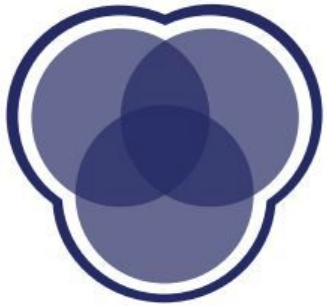


CDS

Cornell Data Science

Data Visualization

Workshop Advertisement

Deep Dive: Parallel Processing

Sept. 22, 2017 Gates Hall G01 5:30-7:30PM

SEP
22

Deep Dive: Parallel Processing

Public · Hosted by Cornell Data Science
Edit

Friday at 5:30 PM - 7:30 PM
4 days from now · 55–81°Scattered Clouds — Edit

G01 Gates Hall — Edit

No tickets — Add Tickets

Big Data.
Big Servers.

Deep Dives into

Parallel Computing

With Spark and Hadoop

9/22 5:30pm - 7:30pm
9/24 6:00pm - 8:00pm
9/29 5:30pm - 7:30pm

Gates G01

Requires: 8gb RAM, 50GB disk space

cornelldata.science

CDS
Cornell Data Science

Sanity Check

- **Did you submit the Quiz?**
- **Are you getting email notification for Piazza Announcement?**
- **Are you in a group of 3-4 people for the project?**
  - **If not, come up to the front after the lecture!**

Jupyter Notebook Demo

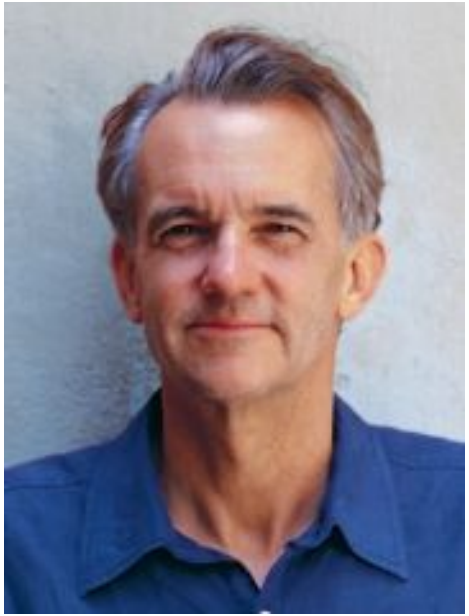
History

Edward Tufte (1942-)

Statistician and Yale professor

Key figure in the field of data visualization

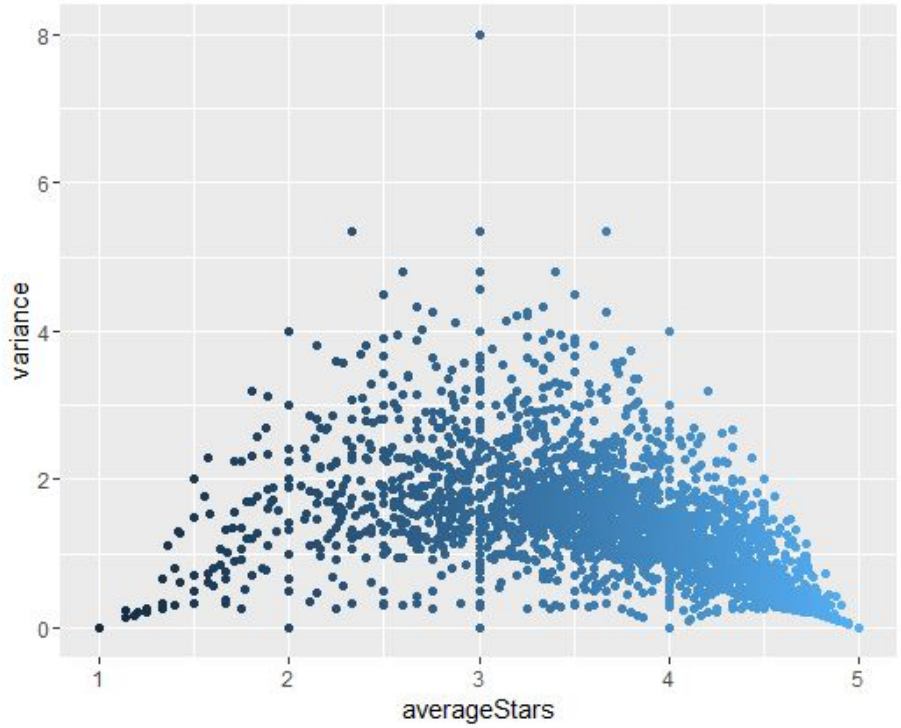
Recommended text: *The Visual Display of Quantitative Information*



