Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Project pitches
- Cleaning and Exploratory data analysis
- HW for Monday

Cleaning data

- Fixing formats
- Missing values?
- Correcting erroneous values
- Standardizing categories

- http://www.kdnuggets.com/2016/03/doing-data-science-kaggle-walkthrough-cleaning-data.html

Fixing formats

- Often when data is saved or translated from one format to another (for example from CSV to Python), some data may not be translated correctly.
- For example a column may contain numbers like 20090609231247 instead of timestamps in the expected format: 2009-06-09 23:12:47.
- A typical job when it comes to cleaning data is correcting these types of issues.

Missing values?

- Structural vs idiosyncratic
    - Structural: which represent measurements that can't be made (e.g., the count of pregnant males) can be safely removed.

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Table 3: The same data as in Table 1 but with variables in columns and observations in rows.

Missing values?

- Ignoring rows with missing values
    - **Only at analysis time!**
- This approach only makes sense if the number of rows with missing data is relatively small compared to the dataset. If you are deleting more than around 10% of your dataset due to rows having missing values, you may need to reconsider.
- **Avoid if at all possible!**

Missing values?

- Fill in missing value
    - If the data is categorical (i.e. countries, device types, etc.), may create a new category that will represent ‘unknown’.
    - It is common to fill the values with the most common value for that column (the mode)
    - May use imputation
- **Don’t fill in missing values!**

Correcting erroneous values

- For some columns, there are values that can be identified as obviously incorrect.
    - This may be a ‘gender’ column where someone has entered a number, or an ‘age’ column where someone has entered a value well over 100.
    - Or it is common to have columns that are temporally dependent such that one could not have happened before another
- These values either need to be corrected (if the correct value can be determined), assumed to be missing, or one can create a special datatype for suspected errors.
- Write scripts to “unit test” these kinds of problems!

Standardizing categories

- In many cases where data is collected from users directly – particularly using free text fields – spelling mistakes, language differences or other factors will result in a given answer being provided in multiple ways.

- For example, when collecting data on country of birth, if users are not provided with a standardized list of countries, the data will inevitably contain multiple spellings of the same country (e.g. USA, United States, U.S. and so on).

- One of the main cleaning tasks often involves standardizing these values to ensure that there is only one version of each value.

Other bits

- Remove duplicates, if any
- Replace values: e.g. replace 1 with 7 for the whole dataset (the dimension doesn't change – only all the 1's are replaced with 7). More generally: regexps.
- Rename index: e.g. change index from 0 to "person1"
- Create new variables: E.g. create a new col "CITIZEN"
- Rename variables: E.g. rename two columns: "ADM_RNO" to something human-understandable

- https://mdl.library.utoronto.ca/technology/tutorials/cleaning-data-python#clean

Tamr example

- Data cleaning and unifying is a *big* problem/ opportunity
- https://www.tamr.com/video/enterprise-data-unification-solution-video/

Exploratory data analysis

- Detect mistakes
- Check assumptions
- Preliminary selection of appropriate models
- Determining relationships among variables
- Assessing the direction and rough size of relationships among variables

- http://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

Exploratory data analysis

- Is there a hypothesis?
- Is there a question being answered?

