



IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

March 8,9 2016

Week 9 Summary Presentation

Week 1 Topic:

What do we mean by Data Analytics?

For this course, we discuss analysis of data in 3 overlapping disciplines; data analytics, data mining, and business intelligence. You will see in the market that most vendor products choose a discipline to focus their marketing efforts. However, dependent on the solution offering, it can have all or one functional offering.

Theoretically, it's all analysis of data to gather information but using these terms adds *purpose* to the activity. We loosely define the 3 disciplines further but as we progress, we will speak more about *why we use certain techniques for what purpose* rather than noodle over what discipline it belongs to.

- ◆ Data Analytics: Further than just visualization of data, the goal of analytics is to support a decision or prove a hypothesis by using quantitative techniques. This confirmatory approach will validate or summarize information to provide the answer.
- ◆ Data Mining: Describes examining data sets to identify undiscovered patterns and uncover hidden relationships. This exploratory approach may use sophisticated models to determine its patterns.
- ◆ Business Intelligence: Describes data analysis efforts that is focused on answering analytical questions about the business, supports business management processes, or provides views of the business whether it be enterprise or departmental.

Week 1 Topic:

More on Data Analytics

We describe more confirmatory, inference driven data analysis activities to data analytics:

- ◆ **Descriptive Statistics:** Use of descriptive statistics like means, medians, and standard deviations
- ◆ **Graph Distributions:** Using distributions of data to summarize groupings and seek out anomalies in data using visual means e.g., histograms, pie charts, bar charts, etc.
- ◆ **String Operations:** Manipulating, parsing, or translating text values to otherwise interpret information that may be coded or concatenated.
- ◆ **Math Functions:** Using counts, sums, percentages, and other math functions to validate or summarize data. Same techniques used in string operations can also be applied to numeric values.
- ◆ Use of filters and sorts to understand data.
- ◆ Logical operations on data e.g., applying finite ranges, $<$, $=$, $>$.
- ◆ Calculating derived values and applying conditions on data.
- ◆ In general, we answer questions or confirm hypothesis through quantitative means that is not tied to sophisticated models. We tie simple math and view manipulations to this activity.

Week 1 Topic:

More on Data Mining

We describe more model driven exploratory data analysis activities with data mining:

- ◆ **Outlier Analysis:** Identification of unusual data records, that might be interesting or data errors that require further investigation.
- ◆ **Correlations:** Association rule learning or dependency modelling – Searches for relationships between variables. This is sometimes referred to as market basket analysis.
- ◆ **Clustering (Segmentation/Summarization):** The act of grouping similar cases together. Discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- ◆ **Classification:** Predicting a discrete categorical value. The task of generalizing known structure to apply to defined categories.
 - ◆ **Forecasting/Regression:** Discovering patterns in data that can lead to reasonable predictions about the future. Attempts to find a function which models the data with the least error.
 - ◆ **Bayesian:** Use of conditional probabilities to infer an outcome.
 - ◆ **Others include use of Neural Networks and Decision Trees**

Week 1 Topic:

More on Business Intelligence

We tie Business Intelligence (BI) with a Data Warehouse (DW) solution. Hence, we focus on OLAP functions as BI analysis techniques. We define OLAP as the ability to join discrete sets of data into a dimensional cube structure that is optimized/aggregated for analysis/reporting. With an OLAP cube, you can:

- ◆ **Slice:** Act of taking a subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension.
- ◆ **Dice:** Creating a subset of a cube for analysis by choosing values from dimensions.
- ◆ **Drill Up/Down:** Allowing a user to navigate through the levels of a dimensional hierarchy up (summarized) and down (more detailed).
- ◆ **Roll Up:** Summarization of data along a dimension e.g., totals or derived.
- ◆ However, BI also encompasses reporting functions similar to the way we defined Data Analytics albeit focused on answering business questions and hypothesis.

Week 1 Topic:

Data Analysis Vendor Landscape

One can perform data analytics in spreadsheets as long as one can acquire the data in the right format which is why Excel is so popular amongst the business users. For more sophisticated work, however, we categorize the following top vendors:

Data Mining (*More Business Use*)

SAS Enterprise Miner

SPSS (IBM)

Stata

Tibco Spotfire

Dell StatSoft

Statistical (*More Academic Use*)

MatLab

Mathematica

Minitab

R

SAS JMP

Business Intelligence – OLAP Included

Cognos (IBM)

BusinessObjects (SAP)

MicroStrategy

Hyperion (Oracle)

Informatica

Business Intelligence - Visualization

Tableau

QlikView

Domo

Sisense

first

Week 1 Topic:

Roles in Data Analysis

Business Analyst

- Supports business users
- Serves as subject matter expert for the business domain
- Usually owns a business application, process and/or function
- Performs business analysis mostly in business application or Excel
- Some super users can model and code as necessary

Quantitative Analyst

- Supports business users and analysts
- Serves as subject matter expert for statistical solutions domain
- PhD types that will use mathematical and statistical programming applications to build and execute model based analysis
- Consumes high volumes of raw data for model data feeds

Reporting Analyst

- Supports business users and analysts by building and supporting reports, dashboards, or BI solutions
- Serves as subject matter expert for the reporting domain
- Usually owns the reporting business process or function and in some cases data steward to the data within reports
- Performs data analysis and validations for data processing and management

Data Analyst

- Supports business users and analysts for ad hoc queries or other data related analysis projects
- Serves as subject matter expert for the data domain and typically serves as data stewards
- Performs data analysis, builds reports as necessary, and supports data processing and management tasks

Week 1 Topic:

Database Vendor Landscape

| Database Software | Language Support |
|---|---|
| RDBMS: <ul style="list-style-type: none"> IBM (Mainframe, DB2, IMS, Cloudant, Informix) Oracle (Latest 12c) Microsoft SQL Server (Latest 2014) SAP (ASE, IQ-columnar) Teradata (columnar) | <ul style="list-style-type: none"> Query: SQL, PL/SQL (Oracle) API exists for various application build languages Some recently extended to support JSON, XML |
| Open Source: PostgreSQL, MySQL, MariaDB Enterprise, Firebird | <ul style="list-style-type: none"> Query: SQL and other scripting languages |
| NoSQL (non-relational database systems with no pre-defined structure): Cassandra, MongoDB, Dynamo | <ul style="list-style-type: none"> Query: Cassandra uses CQL and others use various scripting languages Open distributed file systems. Cassandra resembles RDBMS table structure while MongoDB uses JSON like file structures |
| In-Memory: KDB, EXtremeDB, MemSQL, SAP HANA | <ul style="list-style-type: none"> Query: KDB uses Q and others use direct queries in C/C++, HANA supports Javascripts |
| Hadoop (a distributed file system) – <i>More on this when we cover Big Data</i> | |
| Specialty Appliances: <i>Netezza, XtremeData, Greenplum</i> are optimized for analysis using multi processors, lot of memory, faster networks, large disk space, etc. | |

Week 1 Topic:

Data Management Components

The following is a view into data management concepts and components for an enterprise. Dependent on the role of the data analyst, one can work in any one of the following areas:

| Data Governance | | | |
|-----------------------|-------------------------|---------------------------|-----------------------------|
| Organizational Model | Enablement | Standards & Policies | Processes & Procedures |
| Data Quality | | | |
| Profiling / Analysis | Cleansing | Controls | Enrichment / Enhancement |
| Data Usage | | | |
| Reporting | Analytics (OLAP) | | Data Mining |
| Quantitative Analysis | Scorecard / Dashboards | | Alerts/ Notifications |
| Data Management | | | |
| System of records | Operational data stores | Data warehouse/data marts | Data movement (ETL/EAI/EII) |
| Data protection | Metadata management | Reference data management | Master data management |
| Architecture | | | |
| Conceptual | Logical | Physical / technical | |
| Design patterns | Services | Standards | |

Week 1 Topic:

Data Types – Transaction vs Snapshot

It's important to understand data requirements for analysis as, most likely, you will be defining it for the developers and act as a conduit for the business users' needs. Translating the needs into requirements is not an easy task and in some cases require most time and resources during an implementation. We'll cover data requirements gathering methodologies in the next class. For today, we define what we mean by transaction versus snapshot data:

Transaction Data:

- ◆ Records business events e.g., retail purchases, call detail records, bank deposits/withdrawals, insurance claims, stock trades/quotes, etc.
- ◆ Usually recorded along with date and timestamp for each transaction.
- ◆ Considered raw data or detailed view of data as analysis may be done at a point in time versus looking at each transaction.

