Vizualization in R

Mikhail Stepanov

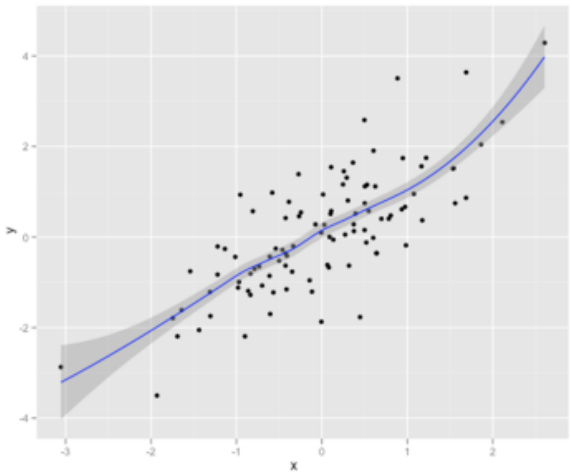
October 4, 2012

Basic stats and Necessity of Visualization

Summary statistics give us some sense of the data:

- Mean vs. Median.
- Standard deviation.
- Quartiles, Min/Max.
- Correlations between variables.

```
summary(data)
       x                    y
 Min.   :-3.05439   Min.   :-3.50179
 1st Qu.:-0.61055   1st Qu.:-0.75968
 Median : 0.04666   Median : 0.07340
 Mean   :-0.01105   Mean   : 0.09383
 3rd Qu.: 0.56067   3rd Qu.: 0.88114
 Max.   : 2.60614   Max.   : 4.28693
```



Visualization gives us a more holistic sense

Median, Quartile, Percentile

- In statistics, a percentile (or centile) is the value of a variable below which a certain percent of observations fall
  - e.g., the 20th percentile is the value (or score) below which 20 percent of the observations may be found
- 25th percentile is also known as the first quartile (Q1)
- 50th percentile as the median or second quartile (Q2)
- 75th percentile as the third quartile (Q3)

Why Visualize?

4 data sets, characterized by the following. Are they the same, or are they different?

Property	Values
Mean of x in each case	9
Exact variance of x in each case	10
Exact mean of y in each case	7.5 (to 2 d.p)
Variance of Y in each case	3.75 (to 2 d.p)
Correlations between x and y in each case	0.816
Linear regression line in each case	Y = 3.00 + 0.500x (to 2 d.p and 3 d.p resp.)

