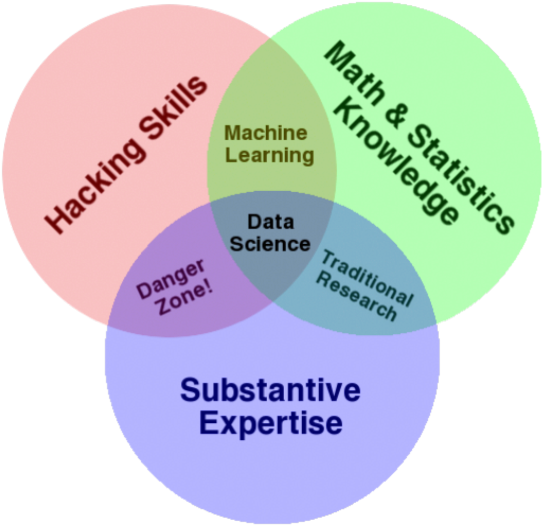


COMM054: Data Science Principles & Practices

UNIVERSITY OF SURREY

Introduction to Data Visualisaton

Dr. Manal Helal

24 October 2019

Outline

- Exploratory Data Analysis
- Distributions & Outliers
- Developing a Visualization Aesthetic
- Chart Types

We will be experimenting with the body measurement data set ***NHANES***, available at:
https://www.statcrunch.com/app/index.php?dataid=1406047

Exploratory Data Analysis

- Answer the following Questions:
  - Who constructed this data set, when, and why?
    - National Health and Nutrition Examination Survey 2009–2010
  - How big is it?
    - This data set has 4978 records (2452 men and 2526 women), each with seven data fields plus gender.
  - What do the fields mean?
    - Which fields are numerical or categorical? – Gender?
    - What units were the quantities measured in? lengths and weight are metric
    - Which fields are IDs or descriptions, instead of data to compute with?

Exploratory Data Analysis - Cont’d

- Look for familiar or interpretable records
- Summary statistics
  - Tukey’s five number summary for numerical values: max, min, median and quartile elements
  - Unique Labels for categorical values, and the frequencies.
- Pairwise correlations
  - Either all pairs, or at least all columns against a dependent variable of interest
  - What is the correlation between height and weight for men vs. women? Can you justify this?
- Class breakdowns
- Plots of distributions

Distributions & Outliers

- You will be given 4 datasets, with identical means, variances, and the exact same correlation between the x and y values.

- First one has a linear trends linear,
- The second looks almost parabolic.
- Two others are almost perfectly linear modulo outliers, but with wildly different slopes.

- Dot Plots makes you instantly see the outliers and the distributions.

Developing a Visualization Aesthetic

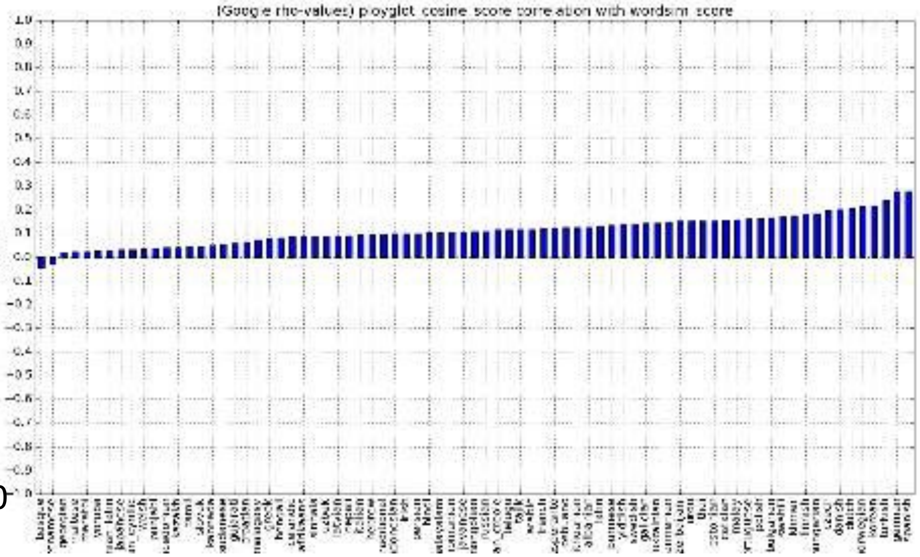
- Maximize data-ink ratio by eliminating background grids, shading, shadows, tic-marks, and use 2D when 3D is not telling more.
- Minimize the lie factor. This could happen if you:
  - Presenting means without variance
  - Presenting interpolations without the actual data
  - Distortions of scale:
    - Golden ratio: width should be 1.6 times the height.
    - 45 degree lines are the most readily interpretable
  - Eliminating tick labels from numerical axes
  - Hide the origin point from the plot

