NYU | Center for Data Science

DS-UA 112
Introduction to Data Science

Week 5: Lecture 2
Tables – Working with Messy Datasets

NYU Center for Data Science

How can we process data to fix inconsistencies particularly missing values?

DS-UA 112
Introduction to Data Science

Week 5: Lecture 2
Tables – Working with Messy Datasets

Adapted from Nolan, Gonzalez, Hug, Lau

Announcements

- Please check Week 5 agenda on NYU Classes
  - Homework 3
  - Lab 5
  - Survey 2
- Remember to post to Piazza

Review

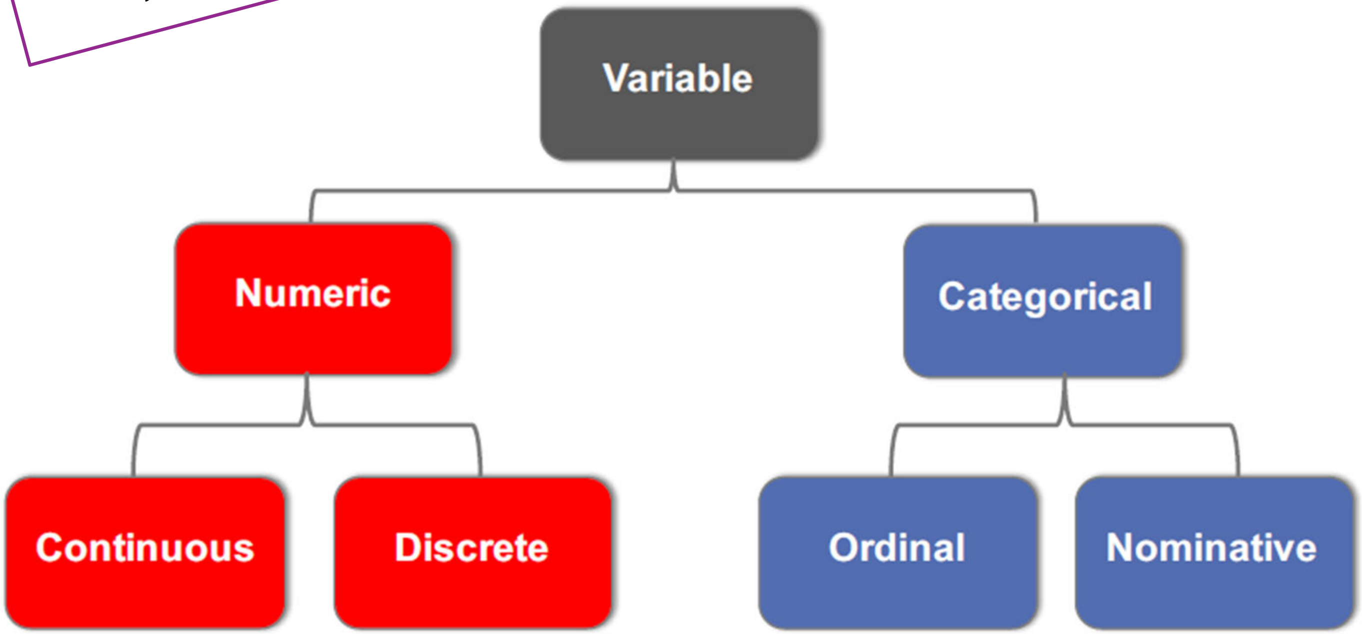
- Compressing Files
- Joining
  - Inner, Outer
  - Left, Right
  - Cross
- Properties of Data in Tables
  - Qualitative or Quantitative