i

x	y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

ii

x	y
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

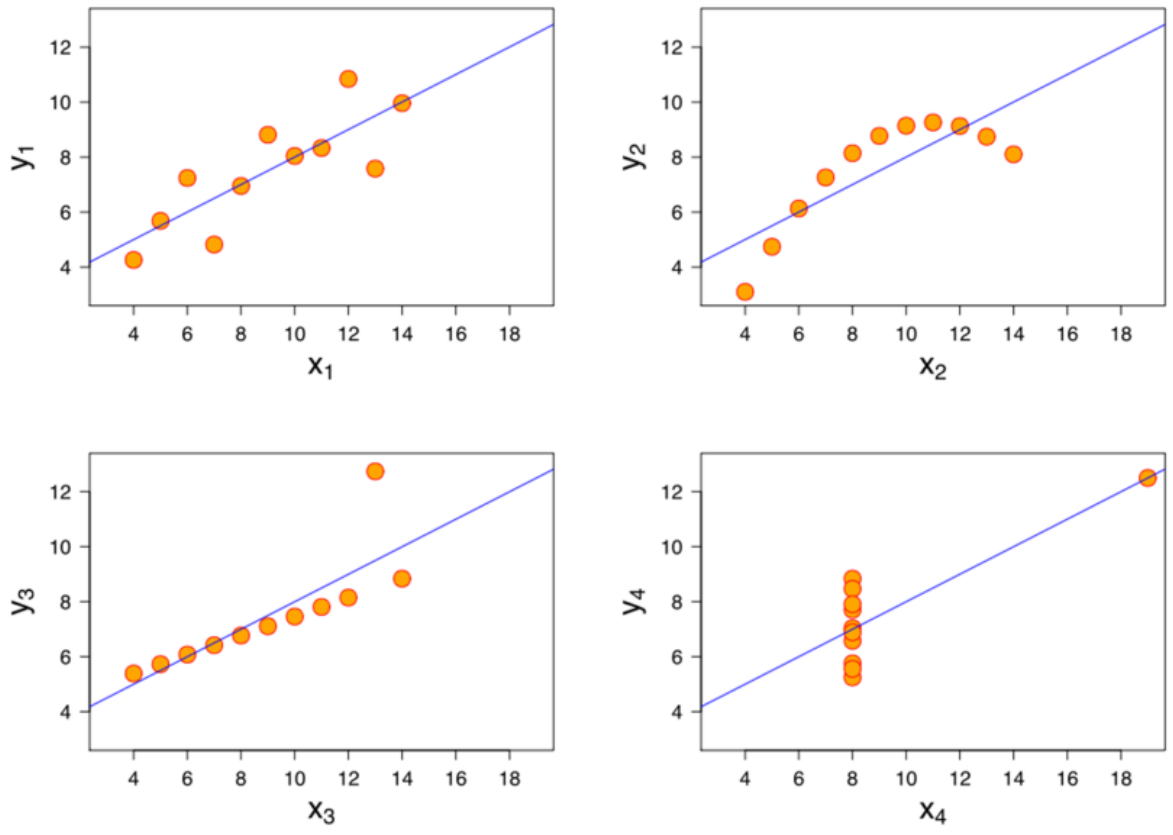
iii

x	y
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

iv

x	y
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

Why Visualize?



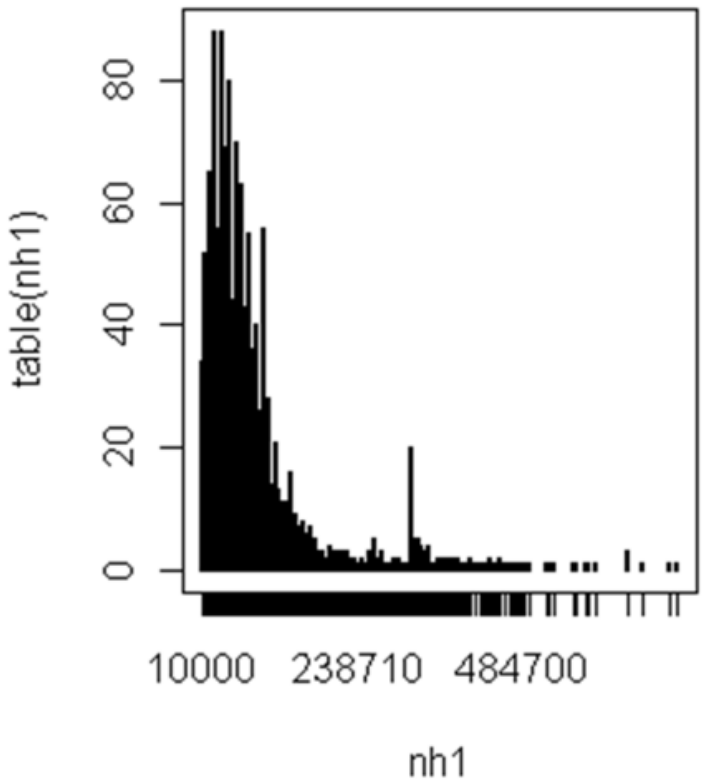

```
>plot(s1)
>plot(lm(s1$y ~ s1$x))
```

Examining the Distribution of a Single Variable

Graphing a single variable

- plot(sort(.)) – for low volume data
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram“

