EE514

Data analysis and machine learning

Ali Intizar

Semester 1
2024/2025

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

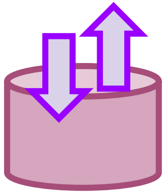
Data Interpretation, Management, Storage, Wrangling and Cleansing

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

Overview

Thinking about and interpreting data

- Datasets
- Data types
- Data as vectors and matrices



Data management and storage

- Storing data in files
- Storing data in databases

Data wrangling and cleaning

- Filtering and transforming
- Imputing missing values
- Fusing multiple data sources

Interpreting data

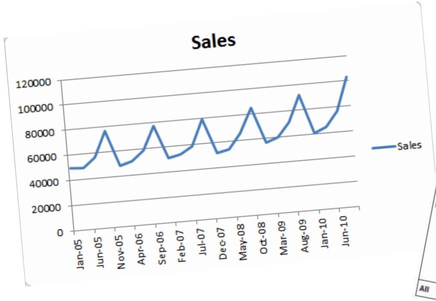
Datasets

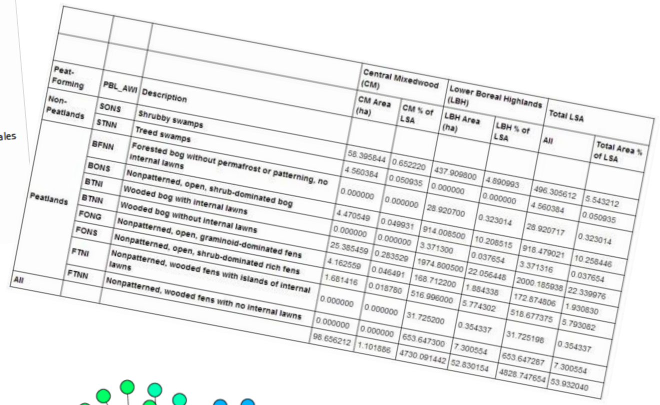
Data can come in a variety of forms

1. Series (1D)
2. Tables
3. Trees
4. Graphs
5. Text
6. Multimedia

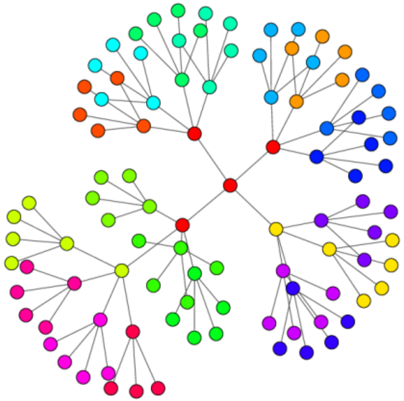
In general, it can always be flattened into tables.

- Tables listing connected nodes in a graph





Peat-Forming	PBL_AWI	Description	Central Mixedwood (CM)		Lower Boreal Highlands (LBH)		Total LSA	
			CM Area (ha)	CM % of LSA	LBH Area (ha)	LBH % of LSA	All	Total Area % of LSA
Non-Peatlands	SONS	Shrubby swamps	58.395844	0.652220	437.909800	4.890993	496.305612	5.543212
	STNN	Treed swamps	4.560384	0.050935	0.000000	0.000000	4.560384	0.050935
Peatlands	BFNN	Forested bog without permafrost or patterning, no internal lawns	0.000000	0.000000	28.920700	0.323014	28.920717	0.323014
	BONS	Nonpatterned, open, shrub-dominated bog	4.470549	0.049931	914.008500	10.208515	918.479021	10.258446
	BTNI	Wooded bog with internal lawns	0.000000	0.000000	3.371300	0.037654	3.371316	0.037654
	BTNN	Wooded bog without internal lawns	25.385459	0.283529	1974.800500	22.056448	2000.185938	22.339976
	FONG	Nonpatterned, open, graminoid-dominated fens	4.162559	0.046491	168.712200	1.884338	172.874806	1.930830
	FONS	Nonpatterned, open, shrub-dominated rich fens	1.681416	0.018780	516.996000	5.774302	518.677375	5.793082
	FTNI	Nonpatterned, wooded fens with islands of internal lawns	0.000000	0.000000	31.725200	0.354337	31.725198	0.354337
	FTNN	Nonpatterned, wooded fens with no internal lawns	0.000000	0.000000	653.647300	7.300554	653.647287	7.300554
All			98.656212	1.101886	4730.091442	52.830154	4828.747654	53.932040



Examining datasets

Two types of information in datasets

- **Metadata**: data about data. Usually contains semantics (meaning)
- **Data**: the values. AKA attributes, features, measurements, variables, variates.

In tables:

- Rows contain **items** (AKA examples, instances, data points)
- Columns contain attributes
- Header contains semantics (metadata)

The titanic dataset

name	survived	sex	age	sibsp	parch	ticket
Allen, Miss. Elisabeth Walton	1	female	29	0	0	24160
Allison, Master. Hudson Trevor	1	male	0.9167	1	2	113781
Allison, Miss. Helen Loraine	0	female	2	1	2	113781
Allison, Mr. Hudson Joshua Creighto	0	male	30	1	2	113781
Allison, Mrs. Hudson J C (Bessie Wa	0	female	25	1	2	113781
Anderson, Mr. Harry	1	male	48	0	0	19952
Andrews, Miss. Kornelia Theodosia	1	female	63	1	0	13502
Andrews, Mr. Thomas Jr	0	male	39	0	0	112050
Appleton, Mrs. Edward Dale (Charlot	1	female	53	2	0	11769
Artagaveytia, Mr. Ramon	0	male	71	0	0	PC 17609
Astor, Col. John Jacob	0	male	47	1	0	PC 17757
Astor, Mrs. John Jacob (Madeleine T	1	female	18	1	0	PC 17757
Aubart, Mme. Leontine Pauline	1	female	24	0	0	PC 17477
Barber, Miss. Ellen "Nellie"	1	female	26	0	0	19877
Barkworth, Mr. Algernon Henry Wils	1	male	80	0	0	27042
Baumann, Mr. John D	0	male		0	0	PC 17318
Baxter, Mr. Quigg Edmond	0	male	24	0	1	PC 17558
Baxter, Mrs. James (Helene DeLaude	1	female	50	0	1	PC 17558

Attribute types

Many different ways we can classify the types of attributes in a dataset

Programmer types:

- String
- Integer
- Float
- Boolean

Mathematical sets:

- Real
- Complex
- Rational
- Integer

Much more useful in data analytics to classify them **according to which operations can be performed on them.**

The titanic dataset

name	survived	sex	age	sibsp	parch	ticket
Allen, Miss. Elisabeth Walton	1	female	29	0	0	24160
Allison, Master. Hudson Trevor	1	male	0.9167	1	2	113781
Allison, Miss. Helen Loraine	0	female	2	1	2	113781
Allison, Mr. Hudson Joshua Creighto	0	male	30	1	2	113781
Allison, Mrs. Hudson J C (Bessie Wa	0	female	25	1	2	113781
Anderson, Mr. Harry	1	male	48	0	0	19952
Andrews, Miss. Kornelia Theodosia	1	female	63	1	0	13502
Andrews, Mr. Thomas Jr	0	male	39	0	0	112050
Appleton, Mrs. Edward Dale (Charlot	1	female	53	2	0	11769
Artagaveytia, Mr. Ramon	0	male	71	0	0	PC 17609
Astor, Col. John Jacob	0	male	47	1	0	PC 17757
Astor, Mrs. John Jacob (Madeleine Ta	1	female	18	1	0	PC 17757
Aubart, Mme. Leontine Pauline	1	female	24	0	0	PC 17477
Barber, Miss. Ellen "Nellie"	1	female	26	0	0	19877
Barkworth, Mr. Algernon Henry Wils	1	male	80	0	0	27042
Baumann, Mr. John D	0	male		0	0	PC 17318
Baxter, Mr. Quigg Edmond	0	male	24	0	1	PC 17558
Baxter, Mrs. James (Helene DeLaude	1	female	50	0	1	PC 17558

Scales of measurement

One of the most useful methods for classifying attribute types is **levels/scales of measurement**

- Defines what operations you are allowed to perform on the variables
  - (<, +, etc)

Stevens' <u>four</u> scales of measurement [Stevens, 1946]

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Stronger assumptions!

Note: values also referred to as **variables, attributes,** and **measurements**

Nominal variables

In nominal measurement the values just **<u>name</u>** the attribute uniquely. No ordering of is implied.

Also called **categorical variables**

Examples:

- { male, female }
- { yes, no }
- { Ireland, UK, France, Spain, Italy, … }
- { Ford, Nissan, Mercedes, … }
- Vocabulary { a, the, person, hat … }

Valid operations:

- $=$, $\neq$

Permissible statistics:

- Counts
- Modes
- Contingency correlation

It **never** make sense to compare nominal variables for order:

- Ireland > France ?

It **never** makes sense to perform arithmetic nominal variables:

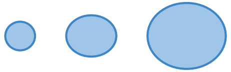
- Yes + No = ?
- AVERAGE([male, male, female])

Ordinal variables

Ordinal attributes (variables) can be **rank-ordered**.

Examples:

- Exam grade { A+, A, A-, B+, B, B-, … }
- Clothing sizes { XS, S, M, L, XL }
- Position in a race { 1st, 2nd, 3rd }

Valid operations: 

- =, ≠, <, >, (≤, ≥)

Permissible statistics:

- Median
- Percentiles
- Spearman correlation

Ordered, so makes sense to compare:

- Large > Medium

Does **not make sense** to find mean, standard deviation, etc.

Quantitative variables

Quantities. Real numbers.

Two subtypes:

- **Interval**: distance between attributes **does** have meaning but there is <u>no absolute zero.</u>
- **Ratio**: same as interval but with a meaningful absolute zero.