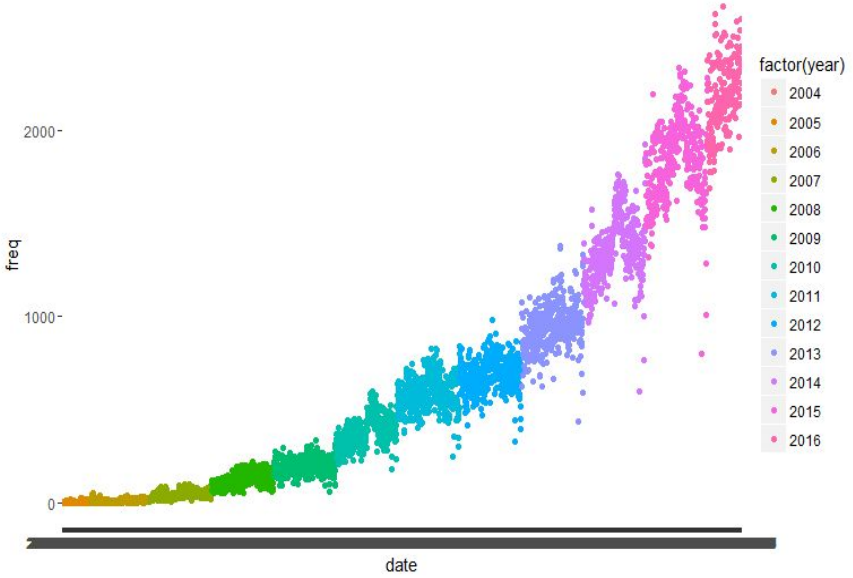
Data Visualization Simple Example: Yelp

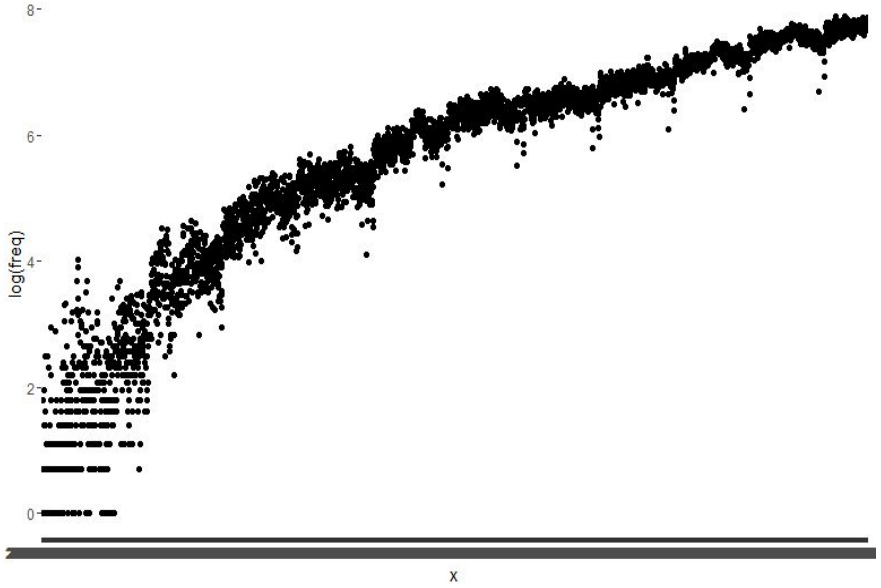
	AVG(stars)	var
AVG(stars)	1.00	-0.43
var	-0.43	1.00

Question: What do you notice? What trends do you see?