Snapshot Data:

- ◆ Records current or past 'state' of a business entity or relationship e.g., customer, account or measures of metric values at a certain point in time.
- ◆ Unlike transaction data, a query will typically want to access only one "time instance" of the snapshot data e.g., balance on the account for month end close.
- ◆ Multiple snapshots can be used for trending or constructing averages over time.

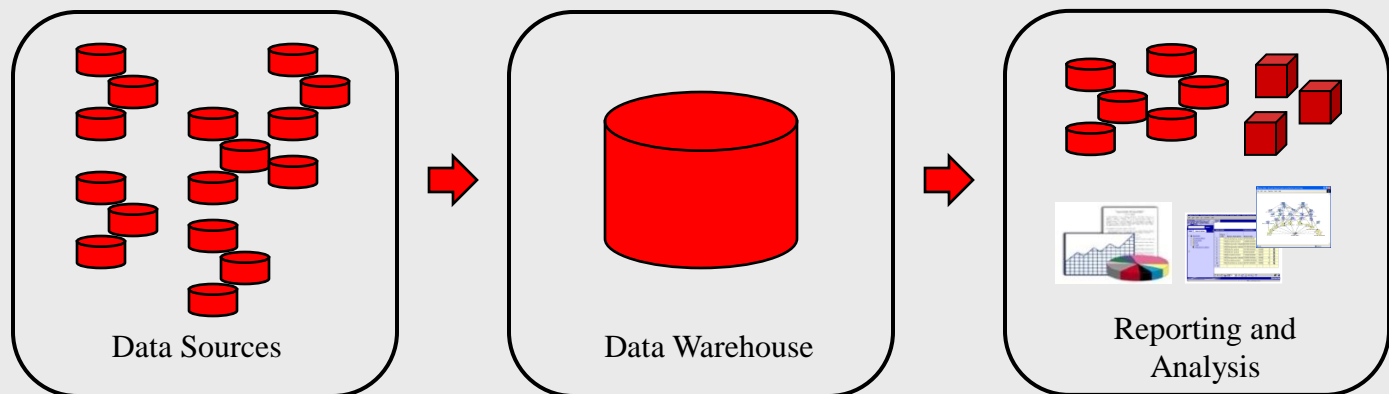
Week 1 Topic:

Data Warehouse: Definition

Most of all, you need data to do data analysis. Most business requirements will require data from multiple data sources with differing data types and varying degrees of data quality. This is where a data warehouse comes in.

Simply put, *a data warehouse is a data repository that collects data from multiple sources into a uniform structure.* Architecture of a data warehouse and its use varies when you start to consider what that uniform structure looks like.

There were two different philosophies on implementing data warehouses in the beginning. Here, we are talking back in the 90s, many years after the concept was first discussed in the 60s.



Week 1 Topic:

Data Warehouse: Inmon versus Kimball

When you start to think about building a data warehouse, one of the first things to consider is how you will model this uniform structure. Of course, this is after you have gathered sufficient business requirements and identified applicable data sources to understand the purpose of the data warehouse in the first place. When we say purpose here, we also mean analytics and its data needs as well.

The two philosophies were:

1. Bill Inmon claimed that this uniform structure is in a 3NF*. Analysis is done off of dimensionally structured data marts** that is produced from this 3NF data warehouse. More of a “top down” approach.
2. Ralph Kimball claimed that this uniform structure is a conglomeration of departmental data marts that may share data through an information bus. Hence, a data warehouse itself may also have a dimensional structure. More of a “bottom up” approach.

*3NF was originally defined by E.F. Codd in 1971. This class assumes that you know basics of data modeling. You will need some data modeling skills to prepare your data sets.

**We will discuss data marts and dimensional data modeling in subsequent slides.

Week 1 Topic:

Data Warehouse: Adoption and Evolution

There was more adoption of Kimball's method as it was considered a "lighter" more practical approach. Investment was easier to justify. This also meant creation of departmental data silos which is another topic all together. He wrote *The Data Warehouse Toolkit* in 1996 which was considered a must read for anyone building a data warehouse at the time.

Following Inmon's approach was a bigger investment with many months spent on design which led to lower adoption by business as it was harder to see benefit in a timely manner. Although, theoretically, it made a lot of sense. He published *Building the Data Warehouse* in 1992. 20+ years after first coining the term.

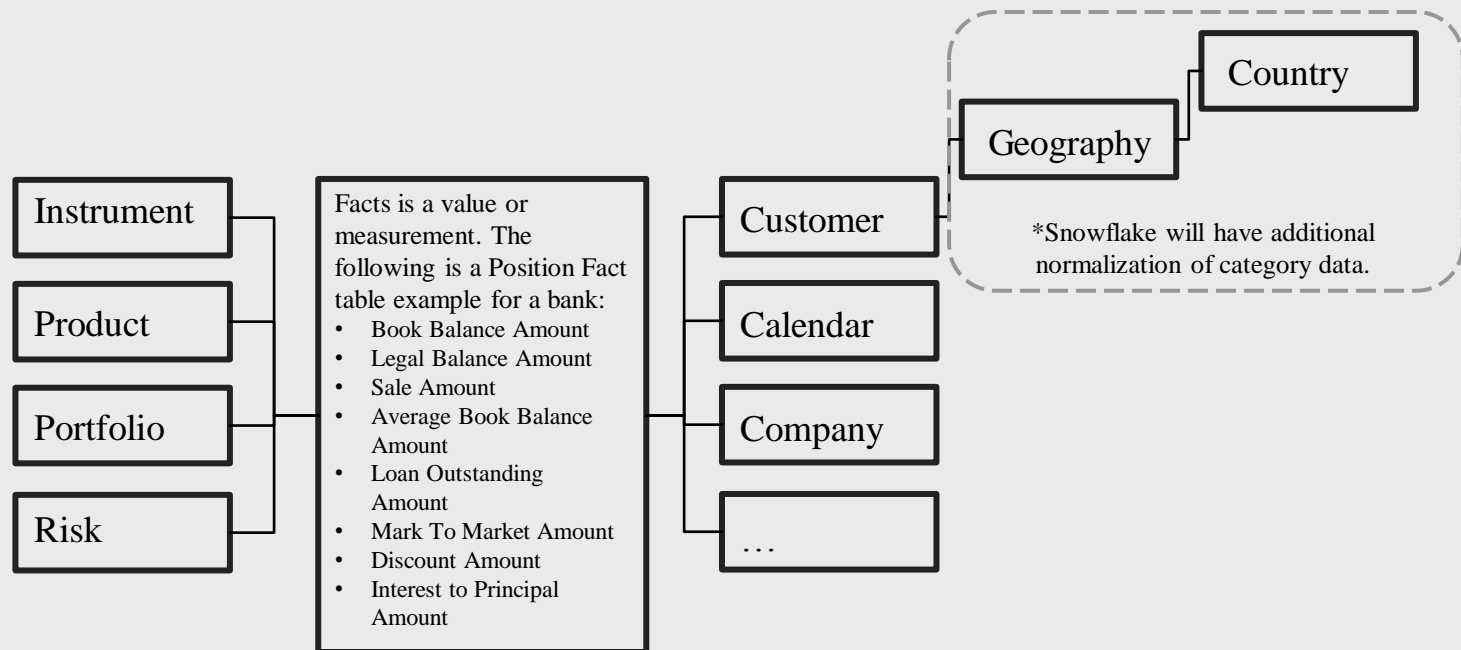
In general, evolution of data warehouses, data analytics, and data management follow technology advancement. Punch cards (used up to about the 70s even though, I still used them in graduate school in the mid 90s) were replaced by magnetic tape (introduced in the 50s) for data storage. Magnetic tapes were replaced with hard disk drives etc. etc.

Today, you can load multiple Terabytes of data into memory for analysis.

Week 1 Topic:

Data Mart: Definition

A data mart is usually dimensional with facts (measures) and dimensions (categories) hence the term, dimensional modeling. We call this data model a star (and/or snowflake*):



A fact table is not normalized but rather designed to respond to queries fast. Dimensional data is normalized and represents a unique descriptive to measures.