Goals

- Zip
- Merge
- Data Types

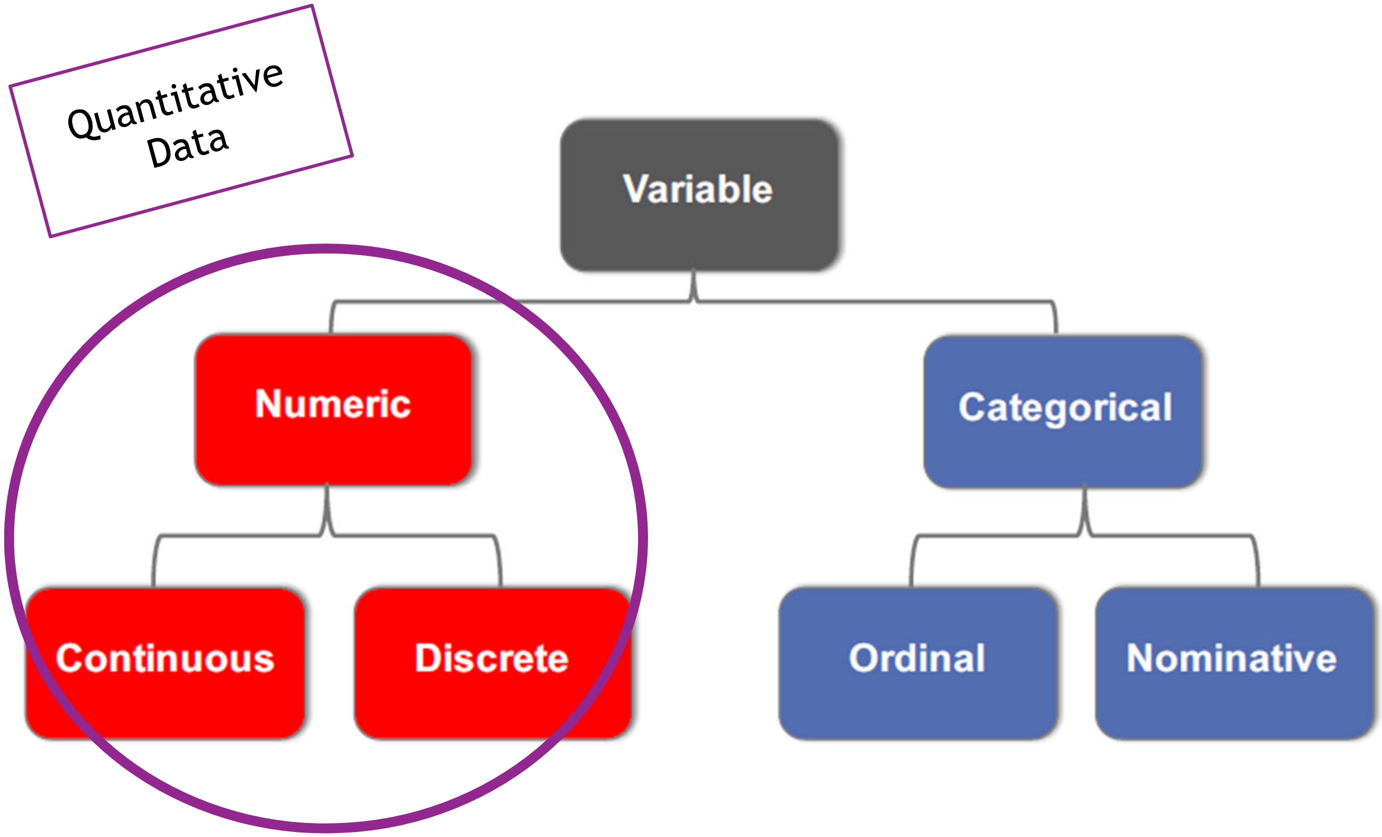
Review

Statistical Data Types



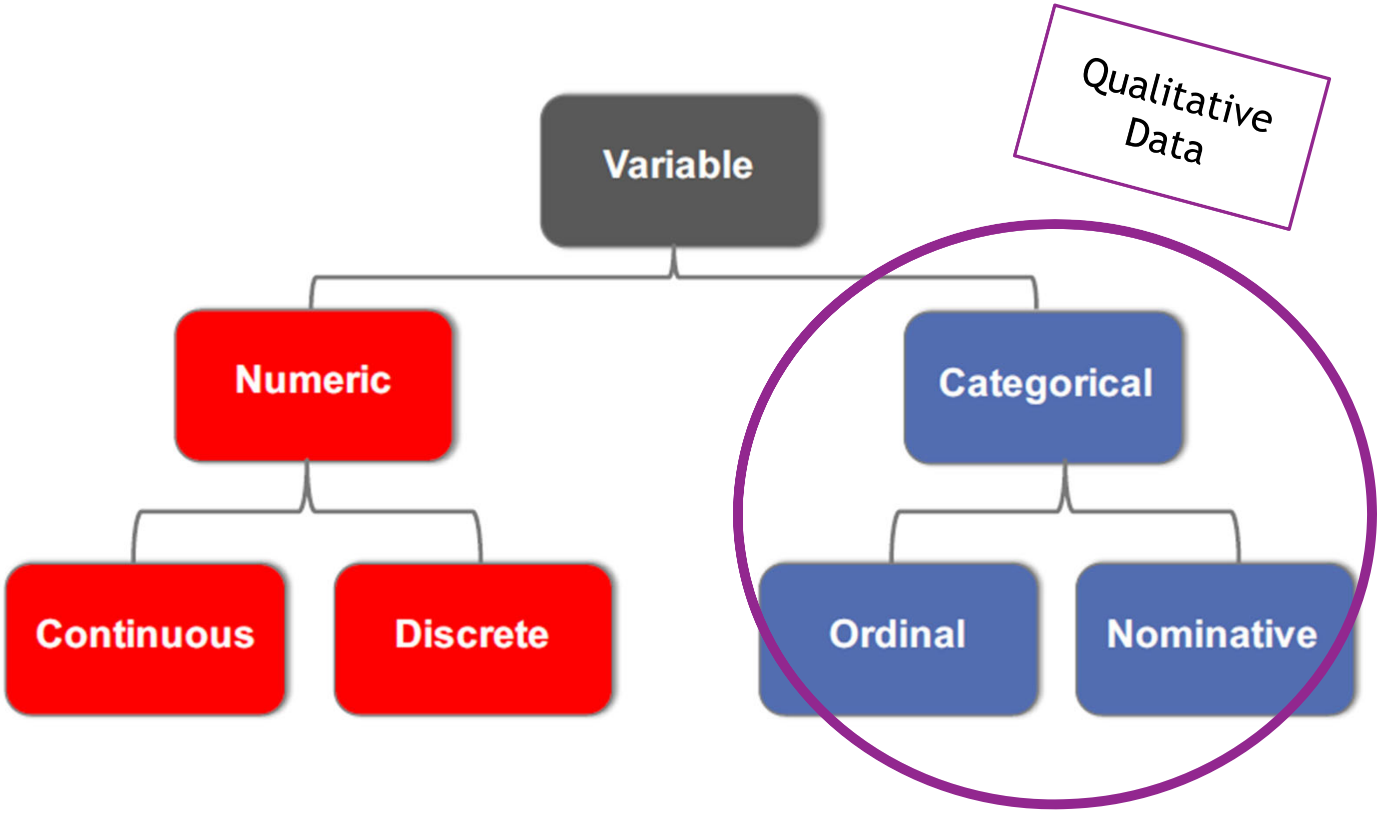
- We study data with different properties.
- Dividing these properties into types helps us to communicate the information behind the data
- We split between properties involving numbers for calculations and non-numbers for labels

Review



- Discrete
  - We can count discrete variables because they can take finitely many values
  - For example 1,2,3
- Continuous
  - We cannot count continuous variables because they can take infinitely many values
  - For example 1.54, 2.43, 3.14

Review



- Nominal
  - We use nominal data for labels to distinguish between different categories
  - For example blue, green
- Ordinal
  - If we can rank nominal data, then we have an order to the variables
  - For example high, medium, low

Agenda

These are properties of data as numbers rather than data as text, images, recordings, …