Interval:

- Date (1 Jan)
- Temperature in degrees F.
- Geometric point

Ratio:

- Length, mass, temperature (in Kelvin)
- Age, height, weight

For both, we can do arithmetic

- Interval: =, ≠, <, >, +, −
- Ratio: =, ≠, <, >, +, −, ×, ÷

Permissible statistics:

- Mean
- Standard deviation
- Pearson correlation

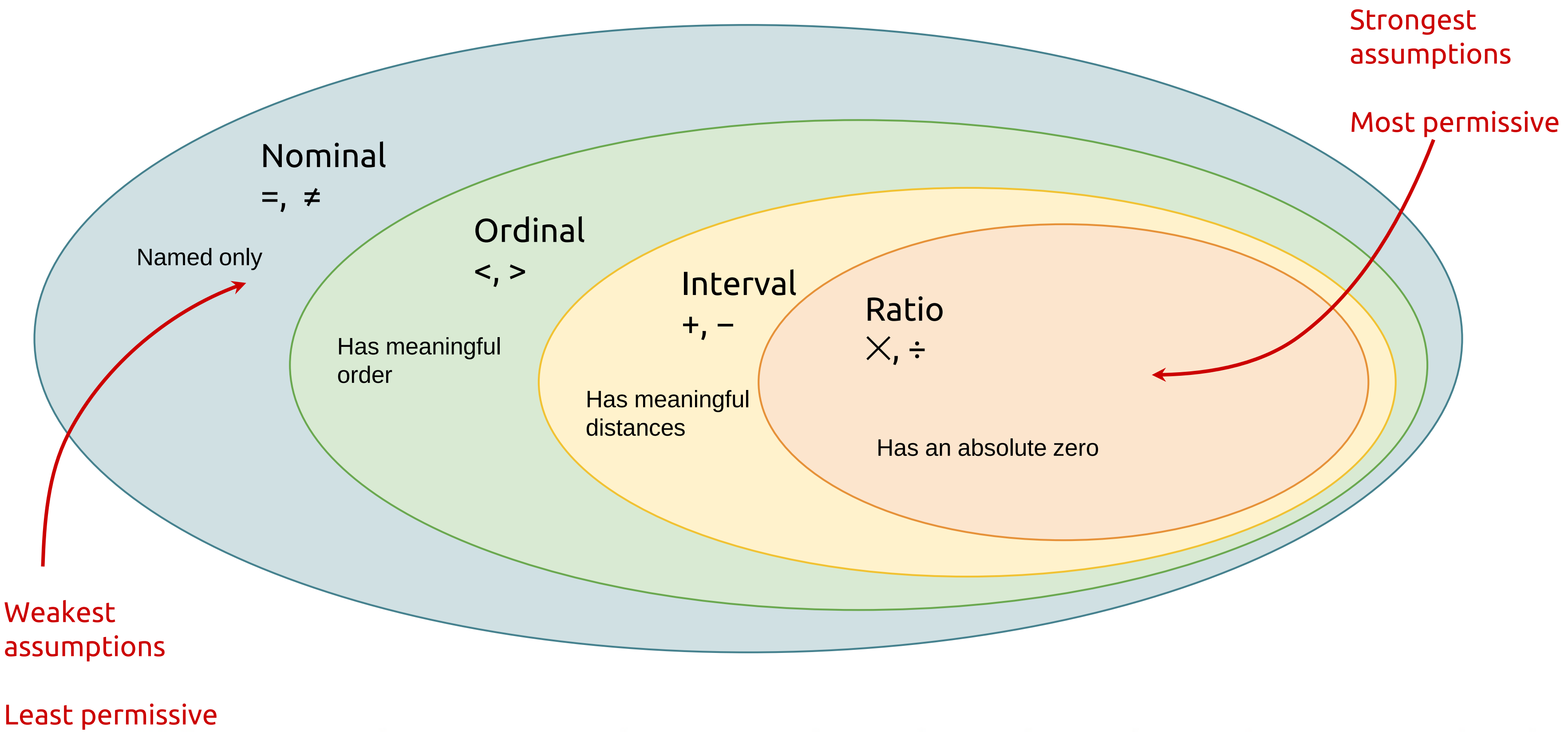
For ratio variables, **ratios make sense**:

- 10 meters is twice 5 meters

For interval variables, **only intervals make sense**

- 6 of Feb is ? to Jan 1?
- Interval: 6 of feb is 37 days after Jan 1

Hierarchy in levels of measurement



name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine Ta	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

Ordinal

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1				0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1				1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine T:	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wasl
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine T	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

Nominal

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	[illegible]	[illegible]	[illegible]	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	[illegible]	[illegible]	[illegible]	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine T	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelan
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine Ta	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

Nominal

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	[illegible]	[illegible]	[illegible]	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	[illegible]	[illegible]	[illegible]	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine T	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48	0	0	19952	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine T	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

Ratio

name	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
Allen, Miss. Elisabeth Walton	1	1	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
Allison, Master. Hudson Trevor	1	1	male	0.9167	1	2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Chesterville
Allison, Miss. Helen Loraine	1	0	female	2	1	2	113781	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Allison, Mr. Hudson Joshua Creighto	1	0	male	30	1	2	113781	151.5500	C22 C26	S		135	Montreal, PQ / Chesterville
Allison, Mrs. Hudson J C (Bessie Wa	1	0	female	25			81	151.5500	C22 C26	S			Montreal, PQ / Chesterville
Anderson, Mr. Harry	1	1	male	48			52	26.5500	E12	S	3		New York, NY
Andrews, Miss. Kornelia Theodosia	1	1	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
Andrews, Mr. Thomas Jr	1	0	male	39	0	0	112050	0.0000	A36	S			Belfast, NI
Appleton, Mrs. Edward Dale (Charlot	1	1	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Queens, NY
Artagaveytia, Mr. Ramon	1	0	male	71	0	0	PC 17609	49.5042		C		22	Montevideo, Uruguay
Astor, Col. John Jacob	1	0	male	47	1	0	PC 17757	227.5250	C62 C64	C		124	New York, NY
Astor, Mrs. John Jacob (Madeleine T:	1	1	female	18	1	0	PC 17757	227.5250	C62 C64	C	4		New York, NY
Aubart, Mme. Leontine Pauline	1	1	female	24	0	0	PC 17477	69.3000	B35	C	9		Paris, France
Barber, Miss. Ellen "Nellie"	1	1	female	26	0	0	19877	78.8500		S	6		
Barkworth, Mr. Algernon Henry Wils	1	1	male	80	0	0	27042	30.0000	A23	S	B		Hessle, Yorks
Baumann, Mr. John D	1	0	male		0	0	PC 17318	25.9250		S			New York, NY
Baxter, Mr. Quigg Edmond	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
Baxter, Mrs. James (Helene DeLaude	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
Bazzani, Miss. Albina	1	1	female	32	0	0	11813	76.2917	D15	C	8		
Beattie, Mr. Thomson	1	0	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MN
Beckwith, Mr. Richard Leonard	1	1	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
Beckwith, Mrs. Richard Leonard (Sal	1	1	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
Behr, Mr. Karl Howell	1	1	male	26	0	0	111369	30.0000	C148	C	5		New York, NY
Bidois, Miss. Rosalie	1	1	female	42	0	0	PC 17757	227.5250		C	4		
Bird, Miss. Ellen	1	1	female	29	0	0	PC 17483	221.7792	C97	S	8		
Birnbaum, Mr. Jakob	1	0	male	25	0	0	13905	26.0000		C		148	San Francisco, CA
Bishop, Mr. Dickinson H	1	1	male	25	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bishop, Mrs. Dickinson H (Helen Wa	1	1	female	19	1	0	11967	91.0792	B49	C	7		Dowagiac, MI
Bissette, Miss. Amelia	1	1	female	35	0	0	PC 17760	135.6333	C99	S	8		
Bjornstrom-Steffansson, Mr. Mauritz	1	1	male	28	0	0	110564	26.5500	C52	S	D		Stockholm, Sweden / Wash
Blackwell, Mr. Stephen Weart	1	0	male	45	0	0	113784	35.5000	T	S			Trenton, NJ
Blank, Mr. Henry	1	1	male	40	0	0	112277	31.0000	A31	C	7		Glen Ridge, NJ
Bonnell, Miss. Caroline	1	1	female	30	0	0	36928	164.8667	C7	S	8		Youngstown, OH
Bonnell, Miss. Elizabeth	1	1	female	58	0	0	113783	26.5500	C103	S	8		Birkdale, England Clevelar
Borebank, Mr. John James	1	0	male	42	0	0	110489	26.5500	D22	S			London / Winnipeg, MB
Bowen, Miss. Grace Scott	1	1	female	45	0	0	PC 17608	262.3750		C	4		Cooperstown, NY
Bowerman, Miss. Elsie Edith	1	1	female	22	0	1	113505	55.0000	E33	S	6		St Leonards-on-Sea, Engla

Numbers of variables

Velocity				
35	50	60	40	20

Univariate data

- You only have one attribute. E.g. time series

lon	lat
-6.258854	53.385381
-6.274475	53.360158
-71.094664	42.359980

Bivariate data

- You have two attributes. E.g. a table of longitude and latitude pairs

Multivariate data

- You have *N* > 1 attributes

name	survived	sex	age	sibsp	parch	ticket	fare	c
Allen, Miss. Elisabeth Walton	1	female	29	0	0	24160	211.3375	B5
Allison, Master. Hudson Trevor	1	male	0.9167	1	2	113781	151.5500	C22 C
Allison, Miss. Helen Loraine	0	female	2	1	2	113781	151.5500	C22 C
Allison, Mr. Hudson Joshua Creighto	0	male	30	1	2	113781	151.5500	C22 C
Allison, Mrs. Hudson J C (Bessie Wa	0	female	25	1	2	113781	151.5500	C22 C
Anderson, Mr. Harry	1	male	48	0	0	19952	26.5500	E12
Andrews, Miss. Kornelia Theodosia	1	female	63	1	0	13502	77.9583	D7
Andrews, Mr. Thomas Jr	0	male	39	0	0	112050	0.0000	A36
Appleton, Mrs. Edward Dale (Charlot	1	female	53	2	0	11769	51.4792	C101
Artagaveytia, Mr. Ramon	0	male	71	0	0	PC 17609	49.5042	
Astor, Col. John Jacob	0	male	47	1	0	PC 17757	227.5250	C62 C

Data as vectors

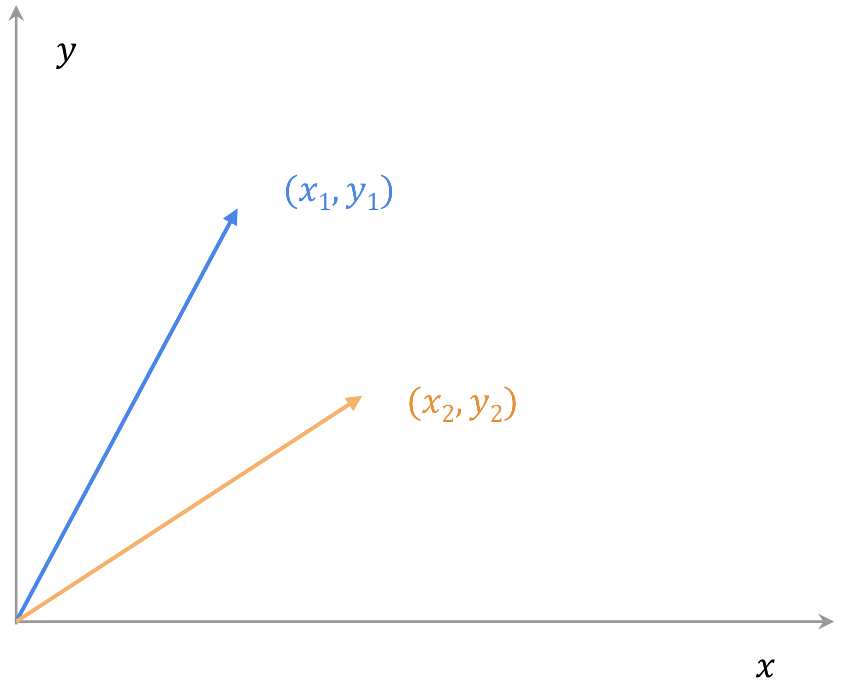
Bivariate data points with **quantitative** variables can be described using vectors in 2D space.

$$\begin{bmatrix}1 & 3\end{bmatrix}^T \quad \begin{bmatrix}1\\2\end{bmatrix} \quad \begin{bmatrix}3\\2\end{bmatrix} \quad \in R^2$$

Multivariate data points can be described using vectors in D-dimensional space

$$\begin{bmatrix}2\\3\\1\\8\end{bmatrix} \in R^D$$

This abstraction is very useful, since it allows us to use linear algebra theory to manipulate data



Datasets as matrices

We can stack quantitative data vectors into matrices.

