How many records should we have?

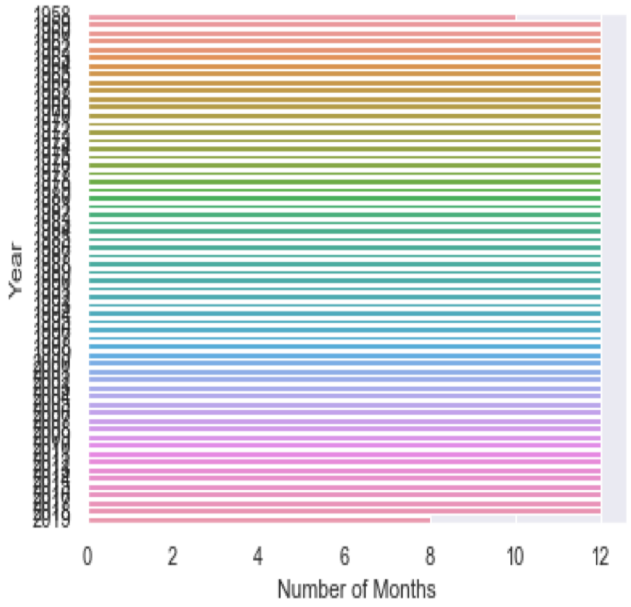
$$12 \times (2019 - 1957) - 4 - 2 = 738$$

- How many records for a month should we have?

$$(2019 - 1957) = 62 \text{ or for some } 61$$

- Are there any/many unusual values?

We saw -99.99 for Avg and -1 for days but we don't know how many there are. Do we care about the -1s?



Bar plot of record count for each year

Plots that we wouldn't normally use for analysis can check data quality

```
checkYr = co2.groupby(["Yr"]).count()
checkYr.tail()
```

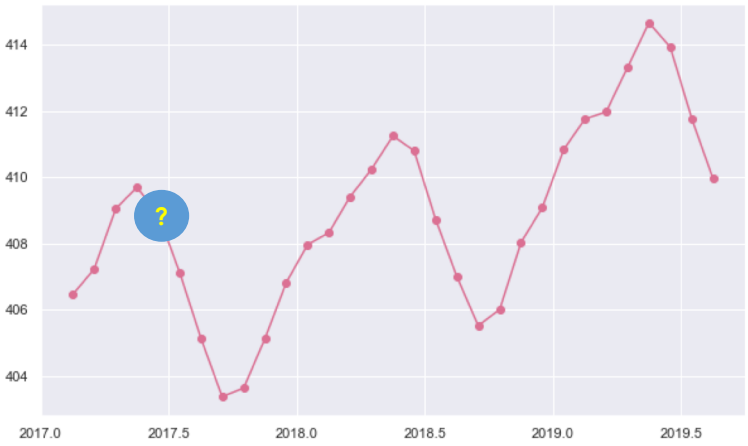
Yr	Mo	DecDate	Avg	Int	Trend	days
2015	12	12	12	12	12	12
2016	12	12	12	12	12	12
2017	12	12	12	12	12	12
2018	12	12	12	12	12	12
2019	8	8	8	8	8	8

Checking data quality

```
co2.describe()
```

	Yr	Mo	DecDate	Avg	Int	Trend	days
count	738.000000	738.000000	738.000000	738.000000	738.000000	738.000000	738.000000
mean	1988.417344	6.491870	1988.916667	350.472087	354.496057	354.483523	18.472900
std	17.768275	3.444944	17.765545	52.214201	28.113985	28.031320	12.200271
min	1958.000000	1.000000	1958.208000	-99.990000	312.660000	314.620000	-1.000000
25%	1973.000000	4.000000	1973.562750	328.587500	328.792500	329.730000	-1.000000
50%	1988.000000	6.000000	1988.916500	351.725000	351.725000	352.380000	25.000000
75%	2004.000000	9.000000	2004.271000	377.000000	377.000000	377.177500	28.000000
max	2019.000000	12.000000	2019.625000	414.660000	414.660000	411.840000	31.000000

Zoom in on a short time period



We see a seasonal component to CO2 measurements

Peak around April/May

Trough around Sep

June 2017 is Missing, What to do?

