Adv. Descriptive Statistics, Visualization, & Simple Regression

Business Analytics

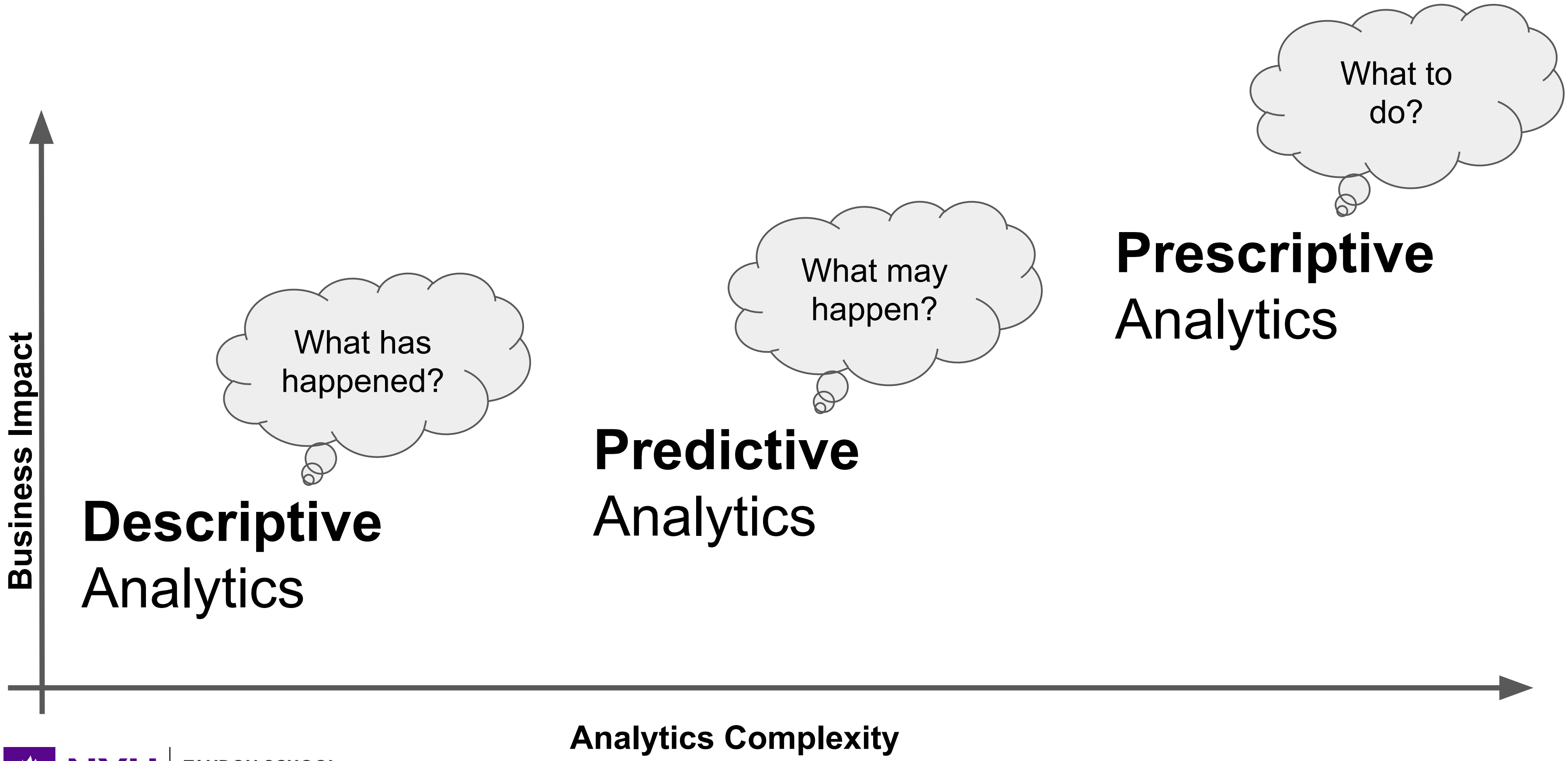
Business Analytics Definition

Business analytics refers to the application of data analysis and modeling techniques for understanding business situations and improving business decisions.

IMPLICATIONS:

- data → past business performance
- methods → statistics + mathematics + computational methods
- business decisions → actionable insight

Types of Analytics



Business Impact

What has happened?

Descriptive Analytics

What may happen?

Predictive Analytics

What to do?

Prescriptive Analytics

Analytics Complexity

Business Problems & Data Science Solutions

- **Classification** attempts to predict, for each individual in a population, which of a (small) set of classes this individual belongs.
- **Regression** (“ value estimation”) attempts to estimate or predict, for each individual, the numerical value of some variable for that individual.
- **Similarity matching** attempts to identify similar individuals based on data known about them.
- **Clustering** attempts to group individuals in a population together by their similarity, but not driven by any specific purpose.
- **Co-occurrence grouping** attempts to find associations between entities based on transactions involving them.
- **Profiling** attempts to characterize the typical behavior of an individual, group, or population.
- **Causal modeling** attempts to identify causal relations between variables of interest, and infer the effects of actions on outcomes.

Review

1. Descriptive Statistics
   a. Measures of central tendency (mean, median, mode)
   b. Measures of spread and variability (range, quartiles, variance, standard deviation)
   c. Measures of association (correlation)
   d. Frequency distributions

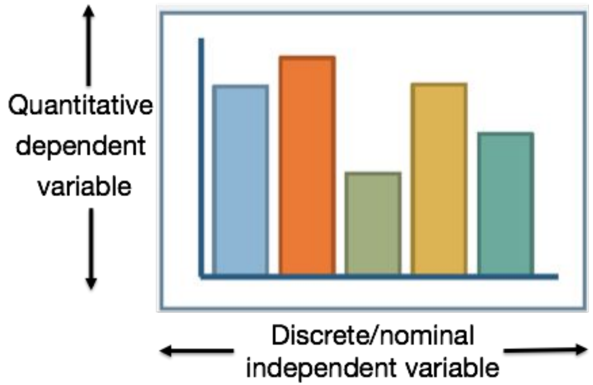
2. Introduction to R
   a. Data variables & basic operations
   b. Loading data and reading data
   c. Summary stats

Lesson Objectives

1. Adv. Descriptive Statistics & Visualization
    a. Shape of Distributions & Statistical Graphics
        i. Histograms
        ii. Scatterplots
        iii. The Box Plot
    b. Z-Scores
    c. Hypothesis testing & statistical significance.

2. Regression
    a. Linear Models
    b. Ordinary Least Squares
    c. Simple Linear Regression

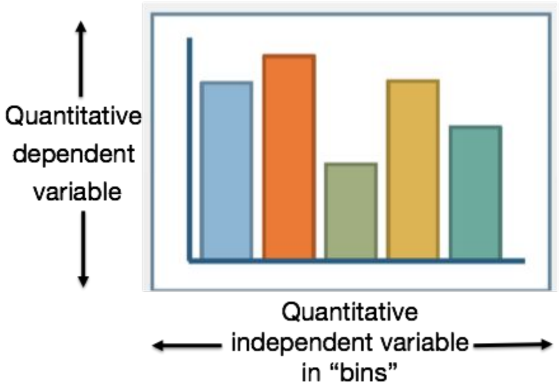
Bar Chart



Quantitative dependent variable

Discrete/nominal independent variable

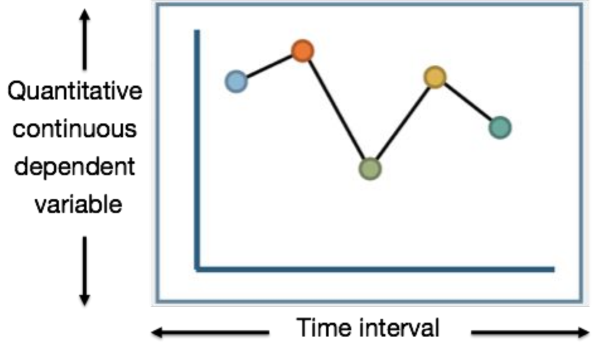
Histogram



Quantitative dependent variable

Quantitative independent variable in “bins”

