Overview & Context Descriptive Statistics

Business Analytics - *Spring 2016*

Meet the Instructors

Nil Simsek

VP, Customer Acquisitions and Line Management Analytics at Citi & Adjunct Professor at NYU

LinkedIn | ns1254@stern.nyu.edu

JeanCarlo Bonilla

Director of Insights at DataKind & Adjunct Professor at NYU

LinkedIn | jb3379@nyu.edu

Course Description

Business analytics is a set of data analysis and modeling techniques for understanding business situations and improving business decisions. This course provides an introduction to business analytics concepts, methods and tools with concrete examples from industry applications.

Throughout the course, we explore the challenges that can arise in implementing analytical approaches within an organization. The course emphasizes that business analytics is not a theoretical discipline: these techniques are only interesting and important to the extent that they can be used to provide real insights and improve the speed, reliability, and quality of decisions.

Pre-Requisites: Probability, Statistics, Linear Algebra, & Excel

Course Description

- In the first part of the course, we will focus on descriptive analytics and exploratory data analysis concepts with a refresher on basic probability and statistics.
- In the second part, we will cover principles, techniques, and techniques for spatial data, time series, and text as data.
- The final part of the course will introduce a project that links business impact and modern data analytics techniques for managerial decision making in functional areas, including finance, marketing, and operations.

Required Materials

Required Textbook

None.

Suggested Textbook by Topic

- Essentials of Business Analytics. Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann, and David R. Anderson. Cengage Learning, 2014.
- *Excell:* Management Science: The Art of Modeling with Spreadsheets, Powell and Baker. Wiley
- *R:* Data Mining and Business Analytics with R, Johannes Ledolter, 1st Edition
- *Business Analytics:* Keeping Up with the Quants: Your Guide to Understanding and Using Analytics, Thomas H. Davenport & Jinho Kim, 2013
- *Decision Models:* Spreadsheet Modeling & Decision Analysis: A Practical Introduction to Management Science, Cliff Ragsdale, 6th edition.
- *Data Visualization:* The visual display of quantitative information, Eduard R. Tufte, 2001
- *Scoping:* Thinking with Data How to Turn Information into Insights, Max Shron, 2014

Grading Policy

- **Weekly Assignments, Quizzes, and In-class Data Dives - 40%**
  - Mostly data analysis and programming assignments. Some assignments will include theoretical aspects to make sure students understand the important mathematical concepts in data analytics.
  - There will be a 10min online quiz during each class testing the understanding of theory reviewed in class as well as reading materials.
  - In-class data dives are hands on sessions around the entire data life cycle. These include project scoping, data manipulation and integration, analysis, visualization, and reporting.

- **Exams – 30%**
  - Two exams on covering theory and applications

- **Team Project –30%**
  - This is the capstone experience of the course where students will form groups consisting of between 3 and 4 people depending upon the size of class. Teams will build a project using a publicly accessible datasets. They will motivate the business problem, do enough explanatory analysis and generate data driven strategic insight. Each team will give a brief class midterm presentation on the project, followed by a final presentation at the end of the course.

Letter Grade	100% Scale	Grade Point Value
A	100-95	4.0
A-	94-90	3.7
B+	89-85	3.3
B	84-80	3.0
B-	79-75	2.7
C+	74-70	2.3
C	69-65	2.0
F	64-0	0

Lett	lue
A	
A-	
B+	
B	
B-	
C+	
C	
F	

NO GRADE NEGOTIATION IN THIS COURSE

Course Web Page

You must have access to the ***NYUClasses*** site (http://classes.nyu.edu/). All announcements and class-related documents (supplemental and suggested readings, discussion questions, etc.) will be posted there.

Some class announcements will be distributed via NYU e-mail. Thus, it is important that you actively use your NYU e-mail account, or have appropriate forwarding set up on NYU Home (https://home.nyu.edu/)

In addition, lecture, data, and code will be posted in the following github repository https://github.com/jcbonilla/BusinessAnalytics

Statement of Academic Integrity

Students are expected to follow standards of excellence set forth by New York University. Such standards include respect, honesty, and responsibility. This class does not tolerate violations to academic integrity including:

- Plagiarism
- Cheating on an exam
- Submitting your own work toward requirements in more than one course without prior approval from the instructor
- Collaborating with other students for work expected to be completed individually
- Giving your work to another student to submit as his/her own
- Purchasing or using solutions or work online or from a commercial firm and presenting it as your own work

Course Schedule

(subject to change)

What is 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.

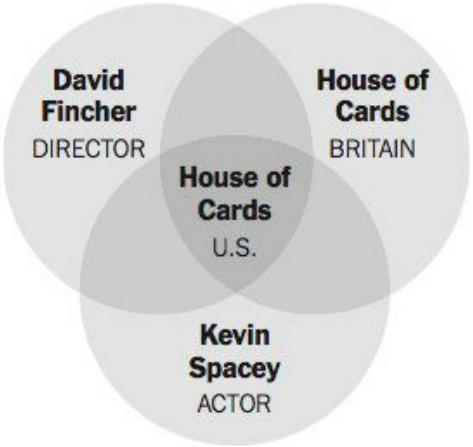
IMPLICATIONS:

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

What analytics examples can you think of?

Circles of Proven Success

Netflix determined that the overlap of these three areas would make "House of Cards" a successful entry into original programming.



THE NEW YORK TIMES


BUSINESS
PLATINUM
Capital One
4000 1234 5678 9010
4000
VALID DATES
00/00 - 00/00/00
SMALL BUSINESS, INC.
LEE M CARDHOLDER
VISA

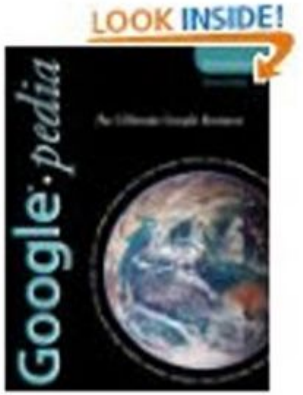
amazon.com

Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop

Google Apps Administrator Guide: A Private-Label Web Workspace



Googlepedia: The Ultimate Google Resource (3rd Edition)

Types of Analytics

Descriptive
Analytics

Descriptive Analytics:
Encompasses the set of techniques that describe what has happened in the past whether that is one minute or one year ago.The vast majority of the statistics we use like sums, averages, percent changes fall into this category.

About 35% of companies surveyed say they do this consistently.

Use Descriptive statistics when you need to understand at an aggregate level what is going on in your company, and when you want to summarize and describe different aspects of your business.

Types of Analytics

Descriptive Analytics

Predictive Analytics

Predictive Analytics:
This type of analytics are about understanding the future and providing estimates about the likelihood of a future outcome. They combine historical data found in ERP, CRM, HR and POS systems to identify patterns and apply models and algorithms to capture relationships between various data sets.