A. Ignore it, and hope it goes away
B. Drop the records with missing values
C. Replace with the average from the previous 6 months
D. Replace with a random June from the previous 6 years

yellkey.com/chair

What to do with the Missing Values?

- Drop the records with missing values
  - We typically selectively drop records for one analysis but not drop them for all analyses. For example, if a variable is not in a model then we don't drop records with missing values for that variable.
- Replace with an average value – mean imputation
  - Typically, we divide the data into subgroups that have the same values for certain variables (e.g., age, sex). Then we impute the missing value with the average value for the group
- Replace with a random value – hot deck imputation
  - Like mean imputation, we divide the data into subgroups. But, we choose a random value from the subgroup and use it for the missing value.

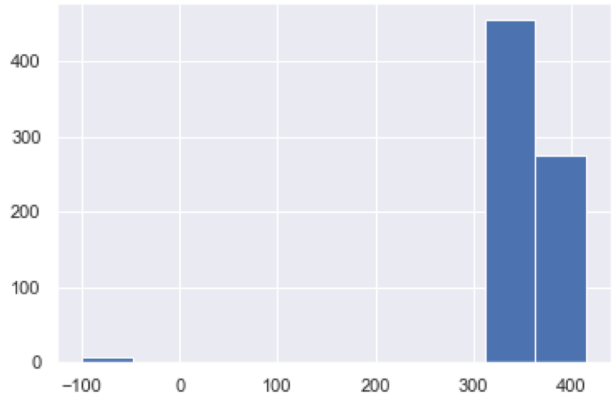
What happens?

- Ideally, the missingness is *at random* – meaning it is not correlated with other variables
- If missingness is correlated, that leads to biased inference
- If too many values for a field are missing, we may need to drop that field from our investigation
- If we impute with averages, the variability is reduced

How does it happen?

- A field of a record may be lost, hidden, removed, replaced, or never entered.
- That record's entity does not have a particular attribute.
  e.g., person without a permanent address.

CO_2 Monthly Average



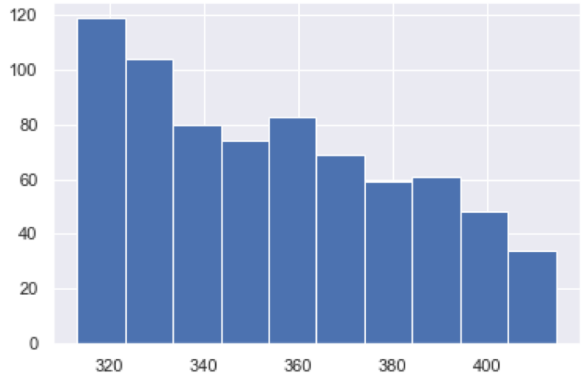

```
co2[co2["Avg"] < 0].count()
```

```
Yr          7
Mo          7
DecDate     7
Avg         7
Int         7
Trend       7
days        7
dtype: int64
```

The "Int" column contains values from "Avg" and interpolated values when "Avg" is missing.

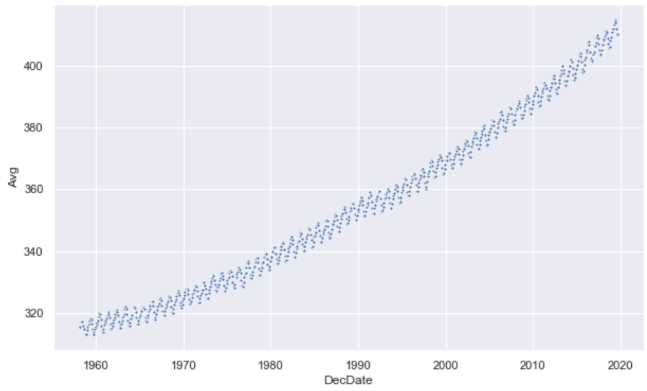
1. The average seasonal cycle in a 7-year window for months around the missing monthly value.
2. The trend is computed after removing the seasonal cycle and linearly interpolates missing months.
3. Missing month =
   avg seasonal cycle + the trend