Usually we stack the items (data points, examples) in the rows, and the attributes (features) in the columns.

- NB: some books/papers do the opposite!

For a dataset X with N items and D attributes, we have a $(N \times D)$ matrix

$$X \in R^{N \times D}$$

Data points

$$\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$$

$$N \text{ items} \left\{ \underbrace{\begin{bmatrix} 0.255 & 0.123 & 0.127 \\ 0.649 & 0.057 & 0.476 \\ 0.379 & 0.184 & 0.943 \\ 0.471 & 0.511 & 0.092 \\ 0.647 & 0.866 & 0.759 \\ 0.475 & 0.345 & 0.858 \end{bmatrix}}_{D \text{ attributes}} \right.$$

One-hot encoding

Sometimes we want a vector encoding for a **nominal** variable

Solution: one-hot encoding

Nominal attribute with K possible values becomes a K-dimensional vector

Transform a categorical variable with K categories into K binary variables

Very useful for encoding text

Categorical attribute with 5 categories

2
3
1
4
5

0	1	0	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	1

One hot-encoding: 5 binary attributes (sparse matrix)

One-hot encoding of text

Can use one-hot encoding of words.

Set of all possible words is called the **vocabulary.**

The **codebook** assigns each word to an integer

"The cat sat on the mat"

Codebook

The	1
cat	2
sat	3
on	4
mat	5

The	1
cat	2
sat	3
on	4
the	1
mat	5

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
1	0	0	0	0
0	0	0	0	1

Bag of words model of text

One way we represent an entire passage of text is called the bag of words model.

Add up the rows of the one-hot encodings.

Note: throws away all information about the ordering of words!

This representation is widely used in search engines and text analysis.

Note: may be useful to exclude uninteresting "stop words" like *a, the, an*, etc., from vocabulary

The	1
cat	2
sat	3
on	4
the	1
mat	5

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
1	0	0	0	0
0	0	0	0	1

BoW encoding	2	1	1	1	1

Why is this useful?

Summary statistics

- Most used words
- Least used words
- Average use of a word

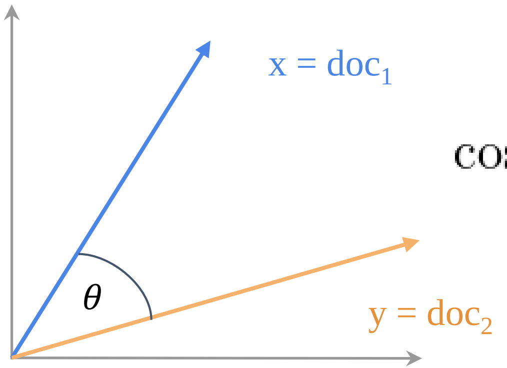
Information retrieval and search

- Encode query using BoW
- Encode all documents using BoW
- Compare query with docs using similarity metric (often cosine similarity)

Similarity

- Measure how similar documents are
- Cluster similar documents
- Topic modelling (PCA/LSI)
- Visualization (t-SNE)

Unsupervised Learning



$$\cos\theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

Multimedia data

Multimedia documents:

- Text, hypertext
- Images
- Video
- Audio

Types of analysis you might want to do:

- Natural language analysis
- Audio transcription, speaker identification
- Face detection, recognition
- Vehicle tracking
- Automatic image tagging
- ...

Fields of study:

- **Computer vision (CV)** is the field concerned with using computer models to understand visual content
- **Natural language processing (NLP)** is concerned with parsing, disambiguating, and understanding language
- **Automatic speech recognition (ASR)** is concerned with using computer models to transform speech to text.

Representing multimedia data

Many different representations used in practice.

Images:

- 3D tensors: $I \in \mathbb{R}^{H \times W \times 3}$
- As functions: $f(x, y) \rightarrow \mathbb{R}^3$
- As vectors: $I \in \mathbb{R}^{3WH}$
- Compressed (JPEG, PNG)
- Using automatically extracted features

Video: similar to images but with **time**. E.g.

$$f(x, y, t) \rightarrow \mathbb{R}^3$$

Digital audio:

- Sampled (44.1 KHz) quantized sound wave
- Single or multi channel
- Single channel sample of length *N* can be represented as a vector in $\mathbf{R}^N$
- Function representation
- Compressed (MP3)

Note: typically we do not store this kind of multimedia data in tables. **Why?**

Summary

Data

- metadata, semantics
- measurements, attributes, values

Stevens' four scales of measurement

- Nominal
- Ordinal
- Interval
- Ratio

Define what operations make sense

Number of variables/attributes

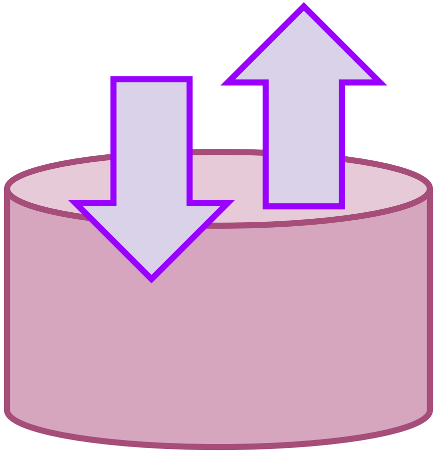
- Univariate
- Bivariate
- Multivariate

Interpreting data as vectors

Datasets as matrices

One-hot encoding

Multimedia



Data storage and I/O

Data storage and I/O

Data can come from files, databases, to be read from streams

Files:
- Can be structured or unstructured
- Good for distribution
- Can be high performance

Databases:
- Good for centralized information and network access
- Can enforce structure via **schemas**
- Multiple readers and writers
- Query languages to filter and search

Streams:
- Real-time processing of live data (e.g. twitter)

Data can be structured, unstructured, or semi-structured

Structured data: includes information on semantics such as relationships and data types. E.g:

- Tables
- Graphs
- Hierarchies
- Relational databases

Semi-structured and unstructured data: semantic information missing or not machine readable. E.g:

- Natural language plain text
- HTML files
- Word docs

File formats

Binary formats

Numeric values stored encoded in binary representation

- 32 bit float (4 bytes)
- 16 bit integer (2 bytes)

Properties:

- Compact
- High performance I/O
- Not human readable
- Need to worry about integer sizes, endiness, signed/unsigned

Plain text formats (ASCII and Unicode)

Numeric values encoded as ASCII or Unicode strings

- float -> "3.1415926"
- int -> "44"

Properties:

- Human readable
- (Somewhat) self-documenting
- Slower I/O
- Less compact

ASCII formats for tabular data

Tables are a common structure with special formats available.

Two very common formats:

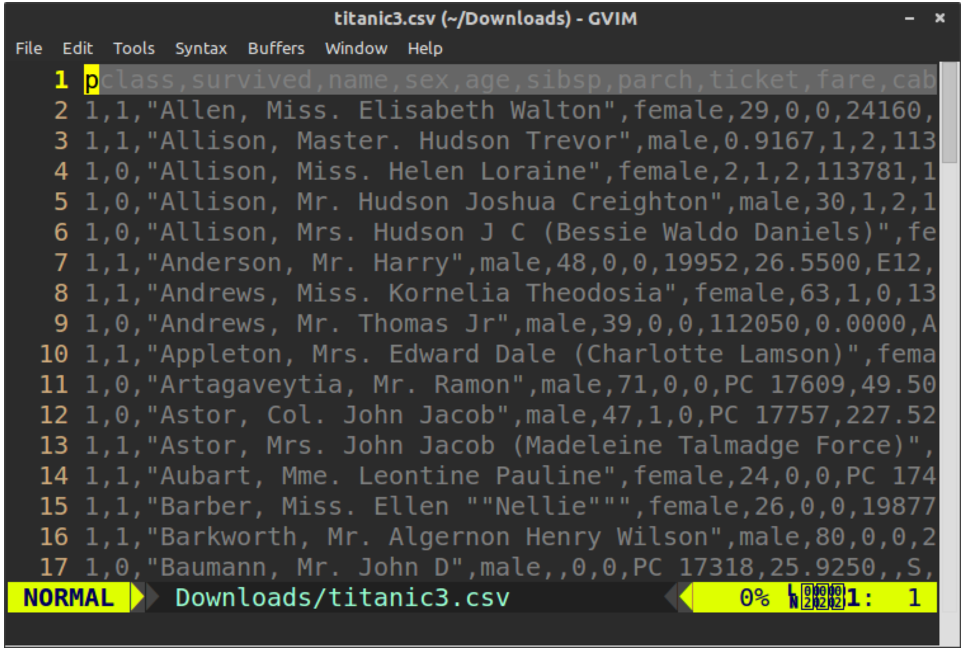
- **CSV**: comma separated values
- **FWF**: fixed width format

	Age						
Preferred cola	18 to 24	25 to 29	30 to 39	40 to 49	50 to 54	55 to 64	65 or more
	%	%	%	%	%	%	%
Coca-Cola	65	41	55	28	46	36	36
Diet Coke	2	10	13	15	8	12	23
Coke Zero	9	23	19	22	28	16	14
Pepsi Light	0	3	0	3	3	6	9
Pepsi Max	16	18	6	10	13	24	14
Pepsi	7	5	7	22	3	6	5
NET	100	100	100	100	100	100	100

CSV

Comma separated values

- Plain text format
- Rows on lines
- Columns separated by commas
- Actually not only commas. Tabs, etc. (TSV)
- Strings containing commas are quoted



CSV

Advantages:

- Can be read by Excel, OpenOffice, Google Sheets, etc.
- Fast to parse/generate
- Can be compressed (`.csv.gz`)
- Do not need to load all into memory (streamable)

Disadvantages:

- Not standardised: many variations
- Bulkier than binary formats esp. when uncompressed
- No types

I/O in Python:

Python has a [built in library for parsing CSV]()

```
import csv

with open('titanic3.csv') as f:
    reader = csv.reader(f)
    for row in reader:
        print(row[0], row[3])

with open('output.csv', 'w') as f:
    writer = csv.writer(f)
    writer.writerow(['A', 1, 'female', 'red'])
```