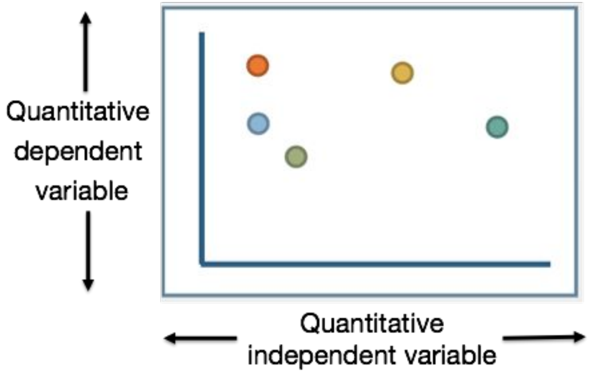
Time Series



Quantitative continuous dependent variable

Time interval

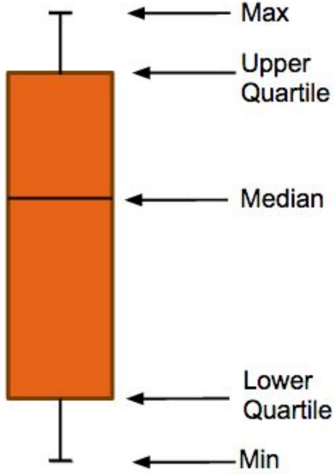
Scatter Plot



Quantitative dependent variable

Quantitative independent variable

Box Plot



Max

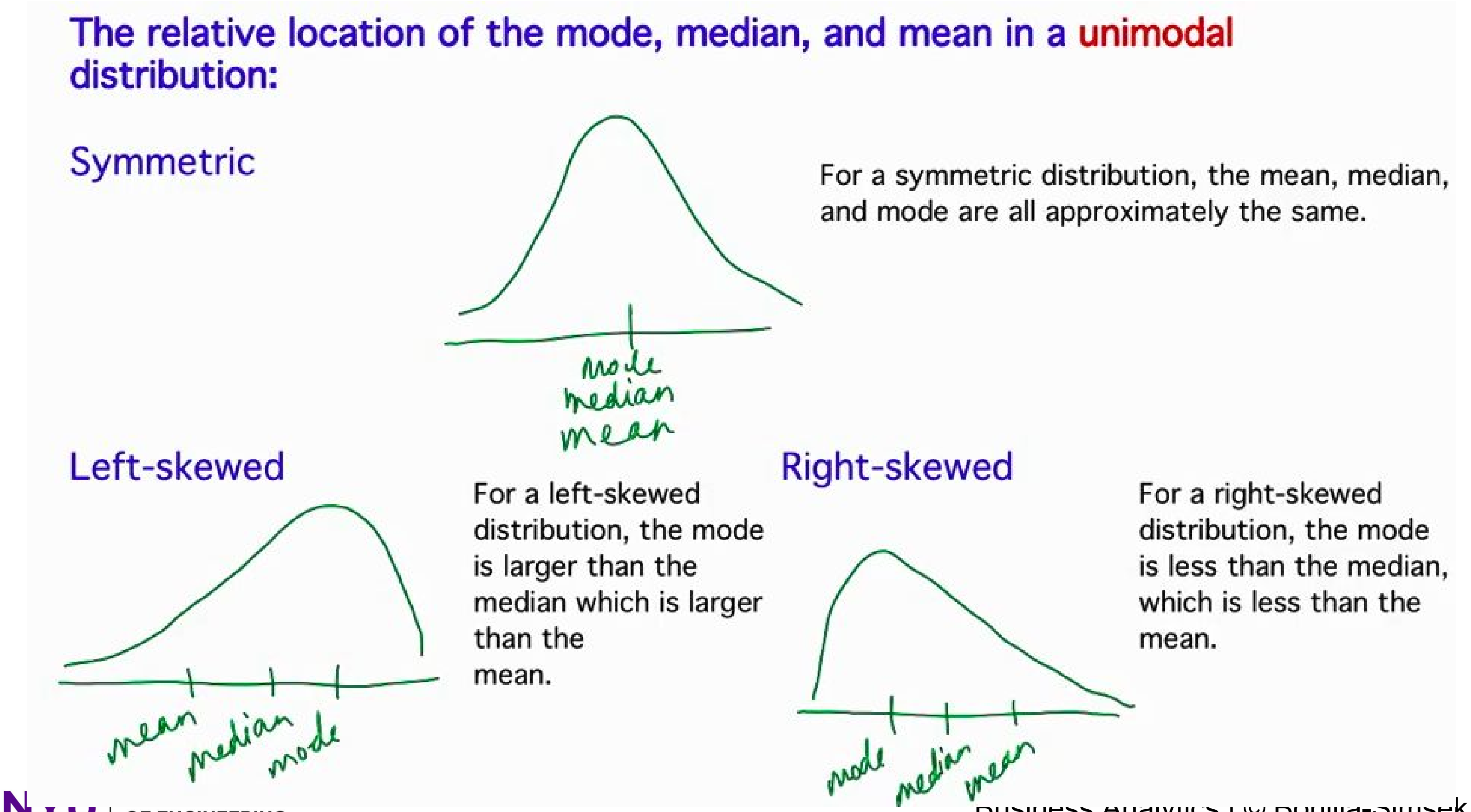
Upper Quartile

Median

Lower Quartile

Min

Shape of Distribution



The relative location of the mode, median, and mean in a unimodal distribution:

Symmetric

For a symmetric distribution, the mean, median, and mode are all approximately the same.

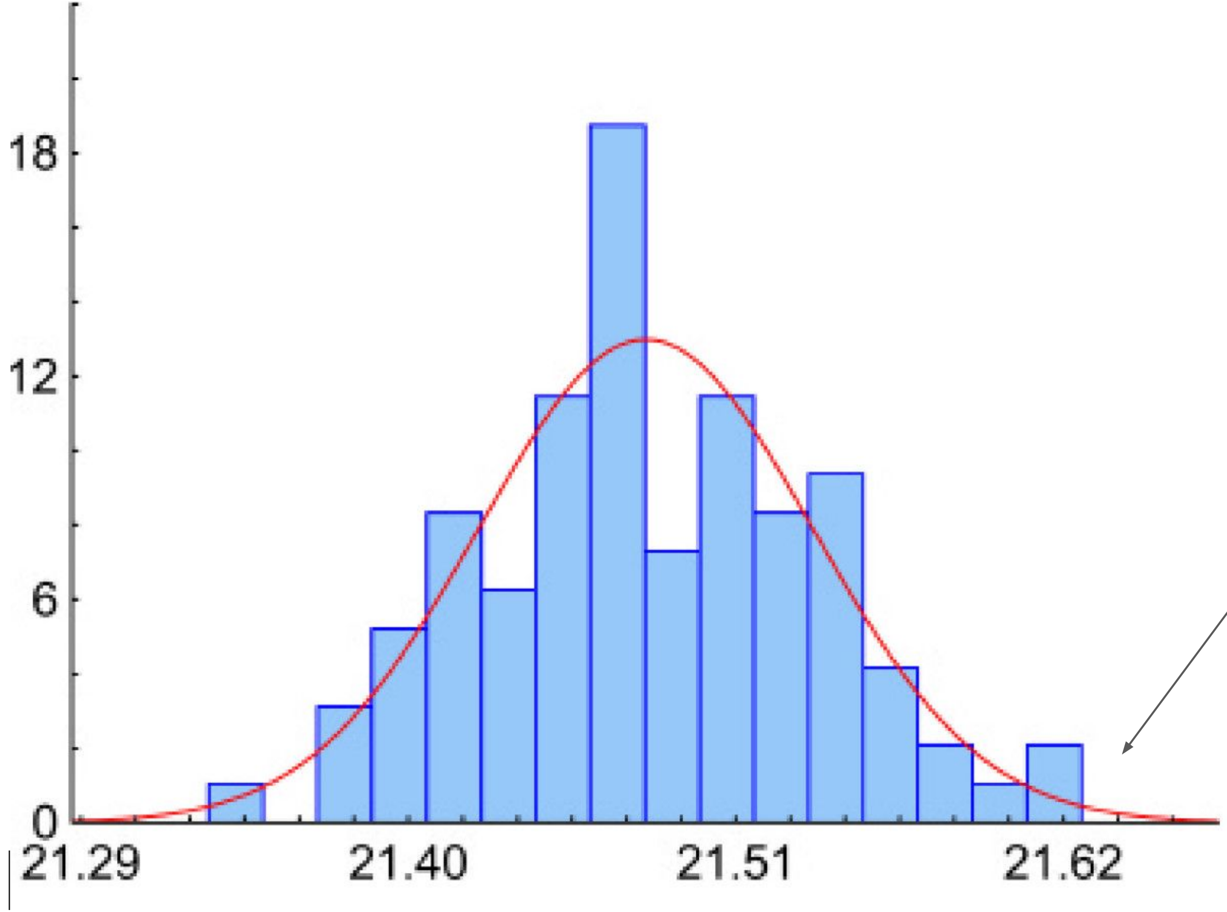
Left-skewed

For a left-skewed distribution, the mode is larger than the median which is larger than the mean.