However, less than 1% of companies surveyed have tried this yet.

Use Predictive analysis any time you need to know something about the future or fill in the information that you do not have.

.

Types of Analytics

Descriptive Analytics

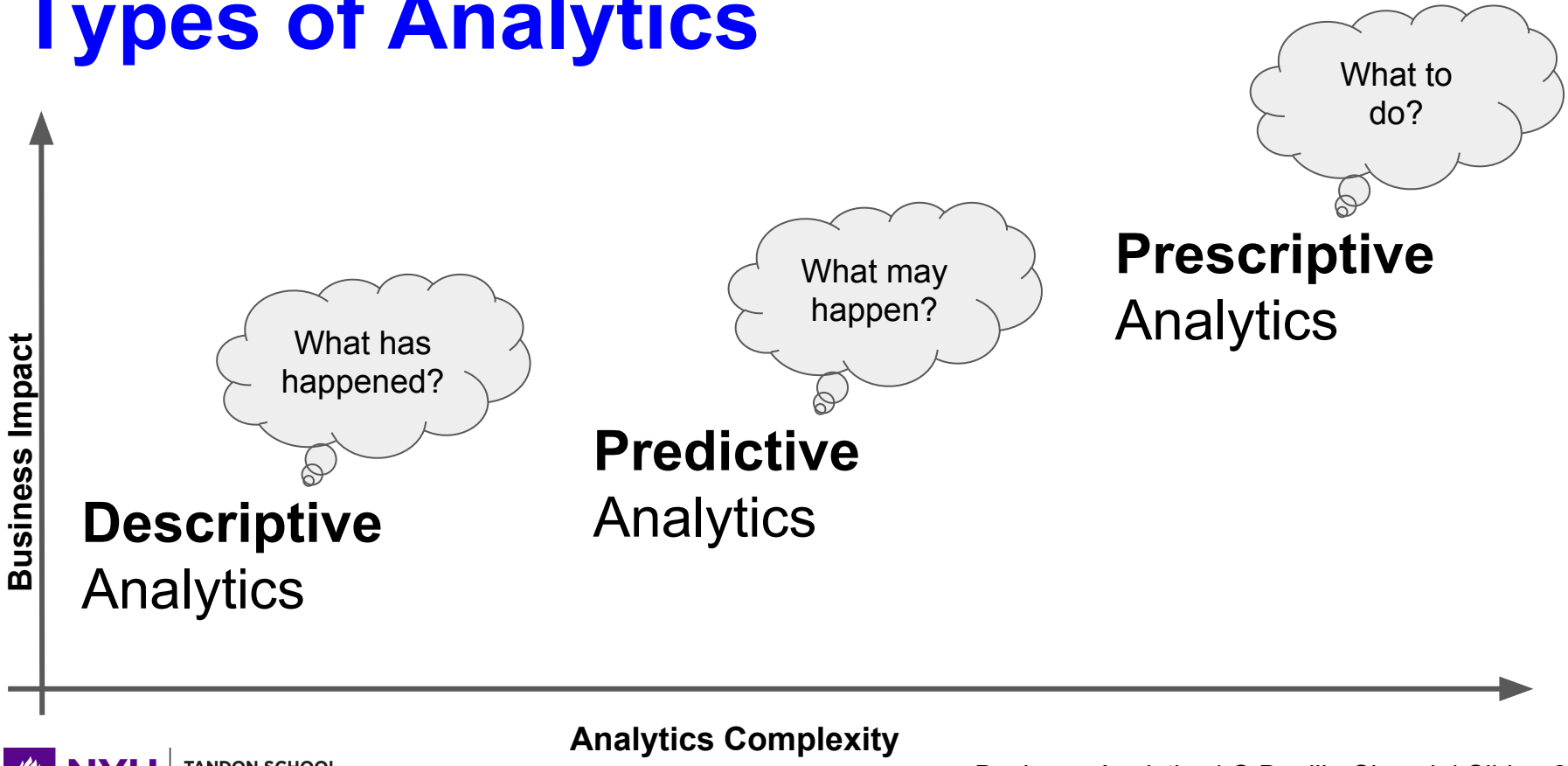
Predictive Analytics

Prescriptive Analytics

Prescriptive Analytics:
Prescriptive analytics attempt to quantify the effect of future decisions in order to advise on possible outcomes before the decisions are actually made. Prescriptive analytics are relatively complex to administer, and most companies are not yet using them in their daily course of business.

Use prescriptive statistics anytime you need to provide users with advice on what action to take.

Types of Analytics



Analytics vs BI vs Data Science

Business Intelligence (BI): tools and systems to gather, store, access and analyze an organization's raw data

- querying and reporting tools, dashboards
- traditionally used to determine trends in historical data

Data Science: involves using automated methods to extract knowledge or insights from structured or unstructured data.

- employs techniques and theories drawn from mathematics, statistics, information science, machine learning, AI, and others.

Analytics vs BI vs Data Science

- Analytics is more a catch-all term that encompasses both BI and Data Science as well as online analytical processing (OLAP), data mining/modeling, and forecasting.
- Analytics hinge upon determining relationships between data that can yield insight, and is increasingly focused on future scenarios -- predicting them and prescribing the most viable option for dealing with them.
- BI actually presents the insights determined by Analytics in reports, dashboards, or interactive visualizations.

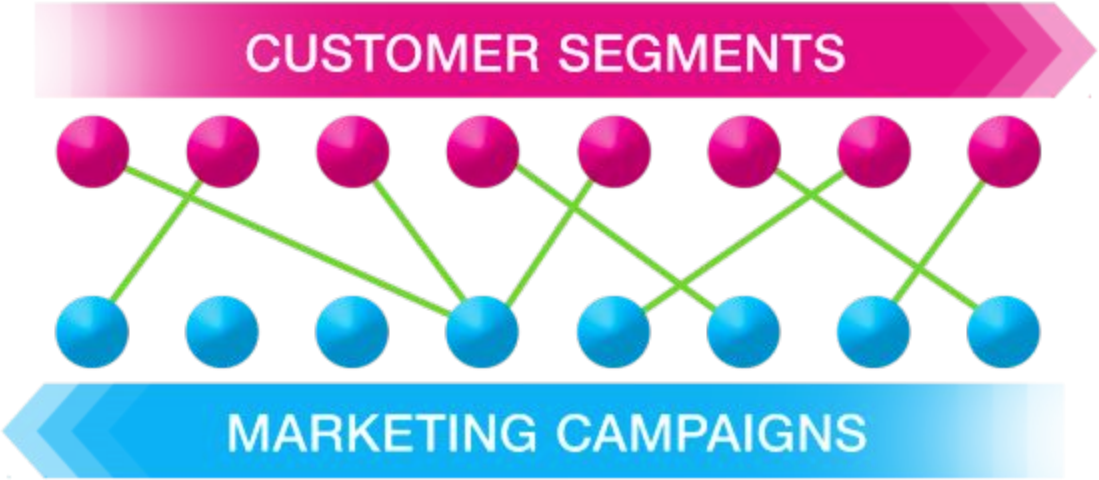
Business Analytics

- Data
- Methods
- Business decisions

Business Problems & Analytics Solutions

Classification attempts to predict, for each individual in a population, which of a (small) set of classes this individual belongs.