- Example
  - Frequency table of household income

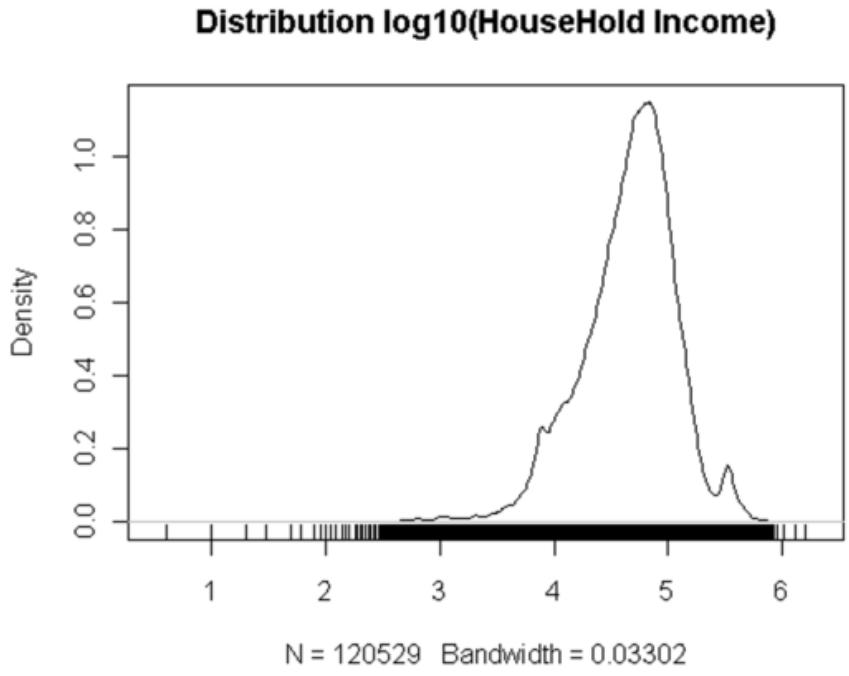


Examining the Distribution of a Single Variable

Graphing a single variable

- plot(sort(.)) – for low volume da
- hist(.) – a histogram
- plot(density(.)) – densityplot
  - A "continuous histogram“