Right-skewed

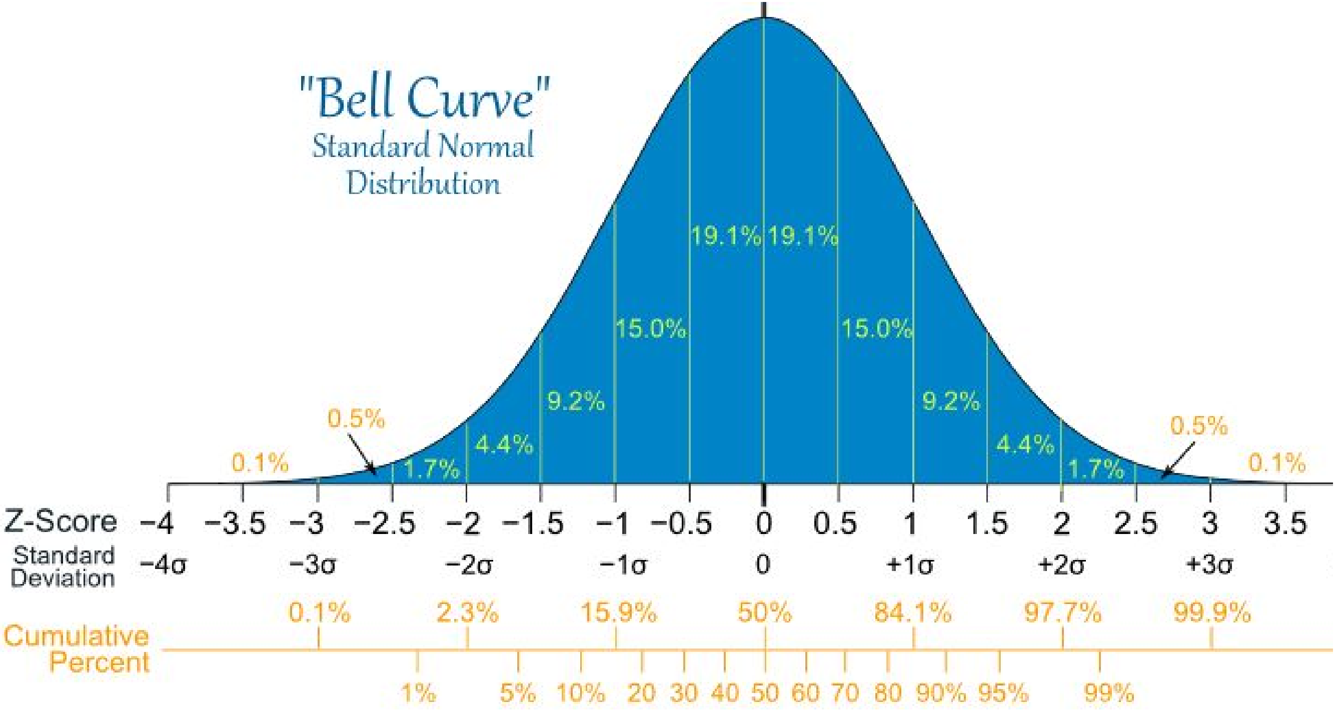
For a right-skewed distribution, the mode is less than the median, which is less than the mean.

Histograms & Distributions



Tail

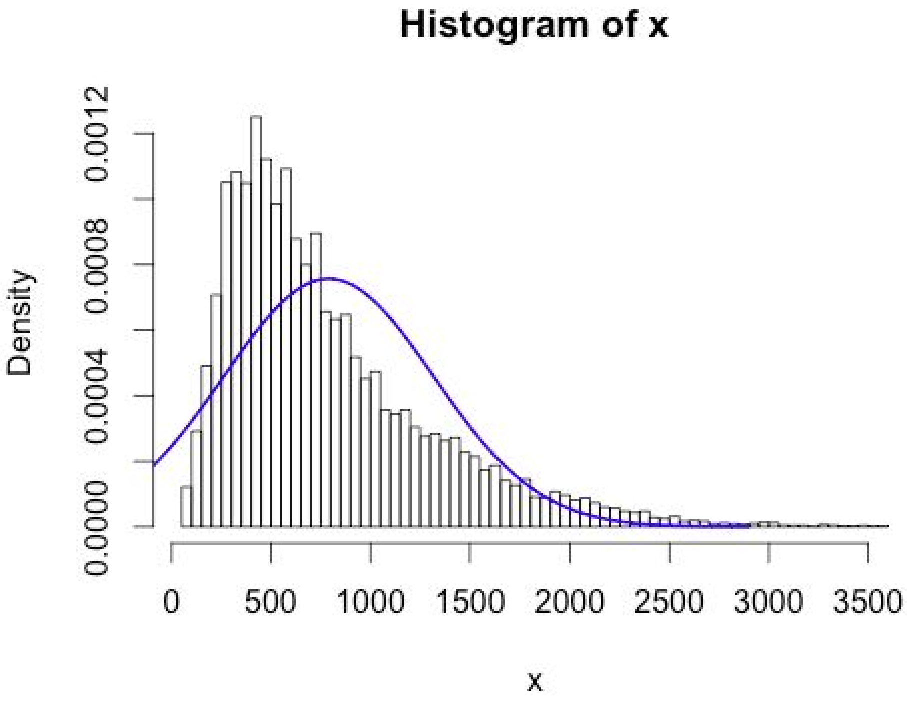
Normal Distribution



Properties:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean
- 68% of values are within 1 standard deviation of the mean
- 95% of values are within 2 standard deviation of the mean
- 99% of values are within 3 standard deviation of the mean

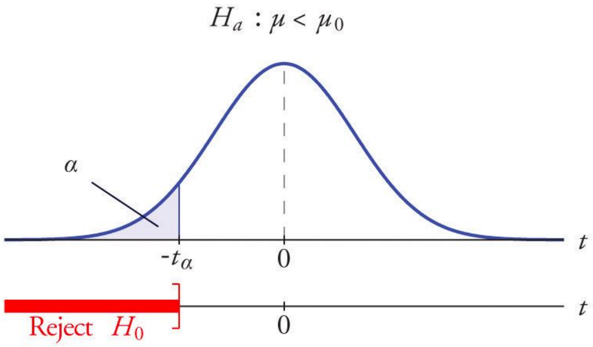
Normalizing Data → Fitting a distribution to data

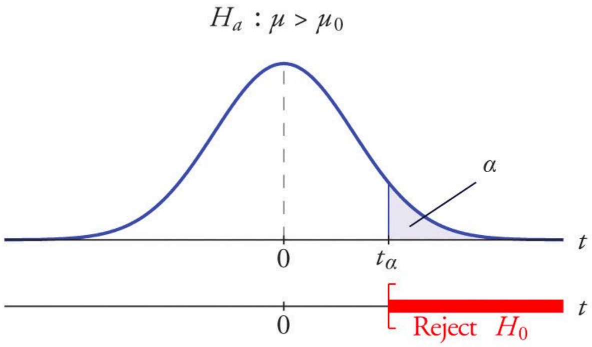


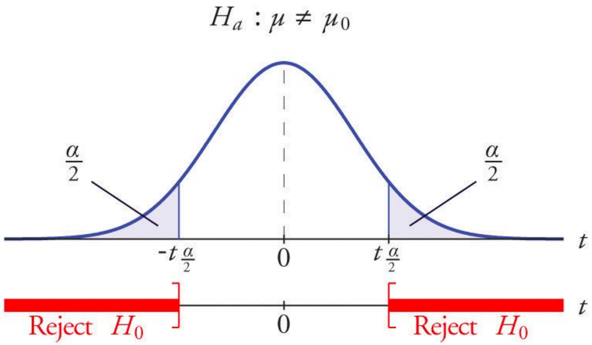
Normal curve centered at mean of data set with standard deviation equal to the deviation of the sample data.

How good is this model?

Significance Testing and Confidence Intervals







Statistical test provides a mechanism for making quantitative decisions about a process or processes.

The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process.

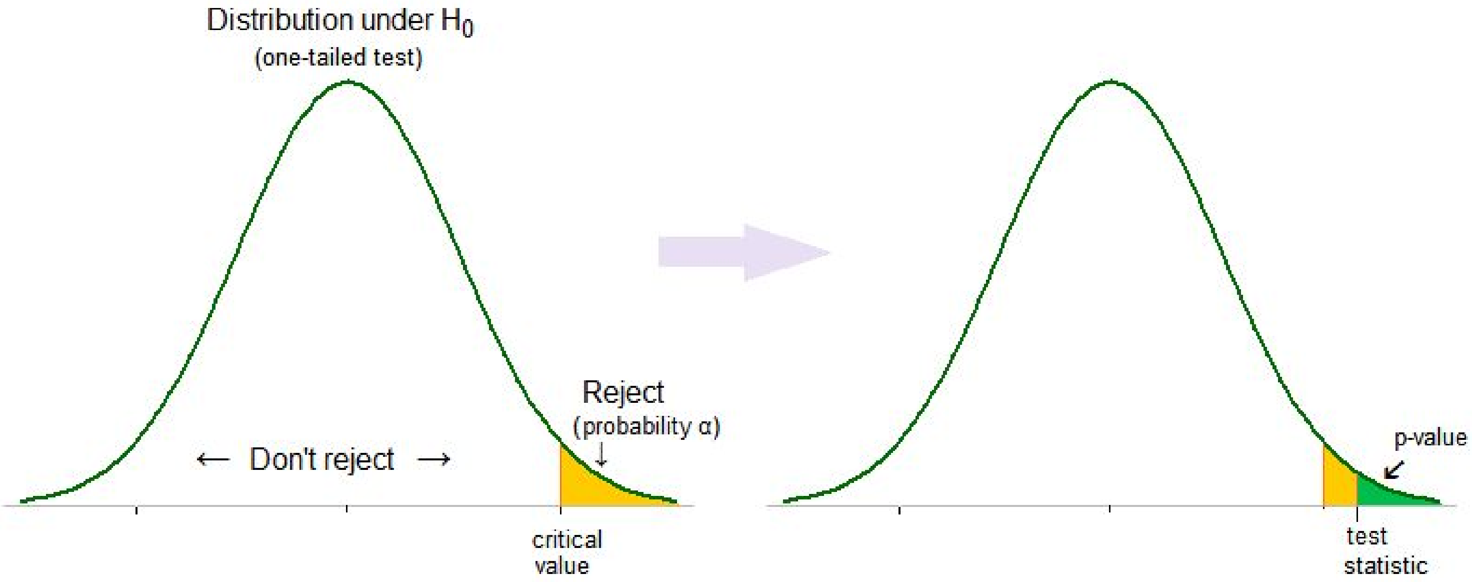
The conjecture is called the **null hypothesis**

Hypothesis Testing

Steps:

1. **State the hypotheses.** This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
2. **Formulate an analysis plan.** The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
3. **Analyze sample data.** Find the value of the test statistic (mean score, proportion, t-score, z-score, etc.) described in the analysis plan.
4. **Interpret results.** Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

Significance Levels (α) & P-Values



Alpha sets the standard for how extreme the data must be before we can reject the null hypothesis.
The P-value indicates how extreme the data are.