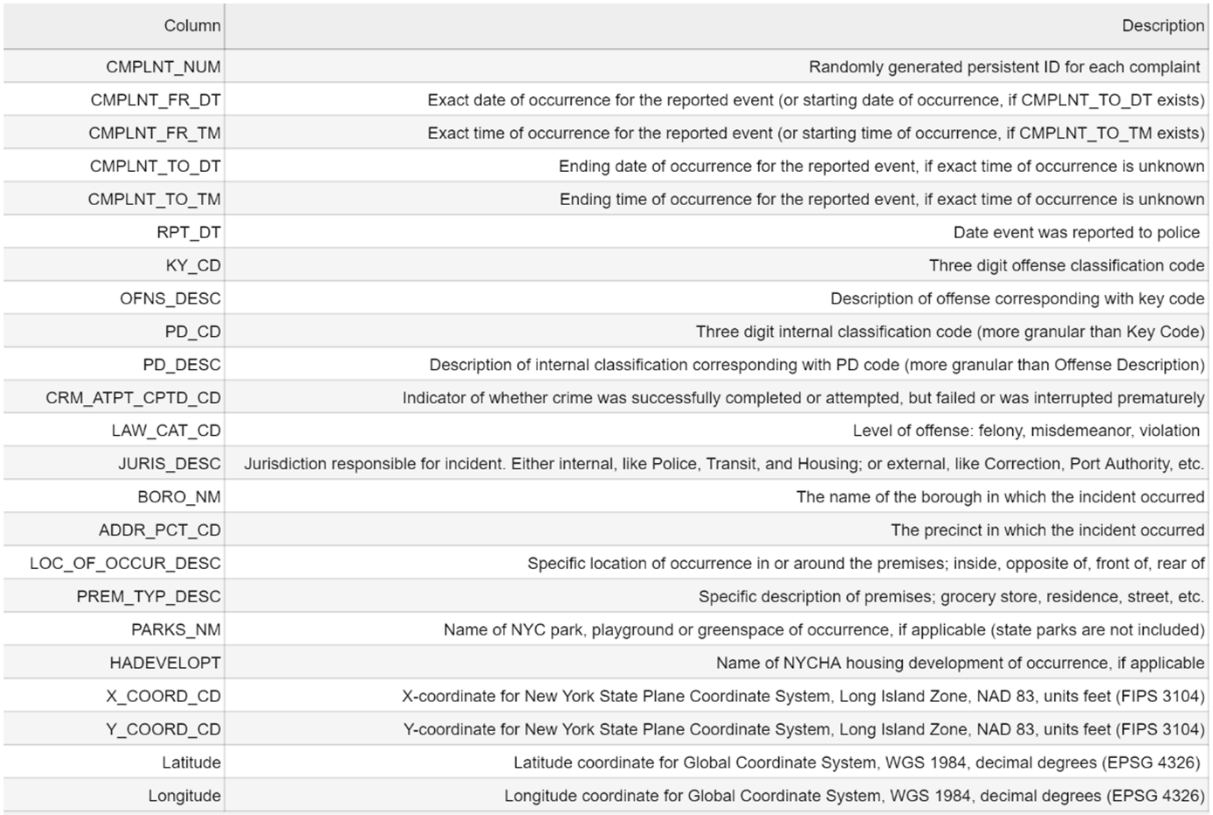
- Properties of Data in Tables
  - Scope
  - Granularity
  - Temporality
  - Faithfulness
- File Size
- Missing Values

References

- Nolan, Lau, Gonzalez (Chapter 5,6)

NYC Police Reports

Properties of Data in Tables



Column	Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

- Granularity
  - Amount of detail in the dataset
- Scope
  - Coverage of the dataset
- Temporality
  - Date and time of the information and the collection of information
- Faithfulness
  - Accuracy of the information

Granularity

Granularity of the data means the information behind the data particularly the amount of detail.

What does a record represent?

- For the NYC dataset each record corresponds to a crime reported to the NYPD

Do records have the same granularity?

- Sometimes we have records with composite data that summarizes many entries
- For the NYC dataset we see INTOXICATED & IMPAIRED DRIVING

If the data was aggregated, then what operation combined the number?

- Compare PD_CD and KY_CD in the NYC dataset

How can we find patterns through matching common entries.

- For the NYC dataset we could group by location to study crime in neighborhoods.

Scope

Scope refers to the coverage of the dataset. Note that coverage depends on the problem.

What does a record represent?

- For the NYC dataset each record corresponds to a crime reported to the NYPD

Do records have the same granularity?

- Sometimes we have records with composite data that summarizes many entries
- For the NYC dataset we see INTOXICATED & IMPAIRED DRIVING

If the data was aggregated, then what operation combined the number?

- Compare PD_CD and KY_CD in the NYC dataset

How can we find patterns through matching common entries.

- For the NYC dataset we could group by location to study crime in neighborhoods.

Temporality

Temporality refers to date and time of the record along with data and time of the record keeping.

What is the meaning of the dates and times?

For the NYC dataset each record has

- CMPLNT_FR_DT, CMPLNT_FR_TM: Exact date and time or starting date and time
- CMPLNT_TO_DT, CMPLNT_TO_TM: Ending date and time granted the exact time is unknown
- RPT_DT: Date of report

What is the representation of the date and time?

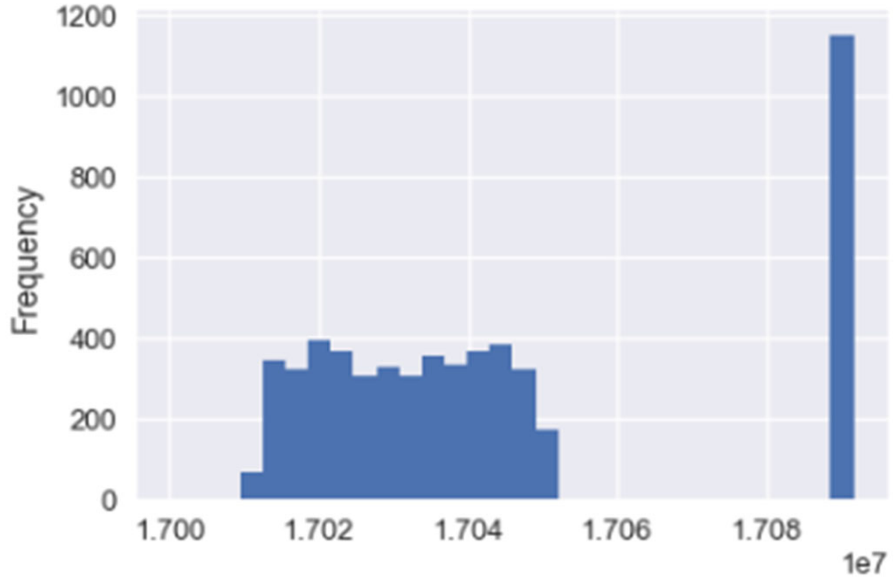
- For example do we have the format MM/DD/YYYY or DD/MM/YYY?

Do the timestamps make sense?

- Entry of 0 for a timestamp can get converted to the earliest date and time in the system.
- For Excel the default date is Jan 1st, 1990

Faithfulness

Faithfulness refers to the accuracy of the dataset. The accuracy depends on different factors such as the source of the data, the collection of the numbers and the entry of the records.



What could indicate inaccuracies in the dataset?

- Unrealistic or Incorrect Values
  - Spreadsheet has 255 columns or 65536 rows
- Inconsistent Values
  - Age and birthday are inconsistent
- Manual Entry
  - Typo such as spelling mistake
- Falsification
  - Zipcode 12345

Exercise

What to do with missing values?

Real world data is messy. Observations are often missing values for some variables. Which of the following approaches may be reasonable for dealing with this issue?