- e.g.: Among all the customers, which are likely to respond to a given offer?"
  - In this example, the two classes could be called "will respond" and "will not respond."
- Scoring or class probability estimation: Instead of a class prediction, a score representing the probability (or likelihood) that individual belongs to each class.

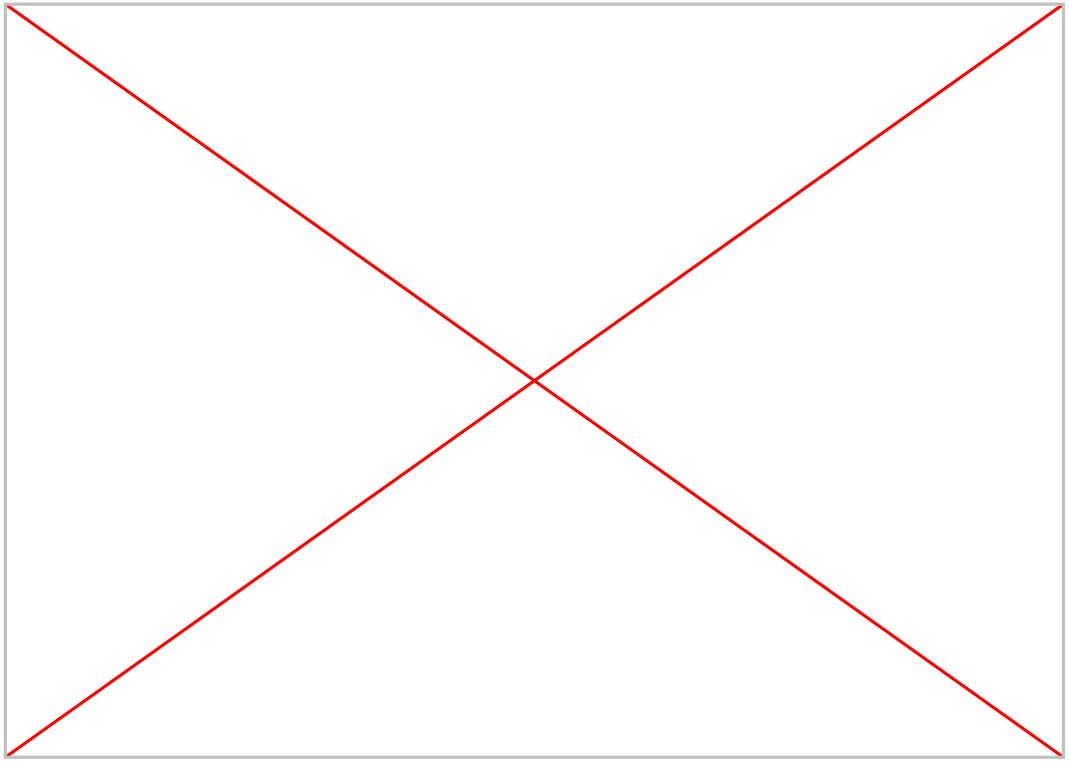

CUSTOMER SEGMENTS
MARKETING CAMPAIGNS

Business Problems & Analytics Solutions

Regression (“ value estimation”) attempts to estimate or predict, for each individual, the numerical value of some variable for that individual.

- e.g.: How much will a given customer use the service?

Classification predicts whether something will happen, whereas regression predicts how much something will happen.



Business Problems & Analytics Solutions

Similarity matching attempts to identify similar individuals based on data known about them.

- e.g.: IBM is interested in finding companies similar to their best business customers in order to focus their sales force on the best opportunities.
- Basis of methods for making product recommendations.

Business Problems & Analytics Solutions

guests who viewed this item ultimately bought

$199.99

reg: $299.99

Nikon Coolpix L840 16.1MP Digtial Camera...

Nikon

spend $25, get free shipping

(11)

$249.99

Canon PowerShot SX400 IS Digital Camera ...

Canon

spend $25, get free shipping

(5)

see low price

reg: $219.99

Sony DSCH300/B 20MP Digital Camera w...

Sony

spend $25, get free shipping

(25)

$129.99

List: $139.99

Sony Cybershot DSCW830 20.1MP Digital Ca...

Sony

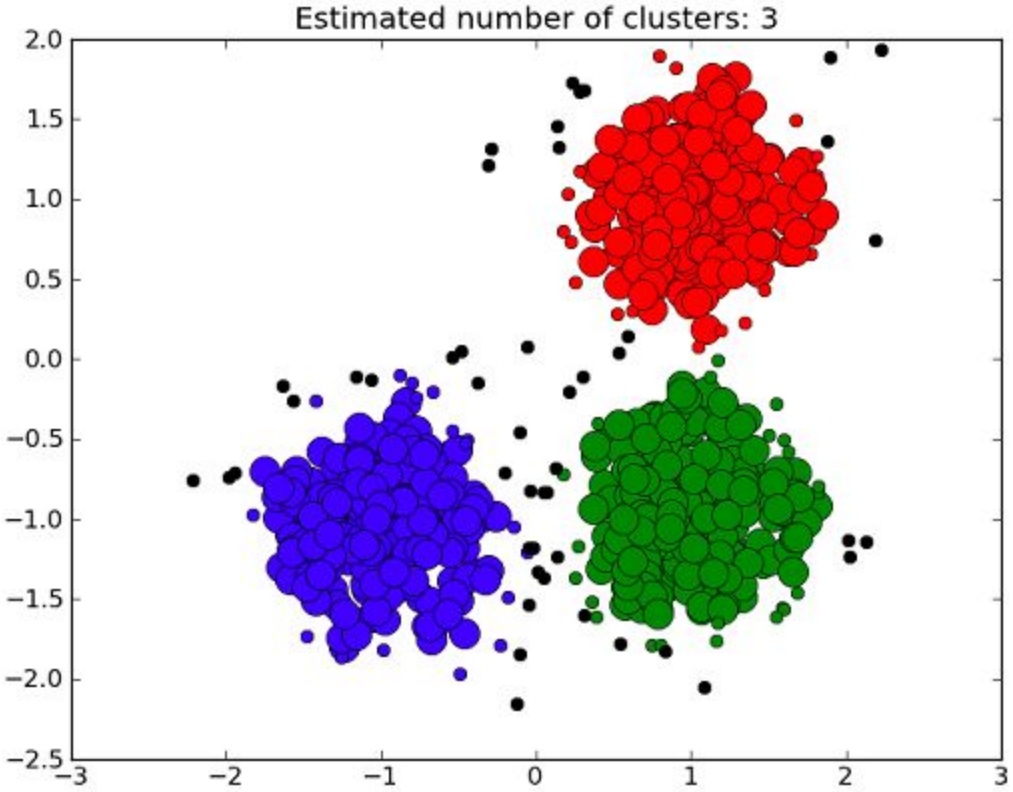
spend $25, get free shipping

(139)

Business Problems & Analytics Solutions

Clustering attempts to group individuals in a population together by their similarity, but not driven by any specific purpose.