- If the p-value is less than or equal to the alpha ($p< \alpha$), then we reject the null hypothesis, and we say the result is statistically significant.
- If the p-value is greater than alpha ($p > \alpha$), then we fail to reject the null hypothesis, and we say that the result is statistically nonsignificant (n.s.)

Statistically Significant Results

- A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- A test result is statistically significant when the sample statistic is unusual enough relative to the null hypothesis that we can reject the null hypothesis for the entire population.
- The common alpha values of 0.05 and 0.01 are simply based on tradition.

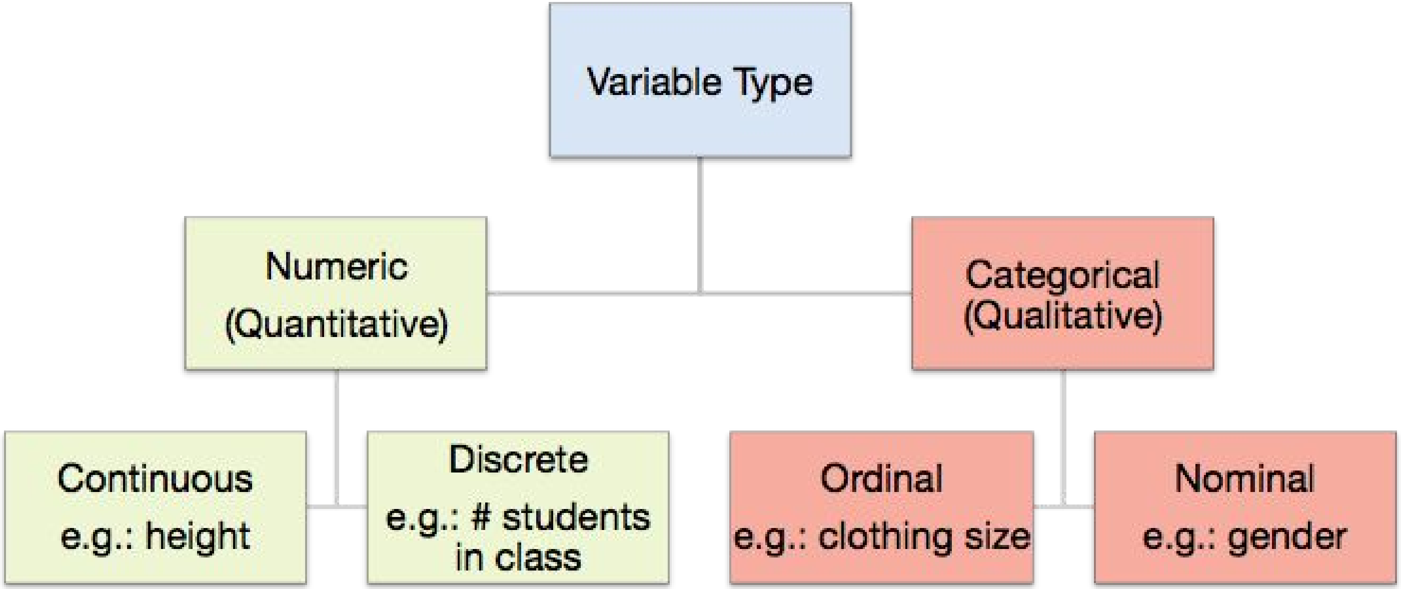
Business Implication

1. Load the Zagat file and run summary statistics study on “service”
2. Are there outliers?
   a. Run the “Outlier Detection & Z-score” Rscript
3. Normalize the data
   a. See section 3.3 on NYUClassess
4. Run a statistical test
   a. Use R function *t.test(x)*

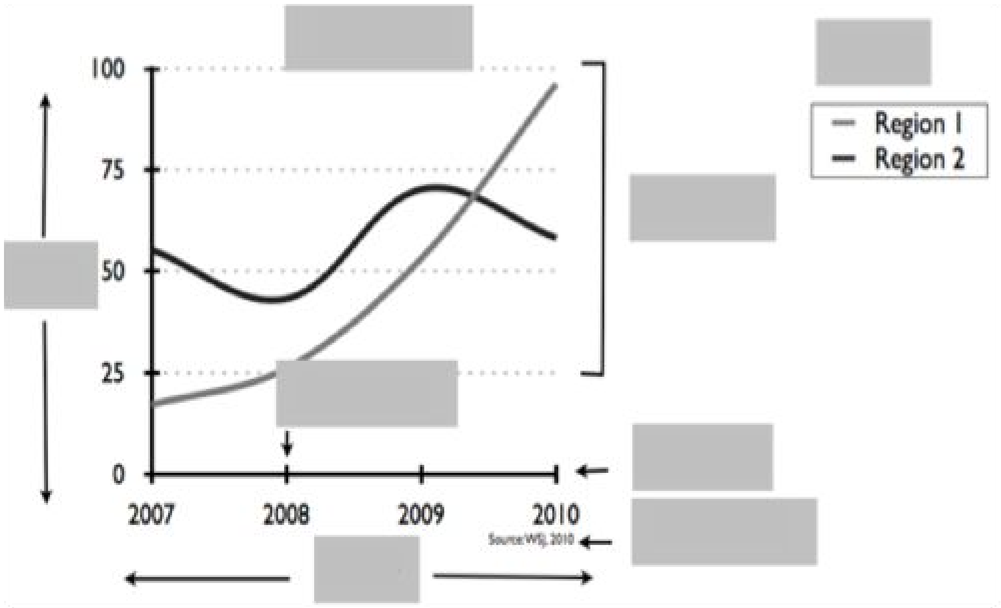
Visualization for Descriptive Analytics

- Data types
- Data transformation - percentages, proportions...
- Chart types - visualizing patterns, relationships, comparisons, or distributions?

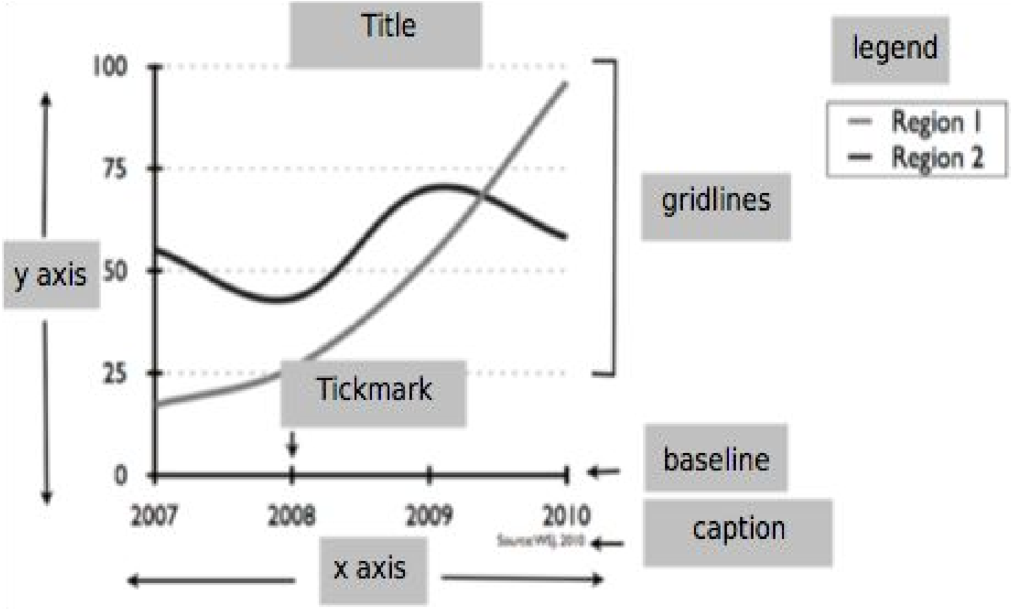
Data Types


Variable Type
Numeric (Quantitative)
Categorical (Qualitative)
Continuous e.g.: height
Discrete e.g.: # students in class
Ordinal e.g.: clothing size
Nominal e.g.: gender