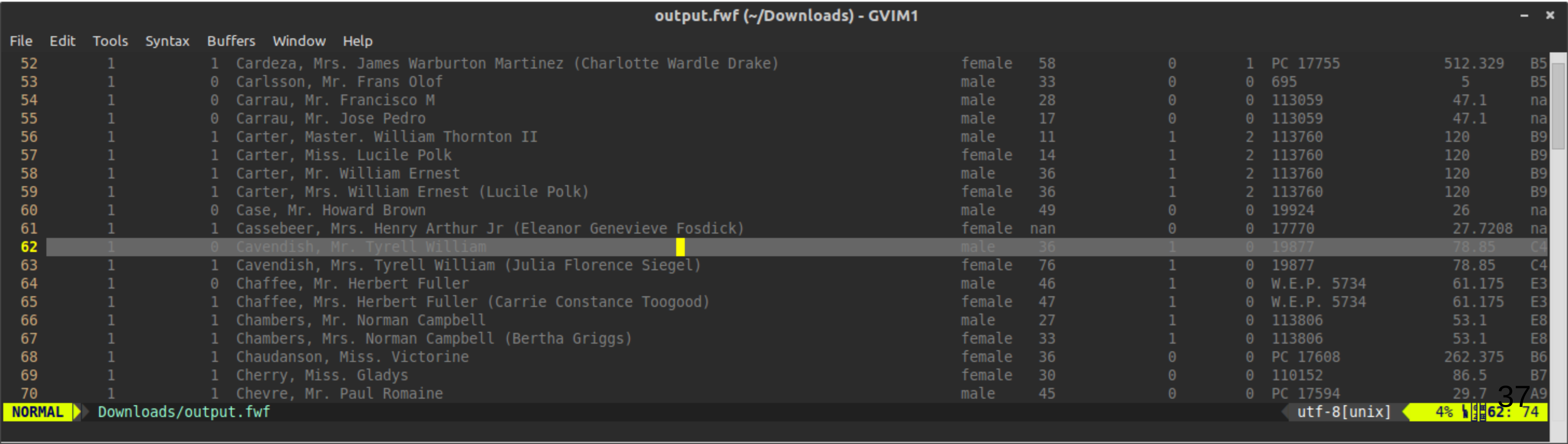
[Pandas]() can also read/write CSV

```
import pandas as pd
df = pd.read_csv('titanic3.csv')
df.to_csv('output.csv')
```

FWF

Fixed width format

- Plain text format
- Rows on lines
- Columns of fixed width (fixed number of characters)
- Columns padded using padding character (usually spaces)



FWF

Advantages:

- Easier to read in plain text than CSV
- Can be read by major spreadsheet programs
- Fast to parse
- Can be compressed (`.fwf.gz`)
- Do not need to load all into memory (streamable)

Disadvantages:

- Not standardised
- Bulkier than CSV (padding characters)
- Need to establish field width before you can write first row
- No types

I/O in Python:

Pandas can also read FWF

```
import pandas as pd
df = pd.read_fwf('titanic3.fwf')
```

To write, you need the **tabulate** package

```
from tabulate import tabulate
content = tabulate(
    df.values.tolist(),
    list(df.columns),
    tablefmt="plain")

open('output.fwf', ‘w’).write(content)
```

Binary formats for tabular data

Binary formats are more compact and performant.

Not human readable.

Common binary formats:

- [HDF5]()
- MATLAB
- NumPy
- Apache Arrow and Feather
- Excel

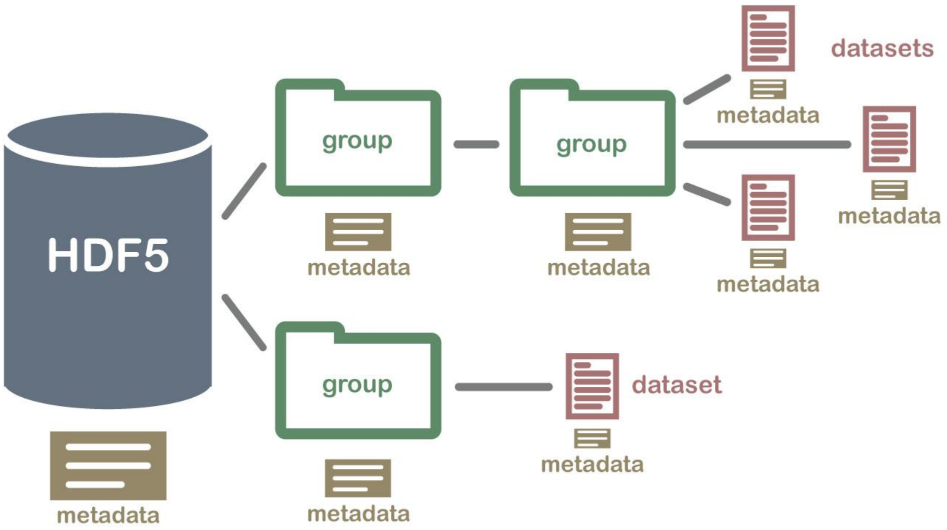
```
0000000 0000 0001 0001 1010 0010 0001 0004 0128
0000010 0000 0016 0000 0028 0000 0010 0000 0020
0000020 0000 0001 0004 0000 0000 0000 0000 0000
0000030 0000 0000 0000 0010 0000 0000 0000 0204
0000040 0004 8384 0084 c7c8 00c8 4748 0048 e8e9
0000050 00e9 6a69 0069 a8a9 00a9 2828 0028 fdfc
0000060 00fc 1819 0019 9898 0098 d9d8 00d8 5857
0000070 0057 7b7a 007a bab9 00b9 3a3c 003c 8888
0000080 8888 8888 8888 8888 288e be88 8888 8888
0000090 3b83 5788 8888 8888 7667 778e 8828 8888
00000a0 d61f 7abd 8818 8888 467c 585f 8814 8188
00000b0 8b06 e8f7 88aa 8388 8b3b 88f3 88bd e988
00000c0 8a18 880c e841 c988 b328 6871 688e 958b
00000d0 a948 5862 5884 7e81 3788 1ab4 5a84 3eec
00000e0 3d86 dcb8 5cbb 8888 8888 8888 8888 8888
00000f0 8888 8888 8888 8888 8888 8888 8888 0000
0000100 0000 0000 0000 0000 0000 0000 0000 0000
*
0000130 0000 0000 0000 0000 0000 0000 0000
000013e
```

HDF5

Hierarchical data format version 5 (.h5)

Industry standard for storing numeric structured tabular data. Widely used in scientific community.

- Can store **multiple datasets** in a single file.
- Organize datasets into a **hierarchical structure**, just like a filesystem
- Include arbitrary metadata in the file using **metadata "attributes"**
- Uses **memory mapping** so entire dataset does not need to be loaded into memory. Memory can be **shared between processes.**
- Library support in many languages.



HDF5

Advantages:

- High performance
- Compact
- Multiple datasets in one file
- Type information
- Metadata
- Parallel reads, shared memory
- Optional compression

Disadvantages:

- Not human readable
- Can be cumbersome when table size is not known in advance

I/O in Python:

Can use h5py library

```
import h5py
f = h5py.File("mytestfile.hdf5", "w")
dset = f.create_dataset(
    "mydataset", shape=(100,), dtype='i')
```

Pandas can also read certain types of .h5 files (using PyTables)

```
pd.read_hdf('store_tl.h5', 'table')
df.to_hdf('store_tl.h5', 'table', append=True)

store = pd.HDFStore('store.h5')
store['mydataset']
```

MATLAB files

Proprietary format used by MATLAB software (.mat)

Common for data exchange in industry and academia.

Can store multiple named arrays and various other structures.

Newer versions of MATLAB (R2006b or later) now store data in HDF5-based files (.mat v7.3)

I/O in Python:
Scipy library includes functions for loading and saving .mat files in the scipy.io package

- `scipy.io.loadmat`
- `scipy.io.savemat`
- `scipy.io.whosmat`

NumPy

NumPy has built in support for two lightweight binary formats for storing NumPy arrays:

- `.npy` files contain single numpy arrays
- `.npz` files contain multiple arrays

`.npz` files are actually zipped archives of .npy files. Arrays in `.npz` files can be named. The format also supports compression.

Advantages and Disadvantages

- Fast, compact
- Supports memory mapping
- Built in support in NumPy
- No metadata
- Not standardized, less portable than H5

I/O in Python:

load(file[, mmap_mode, allow_pickle, ...])
Load arrays or pickled objects from `.npy`, `.npz` or pickled files.

save(file, arr[, allow_pickle, fix_imports])
Save an array to a binary file in NumPy `.npy` format.

savez(file, *args, **kwds)
Save several arrays into a single file in uncompressed `.npz` format.

savez_compressed(file, *args, **kwds)
Save several arrays into a single file in compressed `.npz` format.

Excel format

Very common proprietary format used in Microsoft Excel

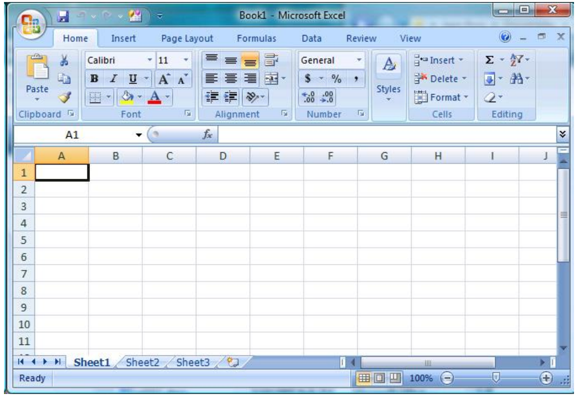
- XLS files, XLSX files

Advantages:

- Works with MS products
- Keeps “formatting” (colors, etc.)

Disadvantages:

- Keeps “formatting”
- Not portable
- Proprietary



I/O in Python:

Pandas has a function to read (`read_excel`) and method to write (`DataFrame.to_excel`)

Also see http://www.python-excel.org/ for info on other libraries for working with Excel

- `xlrd, xlwt`

Common formats for structured data

More complex structures than plain tables (although can be reduced to tables)

Structure: encode data types, attributes, relationships, hierarchies

Common **plain-text formats** for structured data:

- [XML]
- [JSON]
- [YAML]

Binary formats:

- [MessagePack]
- [Google Protobuf]

XML

Extensible Markup Language (XML)

Format for structured data designed to be both human and machine readable

Designed to be a self-describing format: no additional documentation needed

XML document is a hierarchy made up of:

- **Elements:** `<book>`
- **Attributes:** `ISBN="0553212419"`
- **Content:** `Sherlock Holmes...`

```xml
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```

XML

Advantages:

- Self-describing
- "Human readable" (ish)

Disadvantages:

- High overhead (extremely bulky)
- Verbose
- Slow parsers
- Not well suited to big data and large scientific datasets
- XML namespaces are a nightmare
- Separate validation and DTD
- Painful to type manually
- No built-in types for scientific data

I/O in Python:

ElementTree in standard library. Stores entire XML structure in memory

```
import xml.etree.ElementTree as ET
tree = ET.parse('input.xml')
root = tree.getroot()
for child in root:
    print(child.tag, child.attrib)
```

Several other ways in Python

- SAX and Expat: event-driven parsing (callbacks)

See: https://docs.python.org/2/library/xml.html

JSON

[JavaScript Object Notation](#)

Subset of JavaScript for describing data

Value types:

- String
- Number
- Boolean
- Object
- Array
- Null

Type is implicit in syntax

- 35 is a number
- "35" is a string
- { "a" : 1 } is an object
- [1, 2, 3] is an array

```
{
  "employees":[
    {"firstName":"John", "lastName":"Doe"},
    {"firstName":"Anna", "lastName":"Smith"},
    {"firstName":"Peter", "lastName":"Jones"}
  ]
}
```

JSON

Advantages:

- Has types
- Hierarchical
- Human-readable
- Self-describing
- Easy to write by hand
- More compact than XML
- Faster to parse than XML
- Most languages have parsers

Disadvantages:

- Some overhead (plain text)
- Not well suited to big data and large scientific datasets

I/O in Python:

Super simple using built-in json module

```python
import json
data = json.load(open('input.json'))

with open('output.json', 'w') as f:
    json.dump(data, f)
```

JSON now used in place of XML in many applications, especially web apps.

YAML

YAML Ain't Markup Language!

- Like JSON, but easier to write by hand.
- Very useful for configuration files and metadata files.
- Slower than JSON

I/O in Python:
Using pyyaml

```
import yaml
data = yaml.safe_load(open('input.yaml'))
```


```yaml
YAML: YAML Ain't Markup Language

What It Is: YAML is a human friendly data serialization
  standard for all programming languages.