Developing a Visualization Aesthetic – Cont’d

- Minimize chartjunk created by packages
- Use proper scales and clear labeling
- Make effective use of color to highlight data properties
- Exploit the power of repetition

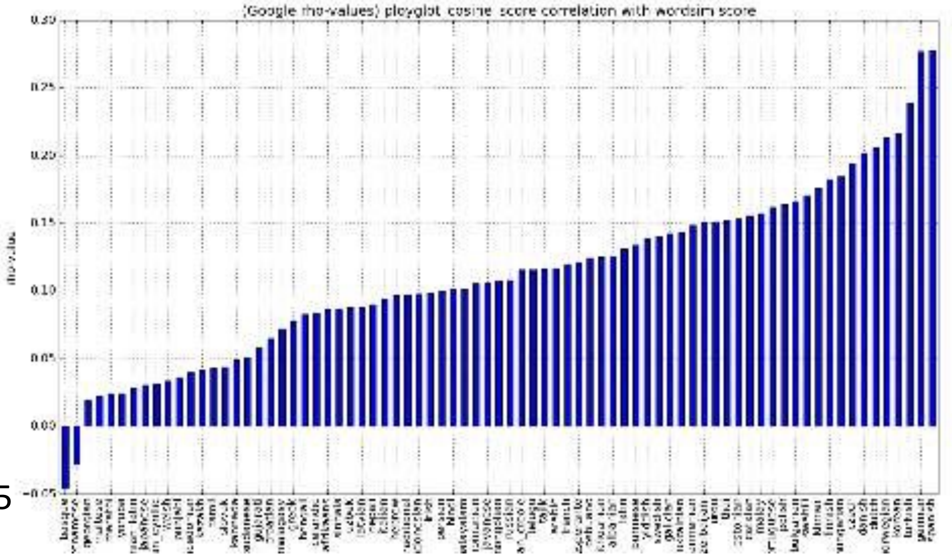
1.00

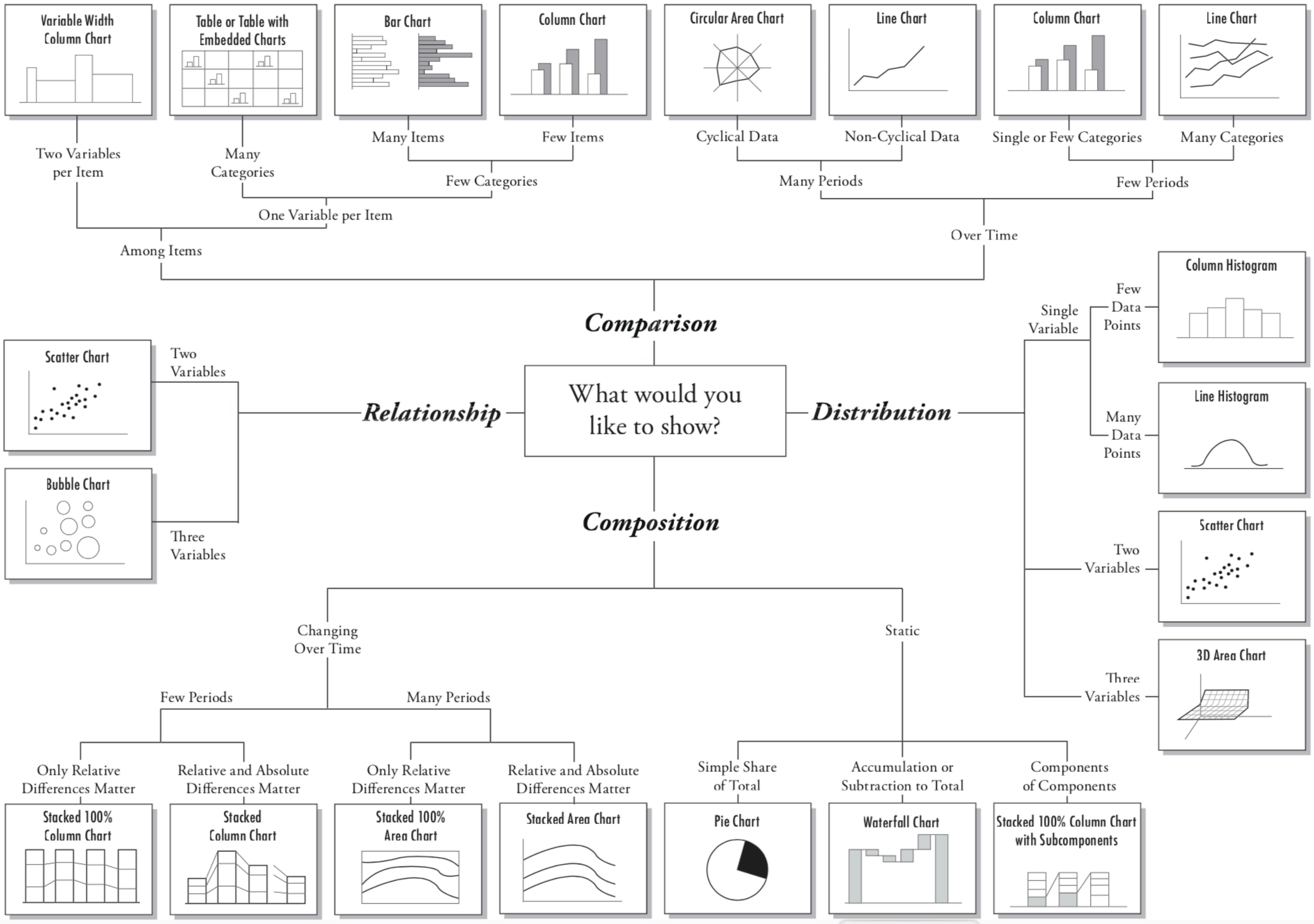
-1.00



0.30

-0.05





What would you like to show?

Comparison

Among Items

Two Variables per Item — Variable Width Column Chart

One Variable per Item

Many Categories — Table or Table with Embedded Charts

Few Categories

Many Items — Bar Chart

Few Items — Column Chart

Over Time

Many Periods

Cyclical Data — Circular Area Chart

Non-Cyclical Data — Line Chart

Few Periods

Single or Few Categories — Column Chart

Many Categories — Line Chart

Relationship

Two Variables — Scatter Chart

Three Variables — Bubble Chart

Distribution

Single Variable

Few Data Points — Column Histogram

Many Data Points — Line Histogram

Two Variables — Scatter Chart

Three Variables — 3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter — Stacked 100% Column Chart

Relative and Absolute Differences Matter — Stacked Column Chart

Many Periods

Only Relative Differences Matter — Stacked 100% Area Chart

Relative and Absolute Differences Matter — Stacked Area Chart

Static

Simple Share of Total — Pie Chart

Accumulation or Subtraction to Total — Waterfall Chart

Components of Components — Stacked 100% Column Chart with Subcomponents

Chart Types

- Tables: precision, scale, multivariate, heterogenous, and compact.
  - Order rows for comparisons, order columns for relatedness
- Dot and line Charts: y = f(x)
  - Show data points, not just fits
  - Show the full variable range if possible
  - Admit uncertainty when plotting averages: standard deviation σ around y as a whisker, showing the interval [y − σ, y + σ].
  - Never connect points for categorical data: Bar Charts are better
  - Use color and hatching to distinguish lines/classes

Chart Types – Cont'd

- Scatter Plots: (x,y) points
  - Scatter the right-sized dots
  - Color or jiggle integer points before scatter-plotting them, or reduce opacity
  - Project multivariate data down to two dimensions, or use arrays of pair- wise plots: USE PCA or SOM, will revisit later
  - Three-dimensional-scatter plots help only when there is real structure to show
  - Bubble plots vary color and size to represent additional dimensions

four dimensions (GDP, life expectancy, population, and geographic region) using x, y, size, and color, respectively