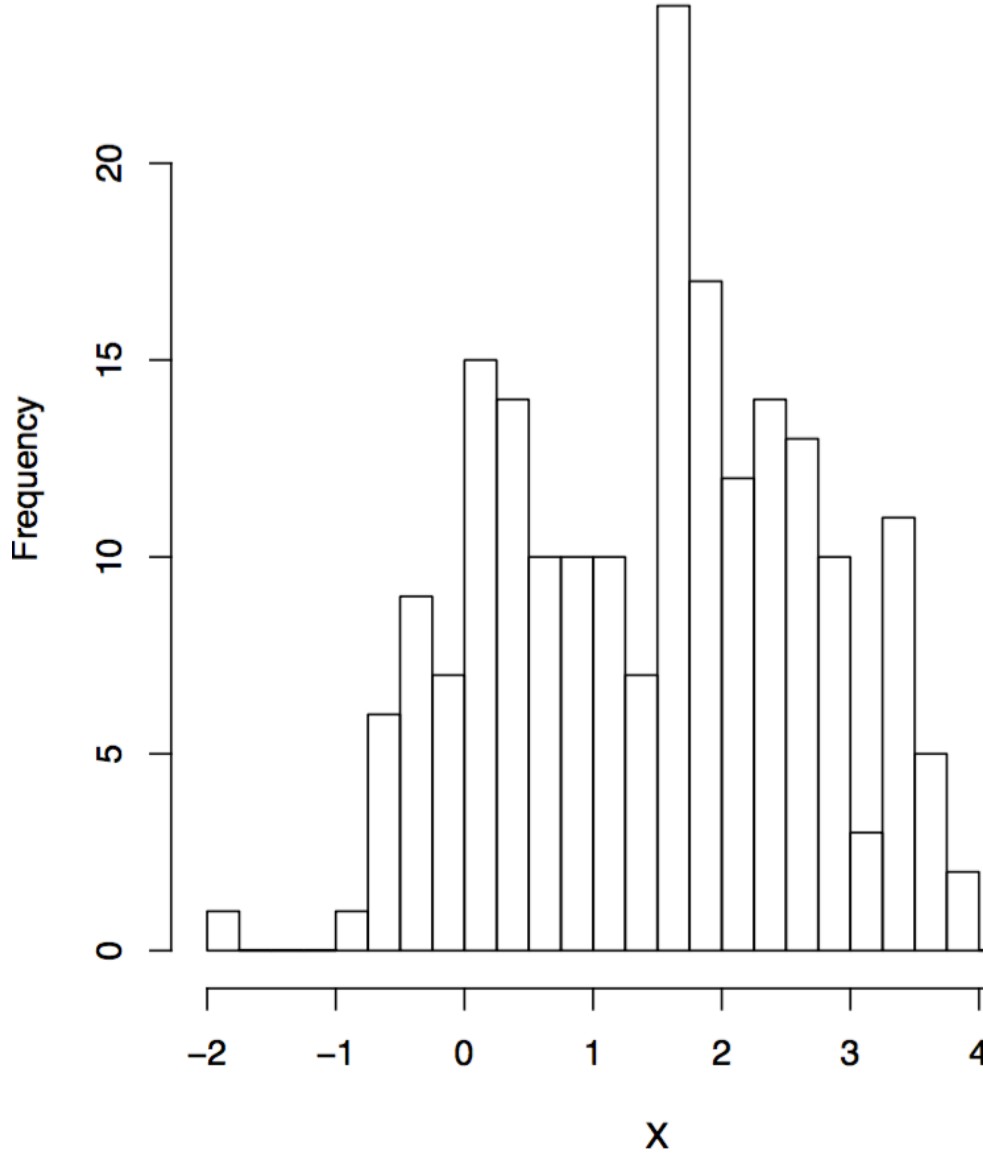
- Four types of EDA:
  - univariate non-graphical,
  - multivariate non-graphical,
  - univariate graphical, and
  - multivariate graphical.

Univariate non-graphical EDA

- Characteristics of quantitative data
    - Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.
    - The characteristics of the population distribution of a quantitative variable are its center, spread, modality (number of peaks in the pdf), shape (including “heaviness of the tails”), and outliers.
    - The characteristics of our randomly observed sample are not inherently interesting, except to the degree that they represent the population that it came from.

Univariate non-graphical EDA

- Characteristics of quantitative data
    - Histogram
    - Modes? Shape? Outliers?



Univariate non-graphical EDA

- Central tendency
  - Mean (arithmetic)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

  - Median
  - Mode
- Robustness?