YAML Resources:
  YAML 1.2 (3rd Edition): http://yaml.org/spec/1.2/spec.html
  YAML 1.1 (2nd Edition): http://yaml.org/spec/1.1/
  YAML 1.0 (1st Edition): http://yaml.org/spec/1.0/
  YAML Issues Page: https://github.com/yaml/yaml/issues
  YAML Mailing List: yaml-core@lists.sourceforge.net
  YAML IRC Channel: "#yaml on irc.freenode.net"
  YAML Cookbook (Ruby): http://yaml4r.sourceforge.net/cookbook/ (local)
  YAML Reference Parser: http://ben-kiki.org/ypaste/

Projects:
  C/C++ Libraries:
  - libyaml       # "C" Fast YAML 1.1
  - Syck          # (dated) "C" YAML 1.0
  - yaml-cpp      # C++ YAML 1.2 implementation
Python:
  - PyYAML        # YAML 1.1, pure python and libyaml binding
  - ruamel.yaml   # YAML 1.2, update of PyYAML with ...
```

MessagePack

MessagePack: binary encoded JSON

Much more compact than JSON

Advantages:

- Very compact
- High I/O performance
- Good for network I/O and databases
- Good for bigger data
- Streamable

Disadvantages:

- Not human readable (but easy to decode)
- Less suited to large numerical arrays

I/O in Python:
Msgpack library

```
>>> import msgpack
>>> msgpack.packb([1, 2, 3])
'\x93\x01\x02\x03'
>>> msgpack.unpackb(_)
[1, 2, 3]
```

Google Protobuf

"Protocol buffers are Google's language-neutral, platform-neutral, extensible mechanism for serializing structured data – think XML, but smaller, faster, and simpler."

Primarily designed for serializing data over the wire, but can also be used as a file format

Protobuf defines a protocol specification language (prototxt) and a wire format.

Notably used for model specification and distribution by the well-known deep learning library developed by Berkeley: **Caffe**

Prototxt example:

```
message Person {
  required string name = 1;
  required int32 id = 2;
  optional string email = 3;
}
```

I/O in Python:
Use Google's protoc compiler to generate "header" files. Then load and save with

```
ParseFromString(data)
SerializeToString()
```

Things to consider when choosing a data format

- ❑ I/O performance
- ❑ Structure support (hierarchies, graphs, etc.)
- ❑ Streamable
- ❑ Scalability
- ❑ Appendable
- ❑ Portability
- ❑ Compactness
- ❑ Metadata
- ❑ Type support
- ❑ Readability
- ❑ Write by hand

Databases

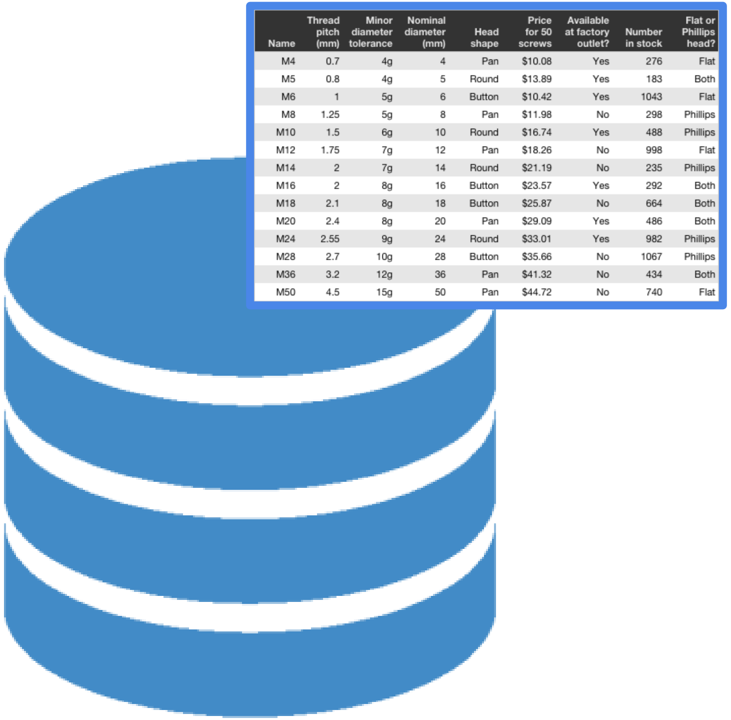
Databases are also common for data acquisition and storage.

Advantages of databases include network access, concurrency, enforced consistency, fast indexes, query languages.

Not necessarily efficient for large binary data (e.g. images, video, audio, sensor data, astro).

Two important types:

- **Relational databases (SQL)**
- **NoSQL databases**



Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Summary of data formats and I/O

Files:

- Plain text formats
- Binary formats
- Structured or unstructured
- Semantics and metadata
- Types

Databases:

- SQL
- NoSQL

Files are good for:

- Data distribution
- Ephemeral or intermediate results
- High-speed processing
- Multimedia data

Databases are good for:

- Centralized network data access
- Concurrent access
- Enforcing consistency
- Subsetting and querying
- (Caching)

Data wrangling

Real-world datasets

Real world datasets are often “dirty”: messy, complicated, inconsistent

Sources of problems:

- Use (or lack of) incomplete standards
- Manual entry errors (typos)
- Measurement errors (equipment and noise)
- Inconsistent notations (naming)
- Redundancies and duplicates
- Missing values (NAs)

Section1All_csv.csv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H	I	J	K
1	Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product										
2	[Percent]										
3	Annual data from 1969 To 2015										
4	Bureau of Economic Analysis										
5	Data published September 29, 2016										
6	File created 9/28/2016 11:41:15 AM										
7											
8	Line			1969	1970	1971	1972	1973	1974	1975	19
9	1	Gross domestic product	A191RL1	3.1	0.2	3.3	5.2	5.6	-0.5	-0.2	5
10	2	Personal consumption expenditures	DPCERL1	3.7	2.4	3.8	6.1	5	-0.8	2.3	5
11	3	Goods	DGDSRL1	3.1	0.8	4.2	6.5	5.2	-3.6	0.7	
12	4	Durable goods	DDURRL1	3.7	-2.7	10	12.4	10.5	-6.4	0.2	12
13	5	Nondurable goods	DNDGRL1	2.8	2.2	1.9	4	2.9	-2.4	0.9	4
14	6	Services	DSERRL1	4.4	3.9	3.5	5.8	4.7	1.9	3.8	4
15	7	Gross private domestic investment	A006RL1	5.6	-6.1	10.3	11.3	10.9	-6.6	-16.2	19
16	8	Fixed investment	A007RL1	5.9	-2.1	6.9	11.4	8.6	-5.6	-9.8	9
17	9	Nonresidential	A008RL1	7	-0.9	0	8.7	13.2	0.8	-9	5
18	10	Structures	A009RL1	5.4	0.3	-1.6	3.1	8.2	-2.2	-10.5	2
19	11	Equipment	Y033RL1	8.3	-1.8	0.8	12.7	18.5	2.1	-10.5	6
20	12	Intellectual property products	Y001RL1	5.4	-0.1	0.4	7	5	2.9	0.9	10
21	13	Residential	A011RL1	3.1	-5.2	26.6	17.4	-0.6	-19.6	-12.1	22
22	14	Change in private inventories	ZZZZZZ1	.....	.....	.....	.....	.....	.....	.....	.....
23	15	Net exports of goods and services	ZZZZZZ1	.....	.....	.....	.....	.....	.....	.....	.....
24	16	Exports	A020RL1	4.9	10.7	1.7	7.8	18.8	7.9	-0.6	4
25	17	Goods	A253RL1	5.2	11.2	-0.1	10.9	24.5	8.5	-2.1	5
26	18	Services	A646RL1	3.9	9	7.3	-0.4	1.7	5.4	5.9	1
27	19	Imports	A021RL1	5.7	4.3	5.3	11.3	4.6	-2.3	-11.1	19
28	20	Goods	A255RL1	5.5	3.9	8.4	13.6	7.1	-2.8	-12.6	22
29	21	Services	A656RL1	6.3	5.2	-2.8	4.2	-3.4	-0.1	-4.3	6

Sheet1

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,19
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,-0.2,5.4,4.6,5.6,3.2,-0.2,2.6,-1.9,4.6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,5.0,-0.8,2.3,5.6,4.2,4.4,2.4,-0.3,1.5,1.4,5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1,1.6,-2.5,1.2,0.7,6.4,7.2,5.3,5.6,1.8,3.7,2.5,0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0.2,12.5,8.8,5.2,-0.5,-8.0,1.0,-0.2,14.3,14.3,10.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.9,4.8,2.3,3.6,2.6,-0.2,1.3,1.0,3.3,4.1,3.0,3.6,1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.6,3.1,1.6,1.7,2.0,5.2,3.9,5.3,3.2,4.5,4.5,3.2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.3,10.9,-6.6,-16.2,19.1,14.3,11.6,3.5,-10.1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,13.6,11.6,5.8,-5.9,2.7,-6.7,7.5,16.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10.8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,      Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.1,14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,      Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,15.5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,      Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0,5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,    Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,22.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,  Change in private inventories,ZZZZZZ1,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,
15,Net exports of goods and services,ZZZZZZ1,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....,....
16,  Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,    Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,    Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,  Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,    Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,    Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,  Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,    National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,    Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,  State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,19
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,-0.2,5.4,4.6,5.6,3.2,-0.2,2.6,-1.9,4.6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,5 [illegible] 5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1, [illegible] 0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0. [illegible] 0.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.9 [illegible] 5,1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.6, [illegible] 3.2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.3,10.9,-6.6,-16.2,19.1,14.3,11.6,3.5,-10.1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,13.6,11.6,5.8,-5.9,2.7,-6.7,7.5,16.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10.8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,      Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.1,14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,      Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,15.5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,      Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0,5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,    Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,22.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,  Change in private inventories,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
15,Net exports of goods and services,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
16,  Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,    Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,    Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,  Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,    Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,    Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,  Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,    National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,    Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,  State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

Headers need to be removed

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,19
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,-0.2,5.4,4.6,5.6,3.2,-0.2,2.6,-1.9,4.6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,5                                 5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1,                             0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0.                             0.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.9                             5,1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.6,                           3.2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.3,10.9,-6.6,-16.2,19.1,14.3,11.6,3.5,-10.1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,13.6,11.6,5.8,-5.9,2.7,-6.7,7.5,16.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10.8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,      Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.1,14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,      Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,15.5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,      Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0,5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,    Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,22.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,  Change in private inventories,ZZZZZZ1,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,
15,Net exports of goods and services,ZZZZZZ1,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,
16,  Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,    Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,    Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,  Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,    Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,    Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,  Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,    National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,    Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,  State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

Blank lines

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,19
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,-0.2,5.4,4.6,5.6,3.2,-0.2,2.6,-1.9,4.6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,5[illegible]5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1,[illegible]0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0.[illegible]0.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.9[illegible]5,1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.6,[illegible]3.2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.3,10.9,-6.6,-16.2,19.1,14.3,11.6,3.5,-10.1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,13.6,11.6,5.8,-5.9,2.7,-6.7,7.5,16.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10.8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,      Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.1,14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,      Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,15.5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,      Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0,5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,    Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,22.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,  Change in private inventories,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
15,Net exports of goods and services,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
16,  Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,    Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,    Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,  Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,    Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,    Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,  Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,    National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,    Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,  State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

Missing column names

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,19
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,-0.2,5.4,4.6,5.6,3.2,-0.2,2.6,-1.9,4.6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,5 [illegible] 5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1, [illegible] 0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0 [illegible] 0.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.9 [illegible] 6,1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.6, [illegible] 3.2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.3,10.[illegible],-6.6,-16.2,19.1,14.3,11.6,3.5,-10.1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,1[illegible].6,11.6,5.8,-5.9,2.7,-6.7,7.5,16.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10[illegible]8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,       Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.[illegible],14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,       Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,1[illegible].5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,       Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0[illegible]5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,     Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,[illegible]2.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,   Change in private inventories,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
15,Net exports of goods and services,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
16,   Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,     Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,     Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,   Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,     Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,     Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,   Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,     National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,     Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,   State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

Subheadings embedded in table

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,19
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,[illegible]6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,[illegible]5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1,[illegible],0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0[illegible]0.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.[illegible],1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.[illegible]2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.[illegible]1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,[illegible]6.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10.8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,      Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.1,14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,      Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,15.5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,      Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0,5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,     Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,22.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,   Change in private inventories,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
15,Net exports of goods and services,ZZZZZZ1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
16,   Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,     Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,     Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,   Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,     Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,     Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,   Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,     National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,     Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,   State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

“Category” is actually a hierarchy of categories implied using whitespace

Data wrangling

Process of transforming “raw” data into data that can be analysed to generate actionable insights

AKA:

- Preprocessing
- Munging
- Cleaning
- Scrubbing
- Preparation
- Transformation

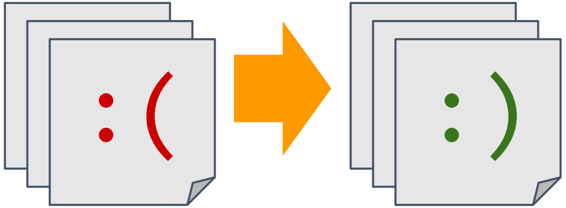
Typical data wrangling tasks

- Fixing ugly and broken formats
- Handling missing values
- Removing redundant attributes and records
- Fixing inconsistencies
- Shaping data
- Fusing data sources
- Scraping and gathering data from external sources
- Extracting information from unstructured sources

Ugly and broken formats

Examples:

- Badly formatted tables
- Broken XML/JSON (syntax errors)
- Hand entered data with syntax errors
- Log files with strange formatting



First step: transform to more machine friendly parsable format

Toolbox: Python **csv** module, text editor like **vim**, custom parsing scripts, regular expressions (**re** module), **tabular**, **pandas**