- e.g.: Do our customers form natural groups or segments?
- Useful as a preliminary step to see which natural groups exist and lead to other questions such as “What products should we offer or develop? “


Estimated number of clusters: 3
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3
−2
−1
0
1
2
3

Business Problems & Analytics Solutions

Co-occurrence grouping attempts to find associations between entities based on transactions involving them.

- e.g.: What items are commonly purchased together?
- While clustering looks at similarity between objects based on the objects' attributes, co-occurrence grouping considers similarity of objects based on their appearing together in transactions.
- The result is a description of items that occur together.

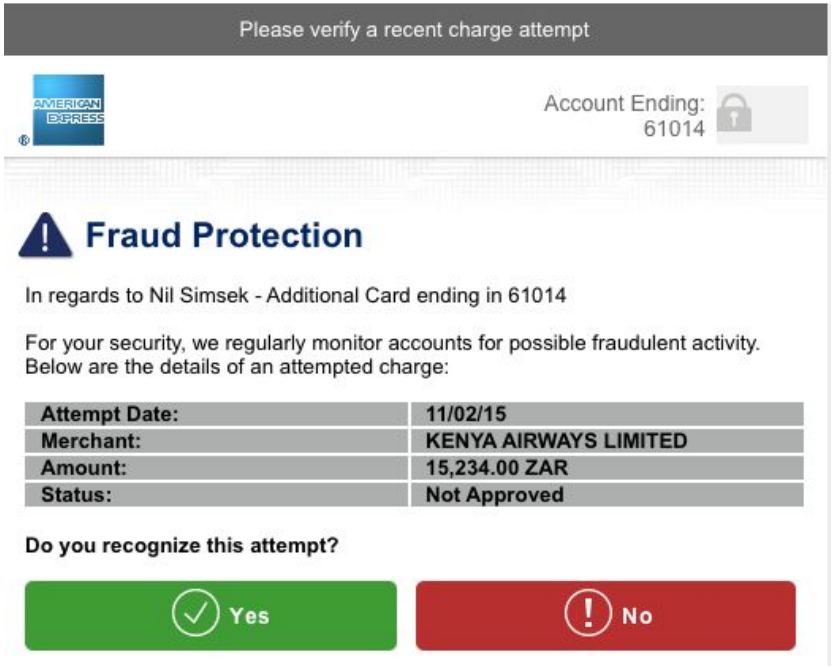
Business Problems & Analytics Solutions

Business Problems & Analytics Solutions

Profiling attempts to characterize the typical behavior of an individual, group, or population.

- e.g: What is the typical cell phone usage of this customer segment?
- What does “normal” behavior look like?

Business Problems & Analytics Solutions

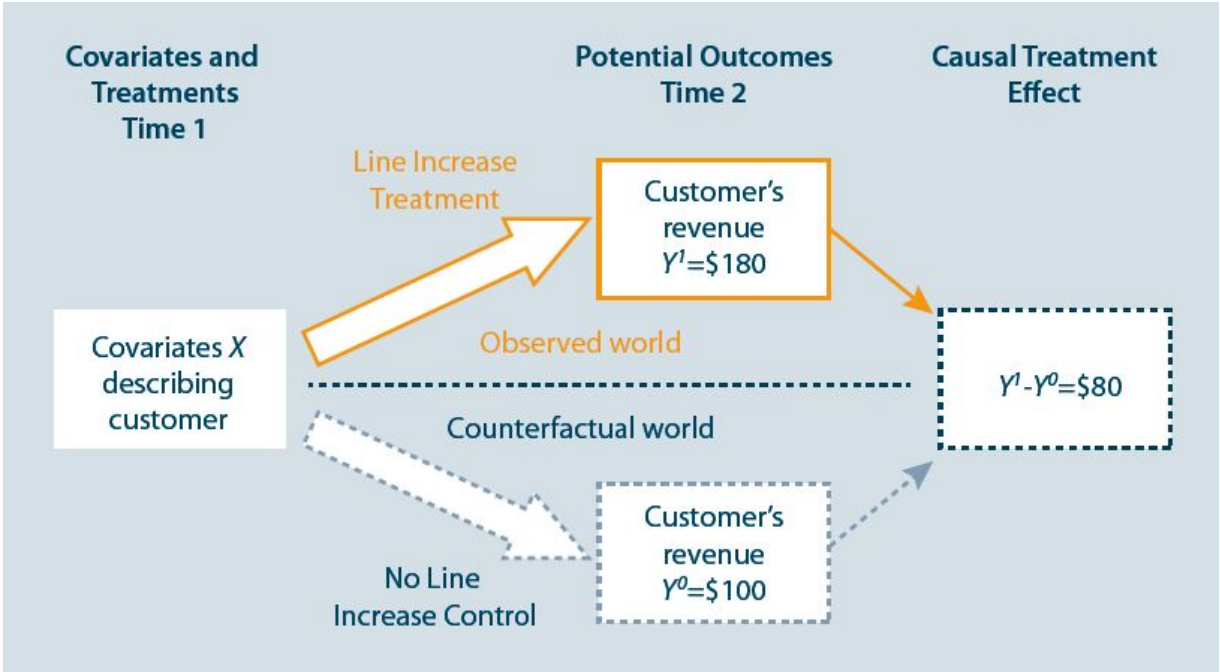


Business Problems & Analytics Solutions

Causal modeling attempts to identify causal relations between variables of interest, and infer the effects of actions on outcomes.

- The model must express more than correlation because correlation does not imply causation.
- e.g.:Does a treatment affect the outcome?