1. Drop the observations with missing values
2. Replace missing values with an average value
3. Replace missing values with comparable values from another dataset
4. Replace missing values with random values
5. Replace a missing value with the last present value
6. Ignore the missing values, they won’t affect our study anyways

Exercise

Imputation means replacing missing data with substitute data

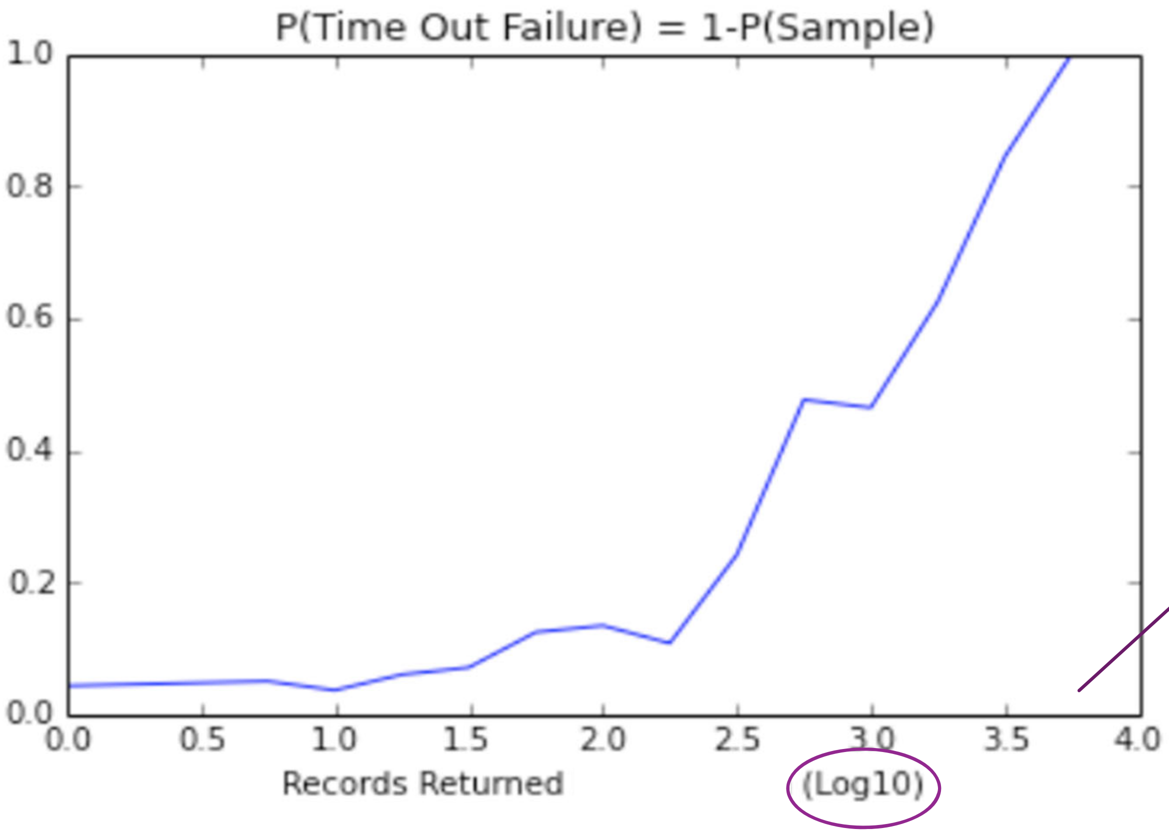
What to do with missing values?

Real world data is messy. Observations are often missing values for some variables. Which of the following approaches may be reasonable for dealing with this issue?

1. Drop the observations with missing values
2. Replace missing values with an average value
3. Replace missing values with comparable values from another dataset
4. Replace missing values with random values
5. Replace a missing value with the last present value
6. ~~Ignore the missing values, they won't affect our study anyways~~

Exercise

Week 3 Lecture 1



Here the probability of inclusion of the record in the sample changes depending on the size of the record

Note log stands for logarithm. We use a logarithmic scale for large numbers. You will study in Lab 5.

Exercise

Fill in the blank: When the missing values in a dataset are ______, it means that being missing is not correlated with other variables.

1. few
2. missing at random
3. missing not at random
4. plentiful

Exercise

Fill in the blank: When the missing values in a dataset are ______, it means that being missing is not correlated with other variables.

1. few
2. [✓ missing at random]
3. missing not at random
4. plentiful

Intuitively, correlated variables point in the same direction.

If one variable goes up then the other variable goes up. If one variable goes down, then the other variable goes down.

NYC Police Reports

Missing Values

- Missing values are omissions from the dataset. However, the omission can be represented in different ways
- Blank
  - Absence of a value have different implications.
  - If the data was a census collected annually, then was the value for that year never collected?
  - If the data was a survey, then did a respondent refuse to answer the question?
- Special Numbers such as 0 or 9999
- Special NULL, #NA, NAN,...

- Missing values are omissions from the dataset. However, the omission can be represented in different ways
- Blank
  - Absence of a value have different implications.
  - If the data was a census collected annually, then was the value for that year never collected?
  - If the data was a survey, then did a respondent refuse to answer the question?
- Special Numbers such as 0 or 9999
- Special NULL, #NA, NAN,...

Missing Values

Missing values are omissions from the dataset. However, the absent values can be represented in different ways.

Blank

- Absence of a value have different implications.
- If the data was a census collected annually, then was the value for that year never collected?
- If the data was a survey, then did a respondent refuse to answer the question?

Special Numbers

- Large numbers like 9999
  - If the dataset contains the age of mothers, then large value should be interpreted as missing value
- Small number like 0
  - 0 for latitude and longitude interpreted as 0°00'00.0"N+0°00'00.0"E island off the coast of Africa
  - 0 date interpreted as 1970-01-01T00:00:00Z

Missing Values

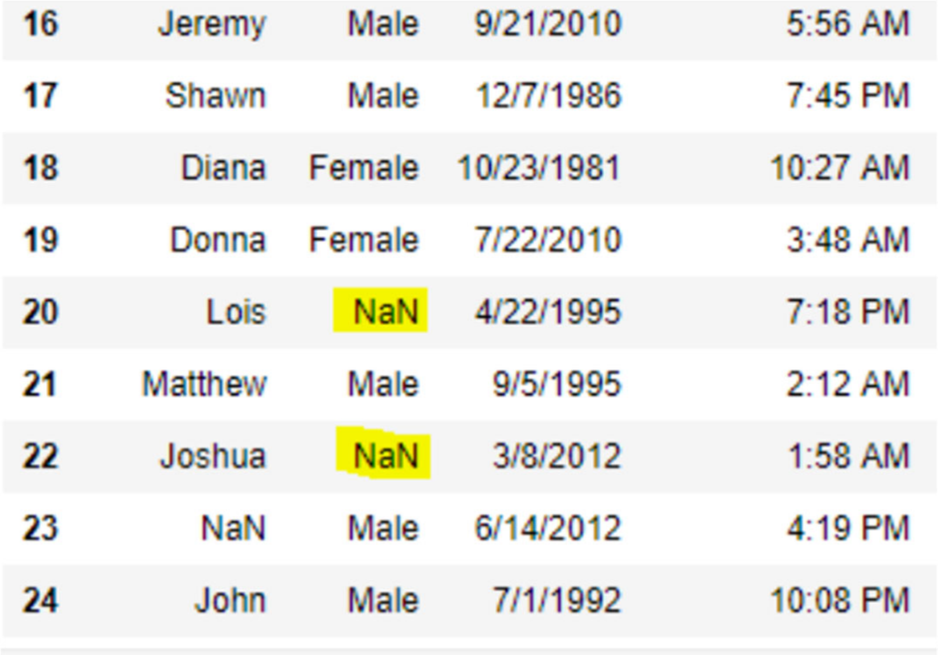
Missing values are omissions from the dataset. However, the absent values can be represented in different ways.