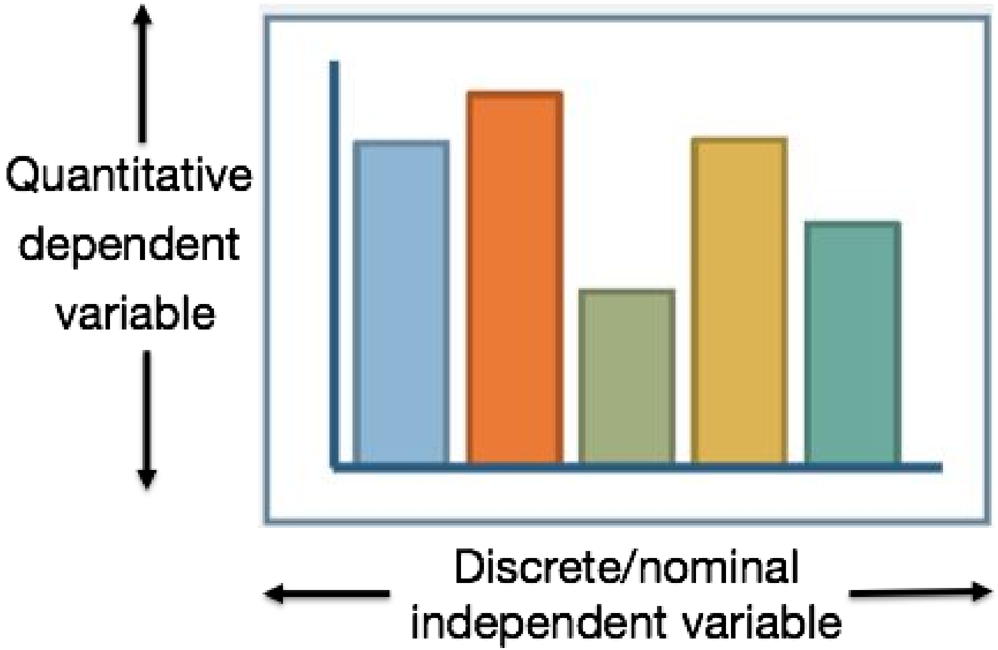
Basic Chart Terminology

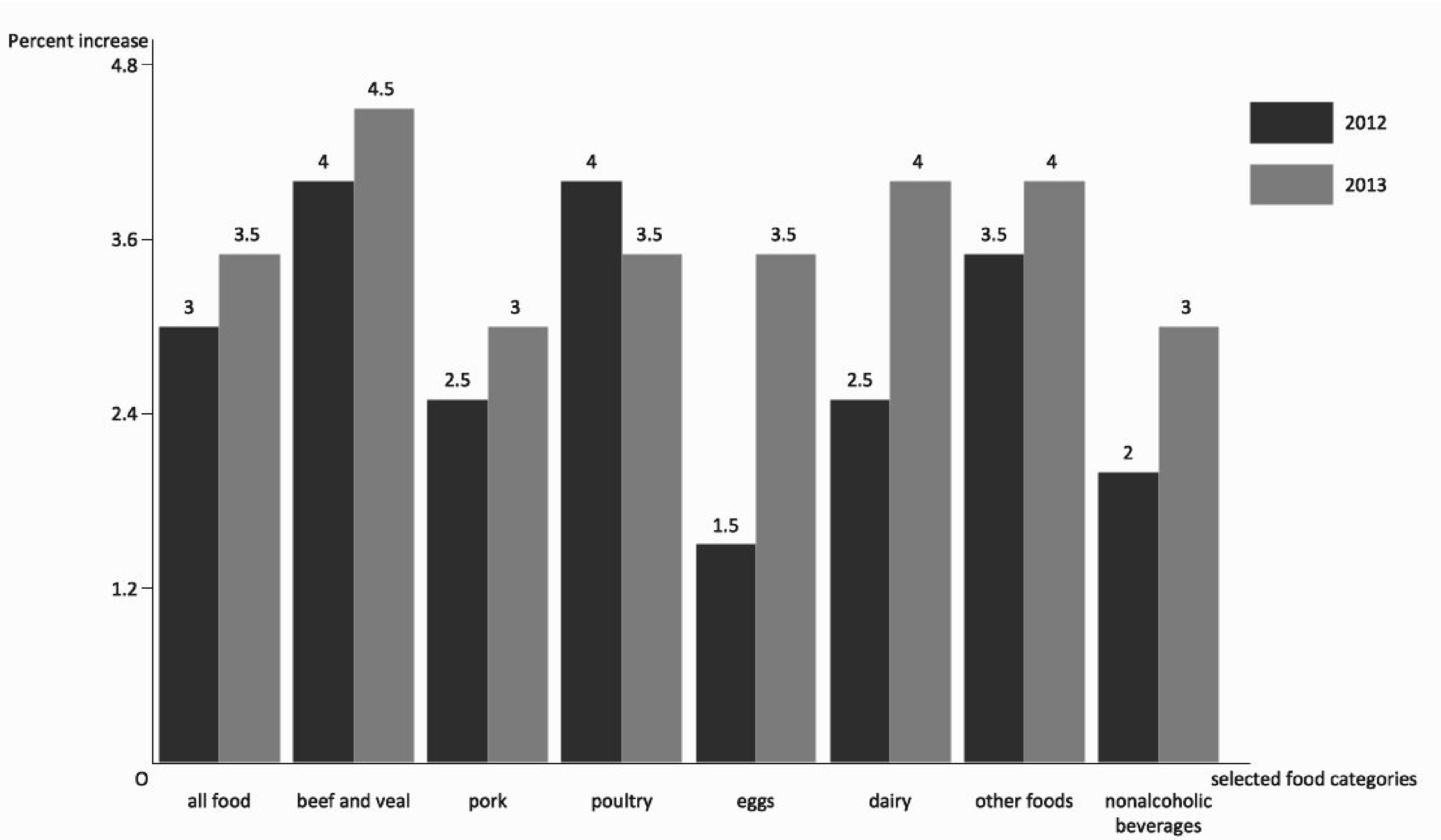


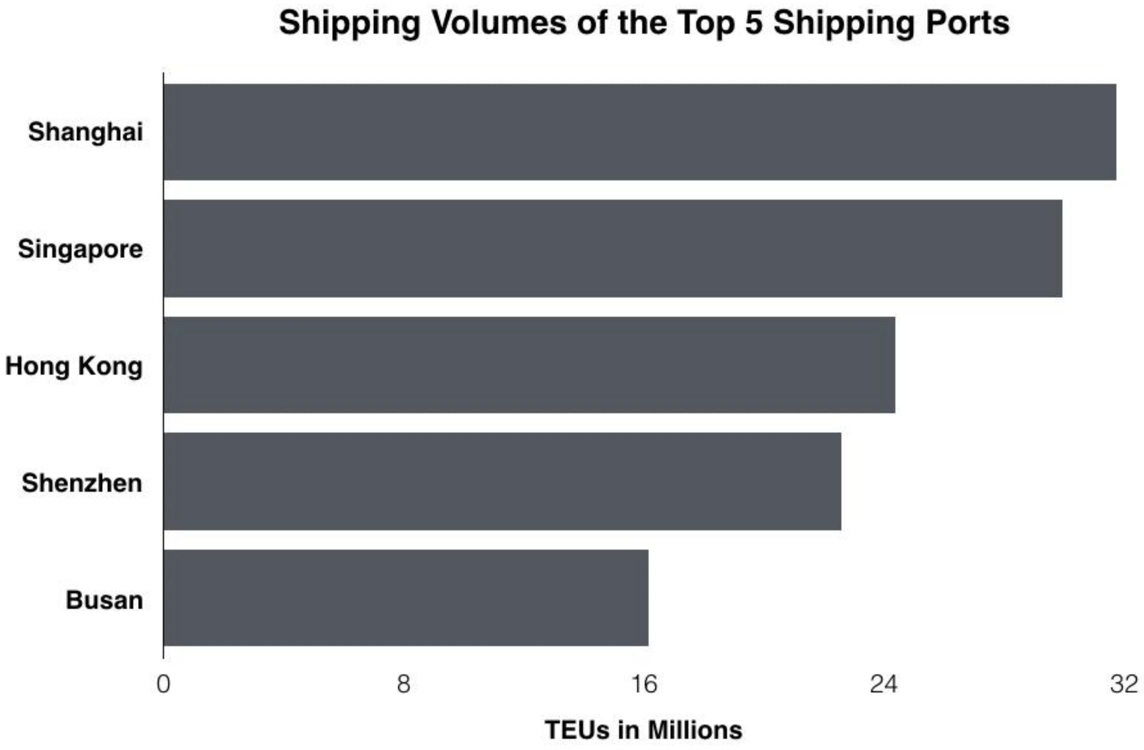
Basic Chart Terminology



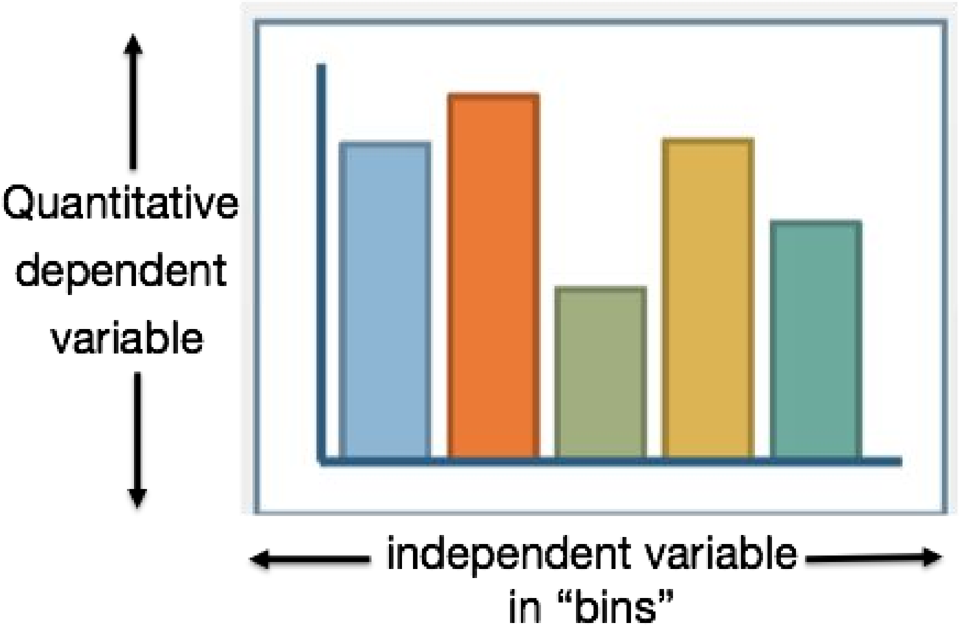
Bar Chart


Quantitative dependent variable
Discrete/nominal independent variable


Percent increase
4.8
3.6
2.4
1.2
0
4.5
4
3.5
3
2.5
4
3.5
1.5
3.5
2.5