Week 1 Topic:

Data Solutions Comparisons

- ◆ Data Warehouse
 - Source of consistent, integrated, enterprise-wide data
 - Mechanism providing the analytical and decision support needs of the enterprise
 - Data is at multiple levels of granularity, including transaction-level and summarization
 - Data is typically retained for 3-7 years
 - Data may be sourced from the Operational environment and/or the ODS
- ◆ Data Mart
 - Mechanism providing the analytical and decision support needs of a business function
 - Data is highly summarized and is usually specific to the business function
 - Data is typically retained for 3-7 years
 - Data may be sourced from the ODS and/or the Data Warehouse
- ◆ Operational Data Store (ODS)
 - Mechanism enabling the collection, cleansing, and integration of operational data for population in either a Data Warehouse or a Data Mart
 - Data is typically at the transactional level of detail
 - Data may be retained for short time spans (90-120 days)

Week 1 Topic:

Data Warehouse: Now and Then

Data Warehouse – *the old days:*

- ◆ Multi-million \$\$, multi-year investment
- ◆ Batch-oriented
- ◆ Limited availability of data – high cost, limited processing power (software and hardware)

Data Warehouse – *new generation:*

- ◆ Speed of data has gotten faster e.g., streaming is not a wish, it's a must
- ◆ Amount of data being generated have increased e.g., in the petabyte range
- ◆ Type of data being processed is more diverse e.g., textural as well as tabular
- ◆ Reclaim of archive data (“dark data”) – more availability
- ◆ More diverse delivery points e.g., handheld devices
- ◆ Cost of technology has plummeted

We will discuss influences of these trends on data management when we discuss Big Data towards the end of the semester.

Week 1 Topic:

People Management

Qualitative items to consider when embarking on a new project is about People & Adoption in the data space. These measures will have an impact on what a developer does and chooses hence should not be ignored especially at initiation phase:

- ◆ Stakeholder Readiness - *understanding opinions*
 - Is the stakeholder sold on the goals, objectives, and value of the project?
 - Does the stakeholder understand or is willing to understand what's involved in implementing the analysis?
- ◆ Organizational Readiness – *understanding situational challenges*
 - Are the needed resources available in the organization? Is the organization receptive to external resources, if not?
 - How long does it take for the organization to adopt new technologies?
- ◆ Financial Readiness
 - How much and how long will the project cost?
 - What are the financial constraints for the project?
- ◆ Data & Technology Readiness
 - What technologies and methods does the organization use currently?
 - Is the data needed for the analysis available? How easily can it be obtained?
 - Is the hardware and software needed for the analysis available?

Week 1 Topic:

Understanding data - retail

Data:

- ◆ (F) Transaction data: purchasing/orders, accounts payable, POS, sales projections, warehouse movements, employee shift records, returns
- ◆ (D) Product data: consumer merchandise, hardware, software, industrial raw materials, any tangible object or service that can be sold or bought along with SKU, EPC, etc.
- ◆ (D) Customer data: name, address, email, phone, demographic, behavioral, financial
- ◆ (D) Store/Branch data: location type, address, manager, size/resources
- ◆ (D) Others include sales reference data (e.g., account type, personnel) and supplier information

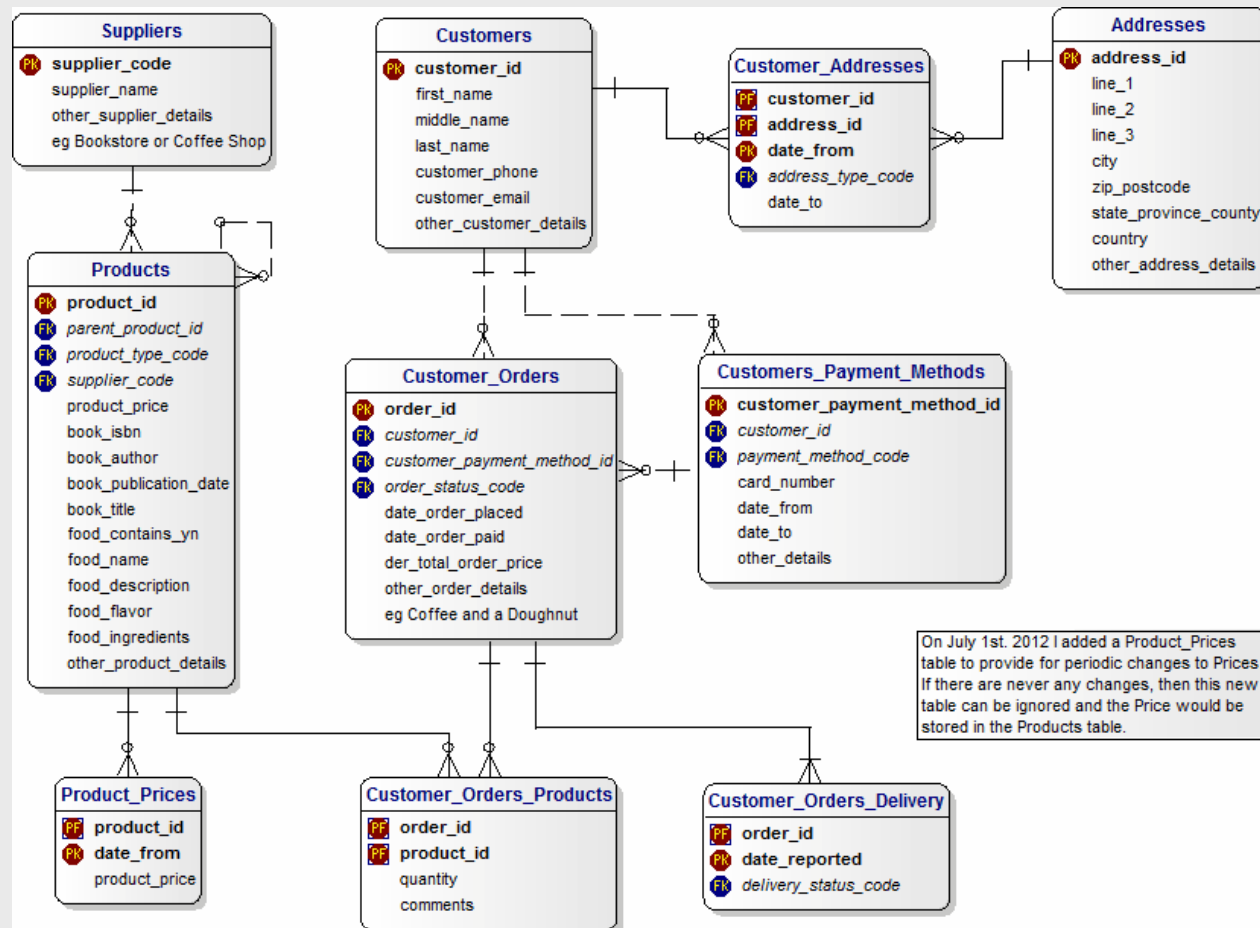
Example Analytics:

- ◆ Perform operational analytics to identify economies of scale, inventory management, cash flow analysis, optimal open hours
- ◆ Identify patterns, trends, and anomalies in transactions to mitigate risk and report fraudulent activities e.g., receipt fraud (falsified, stolen or reused receipts are used to return merchandise), price arbitrage (using higher priced product tags to return lower)
- ◆ Cross sell/Up sell using modeling techniques like market basket analysis or by simple product association e.g., diapers and diaper genie, movies with same actor, etc.
- ◆ Launch market campaigns by segmenting like customer groups together according to set criteria e.g., demographic, geographic, income, etc.