Univariate non-graphical EDA

- Spread
  - Variance

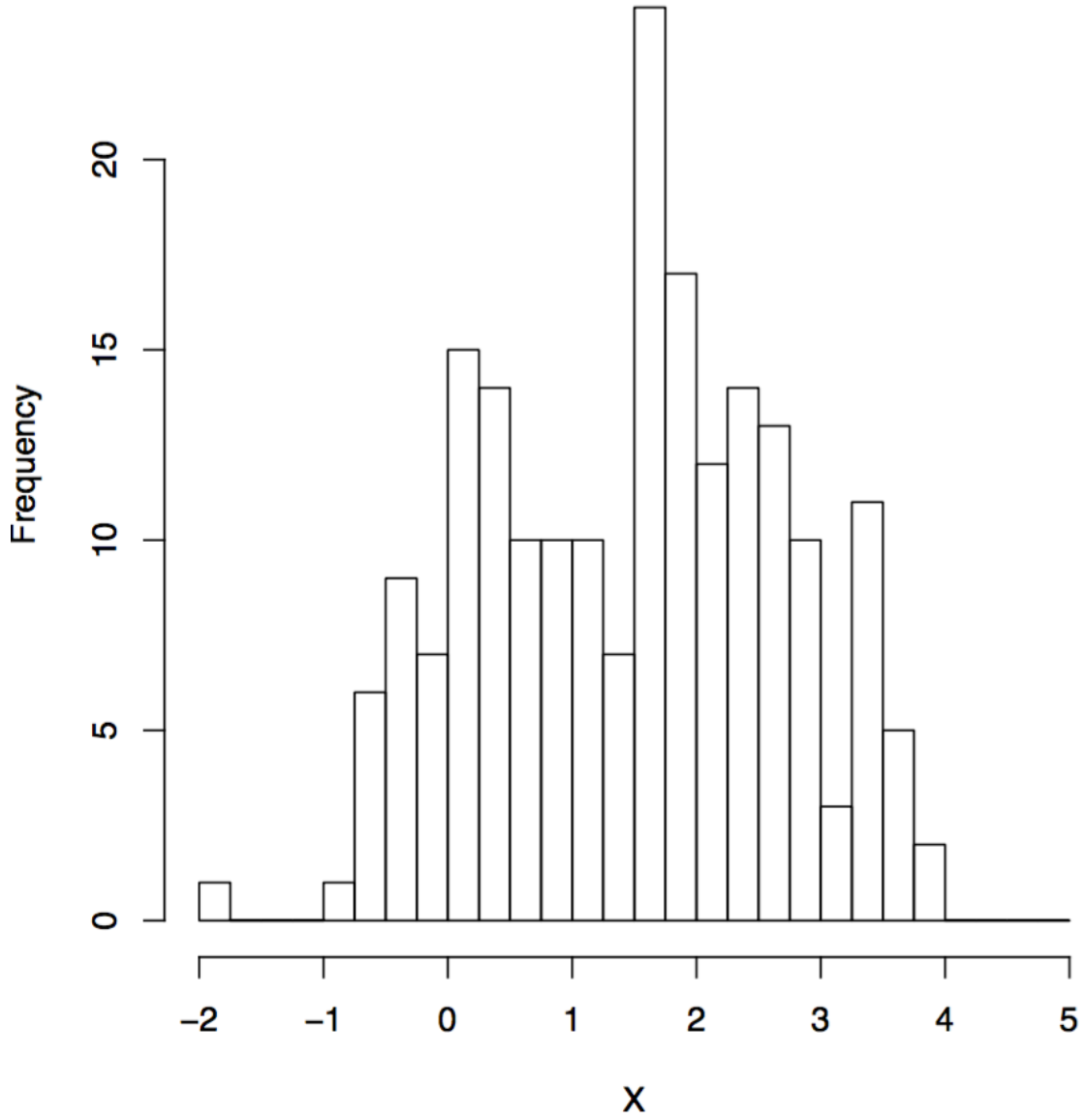
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

  - Inter quartile range (IQR)

- Robustness?