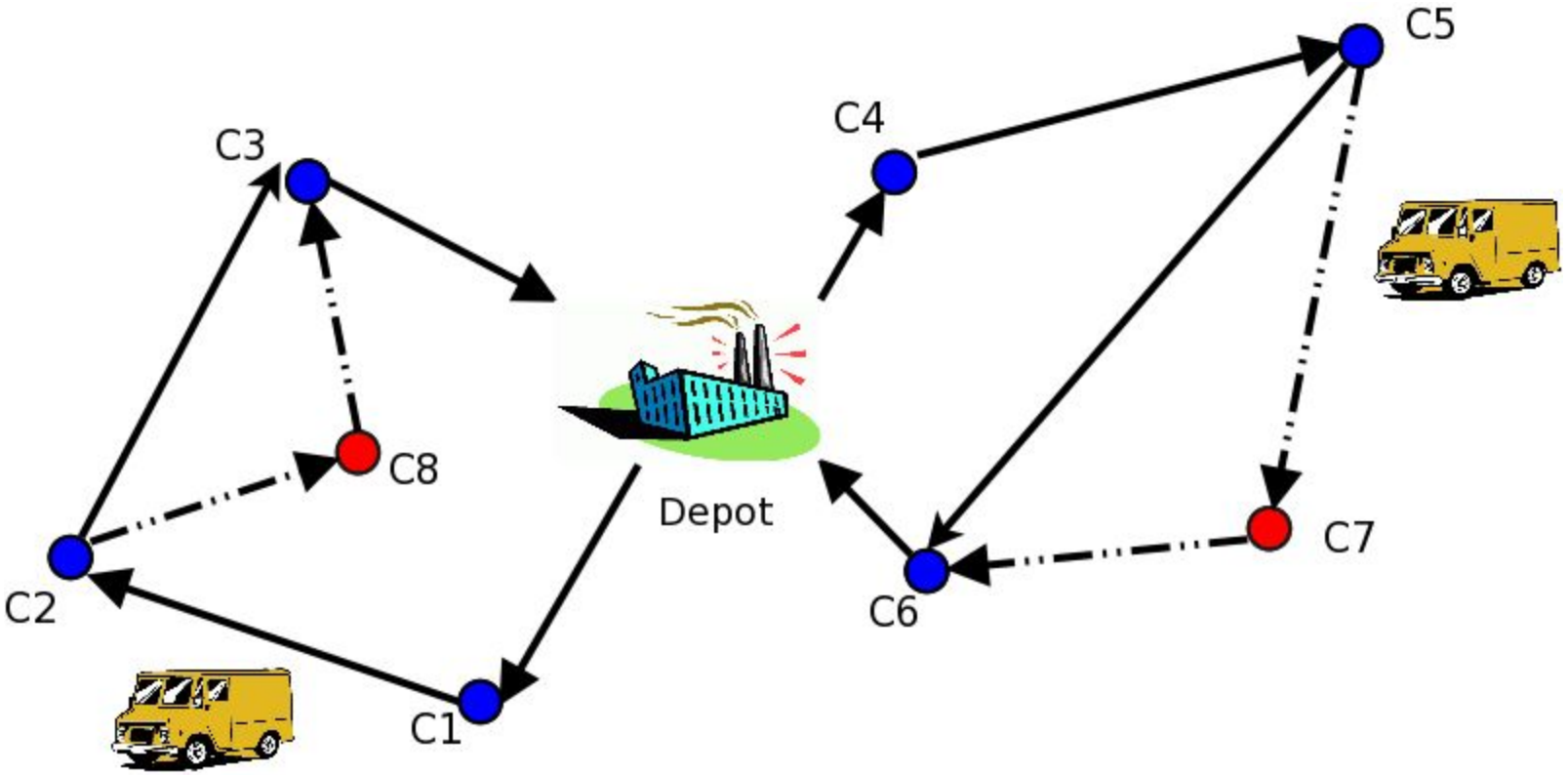
Business Problems & Analytics Solutions



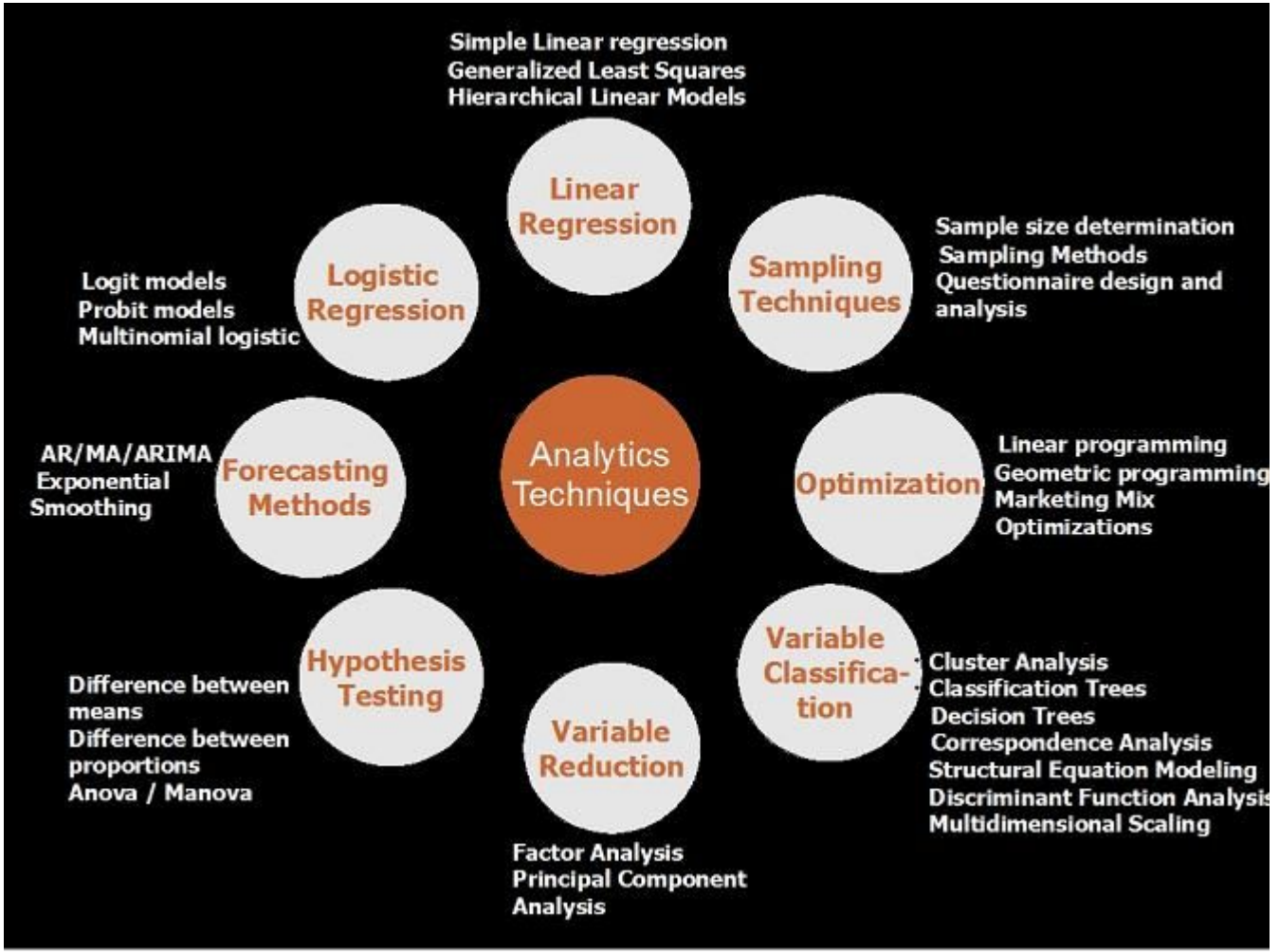
Business Problems & Analytics Solutions

Optimization attempts to find the optimal, most efficient way of using limited resources to achieve the objectives of a business.

- What is the least costly method of transferring merchandise from warehouses to stores?
- How should marketing budget be allocated among different media?


C5
C4
C3
C8
Depot
C7
C2
C6
C1
Static customer (known at t=0)
Dynamic customer (at t>0)
Initial route plan
New route segment

Methods


Simple Linear regression
Generalized Least Squares
Hierarchical Linear Models
Linear Regression
Logistic Regression
Logit models
Probit models
Multinomial logistic
Sampling Techniques
Sample size determination
Sampling Methods
Questionnaire design and analysis
Analytics Techniques
Forecasting Methods
AR/MA/ARIMA
Exponential Smoothing
Optimization
Linear programming
Geometric programming
Marketing Mix Optimizations
Hypothesis Testing
Difference between means
Difference between proportions
Anova / Manova
Variable Reduction
Factor Analysis
Principal Component Analysis
Variable Classifica-tion
Cluster Analysis
Classification Trees
Decision Trees
Correspondence Analysis
Structural Equation Modeling
Discriminant Function Analysis
Multidimensional Scaling

Data Mining vs...

- Data querying (SQL, QBE, other GUI-based querying)
  - Very flexible interface to ask factual questions about data
  - No modeling or sophisticated pattern finding
- Traditional statistical analysis
  - Mainly based on hypothesis testing or estimation/quantification of uncertainty
  - Should be used to follow-up on data mining's hypothesis generation
- Automated statistical modeling(e.g., advanced regression)
  - This is data mining, one type -- usually based on linear models
  - Massive databases allow non-linear alternatives

Examples

- Who (exactly) are my most profitable customers?
  - Database querying
- Do I have confidence that there really is a difference between the blue customers and the red customers?
  - Statistical hypothesis testing
- But who really are these customers? Can I characterize them?
  - Data mining (automated pattern finding)
- Can I predict whether a new customer will be profitable?
  - Data mining (predictive modeling)

Prep for 1st classes:

NYU Classes Tutorials and Prep

- Intro to R (Week 1)
- Descriptive Stats Review (Week 1)
- Adv. Descriptive Statistics & Visualization Theory (week 2)