Week 1 Topic:

Sample retail data model

http://www.databaseanswers.org/data_models/customers_and_orders/:



Week 1 Topic:

Understanding data – banking/trading

Data:

- ◆ (F) Transaction data: trades (price, size), quotes (bid price, ask price, bid size, ask size)
- ◆ (F) Position (financial) data: amount of securities or commodities held e.g., balances of accounts or portfolios
- ◆ (D) Product data (banking): savings, checking, mortgage, credit card, account
- ◆ (D) Instrument data (trading) : tradable assets of any kind e.g., securities, cash
- ◆ (D) Market data: curves, rates, prices, spreads
- ◆ (D) Customer/Obligor/Party data: in addition to the usual CUSIP/SIC/NAIC codes for businesses, SS#, Risk Rating, Obligor (bond issuer, borrower, debtor, contractually/legally obligated entity) Rating

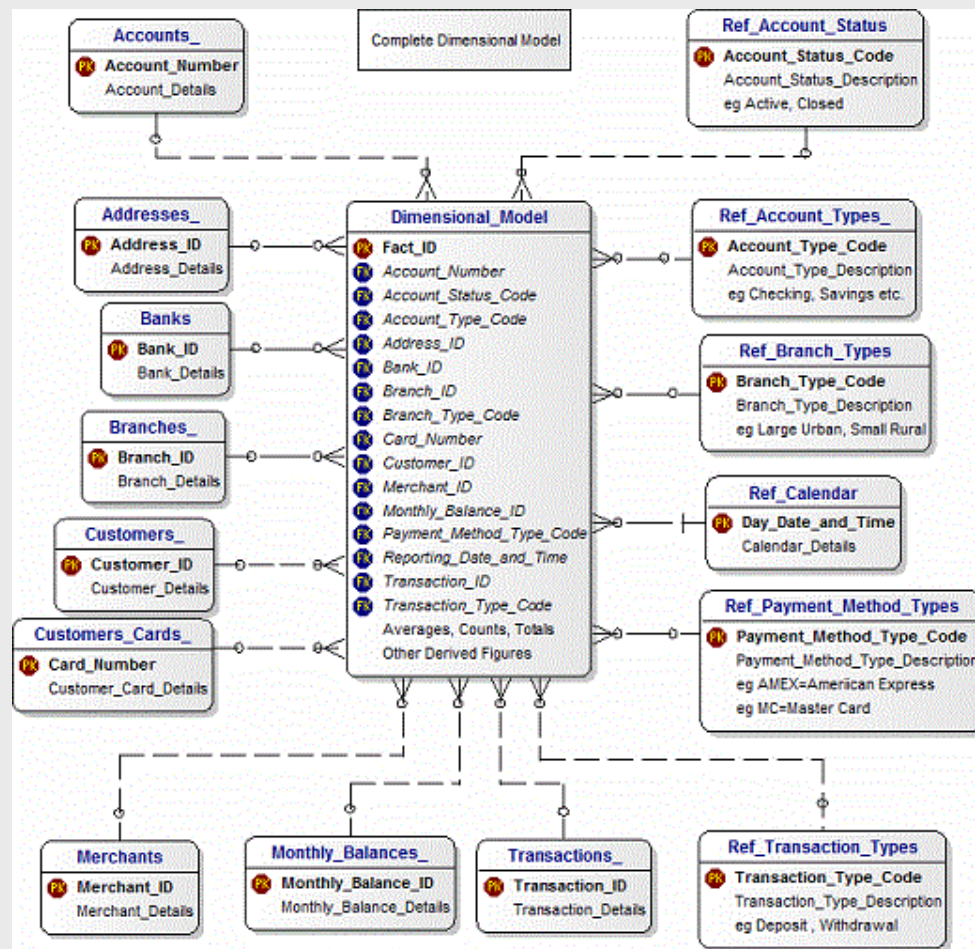
Example Analytics:

- ◆ Banking: Compliance with regulatory measures e.g., AML, KYC, Volcker Rule, Basel, etc.
- ◆ Trading: Generate best execution outlier reports to identify trades that missed best price using transactions
- ◆ Trading: Calculate NBBO (National Best Bid and Offer - this is a regulation that requires brokers to execute customer trades at the best available ask price when buying securities, and the best available bid price when selling securities) matching trades to quotes for a given day

Week 1 Topic:

Sample banking data model

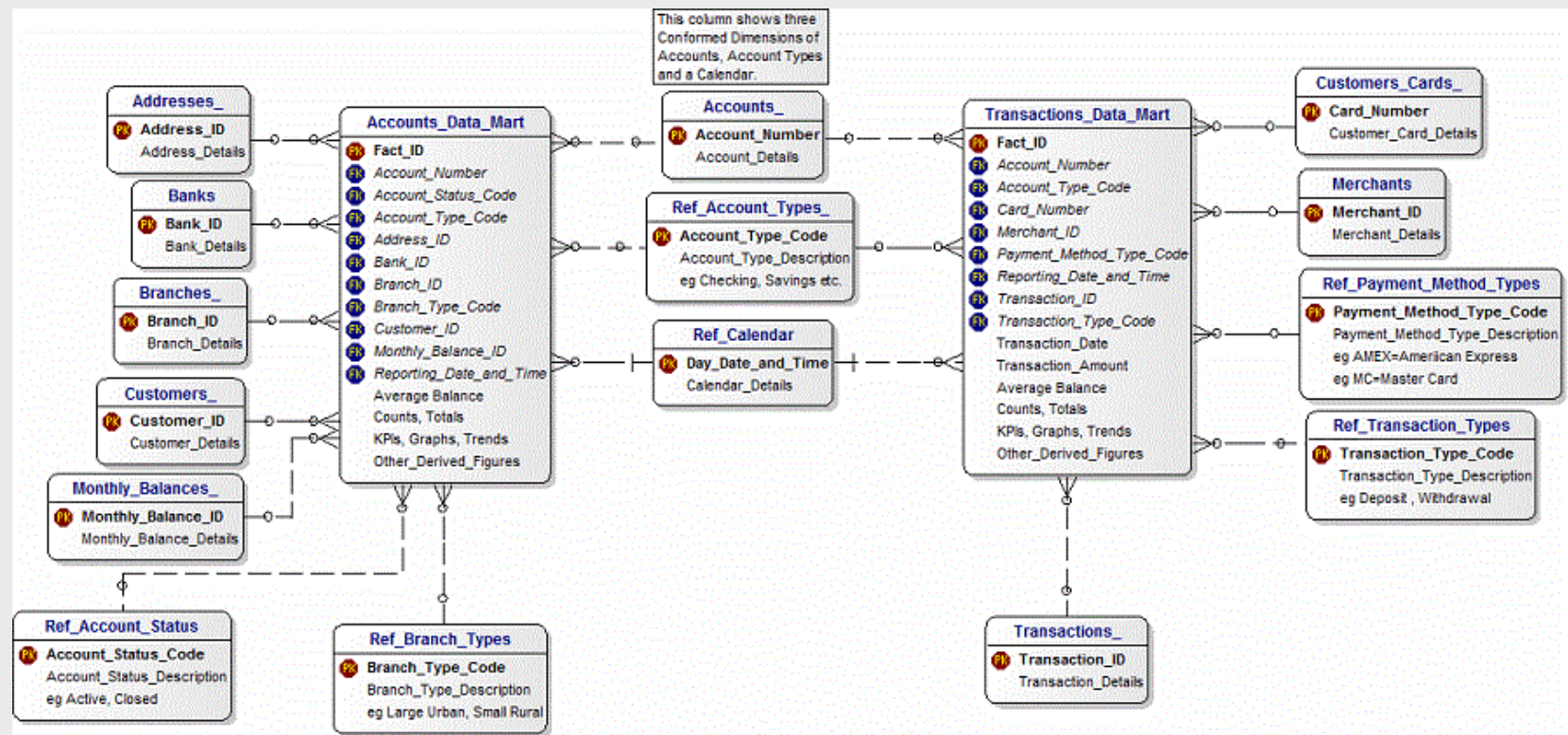
http://www.databaseanswers.org/data_models/retail_banks/:



Week 1 Topic:

Sample banking data model (cont.)

Another example with more facts and dimensions from
http://www.databaseanswers.org/data_models/retail_banks/:



Week 1 Topic:

Data Analysis Methodology

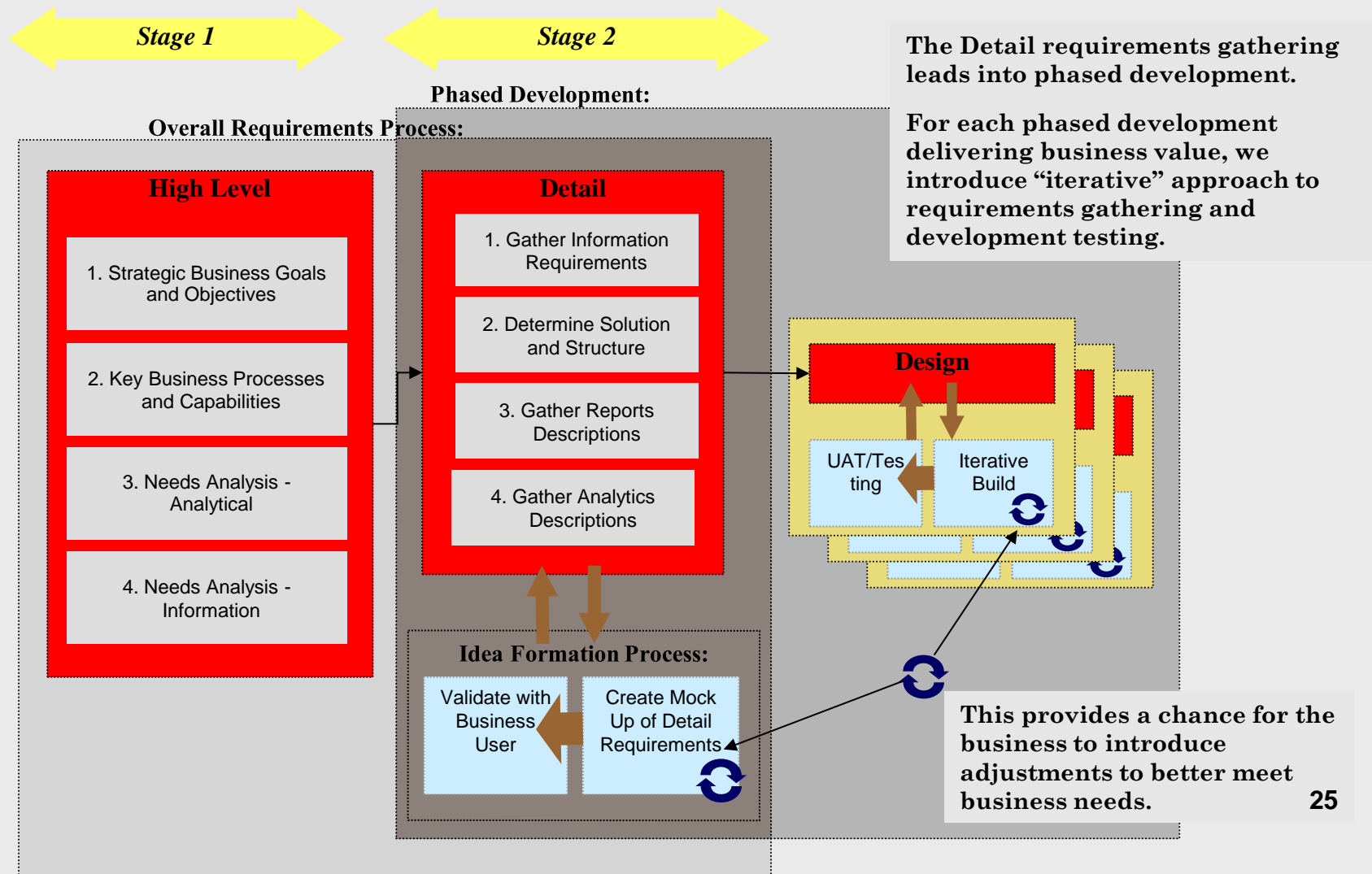
- A. Inspect - All data is inspected and “cleansed”
 - Records are investigated and fixed where appropriate e.g., outliers, type check, range check
 - Consistent default values are assigned to “missing” data
 - Data validation codes are included in the data where appropriate (e.g. invalid zip code) to avoid/compensate/exclude during analysis
- B. Transform - Data is standardized, improved or derived e.g., profitability, scores
 - Promotes consistent analysis using common business rules
 - Reduces analysis “programming” since necessary information is produced prior e.g. calculating months on books, banding score values, computing household product counts
 - Increases understanding of the data
- C. Integrate - All of the required data is in one logical structure e.g., transaction, position, account, customer, household, product, instrument, branch
 - Simplifies data access because all data is located in one location
 - Reduces analysis time since all of the information can be retrieved from a single location through a single query

Week 1 Topic:

Data Analysis Methodology (cont.)

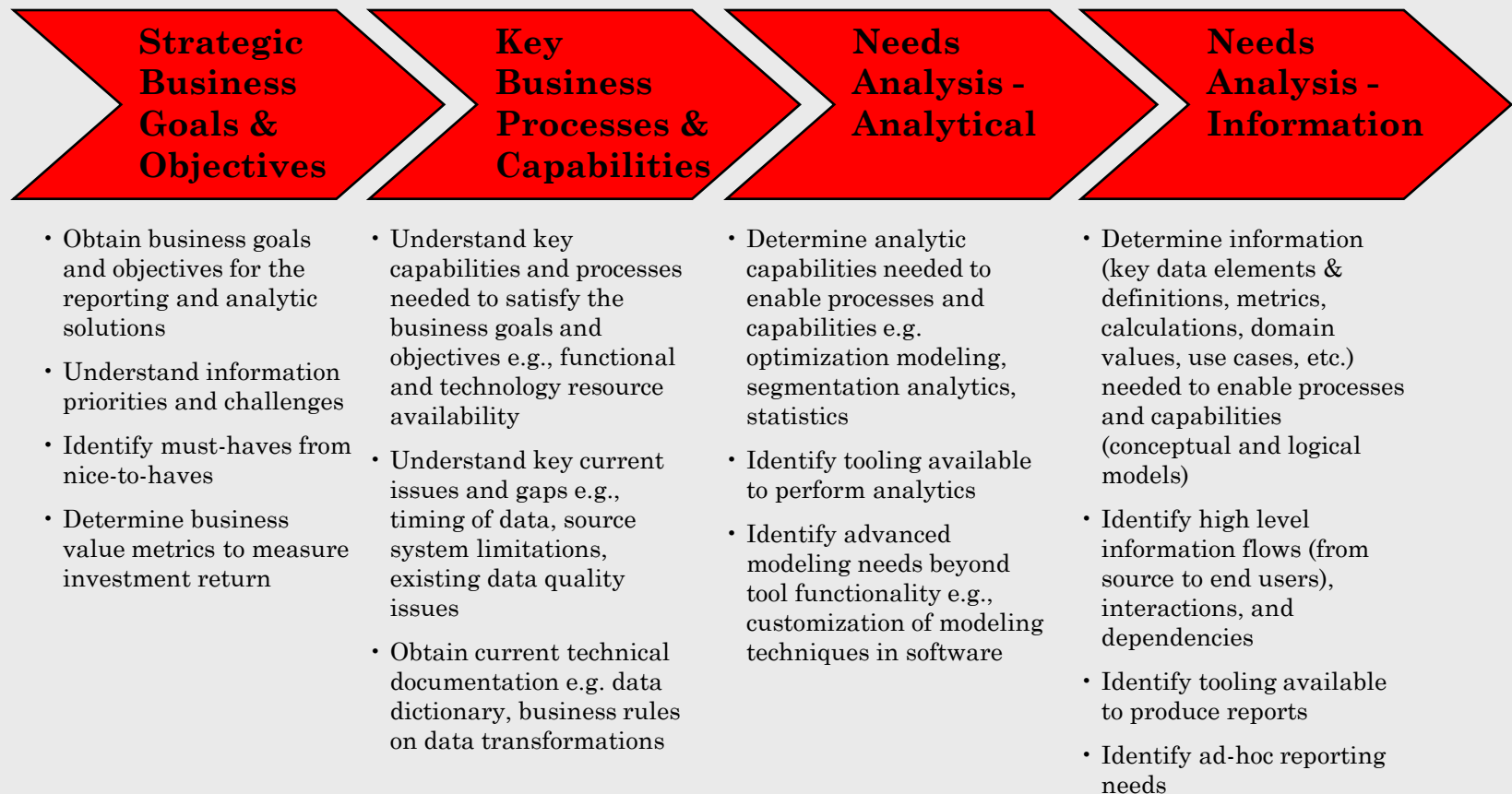
- D. Organize BI - Data is stored dimensionally
 - Simplifies analysis by providing an intuitive, business oriented data design
 - Enables pre-stored aggregations (cubes w/drill through to detail) to be easily developed and managed
- E. Organize A/DM – Data is stored in modeling specific data structure
 - This could be one de-normalized fact table with select dimensional information included in each row of data
 - Since data mining can only uncover patterns already present in the data, the sample should be large enough to contain significant information, yet small enough to process (dependent on resource capability)
- F. Explore - Search for anticipated relationships, unanticipated trends and anomalies:
 - Clustering discovers groups or structures in the data that are similar, beyond the structures known in the data
 - Classification generalizes a known structure to apply to new data, such as classifying a customer as a good or poor credit risk
- G. Document - Data definitions and transformation rules are documented and accessible
 - Able to understand the data and information gathered from the data

Week 1 Topic: Requirements Gathering Process



Week 1 Topic:

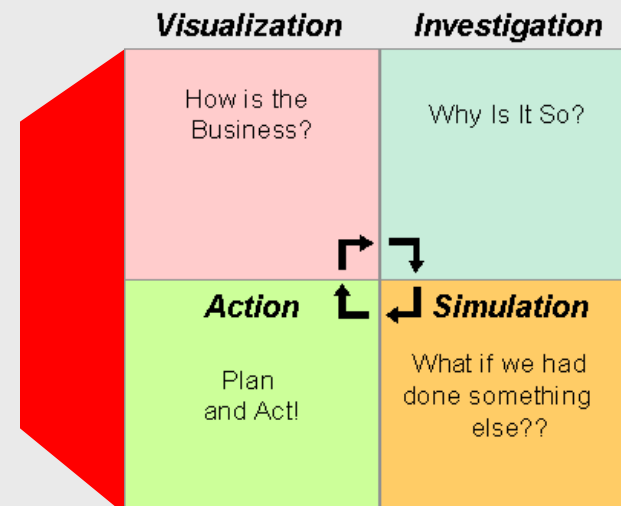
High Level Requirements Gathering



Week 1 Topic:

Understanding that it's iterative

An analytical environment supports not just reporting, but the full range of information usage listed below:



a) Basic Reporting

Periodic reports with the ability to change the report parameters by the users e.g., time period of reporting, metrics reported

b) Ad-hoc Analysis

Ability to access the data in free-form, create new aggregations and report definitions.

c) Custom Applications

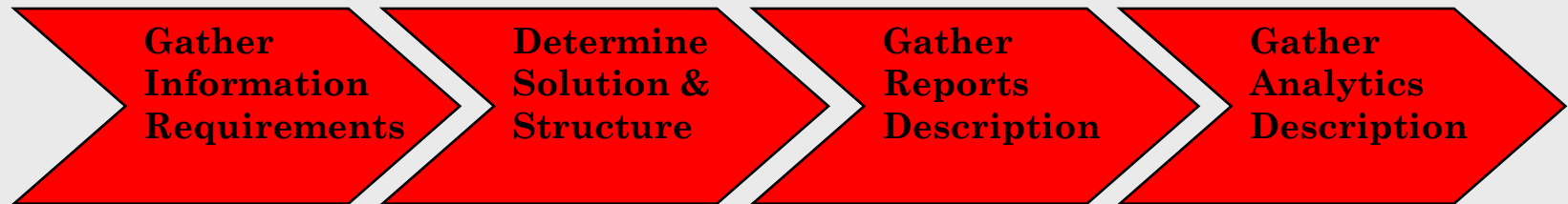
Enterprise application specific functional use e.g., application packages that are used in individual business areas for forecasting, finance, campaign management

d) Intense Analytics

Statistical analysis, modeling and data mining done on an iterative basis to generate segment definitions, offer specifications, credit policies, etc. Requires robust sampling, modeling, scoring and testing capabilities.

Week 1 Topic:

Detailed Requirements Gathering



- Identify reporting and analytic solutions and its data content currently in place, inventory, and leverage existing data sourcing solutions (physical models)
- Describe additional data sets. This may require requesting for additional source files, augmenting current source files, access to external data providers, or derivation of data from existing sources.

- Following the data analysis methodology, produce the standardized data model
- Map source data to target data structure and identify any transformations needed
- Perform functional analysis of toolsets
- Determine operational processes and user access needs

- Develop reports inventory including frequency of updates, data content, user access, query capability, owner, approval process, distribution scheme, etc.
- Prioritize reports for a phased rollout
- Identify whether a current report will satisfy requirements or a new report is needed

- Develop analytic solution inventory including analytic data needs, analytic capability (modeling, statistics, OLAP, etc.), parameters and conditions, etc.
- Understand and determine metrics to meet business goals
- Prioritize needs for a phased rollout

Week 2 Topic:

Overview of Optimization Modeling

Optimization is the process of finding the best values of the variables for a particular criterion or the best decisions for a particular measure of performance.

◆ Components of an Optimization Model:

- Objective – mathematical function to minimize or maximize some measure of performance
- Parameters – Numerical inputs for calculations that may correspond to raw data, estimates, forecasts, or predictions
- Constraints – logical conditions or calculations representing real world limits
- Decision Variables – An unknown quantity or variables to be determined via model run

◆ References for Optimization Modeling:

- Data Smart
- www.solver.com
- <http://opensolver.org>
- Optimization Modeling with Spreadsheets (What's in the PDF handed out. Additional content will be discussed/distributed as needed)
- Step-By-Step_Optimization_S.pdf (in Week 3 Readings)

Week 2 Topic:

Basic Structure of Optimization Modeling

- ◆ The basic structure of a typical mathematical optimization problem formulation is shown here:

min|max objective function


subject to constraints

variable bounds

- ◆ The formulation is easier to understand if it is followed by a description of the decision variables, sets, and parameters.

Week 2 Topic:

Linear Programming Problem

- ◆ Each linear constraint can be either an **inequality** or an **equation** 
- ◆ Bounds can be $\pm\infty$, so x_j can be restricted to be nonnegative ($l_j = 0$ and $u_j = +\infty$) or free ($l_j = -\infty$ and $u_j = +\infty$)
- ◆ Exhibits **proportionality** (contribution from any given decision variable to the objective grows in proportion to its value), **additivity** (contribution from one decision is added to contributions of other decisions), and **divisibility** (fractional decision variable is meaningful).

$$\min | \max \quad f_1 x_1 + \dots + f_n x_n$$

$$\text{subject to} \quad \mathbf{Ax} \{ \leq, =, \geq \} \mathbf{b}$$

$$l_j \leq x_j \leq u_j \quad (j = 1, 2, \dots, n) \quad \img alt="comment icon" data-bbox="775 722 799 755"/>$$

Week 2 Topic:

Nonlinear Programming Problem

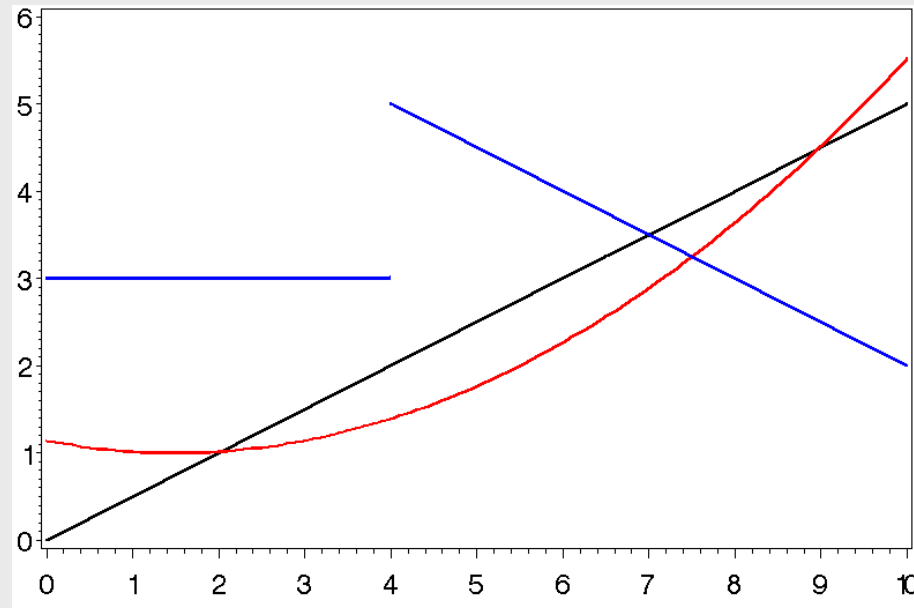
- ◆ $f(\mathbf{x})$ and $c_i(\mathbf{x})$ are continuous functions
- ◆ With the following constraints:
 - **unconstrained:** There are no constraints or bounds.
 - **bound constrained:** There are no constraints other than bounds.
 - **linearly constrained:** All functions $c_i(\mathbf{x})$ are linear.
 - **nonlinearly constrained:** At least one of the functions $c_i(\mathbf{x})$ is nonlinear.

$$\begin{array}{ll}\min | \max & f(\mathbf{x}) \\ \text{subject to} & c_i(\mathbf{x}) \{ \leq, =, \geq \} b_i \quad (i=1,2,\dots,m) \\ & l_j \leq x_j \leq u_j \quad (j=1,2,\dots,n)\end{array}$$

Week 2 Topic:

What's expected using Optimization

- ◆ Able to set up Optimization problems using **add** in function of Excel
- ◆ Understand the three modeling available:
 - GRG Nonlinear: For nonlinear, smooth (red)
 - Simplex LP: For linear, smooth (black)
 - Evolutionary: For nonlinear, non-smooth (blue)