We dropped the missing months



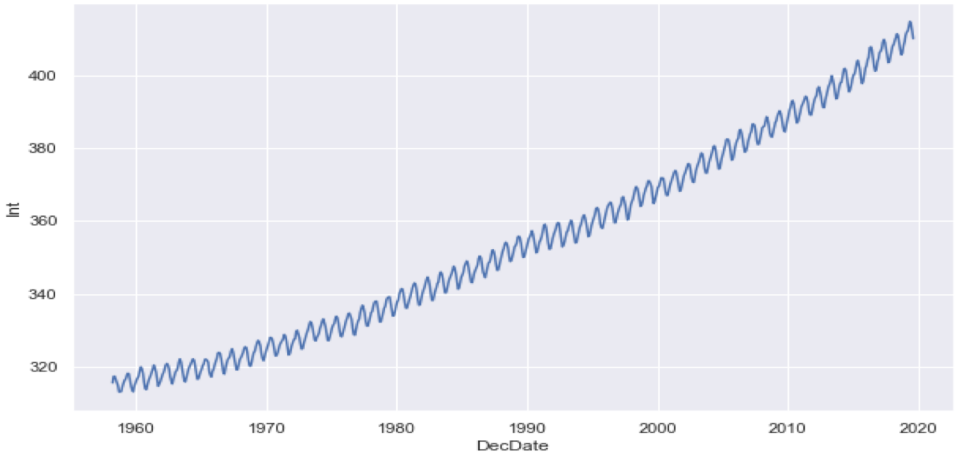
Does this plot make sense?

Line Plot– Pairs: (time, Avg CO_2)

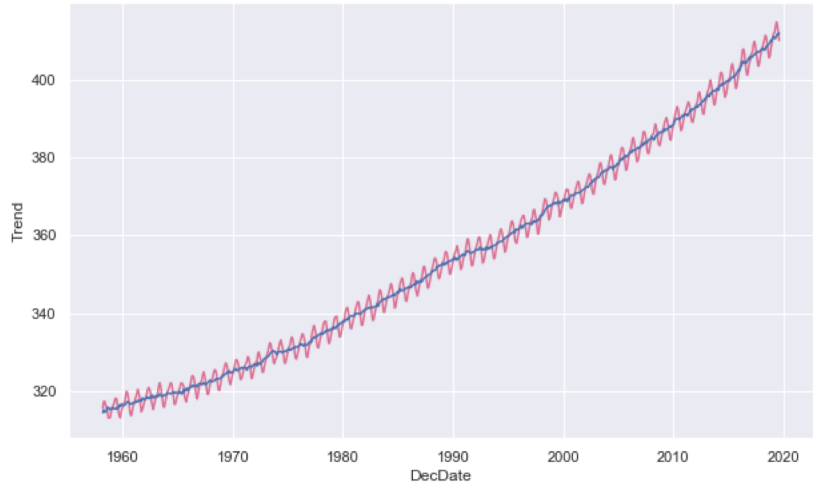


Points are typically not the best way to plot time series

Connect measurements with line segments



Seasonality and long-term Trend



We will return to this plot later, and have a debate around accurate visual representation

Recap Ideas in Data Cleaning

Revisit Sampling Frame

Sampling Frame (and Data)

- How complete/incomplete is the frame (and its data)?
- How is the frame/data situated in time?
- How is the frame/data situated in place?
- How well does the frame/data capture reality?

MLO - CO_2 Data

- Complete/incomplete?
- Situated in time?
- Situated in place?
- Capture reality?