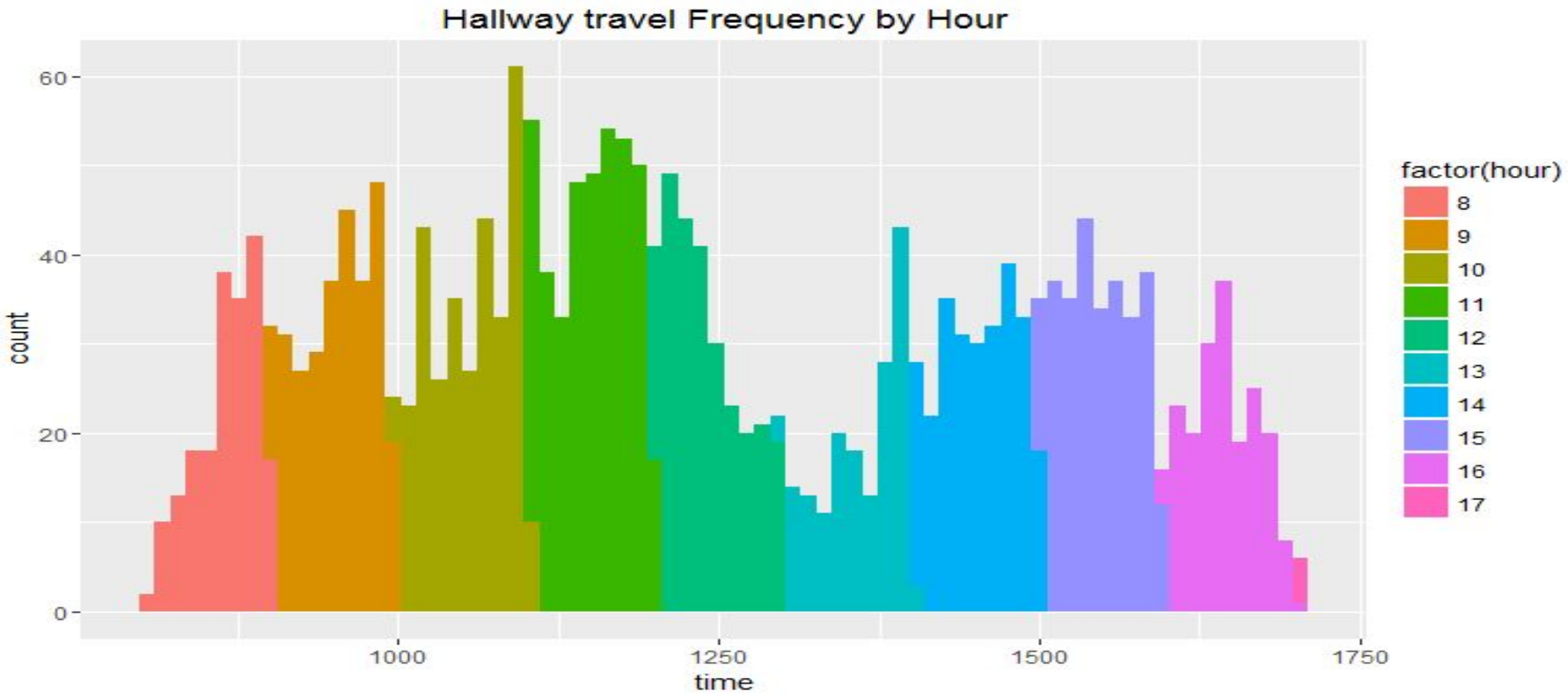


Data Visualization Simple Example: Yelp



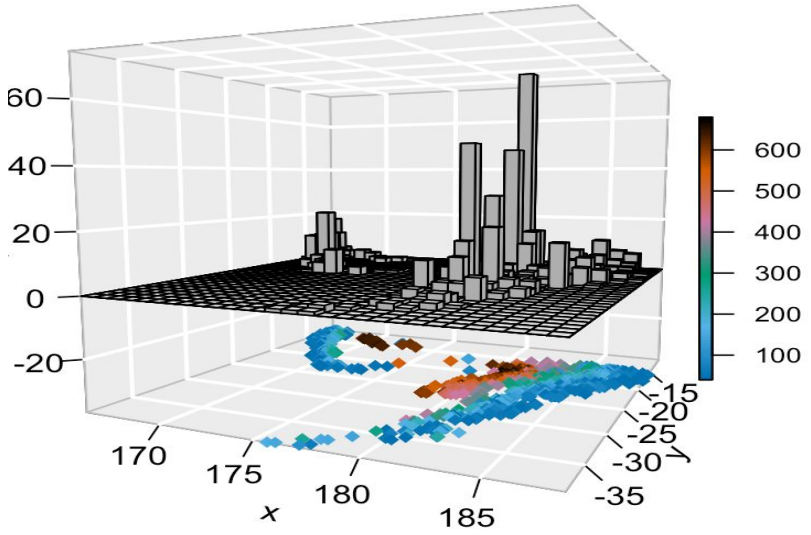


Example: Nurse Hallway Travel Frequency



Why Data Visualization?

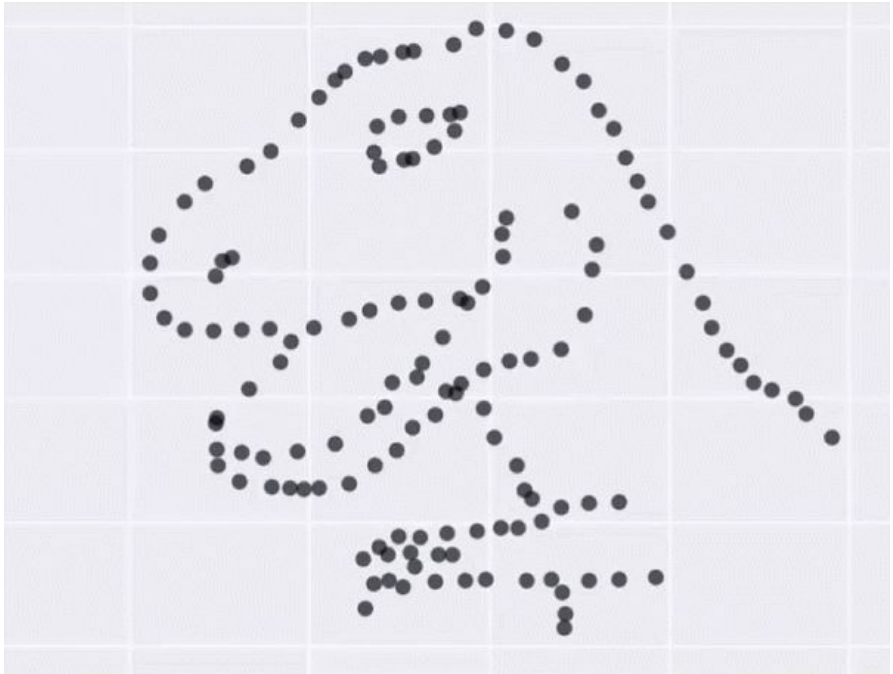
- Understanding a dataset
- Communication of knowledge to an audience



4D Plot For Earthquake Data

Why Data Visualization is Important

- **All Different Datasets**
  They all have same mean, median, mode, variance, line of best fit
- **Same Summary Stat**
  But we need to see how the **actual** data looks



Source

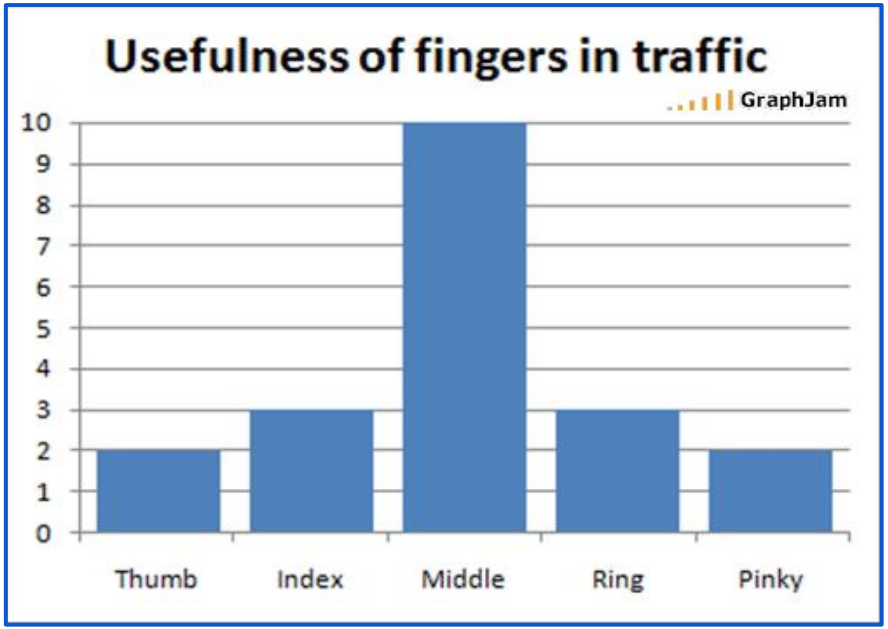
What is matplotlib?

- **Python data visualization package**
  - Capable of handling most data visualization needs
  - Simple object-oriented library inspired from MATLAB
  - Cheatsheet

matplotlib

Let's start with an easy one... a bar graph!