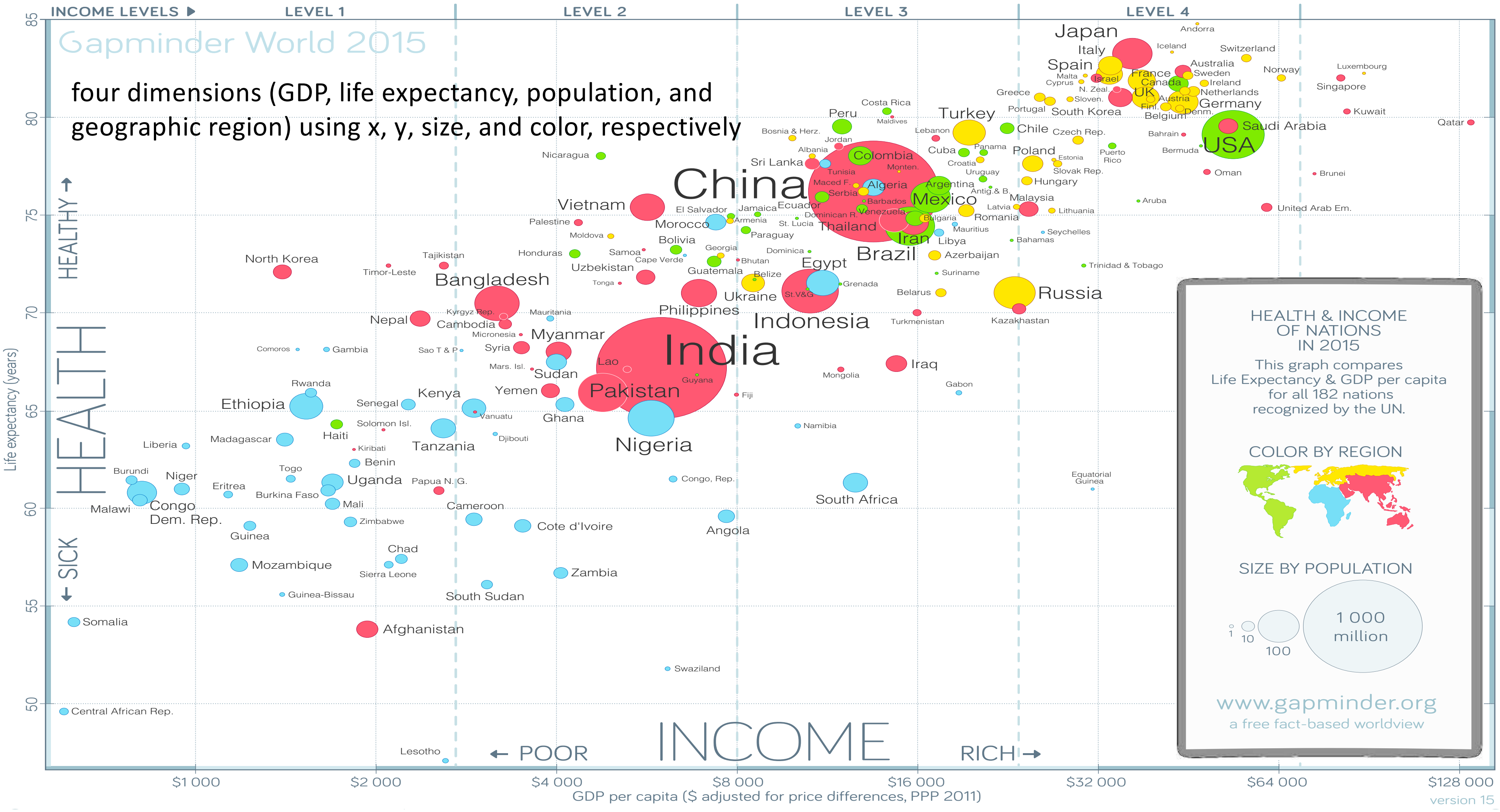
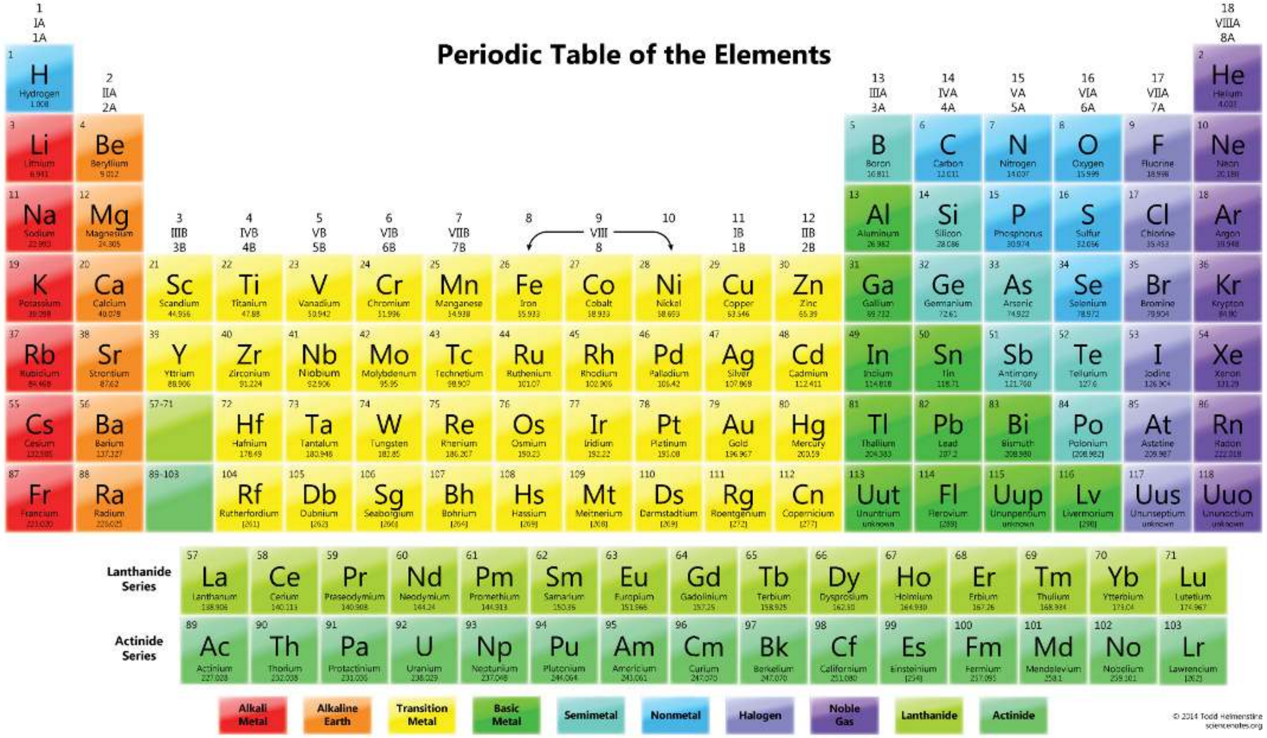