7 records have a missing monthly average. We could use interpolated values or drop these records. – quite complete

Monthly records from Mar 1958 to Aug 2019 from Mauna Loa Observatory.

Unbox the Data

Unbox the Data

- How big is it?
- What is the encoding?
- How is it formatted?
- How is it organized?

MLO - CO_2 Data

- How big is it? 738 records
- Encoding? ASCII – plain text
- Formatting? White space and aligned fields
- Organized? Table with 7 columns

Data File Formats

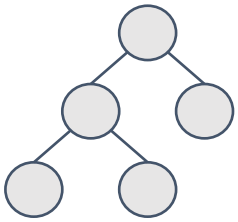
- Delimited values
  - comma, tab, white space
- Fixed-width format
- Key-value pairs

```
name:Tom
sex:m
age:77
ht:70
wt:175
bmi:25.16
overWt:TRUE
```

```
name,sex,age,ht,wt,bmi,overWt
Tom,m,77,70,175,25.16239,TRUE
Maya,f,33,64,124,21.50106,FALSE
Joe,m,79,73,185,24.45884,FALSE
Robert,m,47,67,156,24.48414,FALSE
Sue,f,27,61,98,18.51492,FALSE
Liz,f,33,68,190,28.94981,TRUE
```

```
Tom      m777017525.16239TRUE
Maya     f336412421.50106FALSE
Joe      m797318524.45884FALSE
Robert   m476715624.48414FALSE
Sue      f2761 9818.51492FALSE
Liz      f336819028.94981TRUE
```

JSON, XML, HTML, etc.



There are many formats to represent structured, hierarchical data

- Most formats consist of values, lists, and dictionaries?
- Dictionaries are typically keyed by strings
- Structures vary: list of records, list of columns, tree of documents, etc

Log Files

Sometimes we need to work harder to extract fields from less structured text

```
169.237.46.168 - - [26/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04 HTTP/1.1" 301 328
"http://anson.ucdavis.edu/courses/"  "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)"
```

```
169.237.6.168 - - [8/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04/ HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/" "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)"
```

Identify Granularity

For Tabular Data

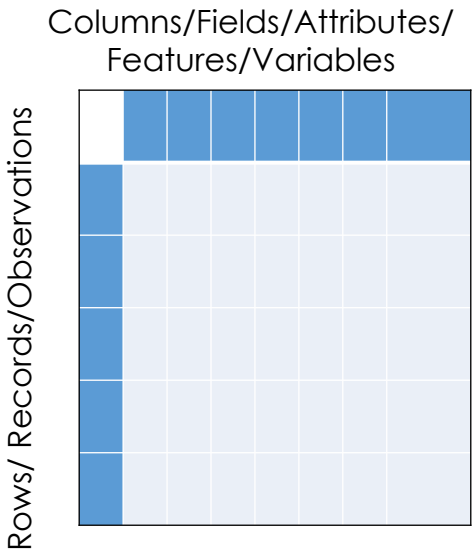
Identify Granularity

- What does a record represent?
  - Do we have the correct number of records?
  - Are there duplicated records?
- Have the data been aggregated?
  - Do the data need to be aggregated?
- Do the data need to be combined from multiple tables?

Understanding and manipulating granularity can help reveal patterns and relationships

Rectangular Data

- Easy manipulate, visualize, and analyze
- Easy to combine multiple tables
- Tabular representation of records and fields is a common paradigm



Rectangular Data

Two main variants

1. **Tables** (a.k.a. data-frames in R/Python and relations in SQL)
   - Named columns with different types
   - Manipulate using data transformation languages
     - map, filter, group by, join, sort
2. **Matrices**
   - All values have the same type
   - Manipulate using multiplication, addition, and element-wise operations
   - Most useful manipulation is linear, described by *linear algebra*

Granularity: Keys

- Often data will appear in multiple tables
- **Primary key:** *the column or set of columns in a table that determine the values of the remaining columns*
  - Primary keys are unique
  - Examples: SSN, ProductIDs, …
- **Foreign keys:** the column or sets of columns that reference primary keys in other tables.

Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Foreign Key →

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

ProdID	Cost
42	3.14
999	2.72

Primary Key →

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Check Quality

Check Quality

- Are the data values reasonable?
- Are there missing or corrupted values?
- Are the value-codings useful for analysis?
- Do we need to extract a feature from a complex value?
- Do field dependencies check out?

Check Quality & Clean

- The process of transforming "raw" data to enable subsequent analysis
- Data cleaning often addresses
  - Formatting
  - Missing values
  - Units
  - String Parsing
  - …
- Data cleaning is a big part of data science

Drug Abuse Warning Network (DAWN)

US Government Survey

SAMHDA

Substance Abuse & Mental Health Data Archive

SAMHSA.gov Contact Us

Search SAMHDA Search

SAMHDA HOME | ABOUT | DATA | LATEST | ANALYZE | FAQS

Drug Abuse Warning Network (DAWN)

The Drug Abuse Warning Network (DAWN) is a nationally represented public health surveillance system that continuously monitors drug-related visits to hospital emergency departments (EDs). A DAWN case is any ED visit involving recent drug use that is implicated in the ED visit. DAWN captures both ED visits that are directly caused by drugs and those in which drugs are a contributing factor, but not the direct cause of the ED visit. Annually, DAWN produces estimates of drug-related visits to hospital EDs for the nation as a whole and for selected metropolitan areas.

DAWN is used to monitor trends in drug misuse and abuse, identify the emergence of new substances and drug combinations, assess health hazards associated with drug abuse, and estimate the impact of drug misuse and abuse on the Nation's health care system. DAWN relies on a longitudinal probability sample of hospitals located throughout the United States.

To be eligible for selection into the DAWN sample, a hospital must be a non-federal, short-stay, general surgical and medical hospital located in the United States, with at least one 24-hour ED. The dataset includes demographics, drugs involved in the ED visit (up to 16 drugs from 2004 through 2008 and up to 22 drugs from 2009 through 2011), toxicology confirmation, route of administration, type of case, and disposition of the patient following the visit.

Prepared DAWN Emergency Department National and Metro data tables are available on the DAWN website. The DAWN website also provides access to DAWN reports.

Studies in this Series

DAWN-2011
Drug Abuse Warning Network

DAWN-2010
Drug Abuse Warning Network

DAWN-2009
Drug Abuse Warning Network

DAWN-2008
Drug Abuse Warning Network

DAWN-2007
Drug Abuse Warning Network

DAWN-2006
Drug Abuse Warning Network

DAWN-2005
Drug Abuse Warning Network

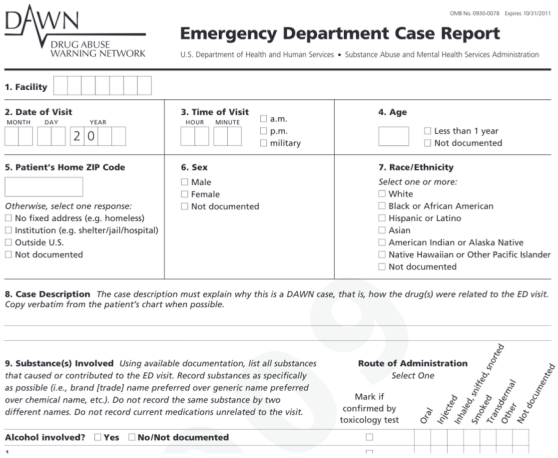
DAWN-2004
Drug Abuse Warning Network

https://www.datafiles.samhsa.gov/study-series/drug-abuse-warning-network-dawn-nid13516

Survey

The DAWN survey takes a probability sample of hospitals.

The hospitals must be:
- non-federal,
- short-stay,
- general surgical and medical hospital
- located in the United States,
- with a 24-hour Emergency Room.