Week 2 Topic:

Usage in business

- ◆ **Production Planning:** Determine which of several possible mixes of products should be produced to achieve the highest profit.
- ◆ **Facility Location:** Find the “best” site for, say, a new factory, in relation to the location of materials suppliers, distribution centers, and so on.
- ◆ **Portfolio Selection:** Maximize ROI, balancing return versus risk.
- ◆ **Personnel Assignment:** Match personnel to work requirements in order to meet current needs and anticipated changes, subject to budget and HR requirements.
- ◆ **Supply Chain Planning:** Find the lowest-cost way to move product from factories to distribution centers to stores, and plan for possible disruptions or expansion.
- ◆ **Promotional Marketing:** Determine the best combination of promotional offers, delivery channels, and customers to maximize the overall return on marketing investment.
- ◆ **Supplier Selection and Evaluation:** Choose which suppliers to deal with in order to satisfy requirements and maximize leverage, rating suppliers using a variety of criteria simultaneously.
- ◆ **Inventory Replenishment:** Set inventory policies (reorder levels and maximum stock levels) to meet customer service goals and minimize costs.
- ◆ **Pricing Decisions:** Establish and maintain optimal everyday prices based on costs, regional demand patterns, and competitive price information.

Week 3 Topic:

Parameters and Statistics

- ◆ Statistics are used to approximate population parameters.
- ◆ *Parameters* are characteristics of populations. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. *Statistics* are quantities calculated from the values in the sample.

| | Population Parameters | Sample Statistics |
|--------------------|-----------------------|-------------------|
| Mean | μ | \bar{x} |
| Variance | σ^2 | s^2 |
| Standard Deviation | σ | s |

Week 3 Topic:

Statistics

Descriptive Statistics:



- ◆ The goals when you are describing data are to
 - screen for unusual sample data values
 - inspect the spread and shape of continuous variables
 - characterize the central tendency of the sample.

Inferential Statistics:



- ◆ The goals for statistical inference are to
 - estimate or predict unknown parameter values from a population, using a sample
 - make probabilistic statements about population attributes.
- ◆ After you select a random sample of the data, you can start describing the data. Although you want to draw conclusions about your population, **you first want to explore and describe your data before you use inferential statistics. Why?**
 - Data must be as error free as possible.
 - Unique aspects, such as data values that cluster or show some unusual shape, must be identified.
 - An extreme value of a variable, if not detected, could cause gross errors in the interpretation of the statistics.

Week 3 Topic:

Parameters and Statistics

- ◆ *Statistics* are quantities calculated from the values in the population.
- ◆ Suppose you have x_1, x_2, \dots, x_n , a sample from some population

$$\bar{x} = \frac{1}{n} \sum x_i$$

The mean is an average, a typical value in the distribution.

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

The variance measures the sample variability.

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

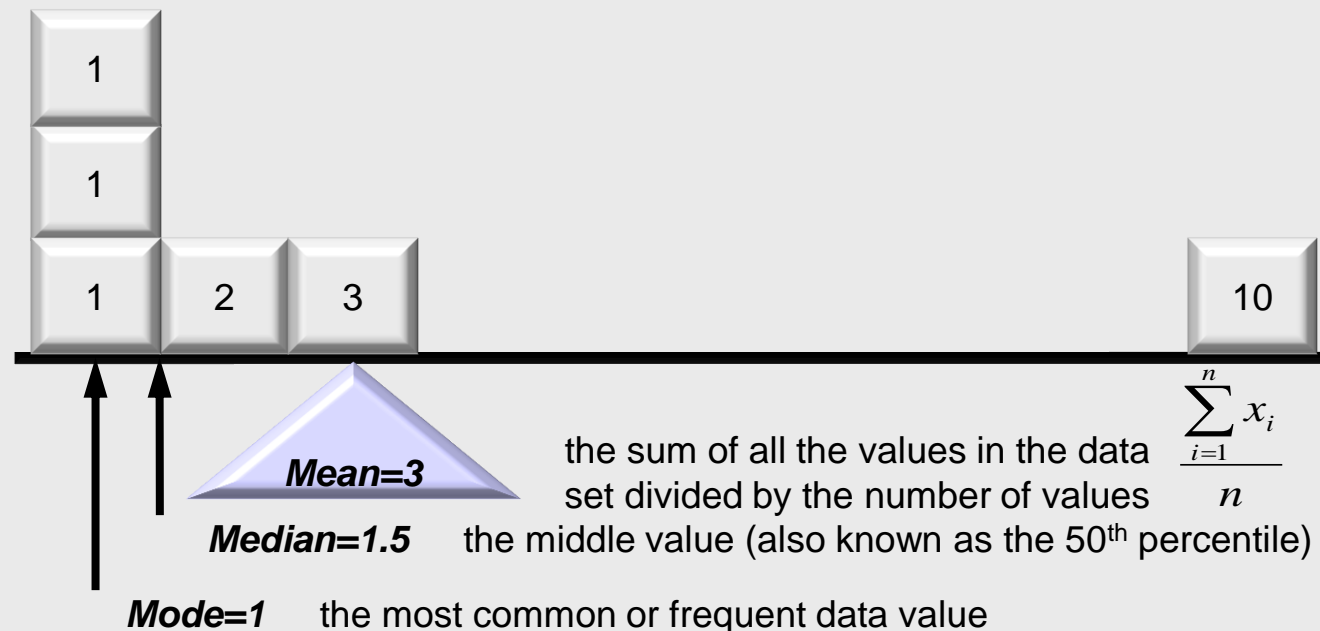
The standard deviation is square root of the variance and also measures variability in data. Usually reported in the same units as the mean.



Week 3 Topic:

Mean, Median, Mode

- ◆ A property of the sample mean is that the sum of the differences of each data value from the mean is always 0, that is, $\sum (x_i - \bar{x}) = 0$.
- ◆ The *mean* is the arithmetic balancing point of your data.
- ◆ The *median* is the data point in the middle of a sorted sequence. It is appropriate for either rank scores (variables measured on an ordinal scale) or variables measured on an interval or ratio scale with a skewed distribution.
- ◆ The *mode* is the data point that occurs most frequently. It is most appropriate for variables measured on a nominal scale. There might be several modes in a distribution.



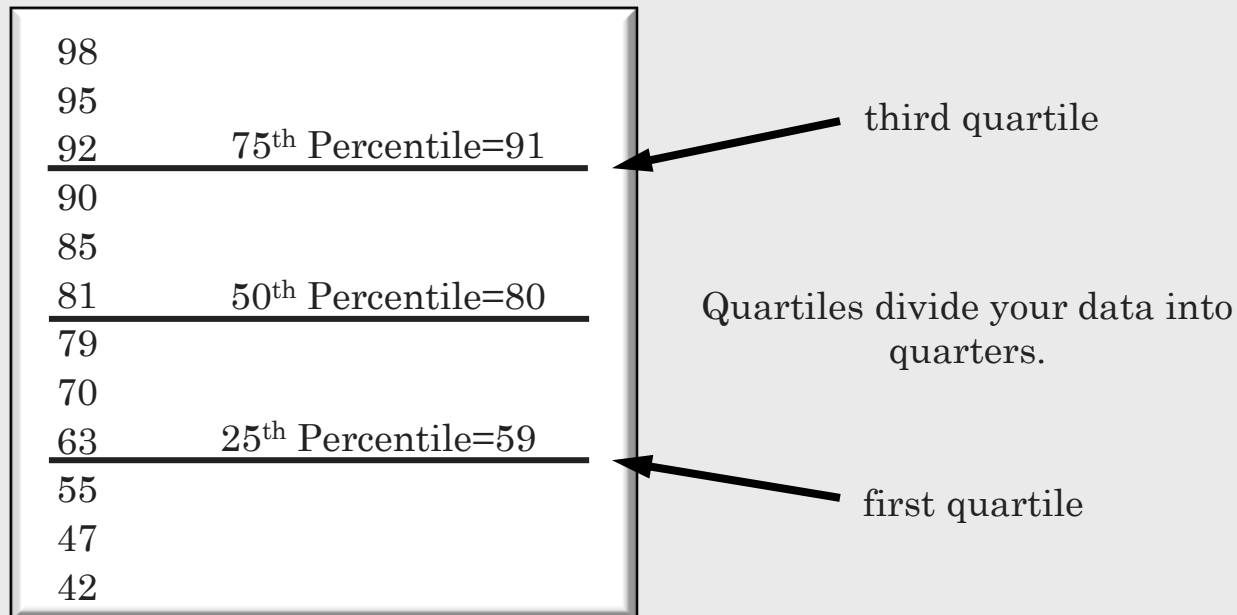
Week 3 Topic:

Distributions

- ◆ A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. For any type of data, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.
- ◆ When you examine the distribution of values in a variable, you can determine the following:
 - the range of possible data values
 - the frequency of data values
 - whether the data values accumulate in the middle of the distribution or at one end
 - Are the values symmetrically distributed?
 - Are any values unusual?
 - What is the best estimate of the average of the values for the population?
 - What is the best estimate of the average spread or dispersion of the values for the population?