- Example
  - Frequency table of household income



One More Way to Examine Distribution

A sense of the data range

- If it's very wide, or very skewed, try computing the log

Outliers, anomalies

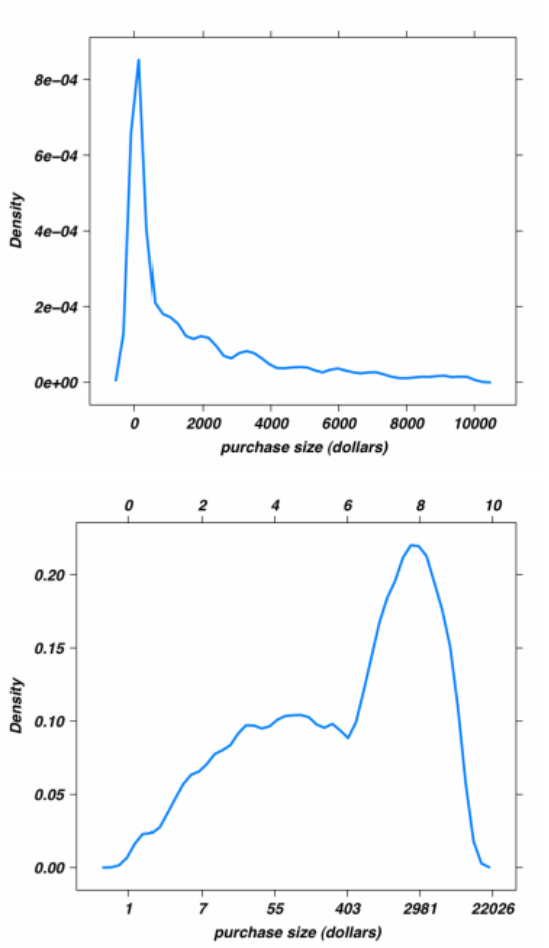
- Possibly evidence of dirty data

Shape of the Distribution

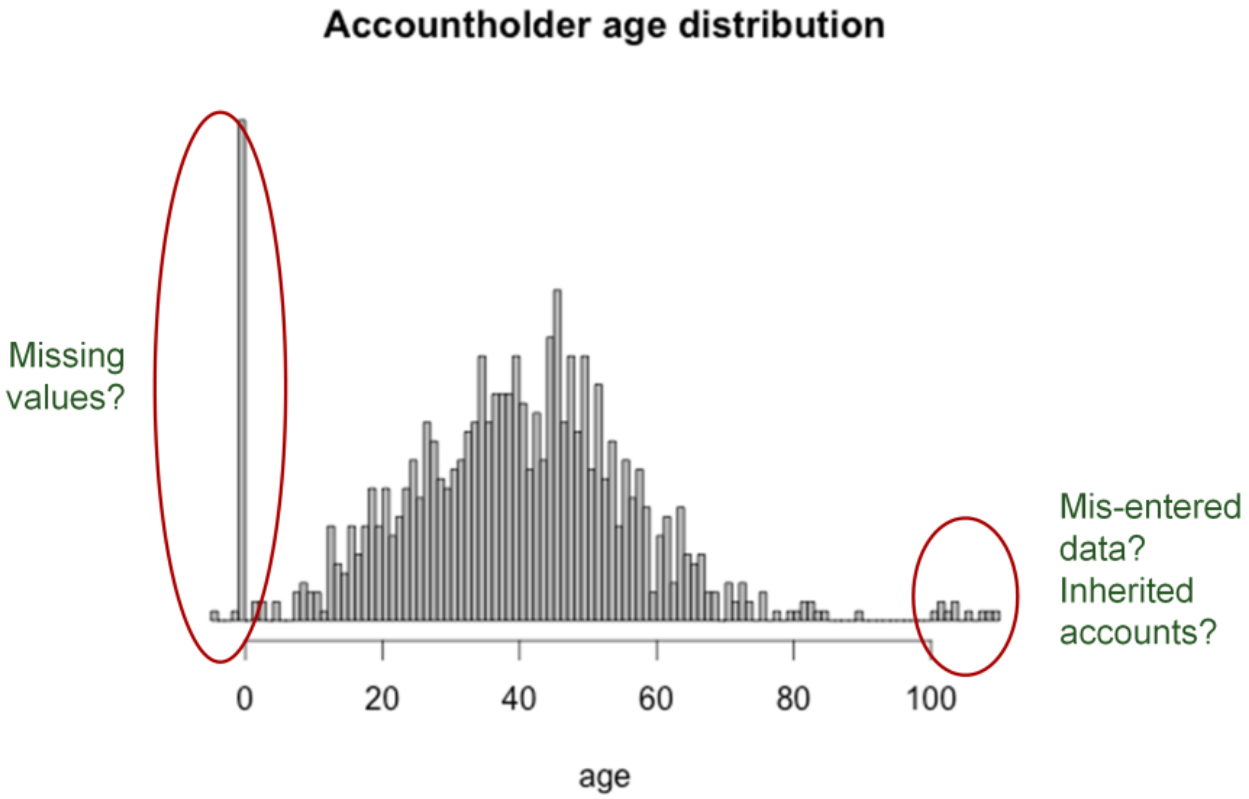
- Unimodal? Bimodal?
- Skewed to left or right?
- Approximately normal? Approximately lognormal?

Example - Distribution of purchase size ($)

- Range from 0 to > $10K, left skewed
- Typical of monetary data
- Plotting log of data gives better sense of distribution
- Two purchasing distributions
  - ~ $55
  - ~ $2900