DAWN
DRUG ABUSE WARNING NETWORK

Emergency Department Case Report

U.S. Department of Health and Human Services • Substance Abuse and Mental Health Services Administration

OMB No. 0930-0078 Expires: 10/31/2011

1. Facility

2. Date of Visit MONTH DAY YEAR 20

3. Time of Visit HOUR MINUTE ☐ a.m. ☐ p.m. ☐ military

4. Age ☐ Less than 1 year ☐ Not documented

5. Patient's Home ZIP Code

Otherwise, select one response:
☐ No fixed address (e.g. homeless)
☐ Institution (e.g. shelter/jail/hospital)
☐ Outside U.S.
☐ Not documented

6. Sex
☐ Male
☐ Female
☐ Not documented

7. Race/Ethnicity
Select one or more:
☐ White
☐ Black or African American
☐ Hispanic or Latino
☐ Asian
☐ American Indian or Alaska Native
☐ Native Hawaiian or Other Pacific Islander
☐ Not documented

8. Case Description *The case description must explain why this is a DAWN case, that is, how the drug(s) were related to the ED visit. Copy verbatim from the patient's chart when possible.*

9. Substance(s) Involved *Using available documentation, list all substances that caused or contributed to the ED visit. Record substances as specifically as possible (i.e., brand [trade] name preferred over generic name preferred over chemical name, etc.). Do not record the same substance by two different names. Do not record current medications unrelated to the visit.*

Route of Administration *Select One*

Mark if confirmed by toxicology test | Oral | Injected | Inhaled, sniffed, snorted | Smoked | Transdermal | Other | Not documented

Alcohol involved? ☐ Yes ☐ No/Not documented

Sampling Frame

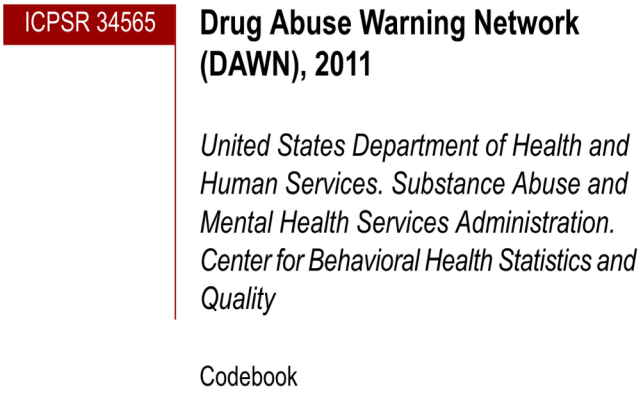
SAMHDA
Substance Abuse & Mental Health Data Archive

- Frame - ER visit for a drug related reason
- Situated in time – 2011
- Situated in place – Emergency Rooms in the US
- Capture reality - *complex sampling scheme based on probability*

Acquiring the

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

- Clean
- Well documented
  - **2356 page codebook!**
- Simple format
- Public Use data



Codebook

CASETYPE **TYPE OF VISIT**

Location: 1214-1214 (width: 1; decimal: 0)

Variable Type: numeric

Value	*Label*	*Unweighted Frequency*	*%*	*Valid %*
1	SUICICDE ATTEMPT:(1)	9033	3.9 %	3.9%
2	SEEKING DETOX:(2)	14841	6.5 %	6.5%
3	ALCOHOL ONLY (AGE < 21):(3)	7421	3.2 %	3.2%
4	ADVERSE REACTION:(4)	88096	38.4 %	38.4%
5	OVERMEDICATION:(5)	18146	7.9 %	7.9%
6	MALICIOUS POISONING:(6)	793	0.3 %	0.3%
7	ACCIDENTAL INGESTION:(7)	3253	1.4 %	1.4%
8	OTHER:(8)	87628	38.2 %	38.2%

Based upon 229211 valid cases out of 229211 total cases.

https://www.datafiles.samhsa.gov/study-series/drug-abuse-warning-network-dawn-nid13516

Here's the Data: 34565-0001-Data.txt

NOW WHAT?