- Represent **magnitude** or **frequency**
- Allows us to compare features



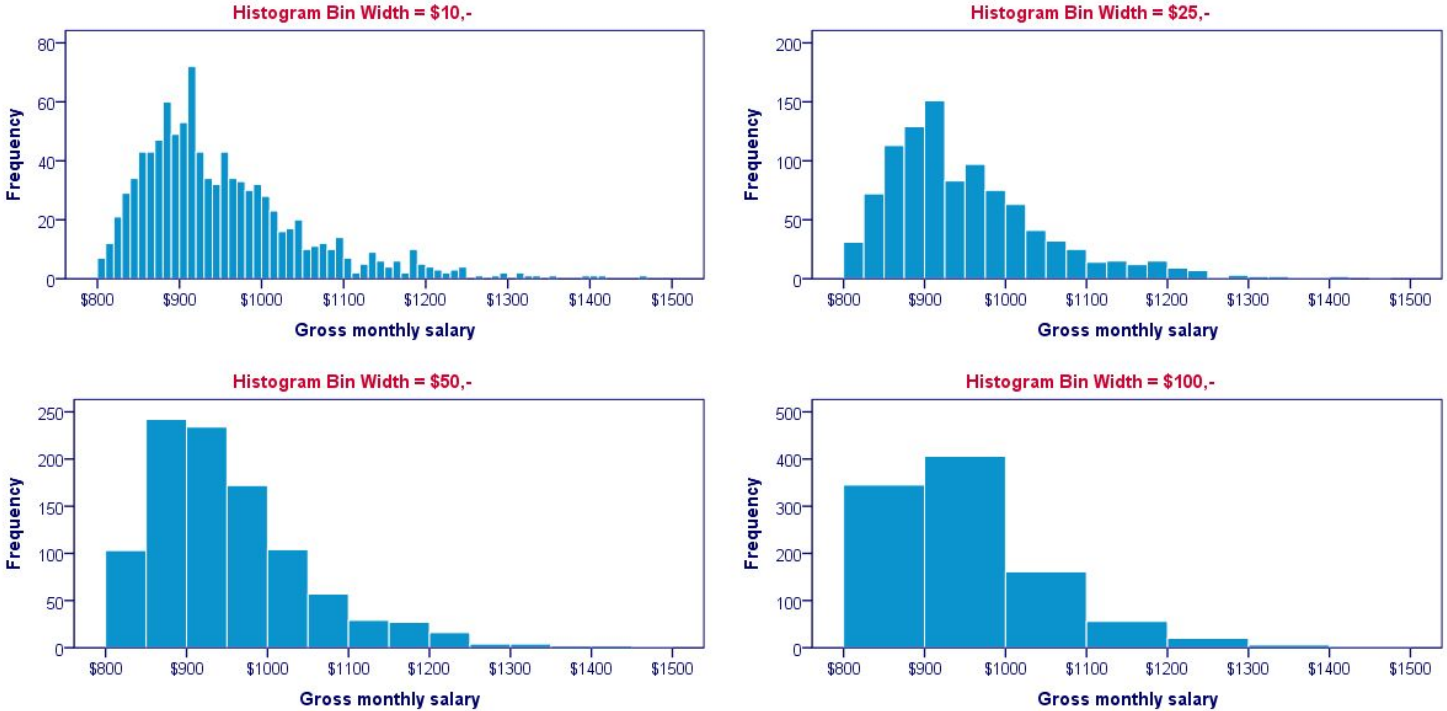
Source

Histograms

Source

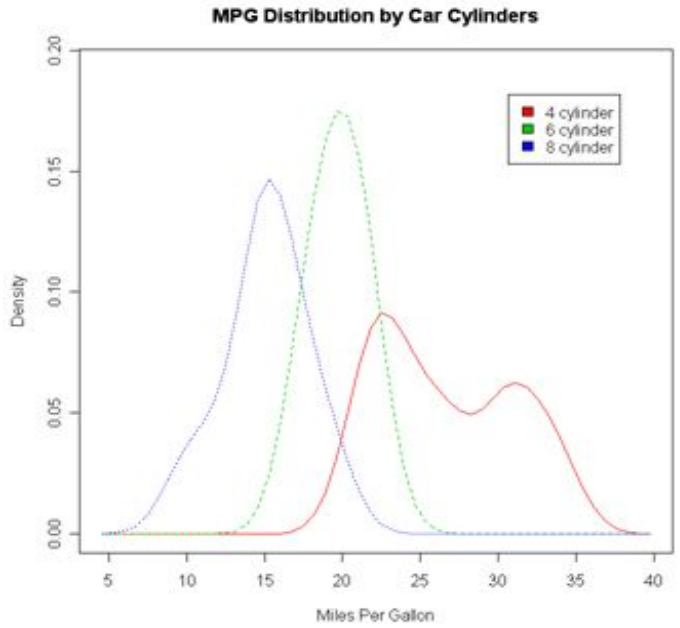
- Used to observe **frequency distribution** of numerical data
- Data split into **bins**