Special Characters

- NULL common for tables in databases
- #NA common for spreadsheets
- NaN common for programming languages
- In Python we use

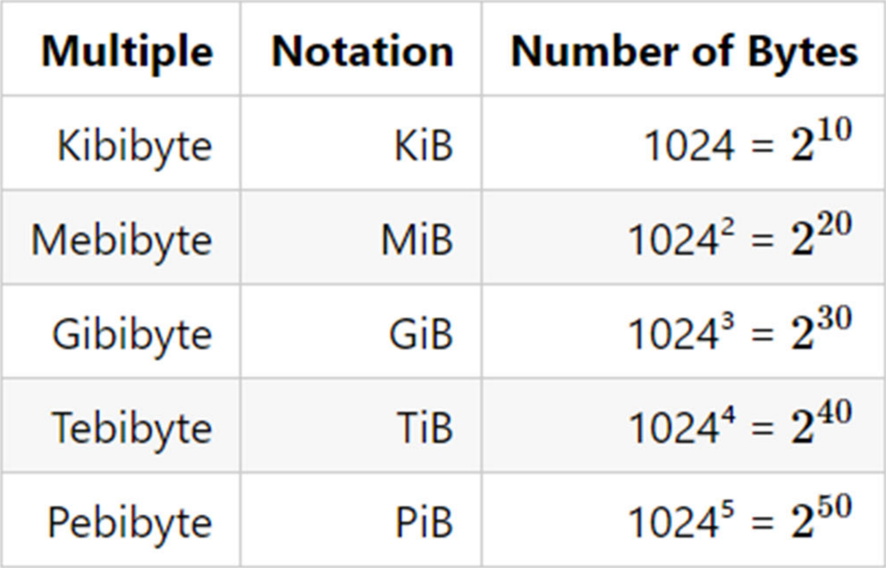
numpy.NaN

to indicate missing values



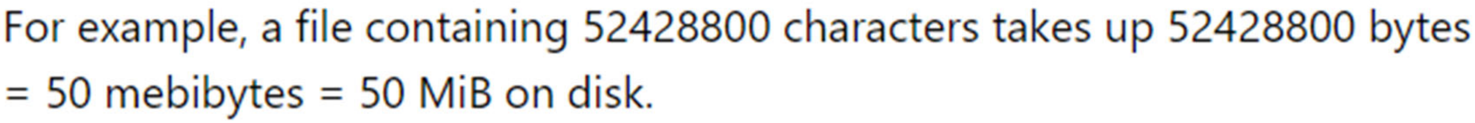
16	Jeremy	Male	9/21/2010	5:56 AM
17	Shawn	Male	12/7/1986	7:45 PM
18	Diana	Female	10/23/1981	10:27 AM
19	Donna	Female	7/22/2010	3:48 AM
20	Lois	NaN	4/22/1995	7:18 PM
21	Matthew	Male	9/5/1995	2:12 AM
22	Joshua	NaN	3/8/2012	1:58 AM
23	NaN	Male	6/14/2012	4:19 PM
24	John	Male	7/1/1992	10:08 PM

File Size



Multiple	Notation	Number of Bytes
Kibibyte	KiB	$1024 = 2^{10}$
Mebibyte	MiB	$1024^2 = 2^{20}$
Gibibyte	GiB	$1024^3 = 2^{30}$
Tebibyte	TiB	$1024^4 = 2^{40}$
Pebibyte	PiB	$1024^5 = 2^{50}$

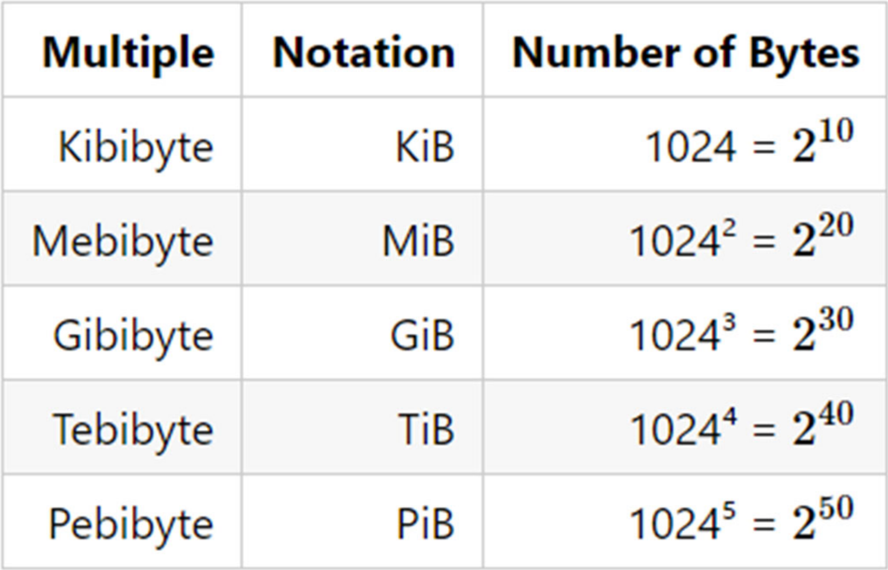
- The size of a file depends on the size of the underlying data. The byte is a unit of information on the computer consisting of eight bits. Each bit is 0 or 1 to indicate off or on in the storage.
- Since we need to represent numbers as a sequence of bits on the computer, we tend to work with the number 2 instead of the number 10



For example, a file containing 52428800 characters takes up 52428800 bytes = 50 mebibytes = 50 MiB on disk.