- How big is it?
- What is the encoding?
- How is it formatted?

The Codebook tells us all of this, but let's take a look

Look at One Record!

```
!head -n 1 data/34565-0001-Data.txt
```

```
     1 2251082    .9426354082   3 4 1 2201141 2 865 105 1102005 1 2 1 2.00-7.00-7.0000-7.0000-
7.00001255 105 1142032 4 1 1 2.50 5.00 5.0100-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.
0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-
7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -
7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.000
0  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-
7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-
7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7
-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7
-7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-
7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.
0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-
7.00-7.00-7.0000-7.0000-7.0000  -7  -7  -7  -7-7-7-7-7.00-7.00-7.0000-7.0000-7.00008 611001
```

Investigation

- Rather than read all of the fields into the data frame, let's focus on age and the type of ER visit.

Read the Data into a Data Frame

```
colspecs =
      [(0,6), (14,29), (33,35), (35, 37), (37, 39), (1213, 1214)]

varNames = ["id", "wt", "age", "sex", "race","type"]

dawn = pd.read_fwf('data/34565-0001-Data.txt',
                   colspecs=colspecs, header=None,
                   index_col=0, names = varNames)
```

Look at the Data

```
dawn.tail()
```

id	wt	age	sex	race	type
29207	4.203385	1	2	12	NaN
292081	4.215246	9	2	12	NaN
29209	4.139613	8	2-	82	NaN
29210	1.601442	1	2	22	NaN
29211	5.261895	0	2	22	NaN

What do you notice?

Look at the Data Again

```
dawn.tail()
```

id	wt	age	sex	race	type
229207	4.203385	11	2	1	4
229208	4.215246	9	2	1	8
229209	4.139613	8	2	-8	4
229210	1.601442	1	2	2	4
229211	5.261895	10	2	2	4

Fix the column specifications. Now how does it look?

CASETYPE **TYPE OF VISIT**

Location: 1214-1214 (width: 1; decimal: 0)

Variable Type: numeric

Value	*Label*	*Unweighted Frequency*	*%*	*Valid %*
1	SUICICDE ATTEMPT:(1)	9033	3.9 %	3.9%
2	SEEKING DETOX:(2)	14841	6.5 %	6.5%
3	ALCOHOL ONLY (AGE < 21):(3)	7421	3.2 %	3.2%
4	ADVERSE REACTION:(4)	88096	38.4 %	38.4%
5	OVERMEDICATION:(5)	18146	7.9 %	7.9%
6	MALICIOUS POISONING:(6)	793	0.3 %	0.3%
7	ACCIDENTAL INGESTION:(7)	3253	1.4 %	1.4%
8	OTHER:(8)	87628	38.2 %	38.2%

Based upon 229211 valid cases out of 229211 total cases.

```
dawn.groupby(['type']).count()
```

type	wt	age	sex	race
1	9033	9033	9033	9033
2	14841	14841	14841	14841
3	7421	7421	7421	7421
4	88096	88096	88096	88096
5	18146	18146	18146	18146
6	793	793	793	793
7	3253	3253	3253	3253
8	87628	87628	87628	87628

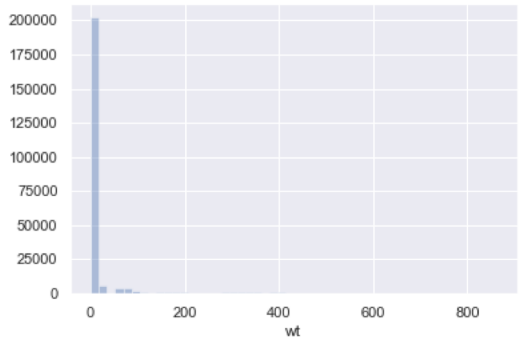
Compare the groupby counts to the codebook….
Nice Work!