Histograms



Source

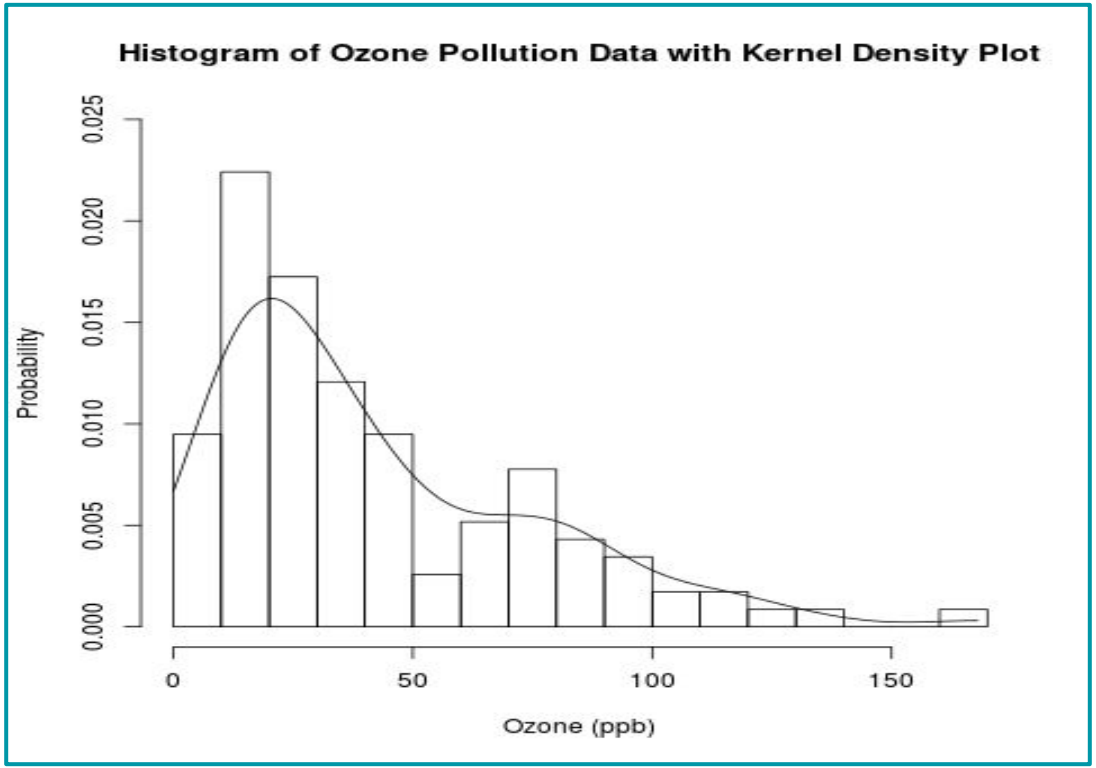
Density Plot



- Like a histogram, but **smooths** the shape of the distribution
- Why is Density Plot important?

Source

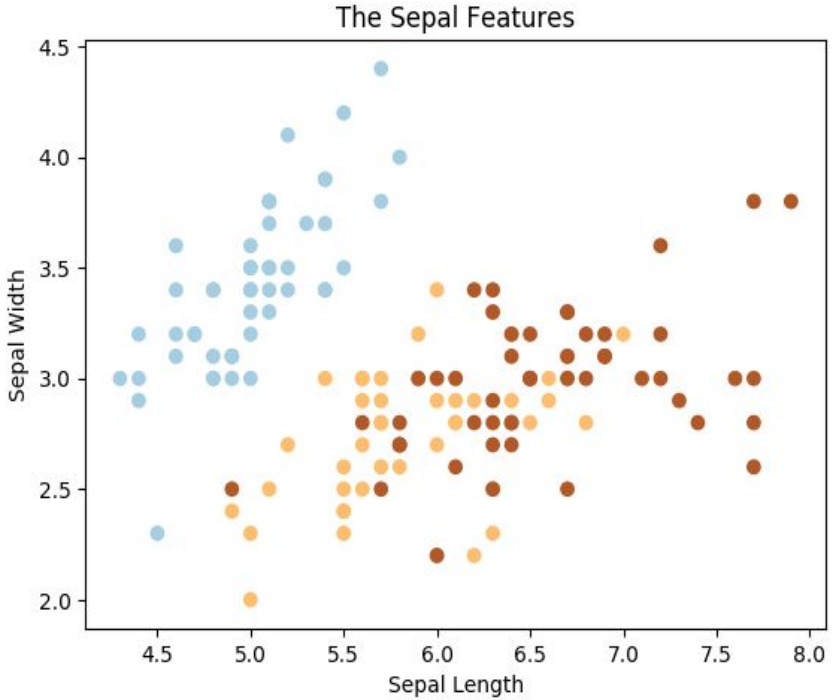
Histogram vs. Density Plot



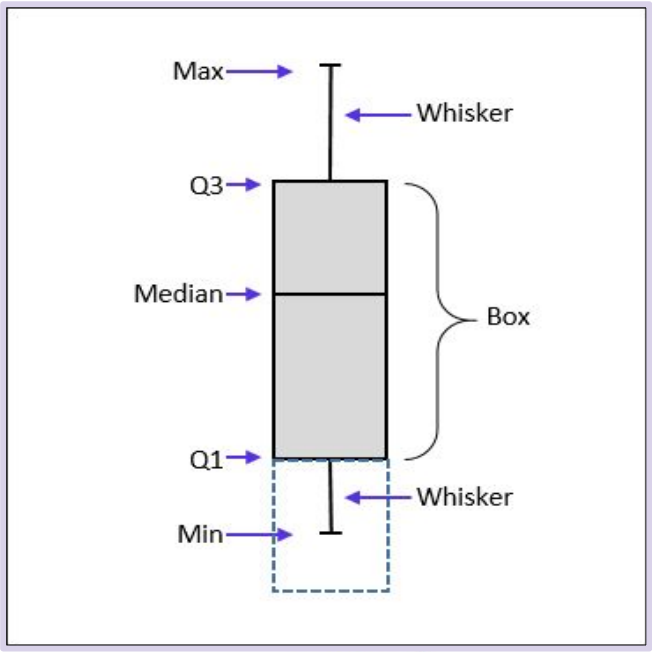
Source

Scatterplot

- See **relationship** between two features
- Can be useful for **extrapolating** information



Boxplot (a.k.a Box-and-whisker plot)