Chart Types – Cont’d

- Bar Plots and Pie Charts: categorical variables
  - Directly label slices of the pie
  - Use bar charts to enable precise comparisons
  - Scale appropriately, depending upon whether you seek to highlight absolute magnitude or proportion
- Histograms: frequency distribution
  - Where is the peak of the distribution, and is the mode near the mean?
  - Is the distribution symmetric or skewed?
  - Where are the tails?
  - Might it be bimodal, suggesting that the distribution is drawn from a mix of two or more underlying populations?

Chart Types – Cont'd

- Data Maps:
  - The map has a story to tell
  - Regions are contiguous, and adjacency means something
  - The squares are big enough to see
  - It is not too faithful to reality:



Libraries Used in the lab

- Scipy stats contains a large number of probability distributions as well as a growing library of statistical functions.
- matplotlib is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
- PyLab belongs to Matplotlib that combines the numerical module numpy with the graphical plotting module pyplot.
- Pygooglechart is a complete Python wrapper for the Google Chart API
- **Xport (did not work, used pyreadstat instead)** contain utility functions for reading the whole binary XPT file and loading the rows into a Python data structure. The to_rows function will simply return a list of rows. The to_columns function will return the data as columns rather than rows.
- xlrd is a library for reading data and formatting information from Excel files, whether they are .xls or .xlsx files
- Re is a python module that provide regular expressions operations.
- There is also a class of third-party systems for building dashboards, like Tableau to build iinteractive Dashboards for the leess technical users of your project.