- Linear Models (week 3)

Introduction to R & Descriptive Statistics

Lesson Objectives

1. Descriptive Statistics
    a. Measures of central tendency
    b. Measures of spread and variability
    c. Measures of association
    d. Frequency distributions

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

Descriptive & Summary Statistics

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way. Specifically it helps describe,

- Patterns that emerge from the data.
- Provide a way to describe our data
- Do not make conclusions or predictions

Descriptive statistics **do not** allow us to make conclusions or predictions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made.

Measures of Central Tendency

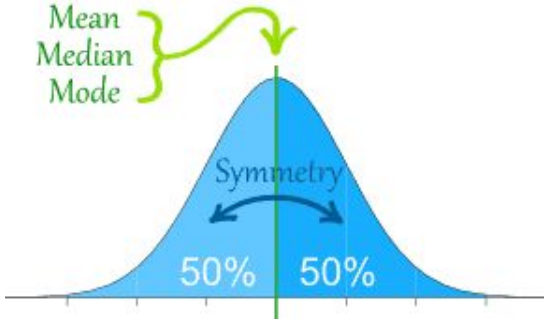
A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data → measures of central location.

1. Mean
    a. Arithmetic mean
    b. Geometric mean
    c. Harmonic mean
    d. Weighted mean
    e. Truncated mean
2. Median
3. Mode

Measures of Central Tendency

For any symmetrically shaped distribution the mean is the point around which the symmetry holds.

What is the single most important distribution pattern occurring in many natural phenomena?



$$Mean = \frac{sum\ of\ all\ values}{total\ number\ of\ values}$$

$$Median = middle\ value\ (when\ the\ data\ are\ arranged\ in\ order)$$

$$Mode = most\ common\ value$$

Note!: There is a difference in notation between the population mean and the sample mean. For population mean we use the Greek lower case letter "mu", denoted as μ, and for sample mean we use the x with bar on top, pronounced x bar

Example

Given this dataset below, find the mean (arithmetic) , media, and mode:

10, 20, 30, 40, 40

Mean = 140/5 =28

Media = 30

Mode = 40

Measures of Central Tendency

- The median is the middle value in the data set above (below) which there are 50% of the values.
  - Usually calculated by sorting the values and looking at the value in the middle position
  - Median household income in US families = $51,939
  - Mean household income = $72,641
- For skewed distributions or when there is concern about outliers, the median may be a better measure to use.
- The mode is the most frequently occurring value.
  - useful when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal)

Example: Zagat

What measures of central tendency would you use?

What is the business problem you are trying to solve?

What does it mean to be above or below the mean?

Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55

Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55
Mean	**23.8**	**17.6**	**20.7**	**53.0**
Median	**23.8**	**17.6**	**20.7**	**53.0**
Mode	**22.0**	**3.0**	**13.0**	**74.0**

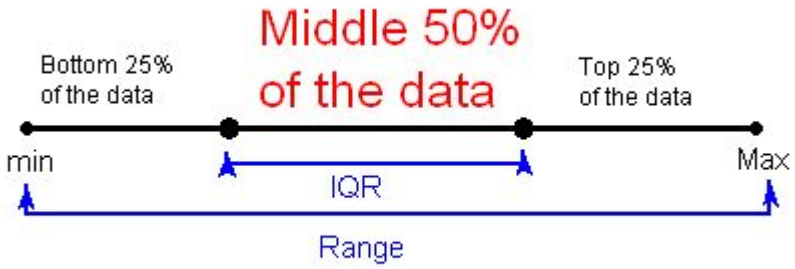
Measures of Spread or Variance

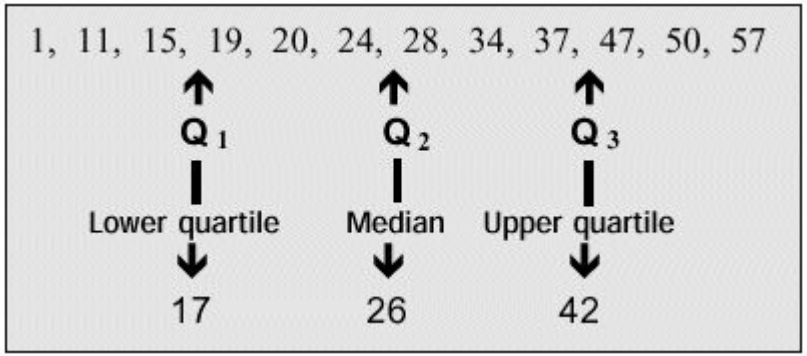
A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population. It is usually used in conjunction with a measure of central tendency, such as the mean or median, to provide an overall description of a set of data.

1. Range
2. Quartiles and Interquartile Range
3. Variance & Standard Deviation
4. Coefficient of Variation

Measures of Spread

- The quartiles of a population divide the observed data into even fourths.
  - The 1st quartile (Q1) is the number below which 25% of the values fall.
  - The 3rd quartile (Q3) is the number below which 75% of the values fall.
  - Interquartile Range (IQR) = Q3 - Q1 → the middle half of the data fall within IQR
  - Useful in detecting outliers





Measures of Spread

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

- A frequently used formal definition for an outlier is any value outside the interval:

  [Q1 − 1.5 x IQR, Q3 + 1.5 x IQR]

Measures of Spread

- The variance of a population is defined as the mean squared deviation.
- This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets.
- The standard deviation is simply the square root of the variance.

Variance → Standard Deviation

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \quad \rightarrow \quad \sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Where

σ^2 = the variance

σ = the standard deviation

$\sum$ = the sum of

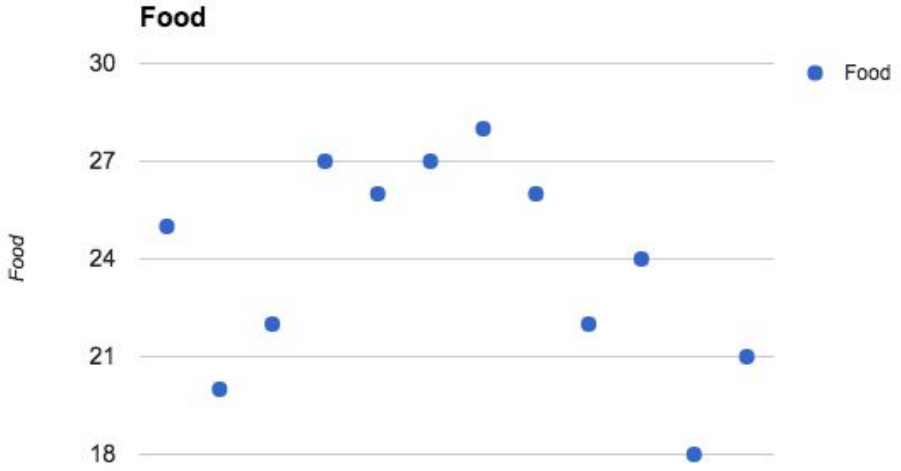
X= a datapoint

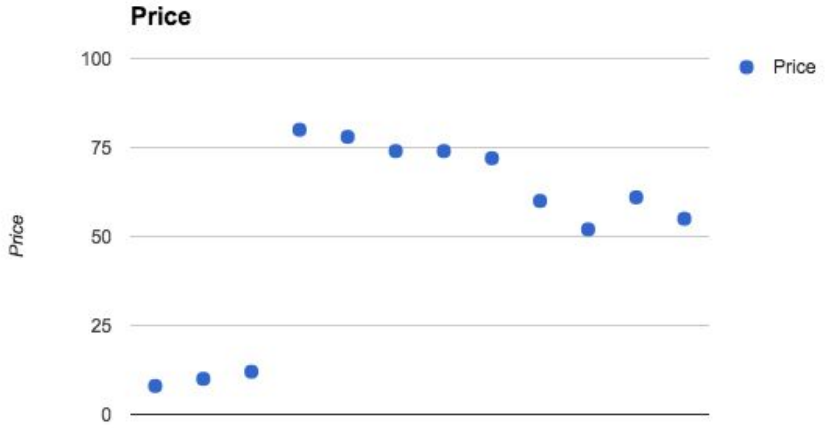
μ= the mean of the data

N = the number of datapoints in the set

Which rating has more variance?

Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55
Mean	**23.8**	**17.6**	**20.7**	**53.0**
Median	**23.8**	**17.6**	**20.7**	**53.0**
Mode	**22.0**	**3.0**	**13.0**	**74.0**