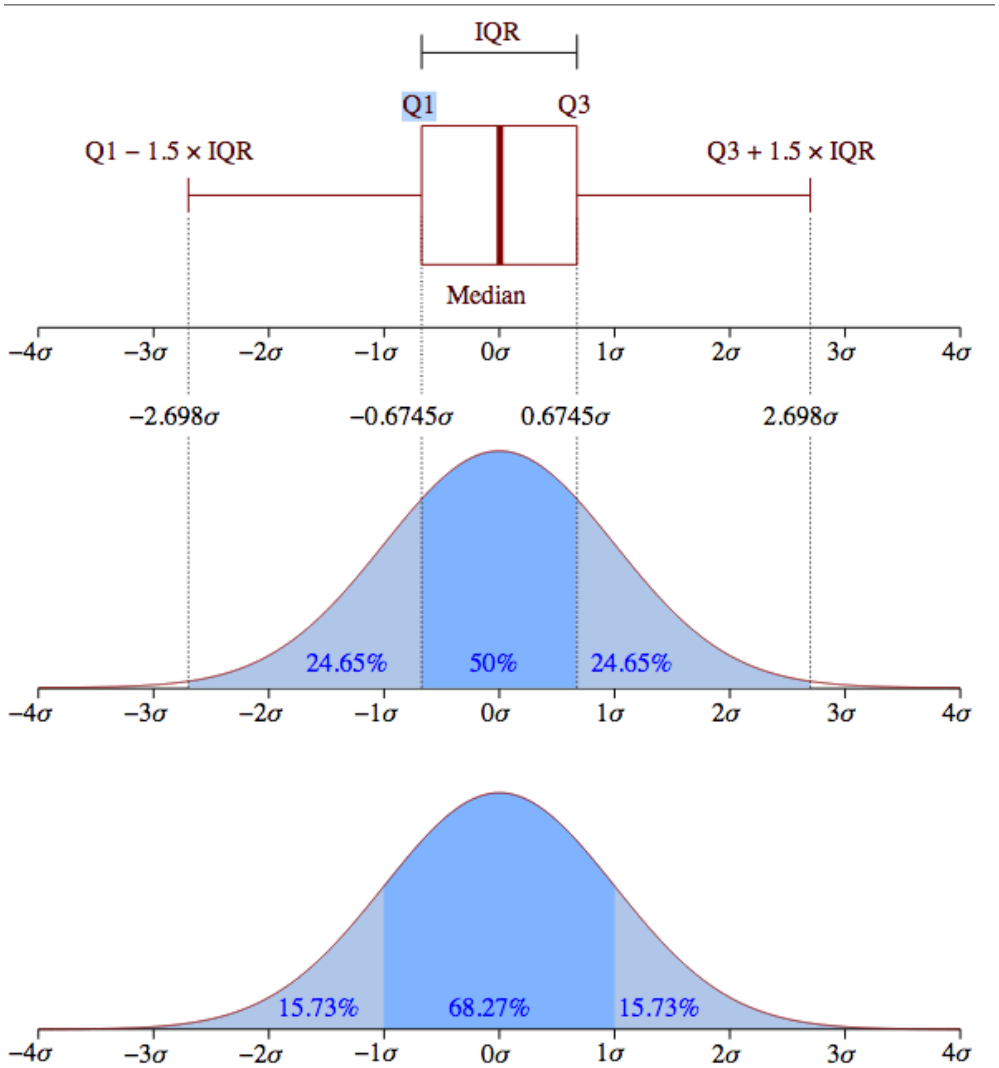


Evidence of Dirty Data




```
hist(age, breaks=100,
main="Accountholder age distribution",
xlab="age", col="gray")
```

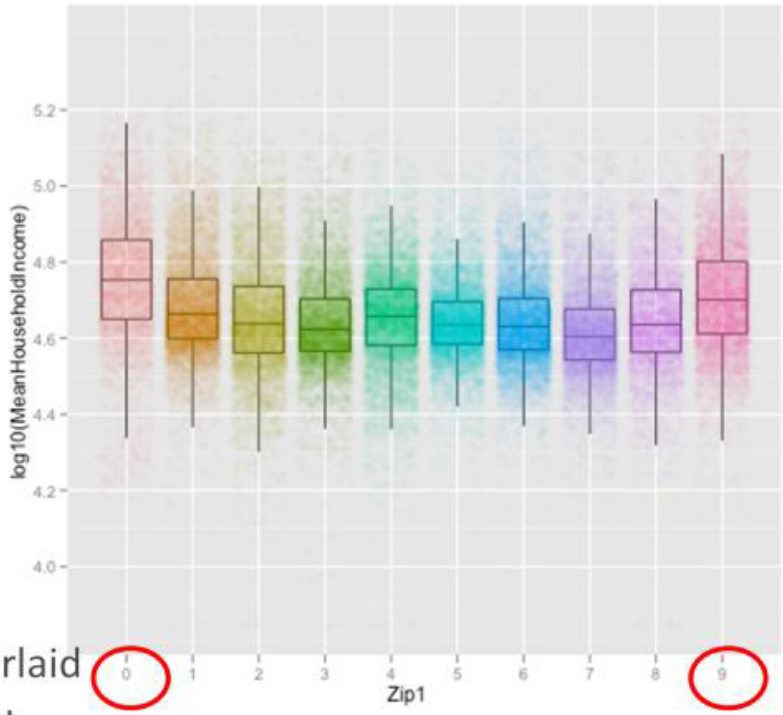
Boxplot and Normal Distribution



Analyzing Relationship Between two Variables

How?

- Two Continuous Variables (or two discrete variables)
  - Scatterplots
  - LOESS (fit smoothed line to the data)
  - Linear models: graph the correlation
  - Binplots, hexbin plots
    - More legible color-based plots for high volume data
- Continuous *vs.* Discrete Variable
  - Jitter, Box and whisker plots, Dotplot or barchart

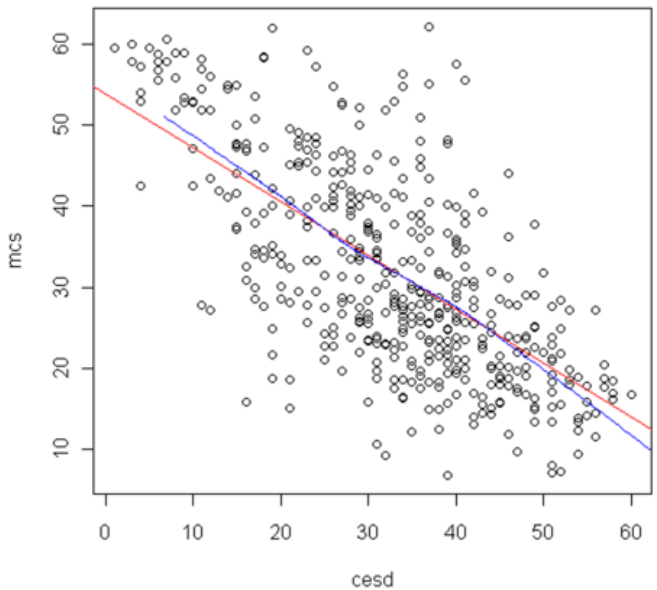


Example:

- Household income by region (ZIP1)
- Scatterplot with jitter, with box-and-whisker overlaid
- New England (0) and West Coast (9) have highest mean household income

Scatterplots and Correlation in Data

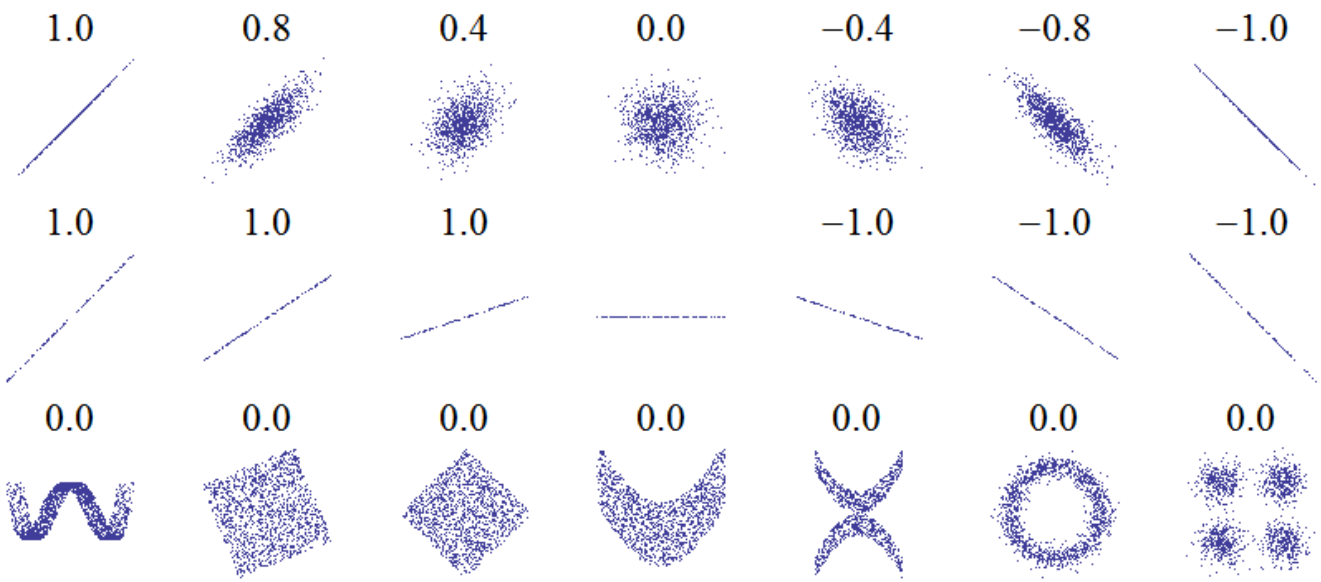
- Is there a relationship between the two variables?
  - Linear? Quadratic?
  - Exponential?
    - Try semi-log or log-log plots
  - Is it a cloud?
    - Round? Concentrated? Multiple Clusters?
- How?
  - Scatterplots
- Example
  - Red line: linear fit
  - Blue line: LOESS
  - Fairly linear relationship, but with wide variance



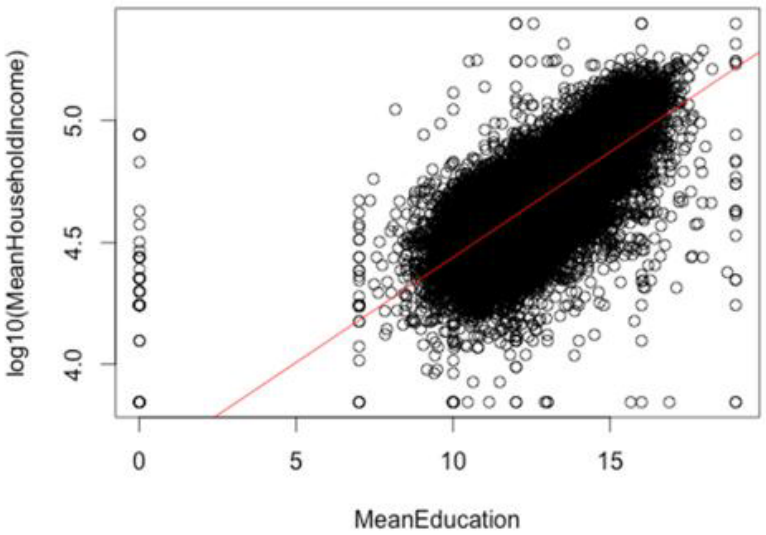

```
with(ds, {
plot(mcs ~ cesd) abline(lm(mcs ~ cesd), lcol="red")
lines(lowess(mcs ~ cesd), lcol="blue") })
```