4
3.5
4
2
3
3
2012
2013
selected food categories
all food
beef and veal
pork
poultry
eggs
dairy
other foods
nonalcoholic beverages


Shipping Volumes of the Top 5 Shipping Ports
Shanghai
Singapore
Hong Kong
Shenzhen
Busan
0
8
16
24
32
TEUs in Millions

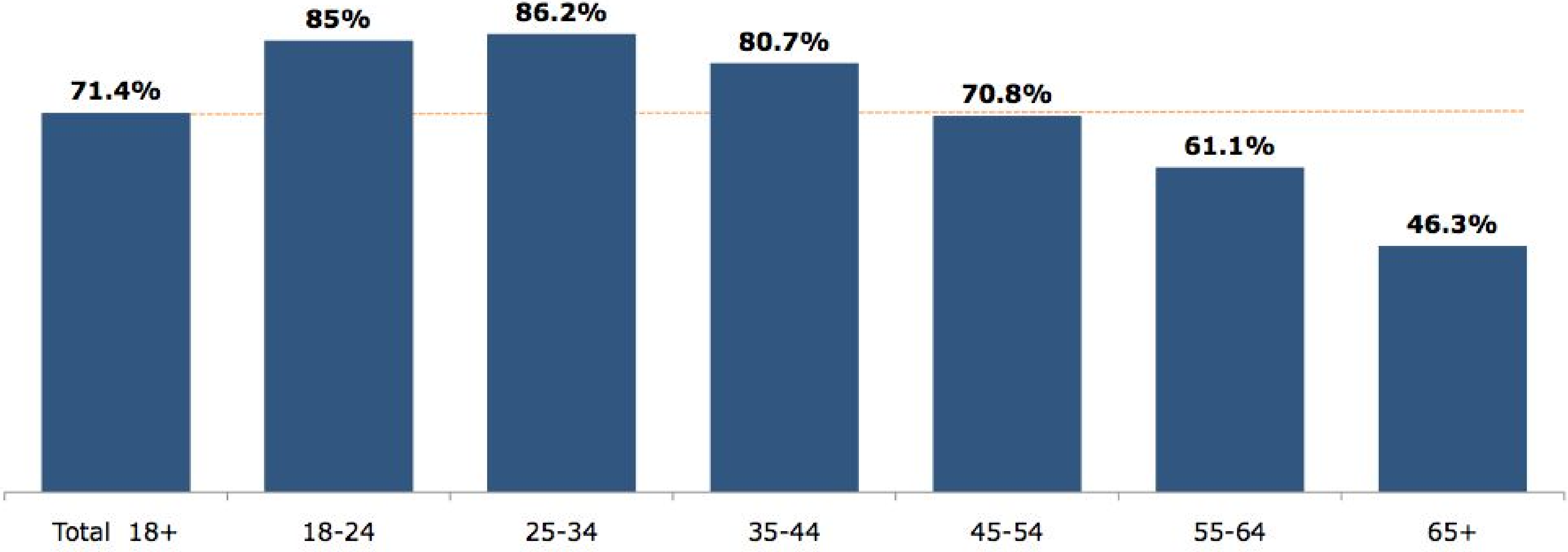
Histogram


Quantitative dependent variable
independent variable in “bins”