Command Line

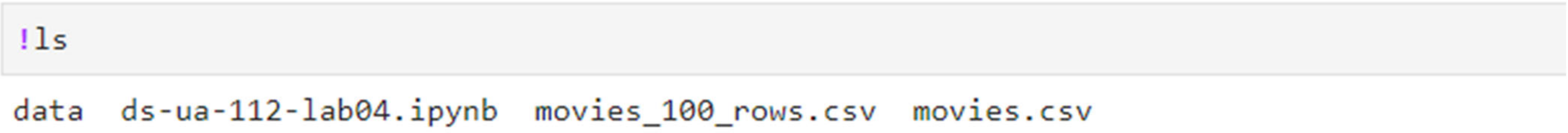
- Computers store data in memory for immediate access.
  - Random-access memory (RAM) is a form of computer memory that can be retrieved or modified in any order
- Programs share memory on the computer
  - Computer with 4 GiB total RAM might have only 1 GiB available RAM
  - With 1 GiB available RAM, pandas will not be able to read in a 1 GiB file.
  - Usually pandas needs twice the size of the file in available memory.



Multiple	Notation	Number of Bytes
Kibibyte	KiB	$1024 = 2^{10}$
Mebibyte	MiB	$1024^2 = 2^{20}$
Gibibyte	GiB	$1024^3 = 2^{30}$
Tebibyte	TiB	$1024^4 = 2^{40}$
Pebibyte	PiB	$1024^5 = 2^{50}$

Command Line

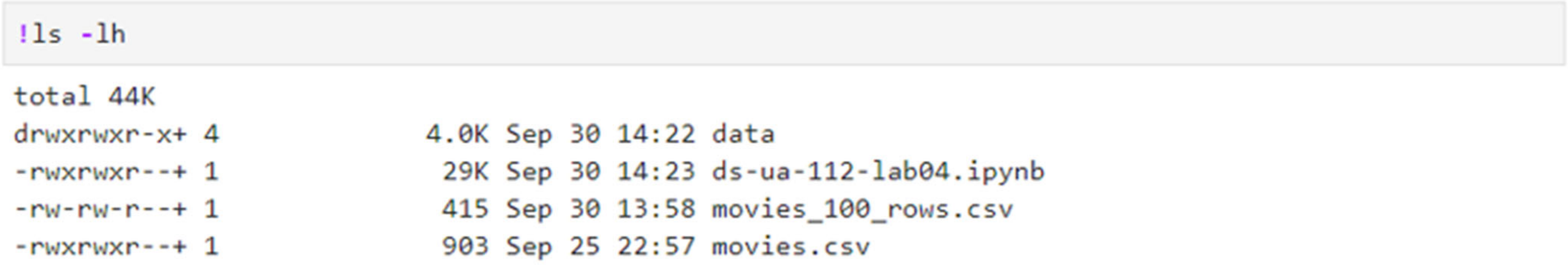
- Computers provide access to a command line interface. Users input commands to perform operations on the computer particularly files.



```
!ls
```

```
data   ds-ua-112-lab04.ipynb   movies_100_rows.csv   movies.csv
```

- We can enter commands in Jupyter notebook using exclamation point



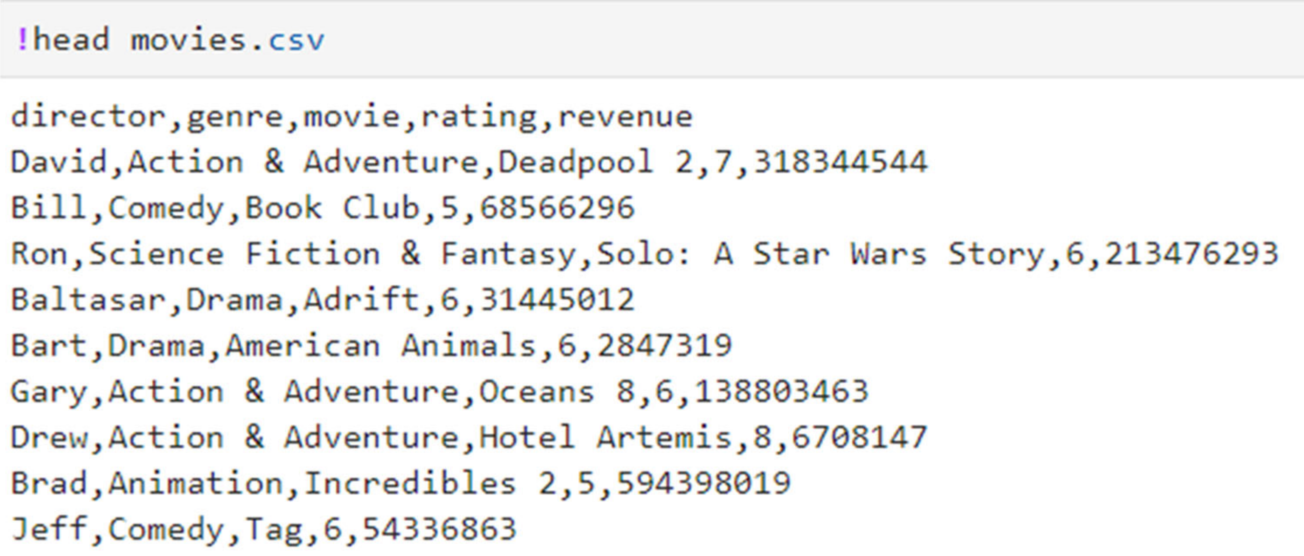
```
!ls -lh
```

```
total 44K
drwxrwxr-x+ 4             4.0K Sep 30 14:22 data
-rwxrwxr--+ 1              29K Sep 30 14:23 ds-ua-112-lab04.ipynb
-rw-rw-r--+ 1              415 Sep 30 13:58 movies_100_rows.csv
-rwxrwxr--+ 1              903 Sep 25 22:57 movies.csv
```

- Note that the commands differ across operating systems. Here we have the commands for the Linux operating system on JupyterHub.

Command Line

- Some commands for accessing files include
  - head
    - Returns the first 10 rows of the file
  - tail
    - Returns the last 10 rows of the file
  - cat
    - Returns all rows of the file