Correlation and Pearson Correlation Coefficient

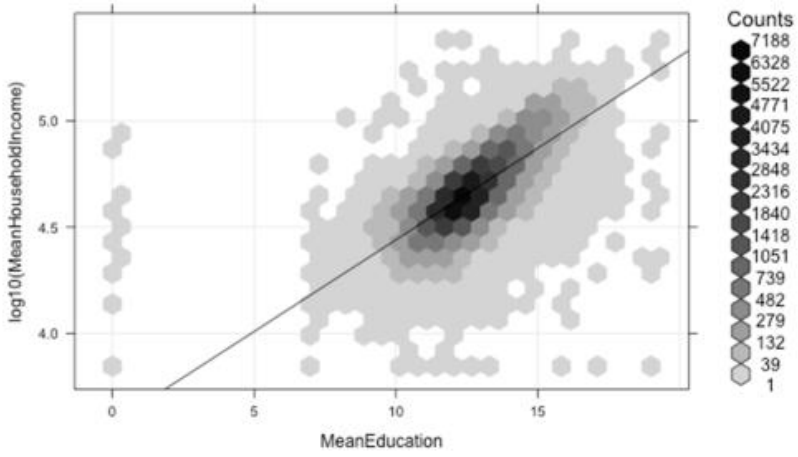
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



And one More About Scatterplots



Scatterplot:
Overplotting makes it difficult to see structure



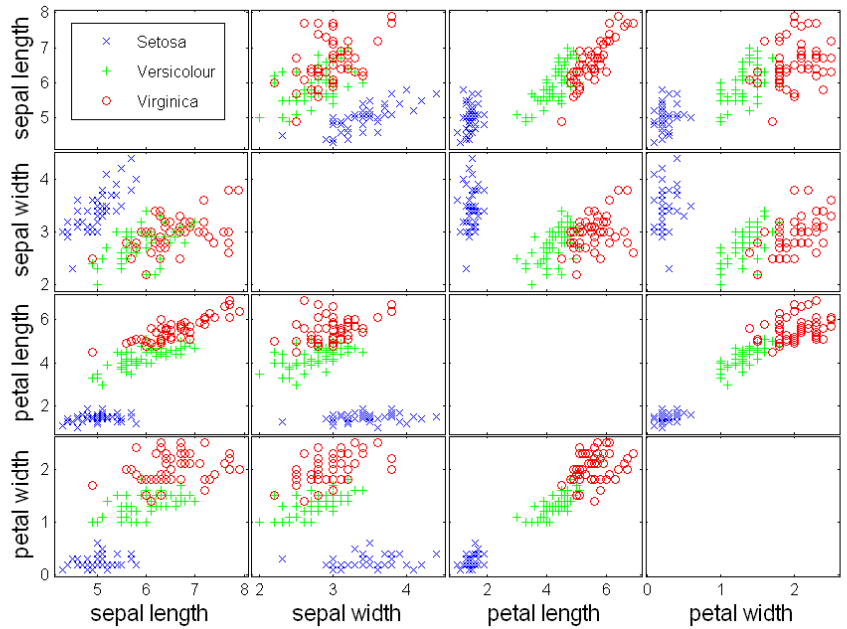
Hexbinplot:
Now we see where the data is concentrated.

```
plot(log10(MeanHouseholdIncome) ~ MeanEducation, data=zcta)
abline(lm(log10(MeanHouseholdIncome) ~ MeanEducation, data=zcta), col='red')

hexbinplot(log10(MeanHouseholdIncome) ~ MeanEducation,
data=zcta, trans = sqrt, inv = function(x) x^2, type=c("g", "r"))
```

Establishing Multiple Pairwise Relationship Between Variables

- Why?
  - Examine many two-way relationships quickly
- How?
  - pairs(ds) can generate a plot of each pairs of variables
- Example
  - Iris Characteristics
    - Strong linear relationship between petal length and width
    - Petal dimensions discriminate species more strongly than sepal dimensions



Timeseries

What?