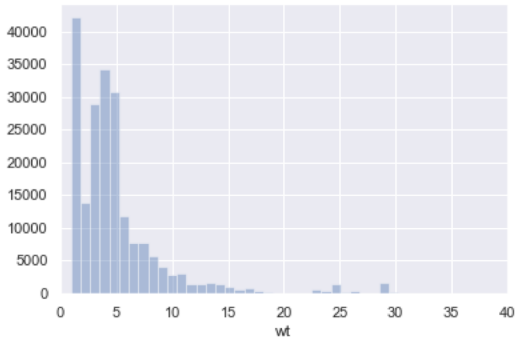
Missing Values – special coding

age				
-8	68	68	68	68
1	8744	8744	8744	8744
2	2102	2102	2102	2102
3	11933	11933	11933	11933
4	17025	17025	17025	17025
5	18268	18268	18268	18268
6	22037	22037	22037	22037
7	19950	19950	19950	19950
8	36918	36918	36918	36918
9	39803	39803	39803	39803
10	23835	23835	23835	23835
11	28528	28528	28528	28528

What do you advise?

Weights





This is a probability sample, so the representation of each individual in the sample can be computed. What do we do with this information?

SEX	GENDER
Location:	36-37 (width: 2; decimal: 0)
Variable Type:	numeric
Range of Missing Values (M):	-8

Value	*Label*	*Unweighted Frequency*	*%*	*Valid %*
1	MALE:(1)	119111	52.0 %	52.0%
2	FEMALE:(2)	110030	48.0 %	48.0%
-8 (M)	NOT DOCUMENTED:(-8)	70	0.0 %	-

Based upon 229141 valid cases out of 229211 total cases.

```
total = dawn['wt'].sum()
total
```

```
5067374.131010554
```

Unweighted avg: 48% female
Weighted avg: 52% female

```
np.average((dawn["sex"] == 2), weights=dawn["wt"])
```

```
0.523468490709998
```

Simple Example of Weighting

	sex	wage	wt
0	1	20	1.0
1	0	5	0.5
2	0	5	0.5
3	0	5	0.5
4	1	20	1.0
5	0	5	0.5
6	0	5	0.5
7	1	20	1.0
8	1	20	1.0
9	0	5	0.5
10	0	50	0.5
11	0	50	0.5

What's the unweighted proportion of females?

What's the weighted proportion?

What's the unweighted median wage?

What's the weighted median wage?

Simple Example of Weighting

	sex	wage	wt
0	1	20	1.0
1	0	5	0.5
2	0	5	0.5
3	0	5	0.5
4	1	20	1.0
5	0	5	0.5
6	0	5	0.5
7	1	20	1.0
8	1	20	1.0
9	0	5	0.5
10	0	50	0.5
11	0	50	0.5

Unweighted proportion of females? 2/3

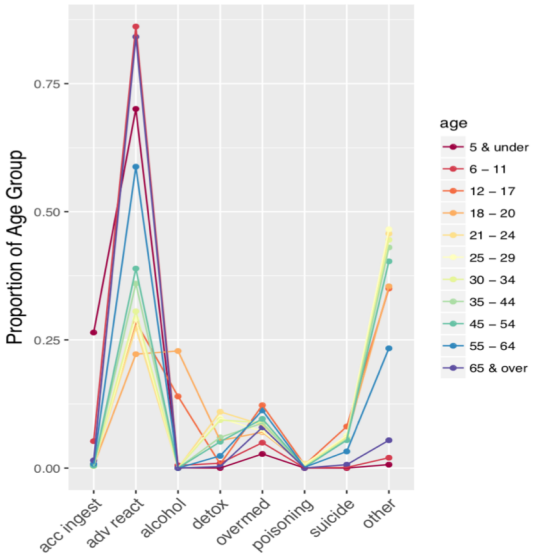
Weighted proportion? 1/2

Unweighted median wage? $12.50

What's the weighted median wage? $20.00

We're headed here –

But first we need to learn more about Visualization



When the Data Are in Multiple Tables