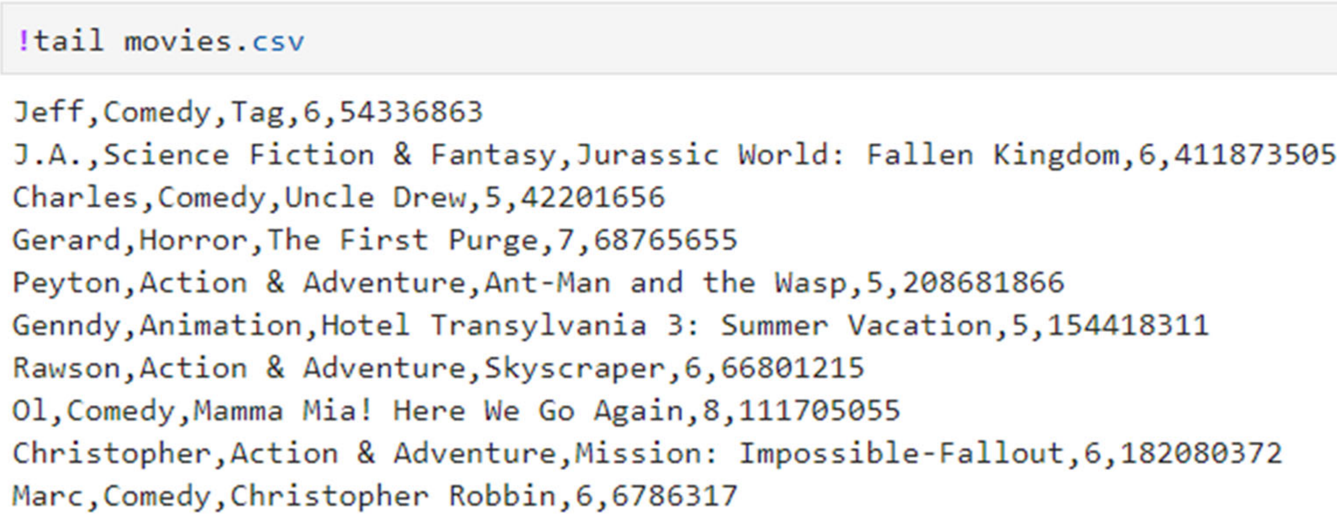


```
!head movies.csv
```

```
director,genre,movie,rating,revenue
David,Action & Adventure,Deadpool 2,7,318344544
Bill,Comedy,Book Club,5,68566296
Ron,Science Fiction & Fantasy,Solo: A Star Wars Story,6,213476293
Baltasar,Drama,Adrift,6,31445012
Bart,Drama,American Animals,6,2847319
Gary,Action & Adventure,Oceans 8,6,138803463
Drew,Action & Adventure,Hotel Artemis,8,6708147
Brad,Animation,Incredibles 2,5,594398019
Jeff,Comedy,Tag,6,54336863
```

Command Line

- Some commands for accessing files include
  - head
    - Returns the first 10 rows of the file
  - tail
    - Returns the last 10 rows of the file
  - cat
    - Returns all rows of the file

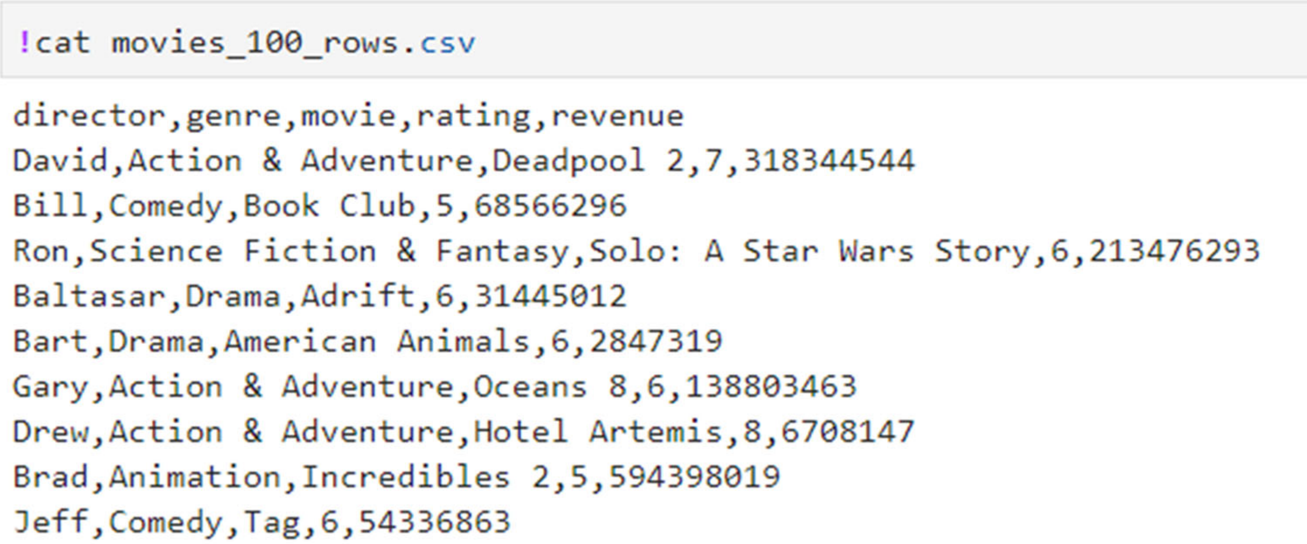


```
!tail movies.csv
```

```
Jeff,Comedy,Tag,6,54336863
J.A.,Science Fiction & Fantasy,Jurassic World: Fallen Kingdom,6,411873505
Charles,Comedy,Uncle Drew,5,42201656
Gerard,Horror,The First Purge,7,68765655
Peyton,Action & Adventure,Ant-Man and the Wasp,5,208681866
Genndy,Animation,Hotel Transylvania 3: Summer Vacation,5,154418311
Rawson,Action & Adventure,Skyscraper,6,66801215
Ol,Comedy,Mamma Mia! Here We Go Again,8,111705055
Christopher,Action & Adventure,Mission: Impossible-Fallout,6,182080372
Marc,Comedy,Christopher Robbin,6,6786317
```

Command Line

- Some commands for accessing files include
  - head
    - Returns the first 10 rows of the file
  - tail
    - Returns the last 10 rows of the file
  - cat
    - Returns all rows of the file

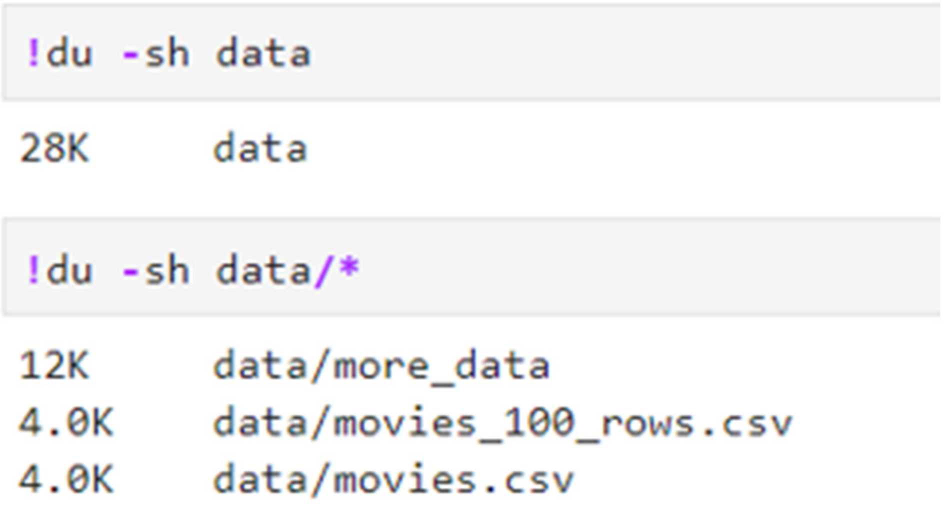


```
!cat movies_100_rows.csv
```

```
director,genre,movie,rating,revenue
David,Action & Adventure,Deadpool 2,7,318344544
Bill,Comedy,Book Club,5,68566296
Ron,Science Fiction & Fantasy,Solo: A Star Wars Story,6,213476293
Baltasar,Drama,Adrift,6,31445012
Bart,Drama,American Animals,6,2847319
Gary,Action & Adventure,Oceans 8,6,138803463
Drew,Action & Adventure,Hotel Artemis,8,6708147
Brad,Animation,Incredibles 2,5,594398019
Jeff,Comedy,Tag,6,54336863
```