```
Table 1.1.1. Percent Change From Preceding Period in Real Gross Domestic Product,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
[Percent],,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Annual data from 1969 To 2015,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Bureau of Economic Analysis,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
"Data published September 29, 2016",,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
File created 9/28/2016 11:41:15 AM,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
Line,,,1969,1970,1971,1972,1973,1974,1975,1976,1977,1978,1979,1980,1981,1982,1983,1984,1985,1986,1987,1988,1989,1990,1991,1992,1993,
1,    Gross domestic product,A191RL1,3.1,0.2,3.3,5.2,5.6,-0.5,-0.2,5.4,4.6,5.6,3.2,-0.2,2.6,-1.9,4.6,7.3,4.2,3.5,3.5,4.2,3.7,1.9,-0.1,
2,Personal consumption expenditures,DPCERL1,3.7,2.4,3.8,6.1,5.0,-0.8,2.3,5.6,4.2,4.4,2.4,-0.3,1.5,1.4,5.7,5.3,5.3,4.2,3.4,4.2,2.9,2.1,
3,  Goods,DGDSRL1,3.1,0.8,4.2,6.5,5.2,-3.6,0.7,7.0,4.3,4.1,1.6,-2.5,1.2,0.7,6.4,7.2,5.3,5.6,1.8,3.7,2.5,0.6,-2.0,3.2,4.2,5.3,3.0,4.5,4
4,    Durable goods,DDURRL1,3.7,-2.7,10.0,12.4,10.5,-6.4,0.2,12.5,8.8,5.2,-0.5,-8.0,1.0,-0.2,14.3,14.3,10.0,9.6,2.0,5.7,2.2,-0.4,-5.4,
5,    Nondurable goods,DNDGRL1,2.8,2.2,1.9,4.0,2.9,-2.4,0.9,4.8,2.3,3.6,2.6,-0.2,1.3,1.0,3.3,4.1,3.0,3.6,1.7,2.6,2.7,1.2,-0.3,1.9,2.5,
6,  Services,DSERRL1,4.4,3.9,3.5,5.8,4.7,1.9,3.8,4.3,4.1,4.6,3.1,1.6,1.7,2.0,5.2,3.9,5.3,3.2,4.5,4.5,3.2,3.0,1.6,4.0,3.1,3.1,3.0,2.9,3
7,Gross private domestic investment,A006RL1,5.6,-6.1,10.3,11.3,10.9,-6.6,-16.2,19.1,14.3,11.6,3.5,-10.1,8.8,-13.0,9.3,27.3,-0.1,0.2,2.
8,  Fixed investment,A007RL1,5.9,-2.1,6.9,11.4,8.6,-5.6,-9.8,9.8,13.6,11.6,5.8,-5.9,2.7,-6.7,7.5,16.2,5.5,1.8,0.6,3.3,3.2,-1.4,-5.1,5.
9,    Nonresidential,A008RL1,7.0,-0.9,0.0,8.7,13.2,0.8,-9.0,5.7,10.8,13.8,10.0,0.0,6.1,-3.6,-0.4,16.7,6.6,-1.7,0.1,5.0,5.7,1.1,-3.9,2.
10,      Structures,A009RL1,5.4,0.3,-1.6,3.1,8.2,-2.2,-10.5,2.4,4.1,14.4,12.7,5.9,8.0,-1.6,-10.8,13.9,7.1,-11.0,-2.9,0.7,2.0,1.5,-11.1
11,      Equipment,Y033RL1,8.3,-1.8,0.8,12.7,18.5,2.1,-10.5,6.1,15.5,15.1,8.2,-4.4,3.7,-7.6,4.6,19.4,5.5,1.1,0.4,6.6,5.3,-2.1,-4.6,5.9
12,      Intellectual property products,Y001RL1,5.4,-0.1,0.4,7.0,5.0,2.9,0.9,10.9,6.6,7.1,11.7,5.0,10.9,6.2,7.9,13.7,9.0,7.0,3.9,7.1,1
13,    Residential,A011RL1,3.1,-5.2,26.6,17.4,-0.6,-19.6,-12.1,22.1,20.5,6.7,-3.7,-20.9,-8.2,-18.1,42.0,14.8,2.3,12.4,2.0,-0.9,-3.2,-8
14,  Change in private inventories,ZZZZZZ1,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...
15,Net exports of goods and services,ZZZZZZ1,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...,...
16,  Exports,A020RL1,4.9,10.7,1.7,7.8,18.8,7.9,-0.6,4.4,2.4,10.5,9.9,10.8,1.2,-7.6,-2.6,8.2,3.3,7.7,10.9,16.2,11.6,8.8,6.6,6.9,3.3,8.8
17,    Goods,A253RL1,5.2,11.2,-0.1,10.9,24.5,8.5,-2.1,5.1,1.9,10.4,10.6,12.3,-0.6,-8.5,-3.2,7.1,3.5,5.4,12.2,17.8,11.4,8.6,6.7,7.5,3.2
18,    Services,A646RL1,3.9,9.0,7.3,-0.4,1.7,5.4,5.9,1.1,4.5,11.2,7.1,4.2,9.8,-4.4,-0.2,11.8,2.8,14.4,7.6,11.9,12.0,9.5,6.4,5.4,3.3,7.
19,  Imports,A021RL1,5.7,4.3,5.3,11.3,4.6,-2.3,-11.1,19.5,10.9,8.7,1.7,-6.6,2.6,-1.3,12.6,24.3,6.5,8.5,5.9,3.9,4.4,3.6,-0.1,7.0,8.6,11
20,    Goods,A255RL1,5.5,3.9,8.4,13.6,7.1,-2.8,-12.6,22.6,12.2,9.0,1.7,-7.4,2.1,-2.5,13.6,24.2,6.3,10.3,4.6,4.1,4.3,2.9,0.5,9.4,10.0,1
21,    Services,A656RL1,6.3,5.2,-2.8,4.2,-3.4,-0.1,-4.3,6.9,5.0,7.1,1.4,-2.2,5.9,5.3,8.1,25.1,7.6,1.1,11.8,3.4,4.8,6.5,-2.6,-2.7,2.7,5
22,Government consumption expenditures and gross investment,A822RL1,0.2,-2.0,-1.8,-0.5,-0.3,2.3,2.2,0.5,1.2,2.9,1.9,1.9,1.0,1.8,3.8,3.
23,  Federal,A823RL1,-2.4,-6.1,-6.4,-3.1,-3.6,0.7,0.5,0.2,2.2,2.5,2.3,4.4,4.5,3.7,6.5,3.3,7.9,5.9,3.8,-1.3,1.7,2.1,0.0,-1.5,-3.5,-3.5,
24,    National defense,A824RL1,-4.1,-8.2,-10.2,-6.9,-5.1,-1.0,-1.0,-0.5,1.0,0.8,2.7,3.9,6.2,7.2,7.3,5.2,8.8,6.9,5.1,-0.2,-0.2,0.3,-1.
25,    Nondefense,A825RL1,3.9,1.0,5.6,7.2,0.2,4.6,3.9,1.6,4.7,6.0,1.7,5.4,1.0,-3.6,4.7,-1.4,5.7,3.1,0.2,-4.3,7.2,7.3,2.4,5.9,0.0,-0.8,
26,  State and local,A829RL1,3.5,2.9,3.1,2.2,2.8,3.7,3.6,0.8,0.4,3.3,1.5,-0.2,-2.0,0.1,1.3,3.8,5.7,5.0,2.2,3.9,4.0,4.1,2.2,2.1,1.2,2.8
```

Missing values

Missing values are common and can have significant effects on analysis and conclusions

Causes:

- Non-response
- Unobserved or unknown values
- Sensor or measurement errors
- Censorship
- Errors in data collection or data entry

Often show up in datasets as:

- Special `NA` values
- `NaN`
- `null` or `None`
- Sentinel values (e.g `age == -1`)
- Blanks

GENDER	AGE	RELIG.	Q1
M	18	CR	...
M	-1	ATH	null
F	22	CR	...
F	36	N/A	...

Important to try understand the reasons for missing values in order to appropriately deal with them…

Missing values

Three types of missing values:

1. **Missing completely at random (MCAR):** missing values are randomly distributed for all observations
2. **Missing at random (MAR):** probability of value being missing depends on other observed variables
3. **Missing not at random (MNAR):** probability of value being missing depends on value of missing variable or another unobserved variable.

GENDER	AGE	RELIG.	Q1
M	18	CR	...
M	-1	ATH	null
F	22	CR	...
F	36	N/A	...

MCAR and MAR assumption is common.

If assumptions are made when dealing with missing values, make them explicit!

Strategies for handling missing values

Three common approaches:

1. Ignore
2. Remove
3. Impute

May introduce bias for MNAR values!

Ignore
Drop missing values when computing summary statistics (e.g. mean, variance).

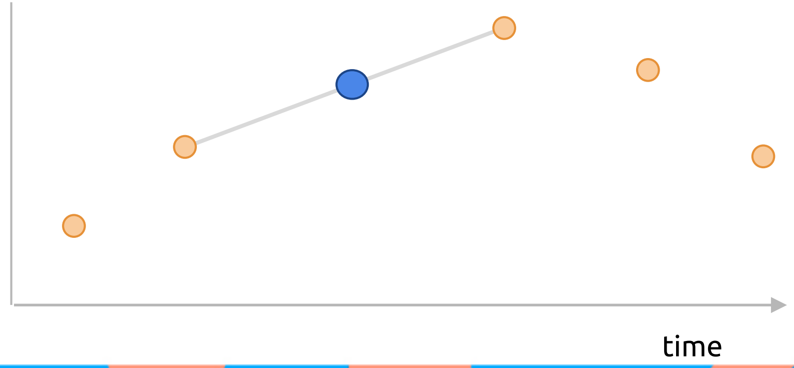
Remove
If plenty of data is available, may be possible to simply ignore rows that have missing values

Impute
Try to “fill in the blanks”

Common imputation techniques:

- Mean/mode substitution
- Predict from other attributes
- Interpolation (e.g. time series)

Note: good idea to include **indicator variable** to state if value was imputed



Redundant attributes

For example:

- Useless attributes
- Duplicated attributes
- Attributes easily derived from other attributes

Can cause problems for some statistical analysis (e.g. regression models).

Eliminate redundancy where possible.

HUMAN	SEX	GENDER	HEIGHT (ft)	HEIGHT (cm)
Y	M	male	5'9	175
Y	M	male	6'4	193
Y	F	female	5'10	178
Y	F	female	6'1	185

Inconsistent categories (nominal attributes)

Ask 10 different people to do the same task and they will do it 11 different ways!

For example:
- Misspellings
- Inconsistent spellings
- Hyphenation
- Inconsistent case
- Inconsistent abbreviations

Techniques:
- Print unique vals and try to detect outliers and splits
- Normalize case and spelling

GENDER	STATE
Male	NY
male	New York
F	"New-york"
fem.	Califrnia, USA

Tools:
- Unix `sort | uniq`
- Python `sort`, `set()`
- `str.lower`, `str.replace`,
- Regular expressions (`re`)

Dates and times

Huge variation in ways dates, times, and timestamps can be represented.

Data cleaning should **standardize** to a single format.

Preferably include **timezone** information.

Standard plain text format: [ISO8601]

- `2016-10-10T16:04:07+00:00`

	A	B
1	Sunday, January 03, 1988	
2	1/3/1988	
3	1/3/88	
4	01/03/88	
5	3-Jan-88	
6	03-Jan-88	
7	Jan-88	
8	January-88	
9	January 3, 1988	
10	1/3/1988	
11	3-Jan-1988	
12		

Parsing dates and times

Python standard library:

```
datetime.strptime(str, fmt)
```

The parsedatetime library:
Will parse almost anything!

```
import parsedatetime
cal = parsedatetime.Calendar()
cal.parse("tomorrow")
```

Also consider using arrow library if you do a lot of date and time manipulation

Input	Parse
19 November 1975	Wed Nov 19 08:41:38 1975
19 November 75	Wed Nov 19 08:41:38 1975
19 Nov 75	Wed Nov 19 08:41:38 1975
tomorrow	Tue Jun 21 09:00:00 2016
yesterday	Sun Jun 19 09:00:00 2016
10 minutes from now	Mon Jun 20 08:51:38 2016
the first of January, 2001	Mon Jan 1 08:41:38 2001
3 days ago	Fri Jun 17 08:41:38 2016

Outliers

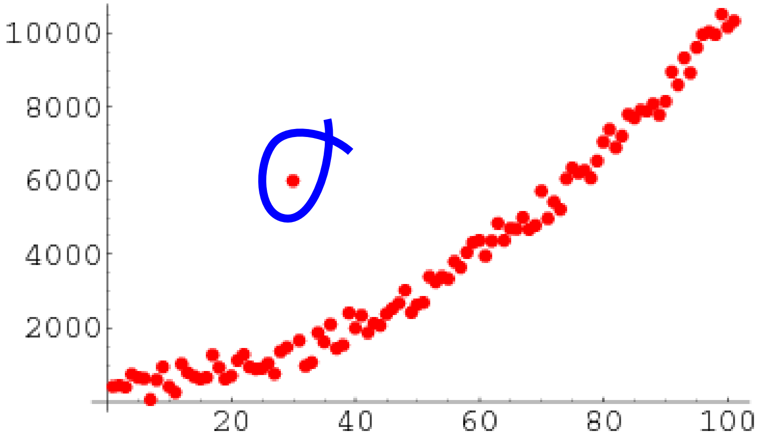
Data points that are extremely unlikely under the data distribution (far from other data points).

Causes:

- Measurement error
- Recording error
- Statistical anomalies (may be interesting!)

Often you'll want to identify outliers prior to further analysis.

- Measure quantity of outliers
- Label outliers
- Completely remove outliers

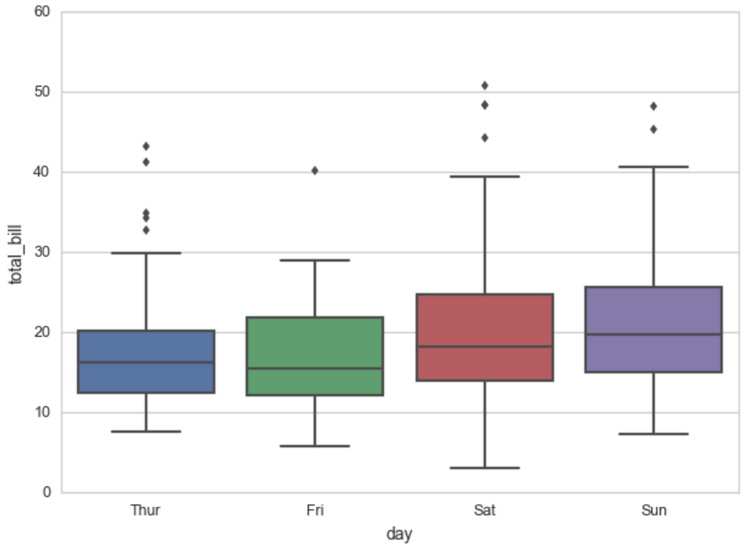


Detecting outliers

Often you can detect outliers by plotting your data and doing some visual inspection.

- Boxplots, jitter plots, histograms
- More details next lecture!

Alternatively, you can make some assumptions about the distribution of the attribute and find items that are unlikely under this distribution.



Detecting outliers

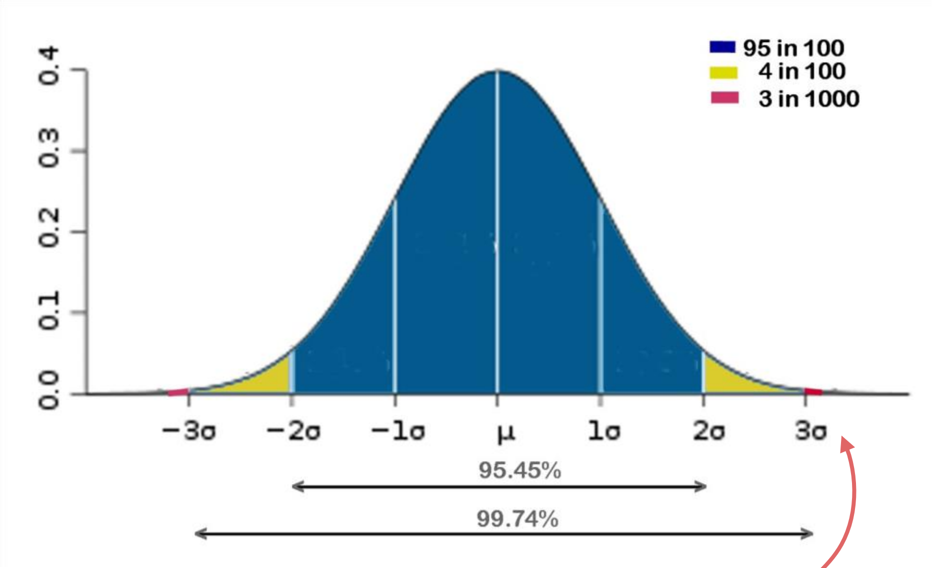
E.g. assume data is **normally (Gaussian) distributed.**

Estimate sample mean and standard deviation from data and compute Z scores.

sample mean

$$z = \frac{x - \mu}{\sigma}$$

sample standard deviation



z=3

Data shaping

Data often stored in "stacked" or record format

date	variable	value
2000-01-03	A	0.469112
2000-01-04	A	-0.282863
2000-01-05	A	-1.509059
2000-01-03	B	-1.135632
2000-01-04	B	1.212112
2000-01-05	B	-0.173215
2000-01-03	C	0.119209
2000-01-04	C	-1.044236
2000-01-05	C	-0.861849
2000-01-03	D	-2.104569
2000-01-04	D	-0.494929
2000-01-05	D	1.071804

Sometimes more convenient to have one "observation" per row with multiple attributes

date	A	B	C	D
2000-01-03	0.469112	-1.135632	0.119209	-2.104569
2000-01-04	-0.282863	1.212112	-1.044236	-0.494929
2000-01-05	-1.509059	-0.173215	-0.861849	1.071804

This can be achieved by reshaping or pivot operations. In pandas:

```
df.pivot(index='date', columns='variable',
values='value')
```

Fusion of data sources

ID	GENDER	LOC	AGE
1	F	COR	32
2	F	DUB	25
3	F	DUB	19
4	M	LIM	43

ID	Q1	Q2	Q3
1	N	Y	1
2	Y	Y	3
3	Y	N	2
4	N	Y	1

JOIN ON ID

ID	GENDER	LOC	Q1	Q2	Q3
1	F	COR	N	Y	1
2	F	DUB	Y	Y	3
3	F	DUB	Y	N	2
4	M	LIM	N	Y	1

Fusion of data sources

Pandas

```
result = pd.merge(left, right, on='key')
```

Lots more options for merging and concatenation.

See docs.

SQL Inner Joins

```
SELECT
      left.A, left.B, left.key,
      right.C, right.D
FROM left
INNER JOIN right
ON left.key = right.key;
```

left

	A	B	key
0	A0	B0	K0
1	A1	B1	K1
2	A2	B2	K2
3	A3	B3	K3

right

	C	D	key
0	C0	D0	K0
1	C1	D1	K1
2	C2	D2	K2
3	C3	D3	K3

Result

	A	B	key	C	D
0	A0	B0	K0	C0	D0
1	A1	B1	K1	C1	D1
2	A2	B2	K2	C2	D2
3	A3	B3	K3	C3	D3

Dealing with unstructured data

So far we've talked mostly about cleaning structured data (even if structure is awful!)

What about semi-structured and unstructured data?

- Natural language plain text
- HTML files

Usually we'll want to **extract some features** from this type of data for analysis

We've already seen one way of encoding text features: **bag-of-words**

```
links%2Cteahouse%7Cext.tmh.thumbnail.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desk
opArticleTarget.noscript%7Cext.wikimediaBadges%7Cmediawiki.legacy.commonPrint%2Cshared%7Cme
iawiki.sectionAnchor%7Cmediawiki.skinning.interface%7Cskins.vector.styles%7Cwikibase.client
init&amp;only=styles&amp;skin=vector"/>
<script async="" src="/w/load.php?
debug=false&amp;lang=en&amp;modules=startup&amp;only=scripts&amp;skin=vector"></script>
<meta name="ResourceLoaderDynamicStyles" content=""/>
<link rel="stylesheet" href="/w/load.php?
debug=false&amp;lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector"/>
<meta name="generator" content="MediaWiki 1.28.0-wmf.21"/>
<meta name="referrer" content="origin-when-cross-origin"/>
<link rel="alternate" href="android-
app://org.wikipedia/http/en.m.wikipedia.org/wiki/Web_scraping"/>
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?
title=Web_scraping&amp;action=edit"/>
<link rel="edit" title="Edit this page" href="/w/index.php?
title=Web_scraping&amp;action=edit"/>
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>
<link rel="search" type="application/opensearchdescription+xml"
href="/w/opensearch_desc.php" title="Wikipedia (en)"/>
<link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?
action=rsd"/>
<link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="canonical" href="https://en.wikipedia.org/wiki/Web_scraping"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-Web_scraping
rootpage-Web_scraping skin-vector action-view feature-footer-v2">       <div id="mw-page-
base" class="noprint"></div>
        <div id="mw-head-base" class="noprint"></div>
        <div id="content" class="mw-body" role="main">
            <a id="top"></a>

                            <div id="siteNotice"><!-- CentralNotice --></div>
                        <div class="mw-indicators">
</div>
            <h1 id="firstHeading" class="firstHeading" lang="en">Web scraping</h1>
                                        <div id="bodyContent" class="mw-body-content">
                                        <div id="siteSub">From Wikipedia, the free
encyclopedia</div>
                                    <div id="contentSub"></div>
                                                <div id="jump-to-nav" class="mw-jump">
                    Jump to:                    <a href="#mw-head">navigation</a>,
<a href="#p-search">search</a>
```

Web scraping: getting data from webpages

Idea: extract structured information from unstructured web pages by automatic downloading and parsing them.

Task	Issues	Toolbox
Getting the data	HTTP requests, cookies, headers, downloads, JavaScript, timing	wget, curl requests, mechanize, selenium
Figuring out which data to get	Link crawling and spidering	scrapy
Extracting structured information	Robust parsing, DOM querying	PyQuery, BeautifulSoup, lxml

Web scraping

wget, curl

Convenient command line tools.

```
$ curl https://curl.haxx.se >> page.html
```

requests

Easily make HTTP requests directly in Python

```
>>> r = requests.get(
        'https://api.github.com/user', auth=('user', 'pass'))
>>> r.status_code
200
>>> r.headers['content-type']
'application/json; charset=utf8'
>>> r.text
u'{"type":"User"...'
>>> r.json()
{u'private_gists': 419, u'total_private_repos': 77, ...}
```

mechanize

Stateful programmatic web browsing. Pretend to be a browser (cookies and headers and all)

```
browser = mechanize.Browser()
browser.open("http://www.example.com/")
response = br.follow_link(text_regex=r"cheese\s*shop", nr=1)
print(browser.title())
print(response.geturl())
print(response.info())  # headers
print(response.read())  # body
```

selenium

Remote control an actual browser!

```
driver = webdriver.Firefox()
driver.get("http://www.python.org")
elem = driver.find_element_by_name("q")
elem.clear()
elem.send_keys("pycon")
elem.send_keys(Keys.RETURN)
```

PyQuery

Parsing and querying HTML with [PyQuery](http://pypi.python.org/pypi/pyquery):

```
>>> from pyquery import PyQuery
>>> html = open("index.html", 'r').read()
>>> pq = PyQuery(html)

>>> pq("title").text()
'PyQuery Test!'

>>> pq("li").eq(1).text()
'DOM Manipulation is EASY!'

>>> for x in pq("a"):
...     print pq(x).text()
...
PyQuery
jQuery

>>> pq("ul").children().eq(0).html()
u'It makes parsing files a <strong>SNAP</strong>!'
```

```html
<!DOCTYPE html>
<html>
  <head>
    <title>PyQuery Test!</title>
  </head>

<body>
  <h1>PyQuery is AWESOME!</h1>
  <p><a
      href="http://pypi.python.org/pypi/pyquery">PyQuery</a>
      is a Python port of the famous
      <a href="http://jquery.com">jQuery</a>
      JavaScript library.
  <h2>What is it Good For?</h2>
  <ul id="pitch">
    <li>It makes parsing files a <strong>SNAP</strong>!</li>
    <li>DOM Manipulation is EASY!</li>
    <li>You <em>never</em>
        have to worry about confusing syntax</li>
  </ul>
</body>
</html>
```

Processing log files

Log files are a good example of semi-structured data. **Log analysis**: making sense and extracting information from log files.

Regular expressions and string partitioning are useful tools:

- Built-in Python [re](re) module
- Python string functions:
    - `str.split`, `str.rsplit`
    - `str.partition`, `str.rpartition`
    - `str.find`, `str.rfind`, `str.replace`
    - `str.strip`, `str.upper`, `str.lower`
    - `str[a:b:c]`
- Unix tools: `sed`, `awk`, `grep`, `vim`

access.log — nginx (git: master)

```
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/angular/angular-cookies.
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/videogular/videogular.js
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/videogular/controls.js H
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/videogular/overlay-play.
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/videogular/buffering.js
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/videogular/poster.js HTT
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/ngDialog/ngDialog.min.js
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/lib/nsPopover/nsPopover.js H
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/js/app.js HTTP/1.1" 304 0 "h
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/js/services.js HTTP/1.1" 304
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/js/controllers.js HTTP/1.1"
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/js/filters.js HTTP/1.1" 304
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/js/directives.js HTTP/1.1" 3
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/js/settings.js HTTP/1.1" 304
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/img/axes45.jpg HTTP/1.1" 304
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/api/user/profile HTTP/1.1" 4
127.0.0.1 - - [16/Feb/2015:16:03:25 +0000] "GET /axes/home/views/asset.html HTTP/1.1" 3
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/api/assets/axes:%2FcAXES%2Fv
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/font/icomoon.woff HTTP/1.1"
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/api/available-services HTTP/
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/api/video-stats/axes:%2FcAXE
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /collections/cAXES/videos/cAXES/v200805
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /collections/cAXES/videos/cAXES/v200805
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/api/related-segments/axes:%2
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/api/related-videos/axes:%2Fc
127.0.0.1 - - [16/Feb/2015:16:03:26 +0000] "GET /axes/home/api/face-tracks/axes:%2FcAXE
127.0.0.1 - - [16/Feb/2015:16:04:02 +0000] "GET /axes/home/api/keyframes/axes:%2FcAXES%
127.0.0.1 - - [16/Feb/2015:16:04:04 +0000] "GET /axes/home/api/transcript/axes:%2FcAXES
127.0.0.1 - - [16/Feb/2015:16:04:30 +0000] "GET /axes/home/api/home-topics HTTP/1.1" 20
127.0.0.1 - - [16/Feb/2015:16:04:31 +0000] "GET /axes/home/api/interesting-items HTTP/1
127.0.0.1 - - [16/Feb/2015:16:04:32 +0000] "GET /axes/home/api/assets/axes:%2FcAXES%2Fv
127.0.0.1 - - [16/Feb/2015:16:04:32 +0000] "GET /collections/cAXES/videos/cAXES/v200807
127.0.0.1 - - [16/Feb/2015:16:04:32 +0000] "GET /collections/cAXES/videos/cAXES/v200807
```

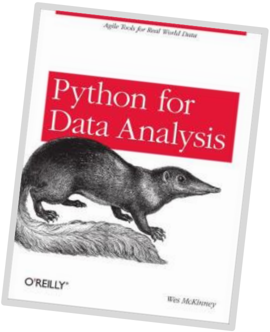
Line: 1 | Plain Text | Soft Tabs: 2

Data wrangling best practices

- ❑ **Keep raw data separate from cleaned data**. Never overwrite raw data. You may need it again. Keep backups!
- ❑ **Script data wrangling steps as much as possible.** If you need to change something later on, it's far easier if the steps are scripted.
- ❑ **Document all the transforms carried out and assumptions made**. Distribute this documentation with the cleaned dataset. Try make the wrangling process as reproducible as possible.
- ❑ **For large datasets, start with a small random sample.** Iterate faster: once you have perfected your cleaning steps on the sample, apply to full dataset.

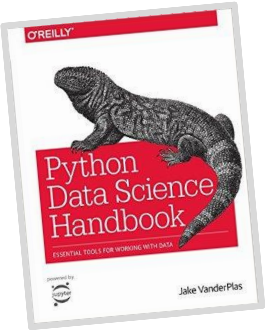
Questions?

Further reading



Python for Data Analysis, Wes McKinney (DCU library)

- Chapter 5: Getting Started with Pandas
- Chapter 6: Data Loading, Storage, and File Formats
- Chapter 7: Data Wrangling: Clean, Transform, Merge, Reshape



Python for Data Science Handbook, Jake Vanderplas (Available online)

- Chapter 3: Data Manipulation with Pandas