US Smartphone Penetration Rate, by Age Group

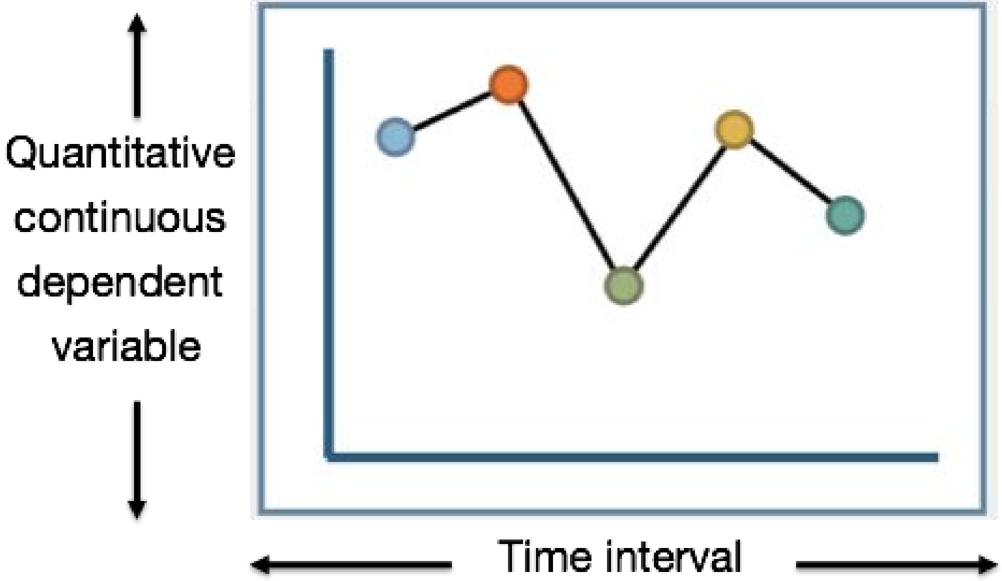
among mobile subscribers in the US

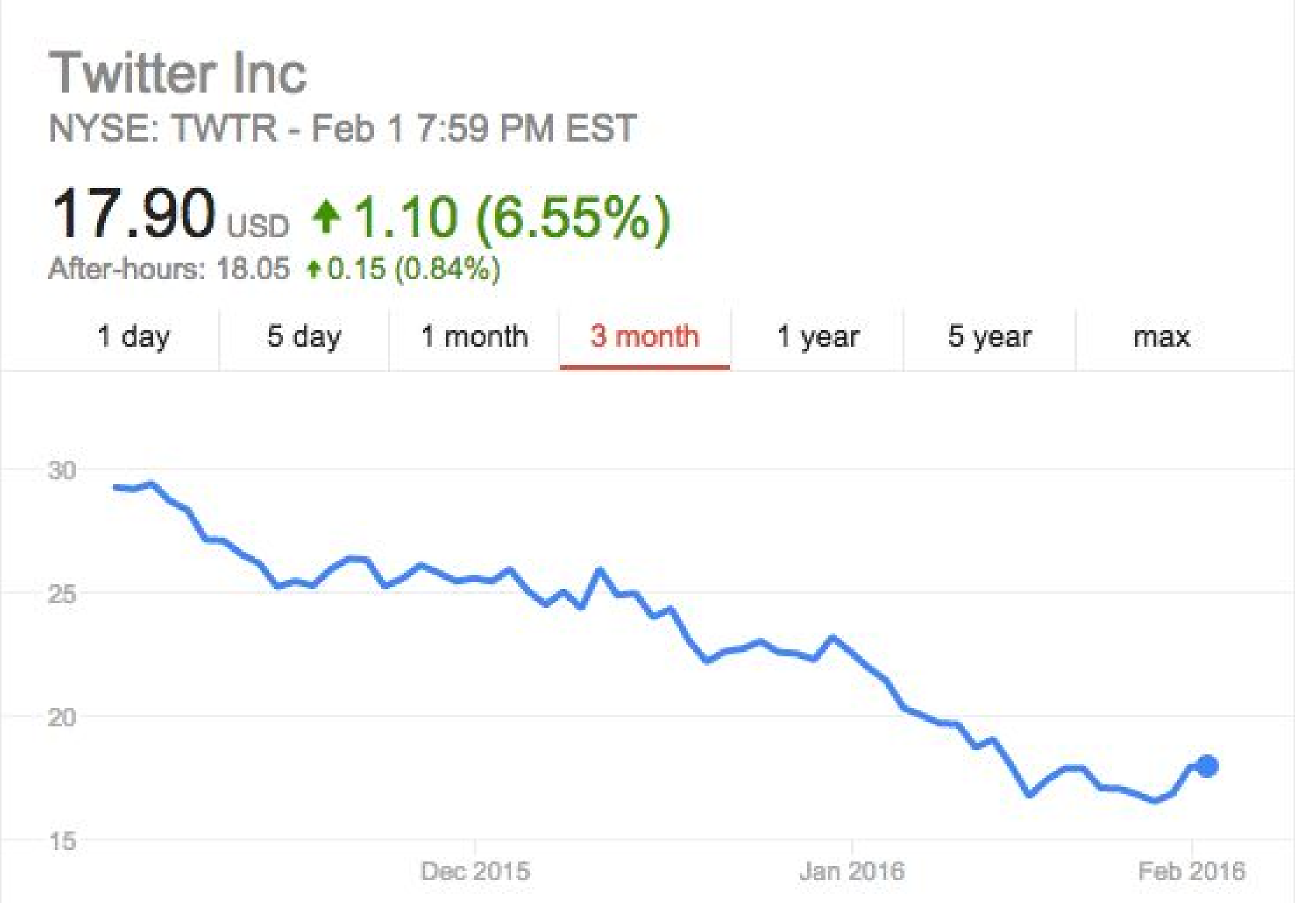
During Q2 2014



MarketingCharts.com | Data Source: Nielsen

Time Series


Quantitative continuous dependent variable
Time interval



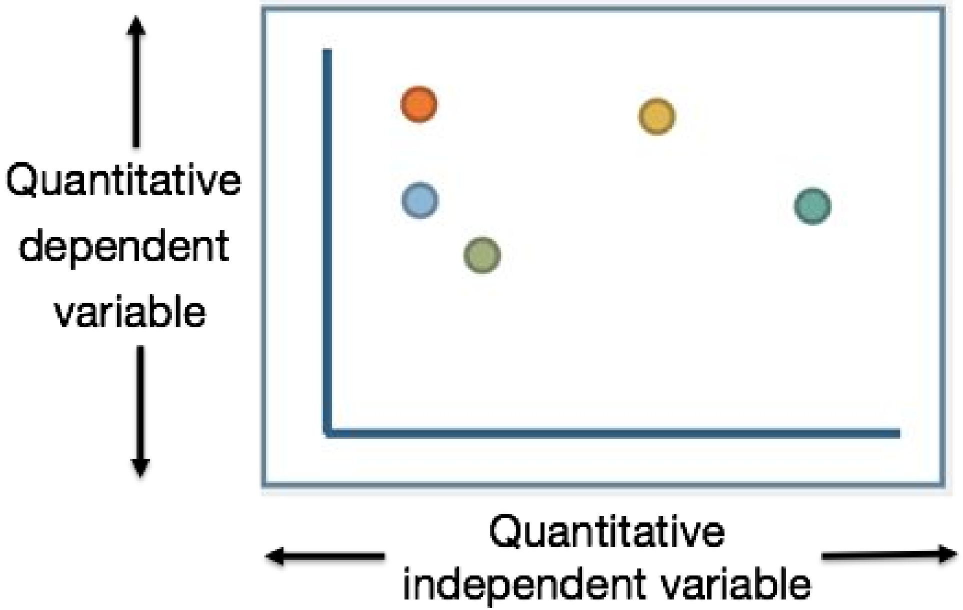
Twitter Inc

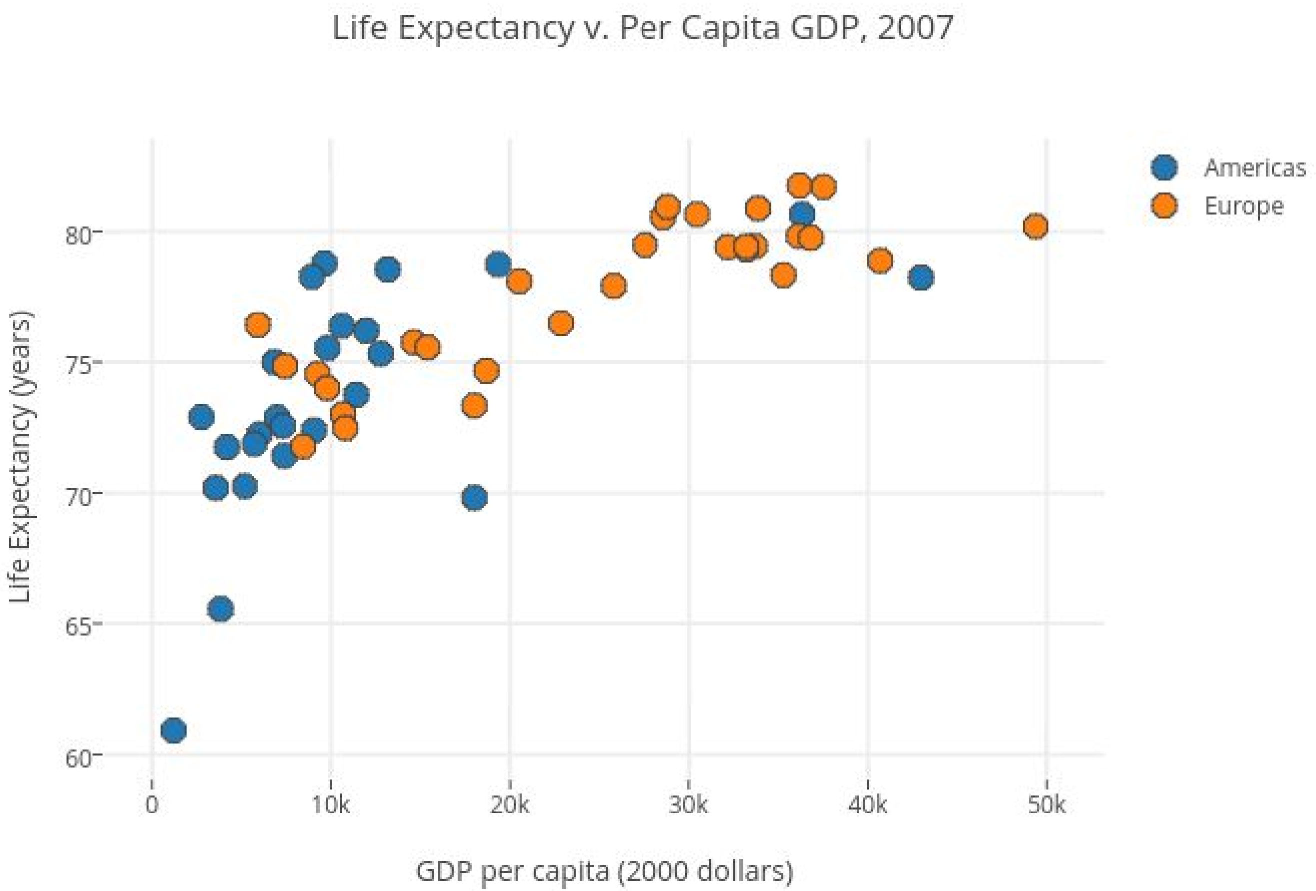
NYSE: TWTR - Feb 1 7:59 PM EST

17.90 USD ↑1.10 (6.55%)

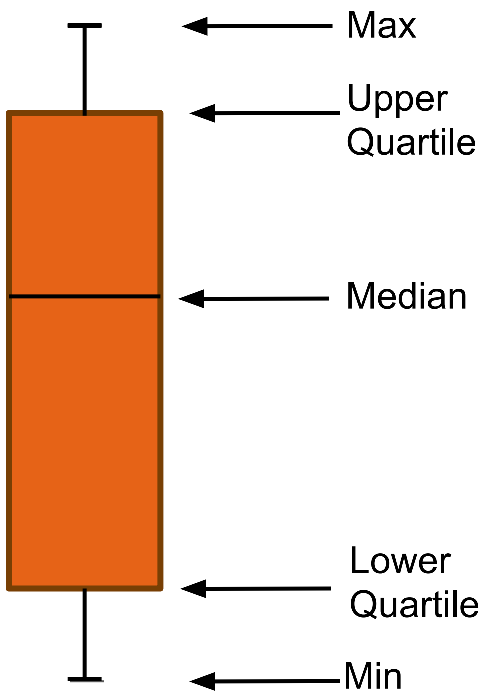
After-hours: 18.05 ↑0.15 (0.84%)