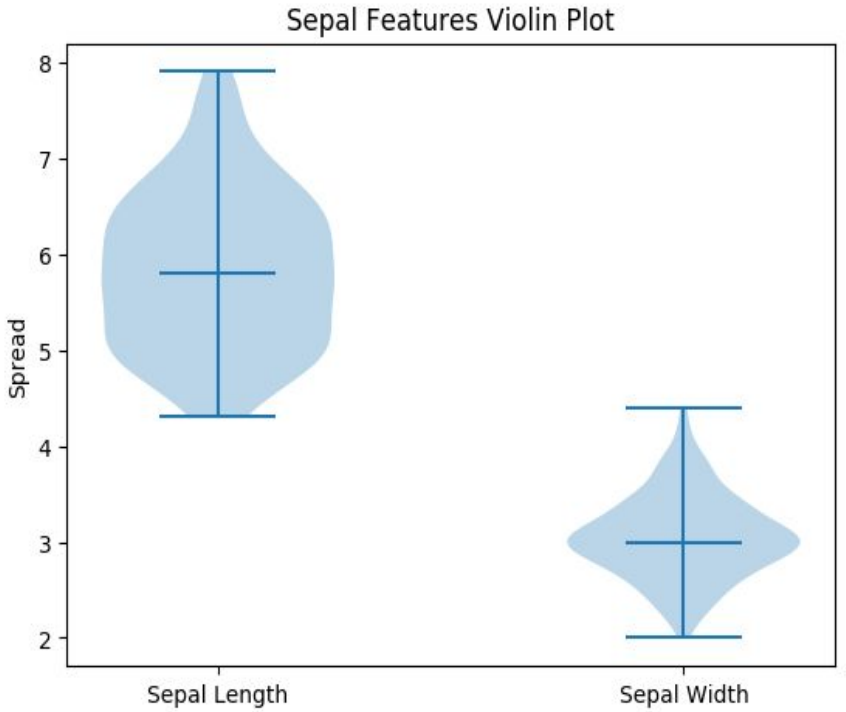


- Summary of data
- Shows **spread** of data
- Gives range, interquartile range, median, and outlier information

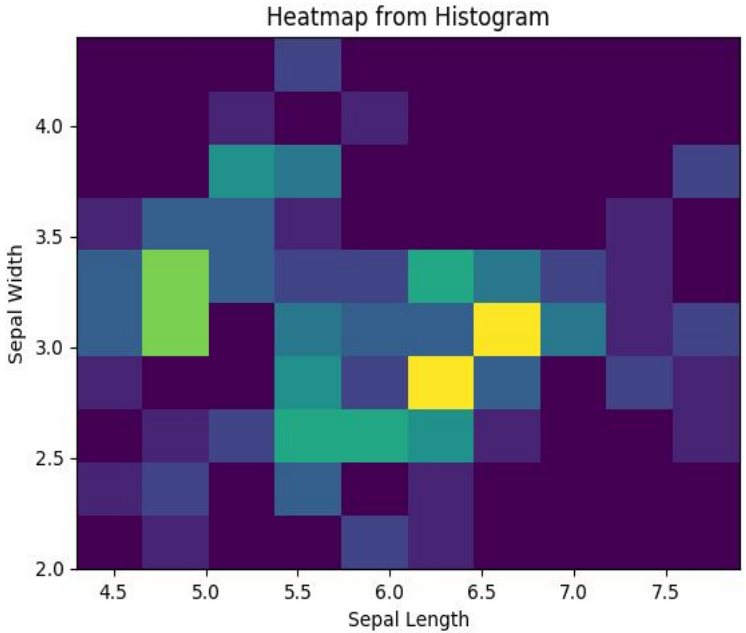
Source

Violin Plot

- Combination of **boxplot** and **density plot** to show the **spread** and **shape** of the data
- Can show whether the data is **normal**



Heatmaps

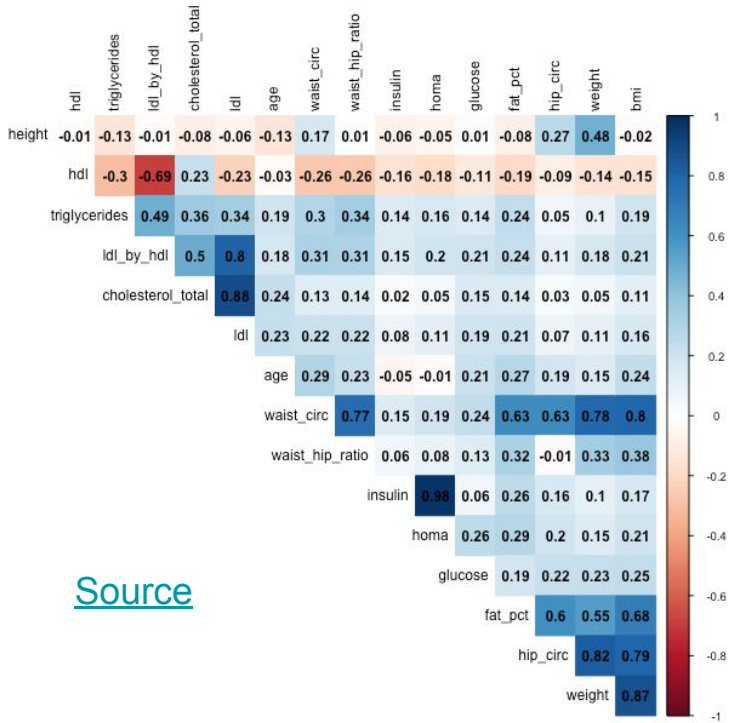


- Varying degrees of one metric are represented using **color**[1]
- Especially useful in the context of **maps** to show geographical variation