Food
30
27
24
21
18
Food
Food

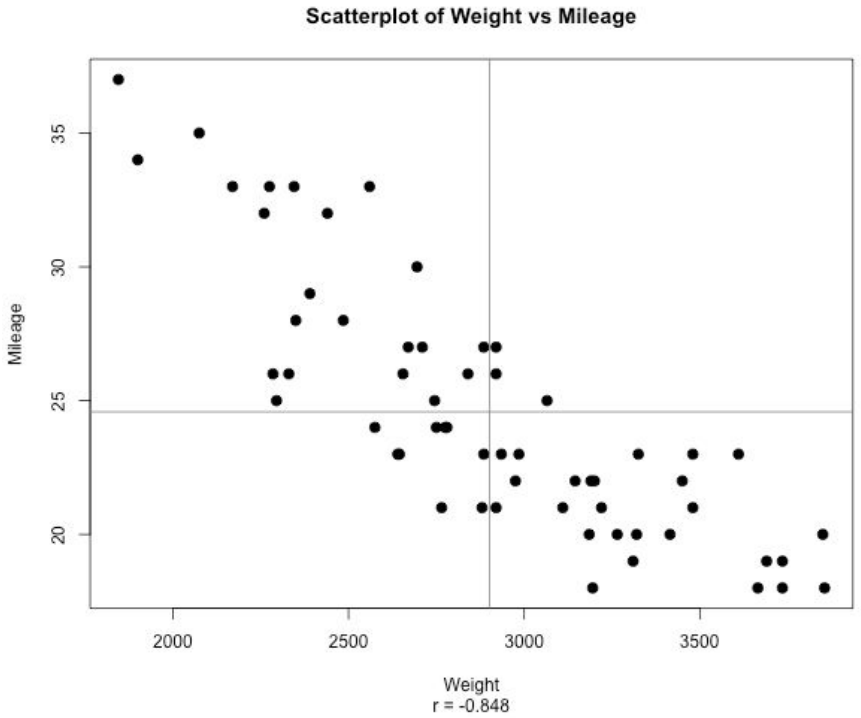

Price
100
75
50
25
0
Price
Price

Name	Food	Decor	Service	Price
Magnolia Bakery	25	10	13	8
Vinnie's pizza	20	3	13	10
big wong	22	3	11	12
Veritas	27	22	26	80
Four Season	26	27	26	78
Il mulino	27	18	24	74
Nobu	28	23	24	74
Union Pacific	26	26	25	72
Morton's of Chicago	22	19	22	60
Nanni	24	13	22	52
Oak room	18	24	21	61
Osteria del circo	21	23	21	55
Mean	**23.8**	**17.6**	**20.7**	**53.0**
Median	**23.8**	**17.6**	**20.7**	**53.0**
Mode	**22.0**	**3.0**	**13.0**	**74.0**
Variance	10.2	71.4	28.4	751.8
Standard Deviation	3.2	8.4	5.3	27.4

Measures of Association

The measures of association refer to a wide variety of coefficients that measure the statistical strength of the relationship on the variables of interest.

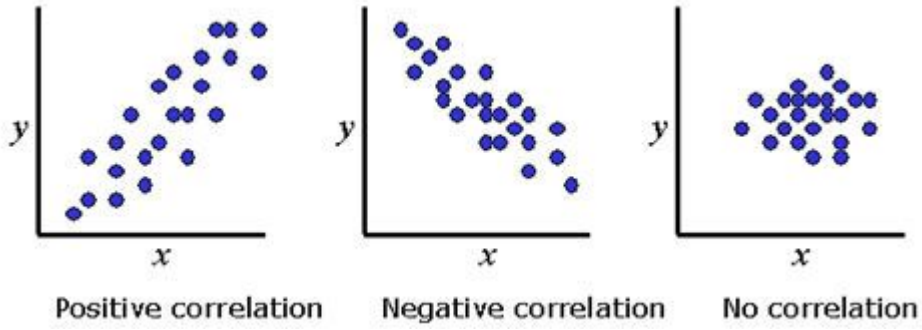
The term "association" is closely related to the term "correlation" → two or more variables vary according to some pattern.



Correlation

Correlation refers to any of a broad class of statistical relationships involving dependence.

Sometimes the pattern of association is a simple linear relationship → "the correlation coefficient"



Correlation Examples

Which of this are correlations in business

- Price & Revenue
- Salary & Education
- Gold & Oil

Classify as negative or positive correlation

- A student who has many absences has a decrease in grades
- As one increases in age, often one's agility decreases.
- When workers get a raise, morale improves.

~ http://www.tylervigen.com/spurious-correlations

Outliners and Z-Scores

$$z = \frac{x - \mu}{\sigma}$$

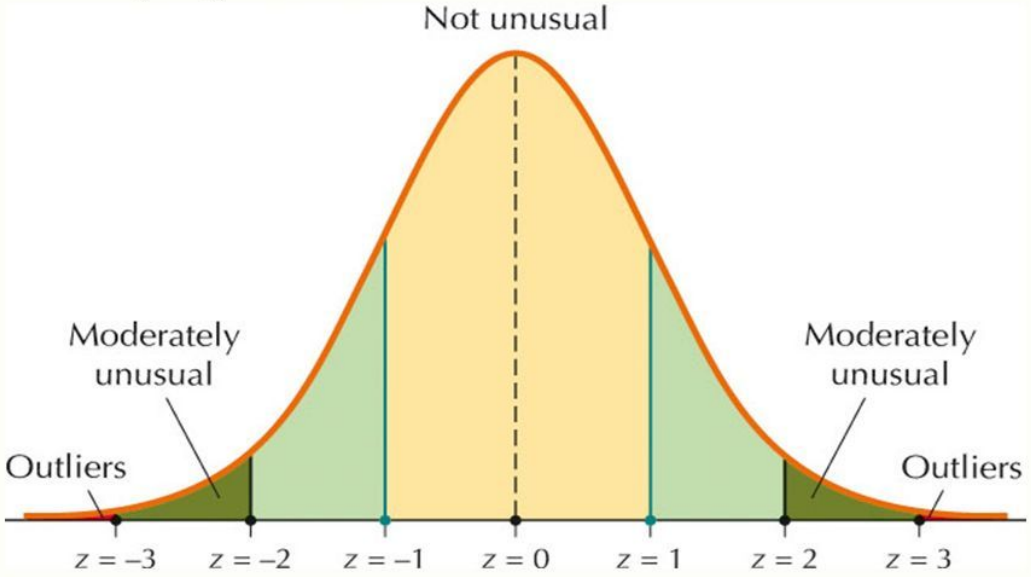
μ = Mean

σ = Standard Deviation

Detecting Outliers with *z*-Scores

28

An **outlier** is an extremely large or extremely small data value relative to the rest of the data set. It may represent a data entry error, or it may be genuine data.



- Z-score = 0 (observation is equal to the mean)
- Z-score = x (observation is x standard deviations away from mean)
- Z-scores between -3 and +3 (detects 99% of your information in dataset)
- Z-scores between -2 and +2 (detects 95% of your information in dataset)
- Z-scores between -1 and +1 (detects 68% of your information in dataset)

Introduction to R

R is a language and environment for statistical computing and graphics. It was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

You will need to install **RStudio**. RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

R Prep:

lynda.com

Code School
a Pluralsight company

In class quiz (7min)

NYU Classes>>Test & Quizzes

Understanding Your Data

Exploratory analysis of data is useful for:

- understanding data properties
- detecting errors, ensuring data quality
- finding patterns in data
- determining relationships among variables
- checking assumptions
- mapping business problems into data mining tasks and suggesting modeling strategies

Homework

Citibike Descriptive Analytics

Analytics Questions:

- Compute summary statistics for tripduration
- Compute summary statistics for age
- Compute summary statistics for tripduration in minutes (Need to transform tripduration from seconds to minutes)
- Compute the correlation between age and tripduration
- Plot the histograms and box plots for tripduration by gender

Business Questions:

- What is the total revenue assuming all users riding bikes from 0 to 45 minutes pay $3 per ride and user exceeding 45 minutes pay an additional $2 per ride.
- Looking at tripduration in minutes, what can you say about the variance in the data.
    - What does this mean for the pricing strategy?
    - What does this mean for inventory availability?
- A business manager wants to reallocate the $5M marketing budget using a gender segmentation strategy. Specifically, the manager is asking you to create two models:
    - A model that use % of male vs females in the dataset
    - A model based on average trip duration by gender