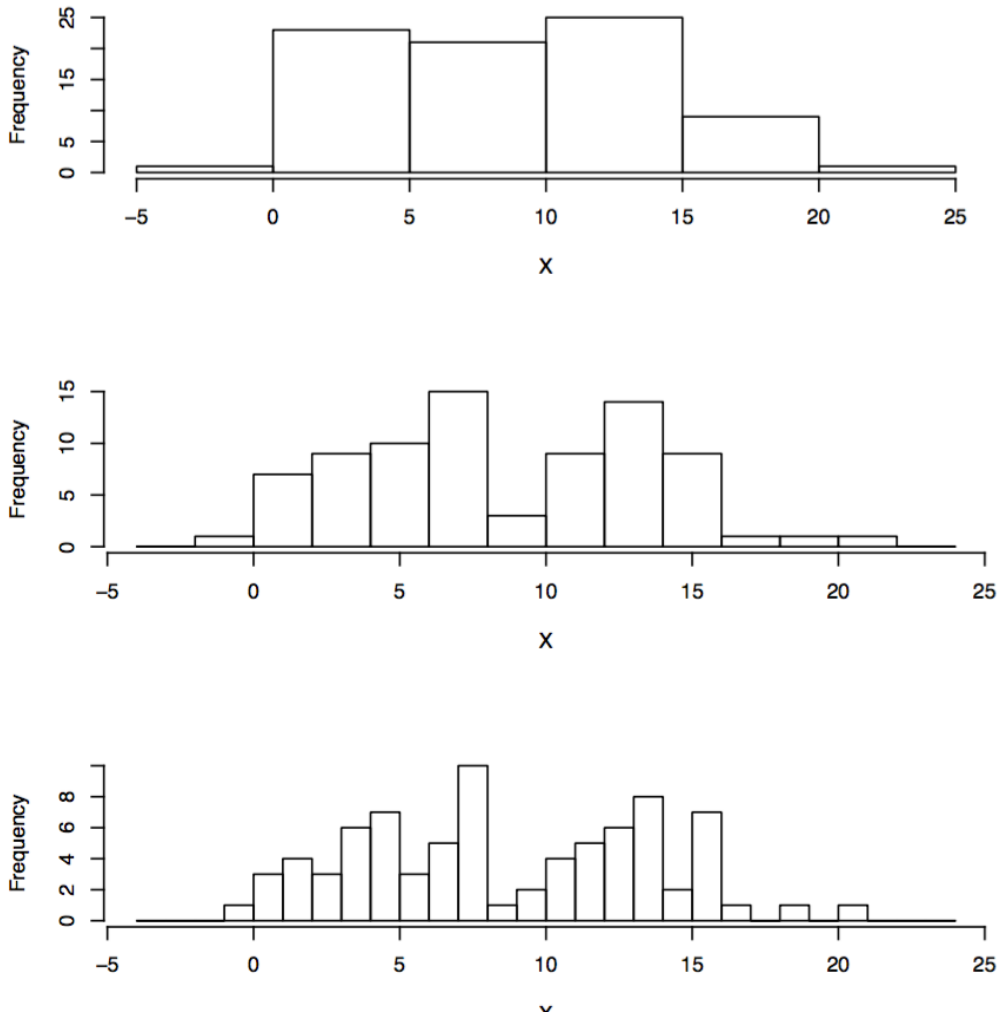
Univariate graphical EDA

- Histogram



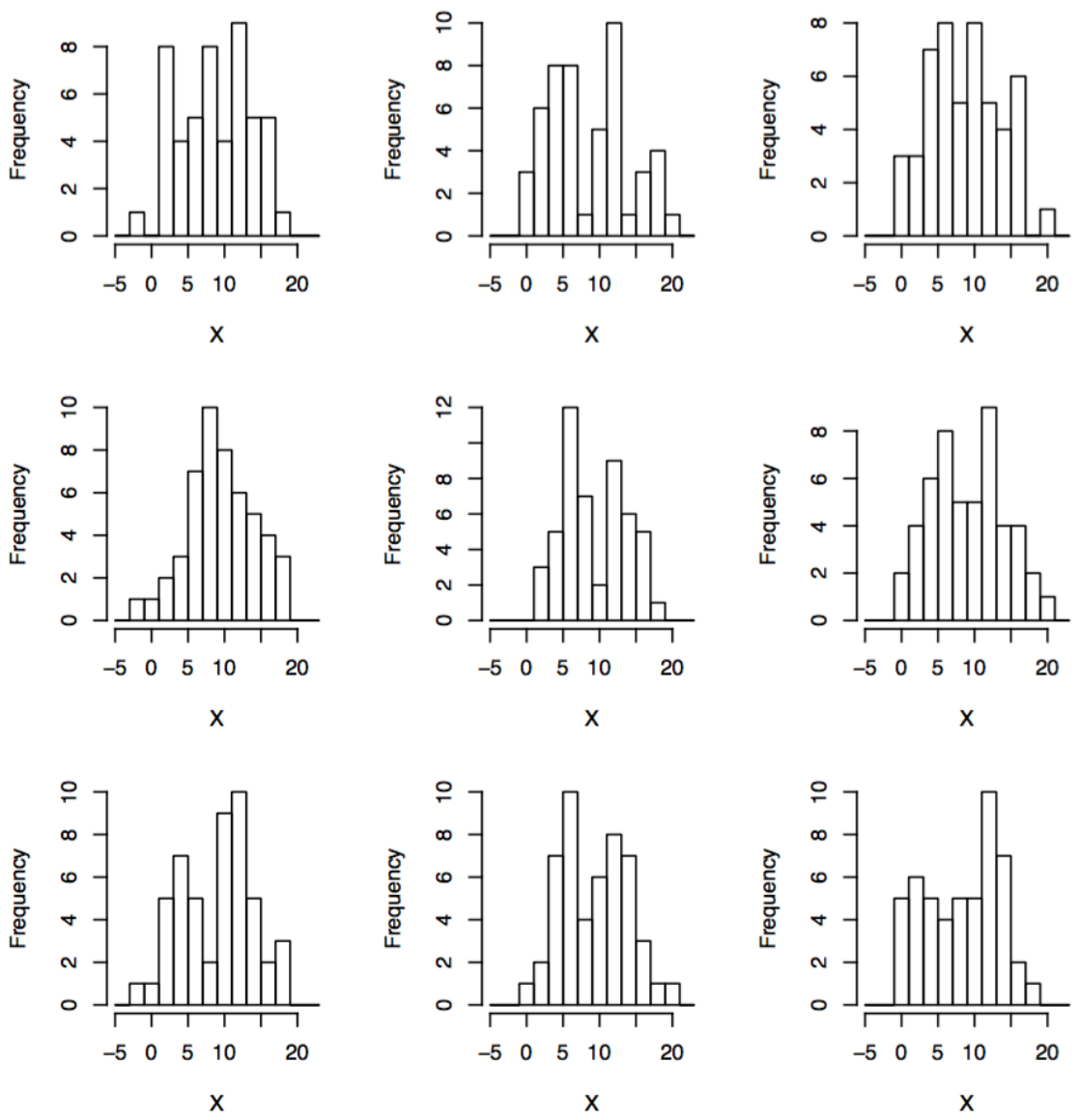
Univariate graphical EDA

- Histogram
  - Bin widths matter!



Univariate graphical EDA

- Histogram
  - Variability is expected!