1 Defined by https://www.marketingterms.com/dictionary/heatmap/

Correlation Plot

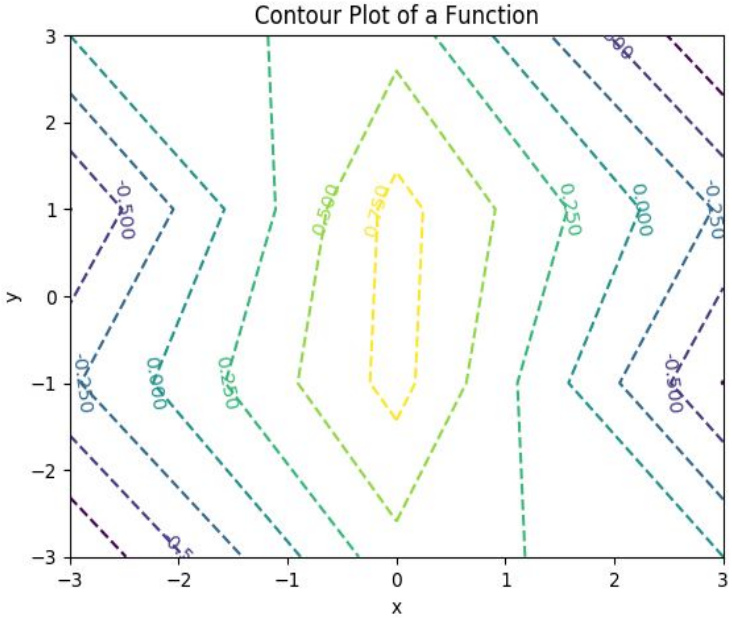
- 2D matrix with all variables on each axis
- Entries represent the **correlation coefficients** between each pair of variables



Source

```
[[ 1.          -0.10936925  0.87175416  0.81795363]
 [-0.10936925  1.          -0.4205161  -0.35654409]
 [ 0.87175416 -0.4205161   1.           0.9627571 ]
 [ 0.81795363 -0.35654409  0.9627571   1.        ]]
```

Contours

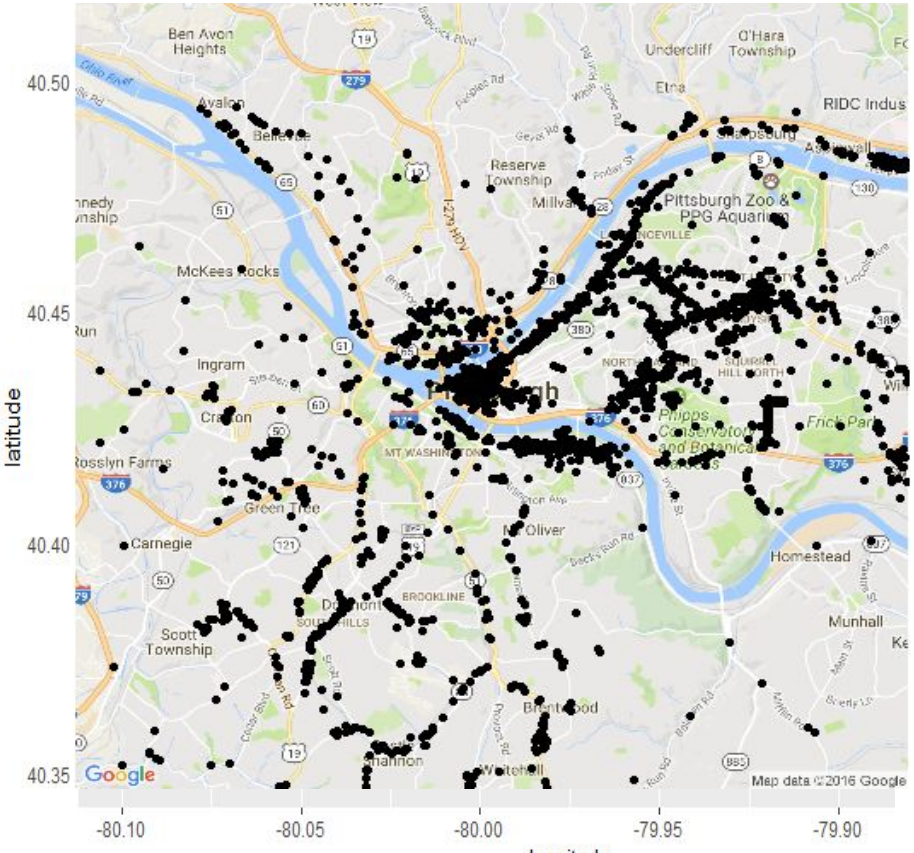


- Used to show **distribution** of the data or a function
- Observe variation among portions of data
- In maps, they indicate the shape of the land