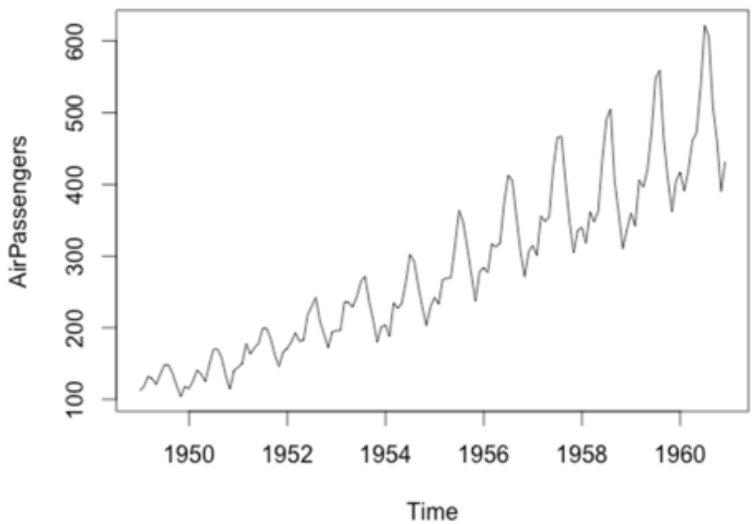
- Looking for ...
  - Data range
  - Trends
  - Seasonality

How?

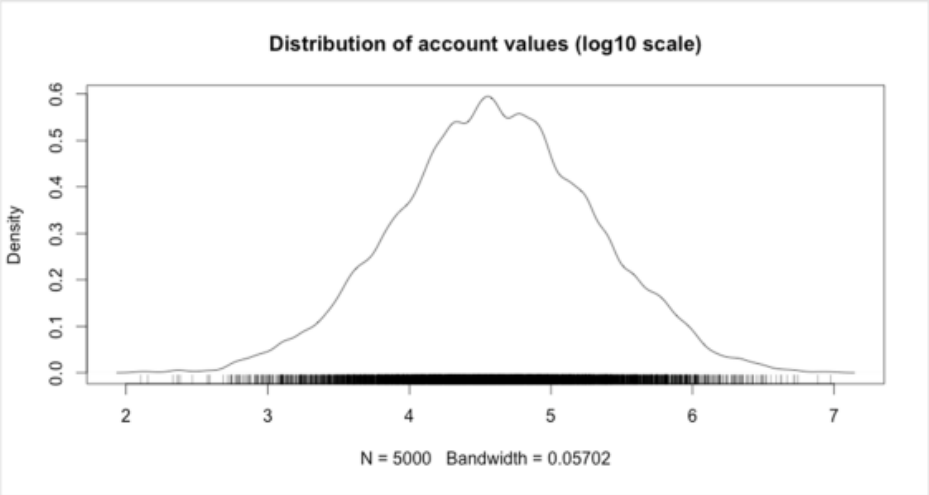
- Use time series plot

Example

- International air travel (1949-1960)
- Upward trend: growth appears superlinear
- Seasonality
  - Peak air travel around Nov. with smaller peaks near Mar. and June

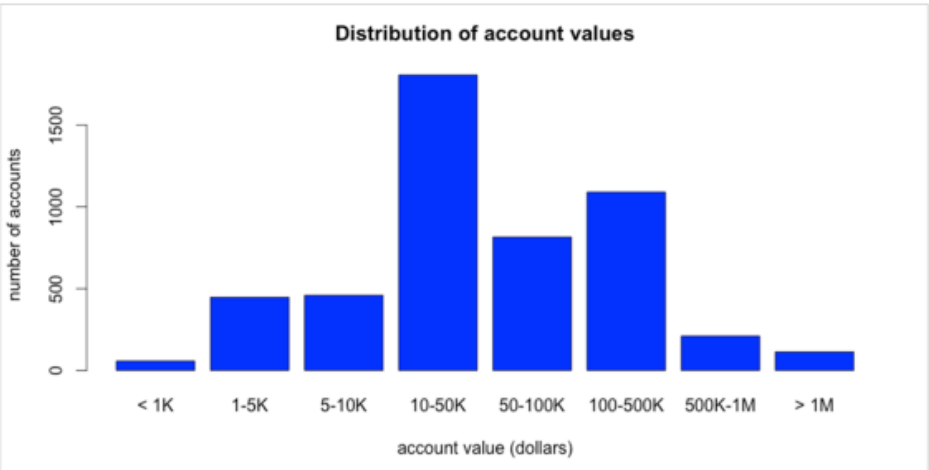


Data Exploration vs Data Presentation



Data Exploration:

This tells you what you need to know.



Presentation:

This tells the stakeholders what they need to know.