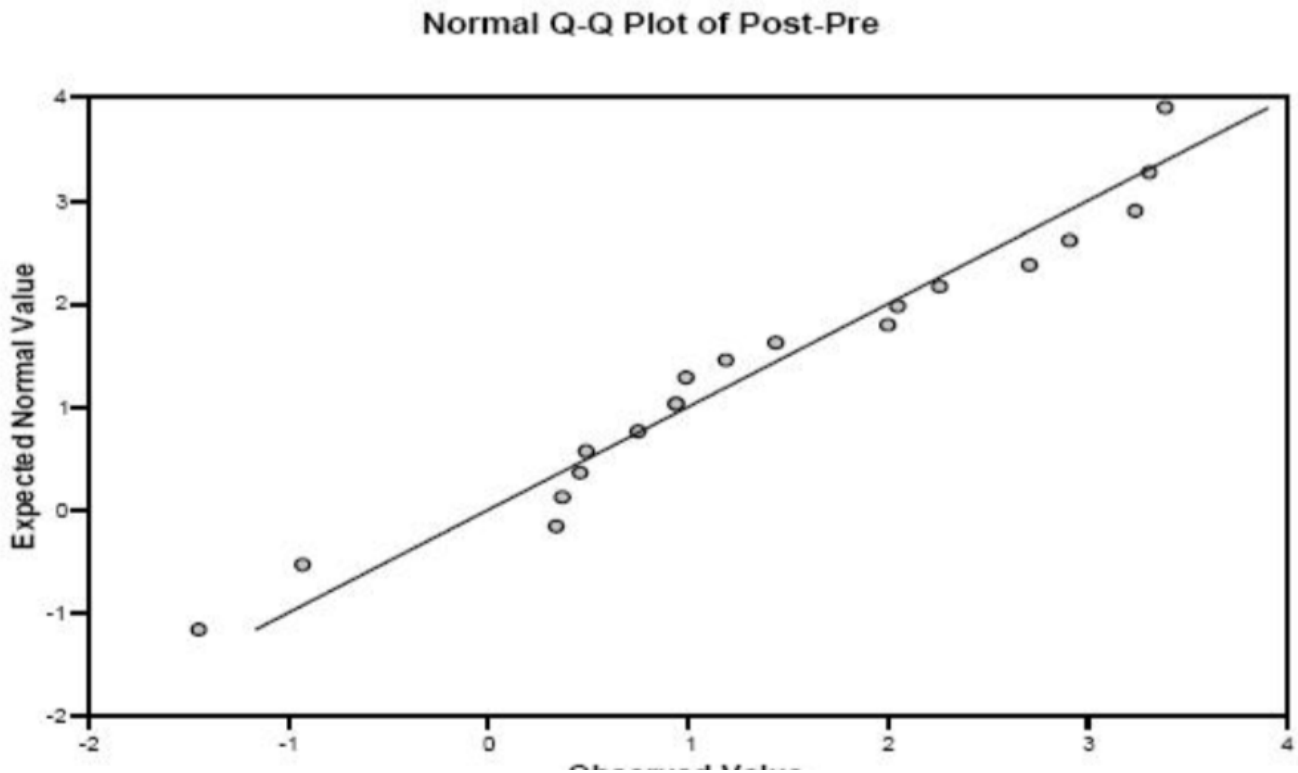


Univariate graphical EDA

- Also, box plots, violin plots

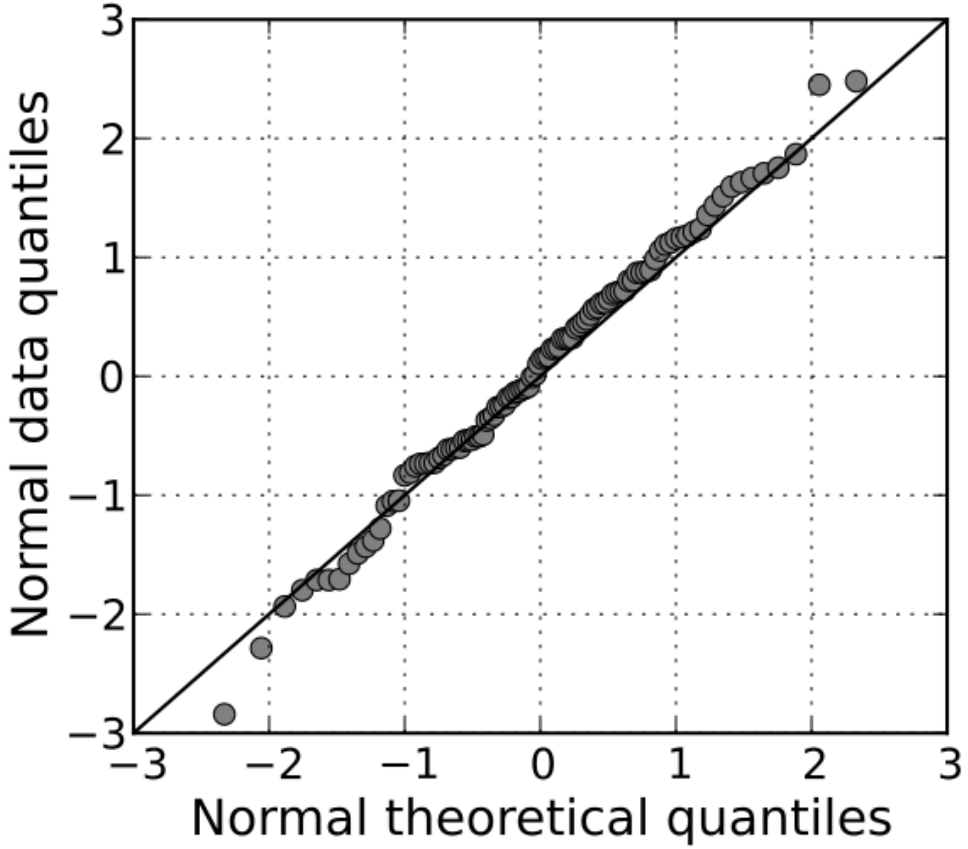
Univariate graphical EDA

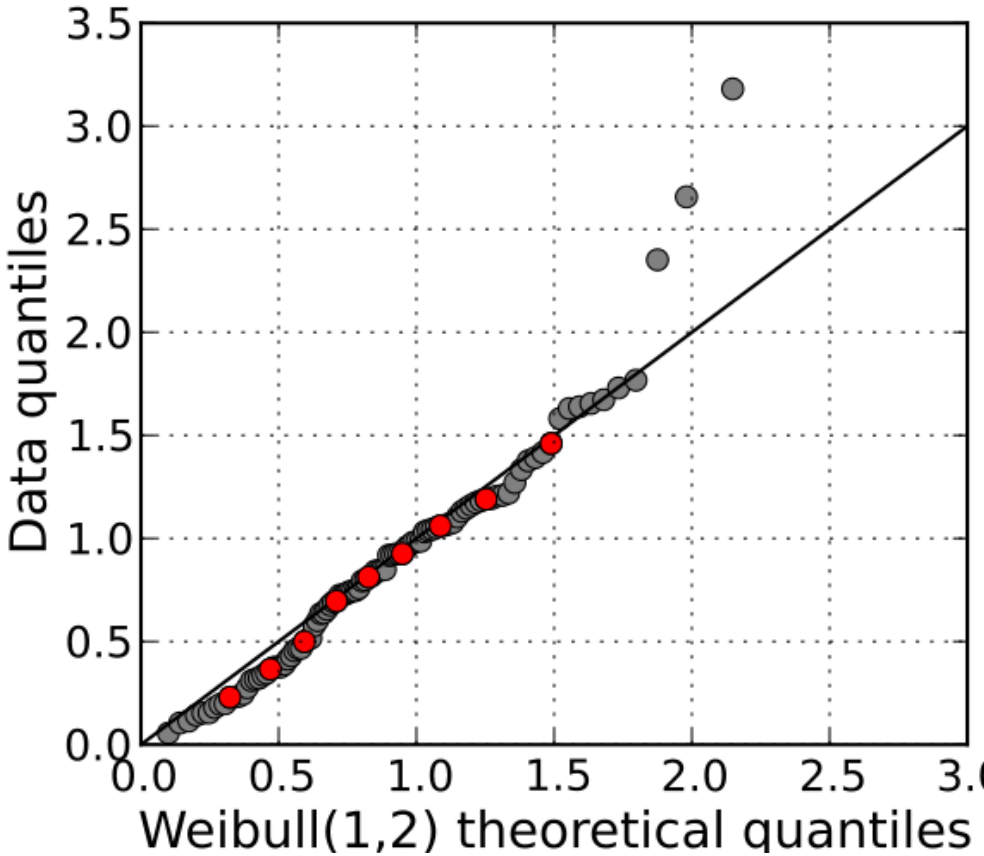
- QQ plot: Compare observed values to some expected or observed values to assess whether they are from the same distribution. The comparison distribution is commonly a Normal (Guassian)



Univariate graphical EDA

- QQ plot:





Multivariate non-graphical EDA

- Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

- Cross-tabs

  - For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels

Multivariate non-graphical EDA

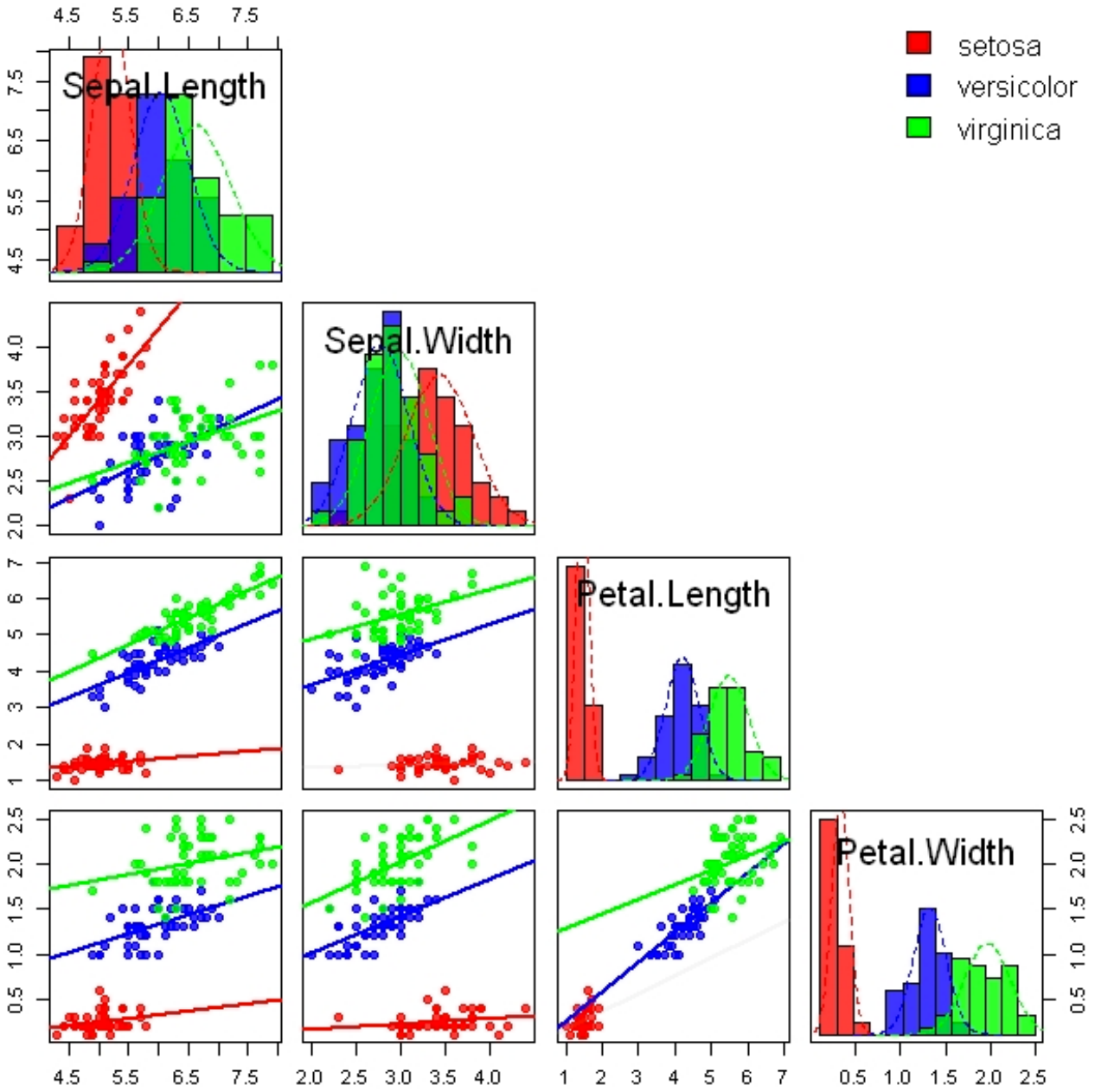
Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Data

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Crosstab

Scatter matrix



- **HW: Pick one data set, write notebook that downloads and cleans the data (for general purpose analyzing), with explanations**

- NYC open data
    - https://opendata.cityofnewyork.us/data/#datasetscategory
    - Examples:
        - http://blog.nycdatascience.com/student-works/r-shiny/noise-coming-case-study-nycs-311-noise-complaints/
        - http://blog.nycdatascience.com/student-works/new-york-city/