Command Line

- Some commands for determining the size of files include
  - ls -lh
    - Will list all files along with properties
  - du -sh
    - Will calculate the size of files or folders
- Note that du is more accurate than ls



```
!du -sh data
```

```
28K	data
```

```
!du -sh data/*
```

```
12K	data/more_data
4.0K	data/movies_100_rows.csv
4.0K	data/movies.csv
```

Summary

- Properties of Data in Tables
  - Scope
  - Granularity
  - Temporality
  - Faithfulness
- File Size
- Missing Values

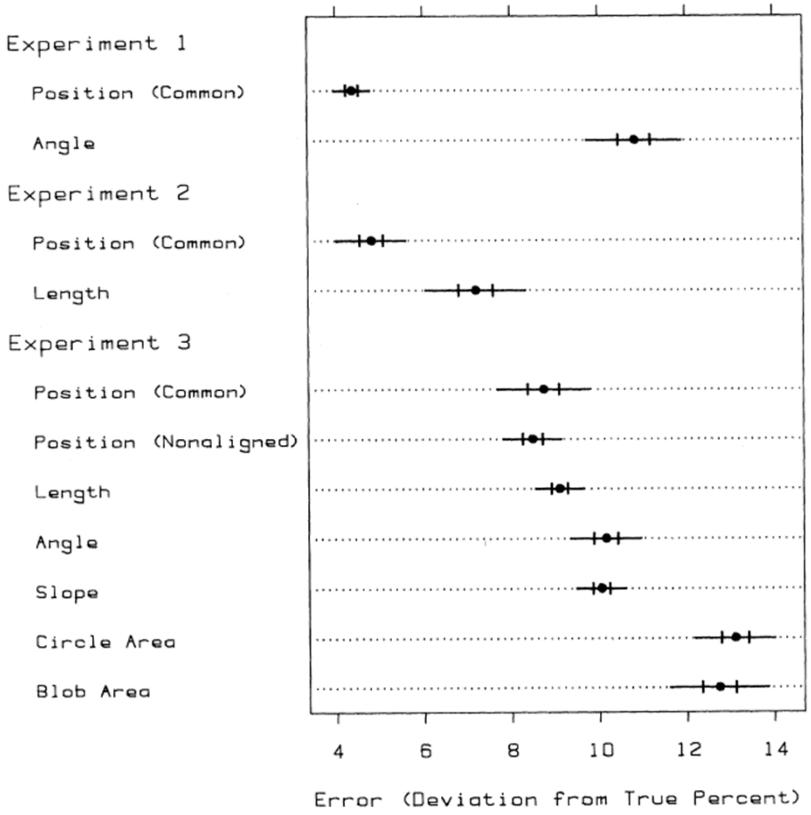
Goals

- Describing Tables
- Working with Messy Data
- Assessing File Size

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

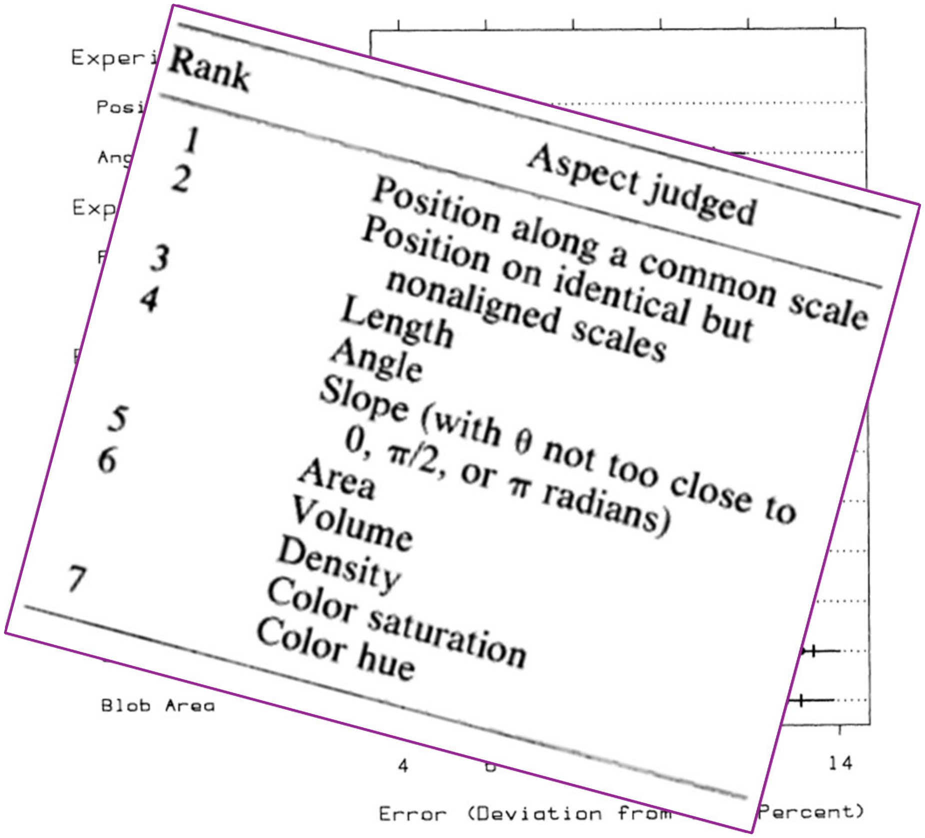
What aspects of charts are most understandable to you?



Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

What aspects of charts are most understandable to you?


Experi
Posi
Ang
Exp
Blob Area
4
6
14
Error (Deviation from
Percent)

Rank	Aspect judged
1	Position along a common scale
2	Position on identical but nonaligned scales
3	Length
4	Angle Slope (with θ not too close to 0, π/2, or π radians)
5	Area
6	Volume Density Color saturation
7	Color hue