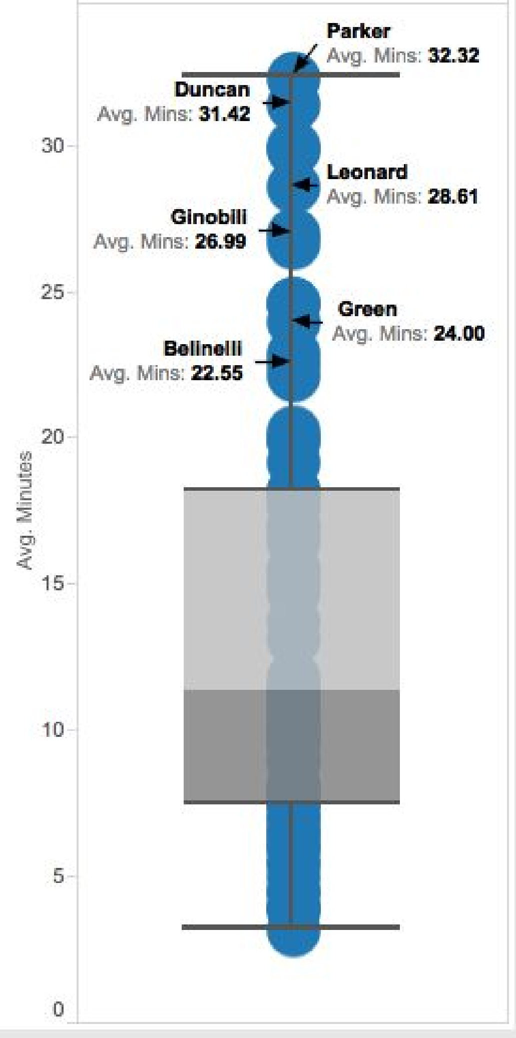
Scatter Plot


Quantitative dependent variable
Quantitative independent variable


Life Expectancy v. Per Capita GDP, 2007
Americas
Europe
Life Expectancy (years)
80
75
70
65
60
0
10k
20k
30k
40k
50k
GDP per capita (2000 dollars)

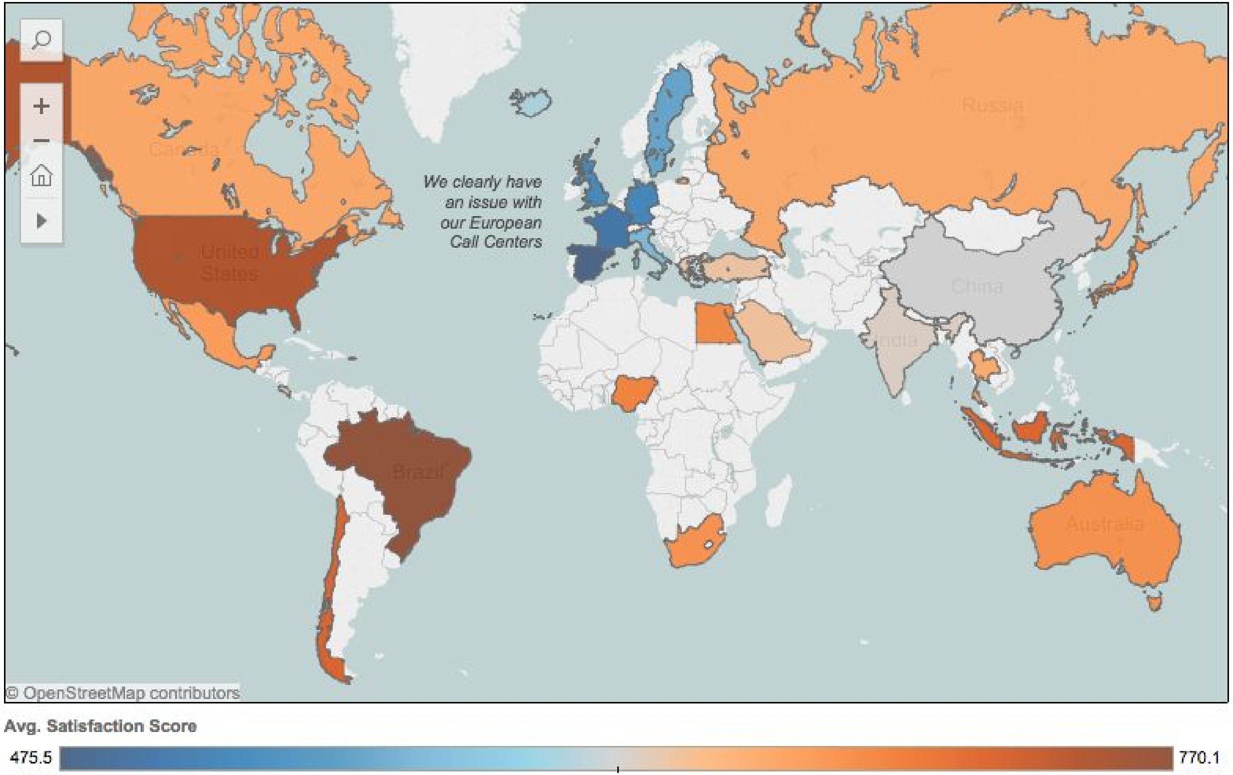
Box Plot (Box & Whisker diagram)



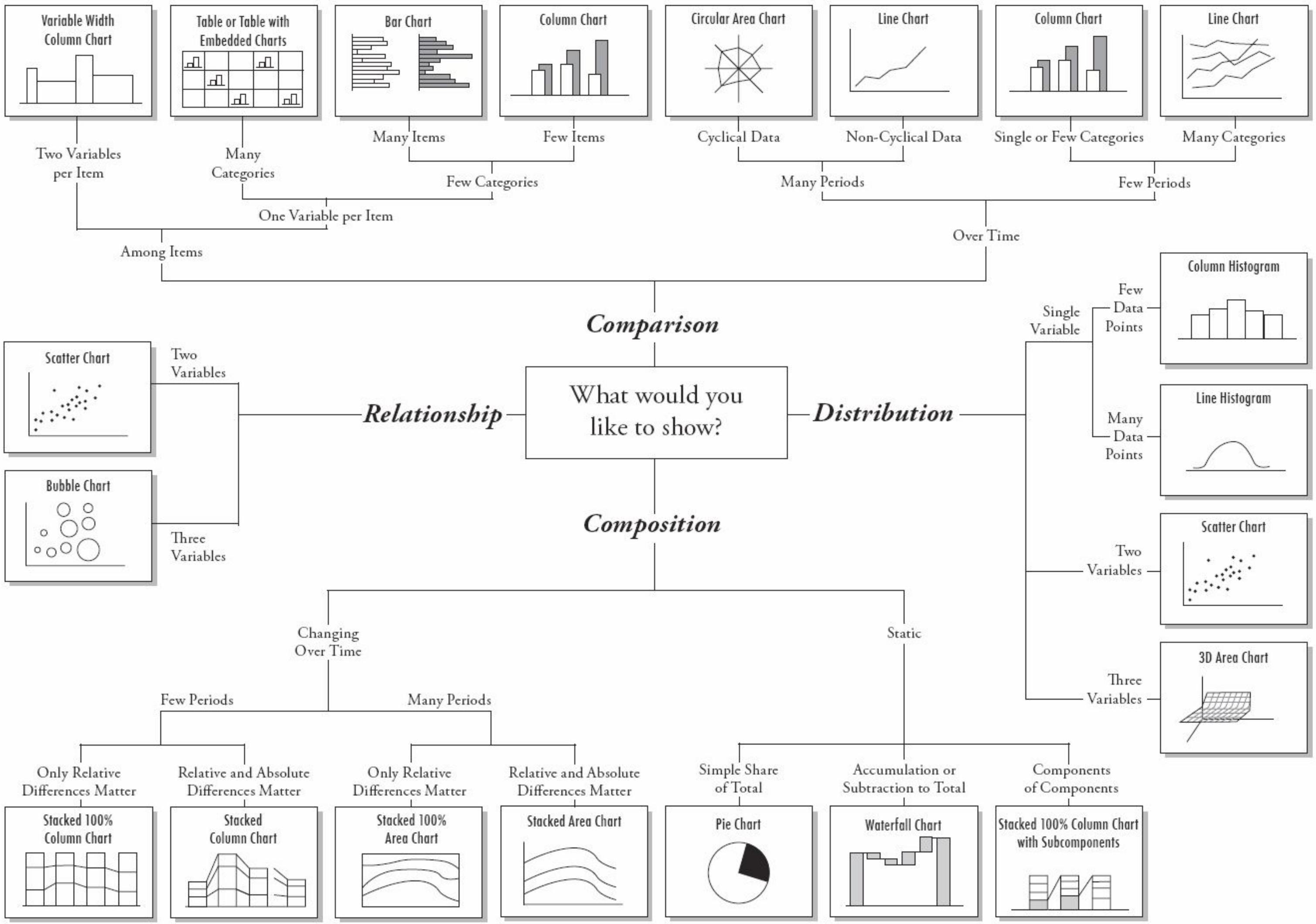

Parker
Avg. Mins: 32.32
Duncan
Avg. Mins: 31.42
Leonard
Avg. Mins: 28.61
Ginobili
Avg. Mins: 26.99
Green
Avg. Mins: 24.00
Bellnelli
Avg. Mins: 22.55
Avg. Minutes
30
25
20
15
10
5
0

Geo-spatial Map

Customer Call Center Satisfaction



We clearly have an issue with our European Call Centers



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

Distribution

Single Variable

Few Data Points

Column Histogram

Many Data Points

Line Histogram

Two Variables

Scatter Chart

Three Variables

3D Area Chart

Relationship

Two Variables

Scatter Chart

Three Variables

Bubble 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