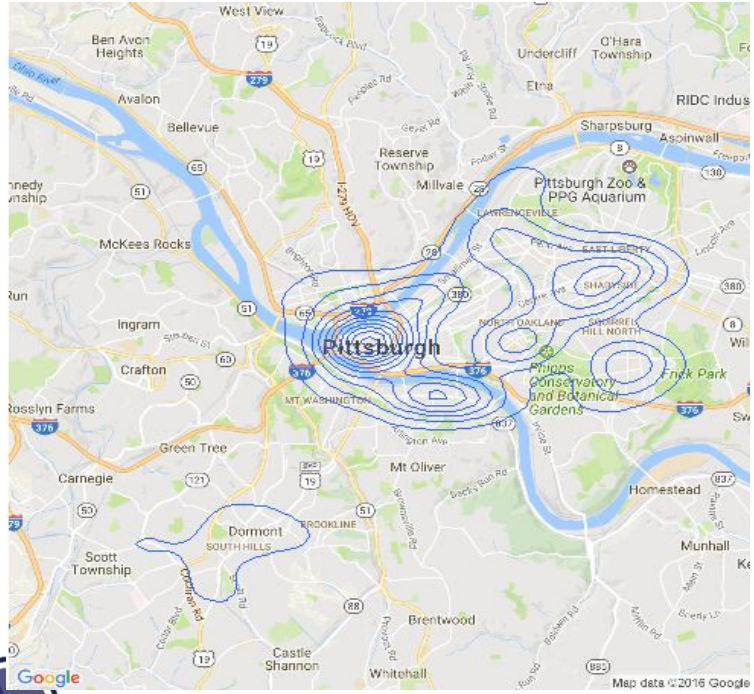
Using Maps

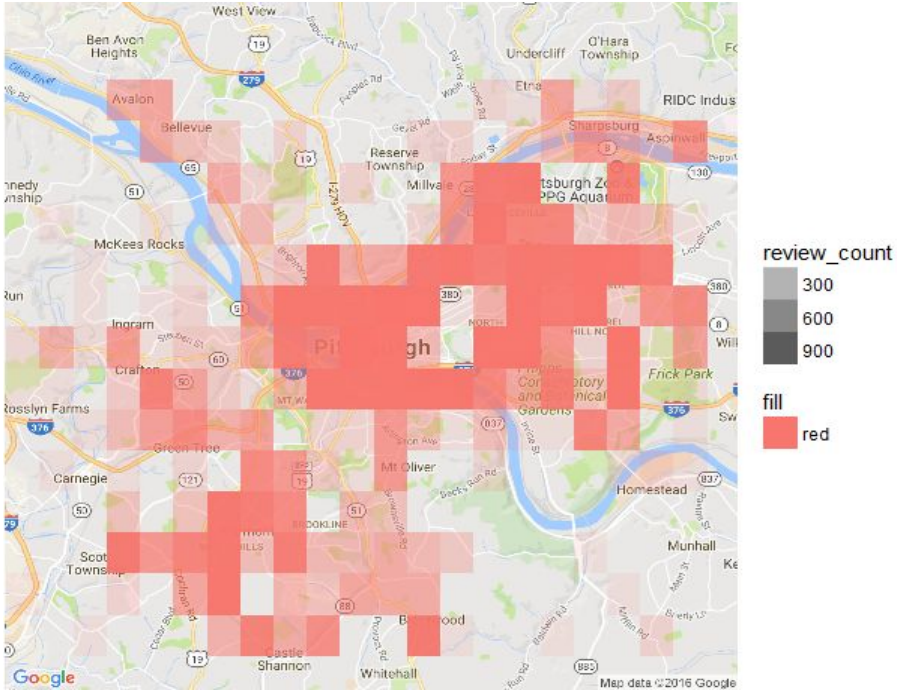
- **Map visualization -> contextual information**
  - **Trends are not always apparent in the data itself**
  - Ex) Longitudes and Latitudes in your data
    - *Geographical Visualization*

Example: Pittsburgh Data

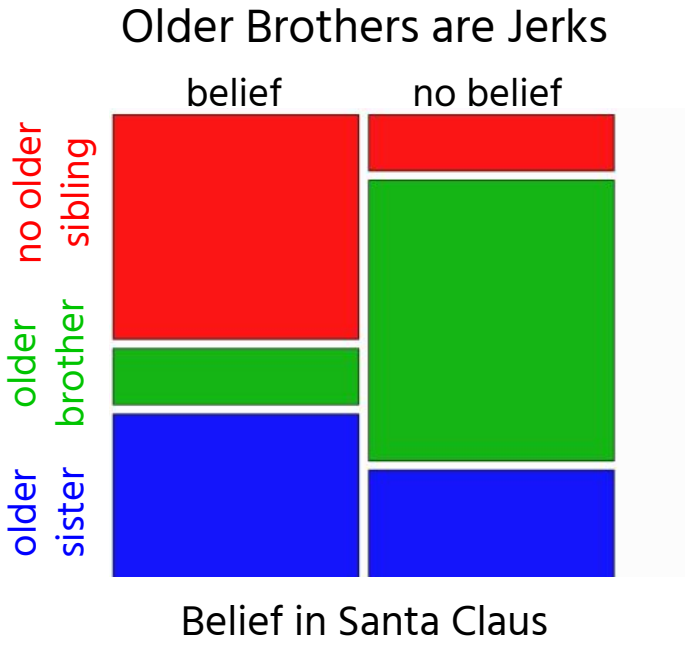


Applications for Contour Map and Heatmap





Mosaic Plot



- Represents **two-way frequency**
- Horizontal dimension represents the frequency of one variable while the vertical dimension represents the other

Source

Challenges of Visualization

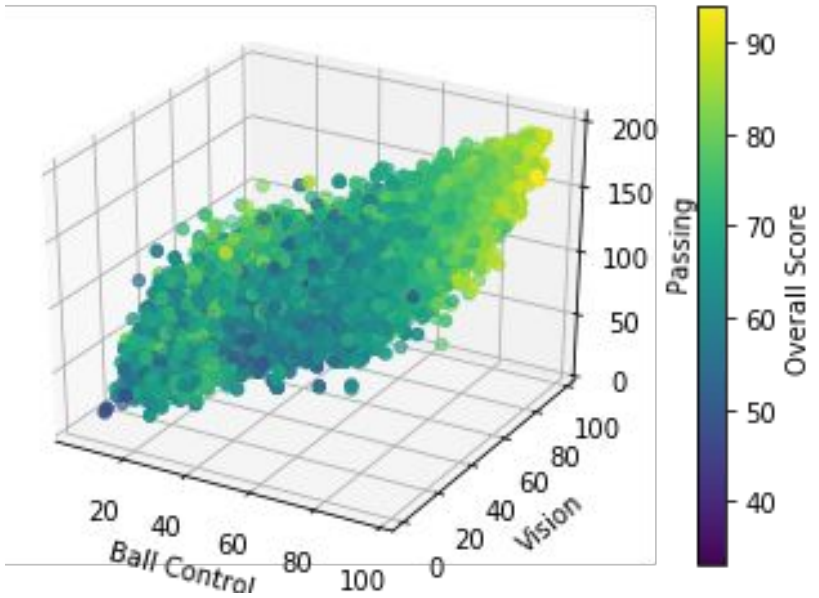
Higher Dimension

Non-Trivial

Time Consuming

Hard to show Uncertainty

Higher Dimensional Data



- **Color**, **time animations**, or **point shape** can be used for higher dimensions
- There is a limit to the number of features that can be displayed

Coming Up

Your problem set: Unleash your creativity by visualizing a data set

Next week: Introduction to Supervised Learning

See you then!