Week 3 Topic: Percentiles

- ◆ *Percentiles* locate a position in your data larger than a given proportion of data values.
- ◆ These are commonly reported percentile values:
 - the 25th percentile, also called the first quartile
 - the 50th percentile, also called the median
 - the 75th percentile, also called the third quartile



Week 3 Topic:

Spread of Distribution - Dispersion

- ◆ Measures of dispersion enable you to characterize the dispersion, or spread, of the data.
- ◆ A value better suited to reflect dispersion is the *interquartile range*. The interquartile range shows the range of the middle 50% of data values. 

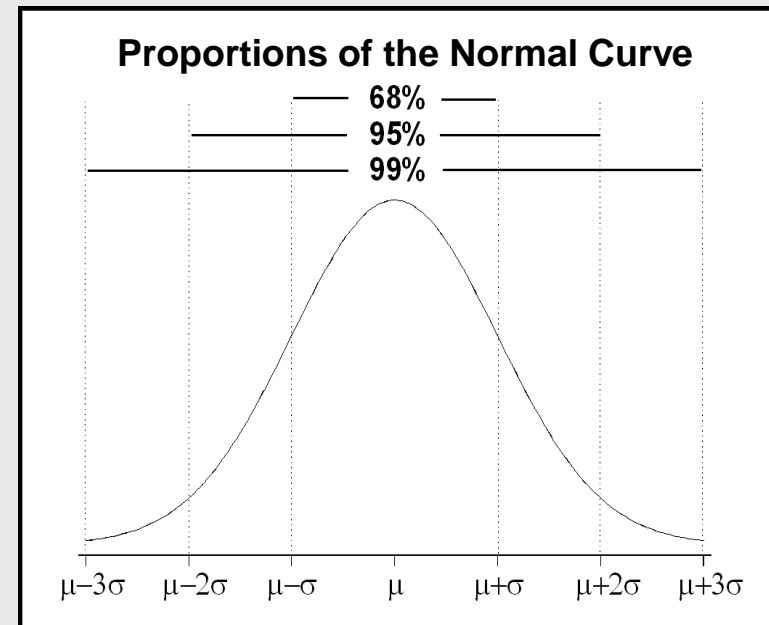
| Measure | Definition |
|---------------------|---|
| Range | the difference between the maximum and minimum data values |
| Interquartile Range | the difference between the 25th and 75th percentiles |
| Variance | a measure of dispersion of the data around the mean |
| Standard Deviation | a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance) |

Week 3 Topic:

Normal Distribution

A normal distribution

- ◆ is symmetric. If you draw a line down the center, you get the same shape on either side.
- ◆ defined by two parameters, μ (the population mean) and σ (the population standard deviation).
- ◆ is bell shaped and has mean = median = mode which describes the midpoint of the distribution
- ◆ Another name for the normal distribution is the Gaussian distribution.
- ◆ The standard normal curve has $\mu=0$ and $\sigma=1$. The area under the curve between any two values can be calculated.
- ◆ Approximately 68% of the total area lies within 1 standard deviation of the mean.
- ◆ Approximately 95% of the total area lies within 1.96 standard deviations of the mean.
- ◆ Approximately 99.7% of the area lies within 3 standard deviations of the mean.



Week 3 Topic:

Normal Distribution (cont.)

- ◆ Often in analysis, although not always, a normal distribution is assumed.
- ◆ The normal distribution is a mathematical function. The height of the function at any point on the horizontal axis is the “probability density” at that point. Normal distribution probabilities (which can be thought of as the proportion of the area under the curve) tend to be higher near the middle.
- ◆ The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ . A larger standard deviation implies a wider normal distribution. The mean locates the distribution (sets its center point) and the standard deviation scales it.
- ◆ Often, values that are more than two standard deviations from the mean are regarded as unusual. Only about 5% of all values are at least that far away from the mean.
- ◆ You use this information later when you discuss the concepts of confidence intervals.

Week 3 Topic:

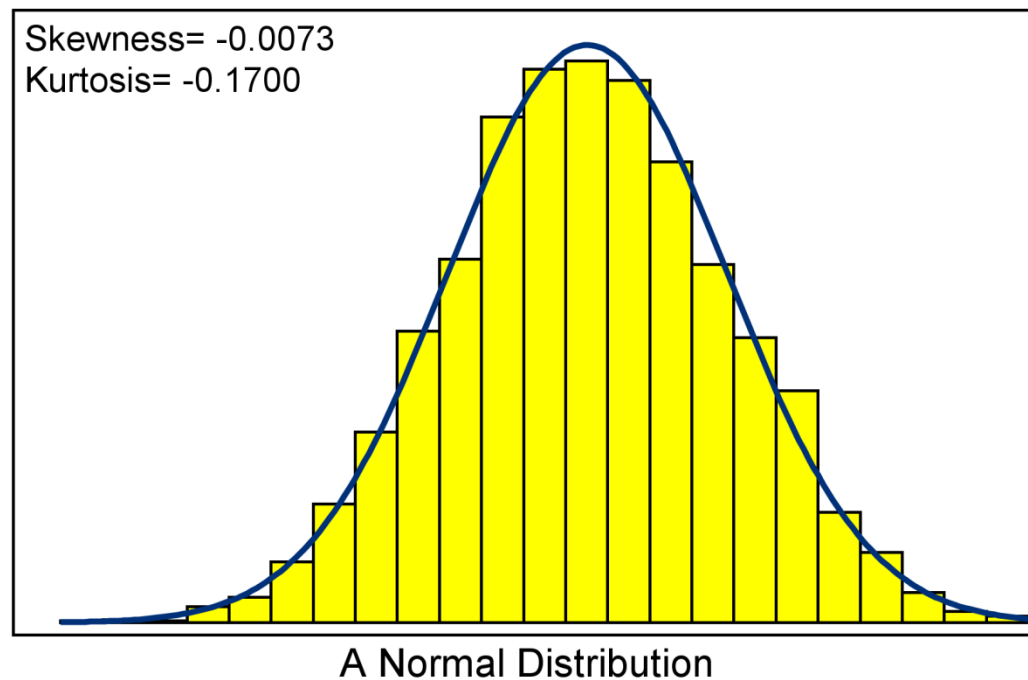
Distributions compared to Normal

- ◆ The distribution of your data might not look normal. There are an infinite number of ways that a population can be distributed. When you look at your data, you might notice the features of the distribution that indicate similarity or difference from the normal distribution.
- ◆ When you evaluate distributions, it is useful to look at statistical measures of the shape of the sample distribution compared to the normal.
- ◆ Two such measures are skewness and kurtosis.



Week 3 Topic: Normal Distribution

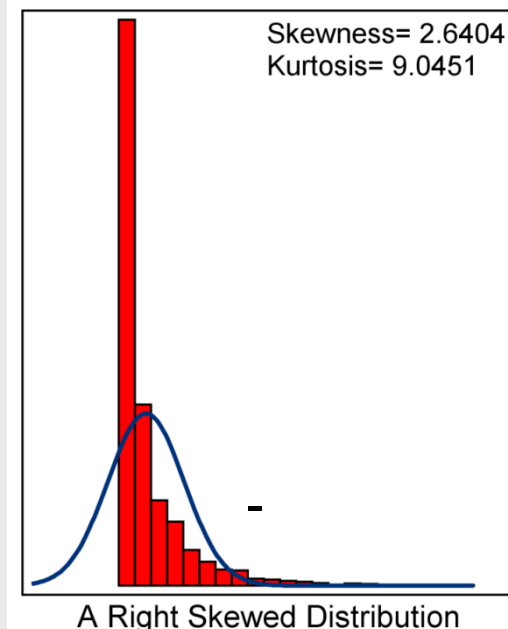
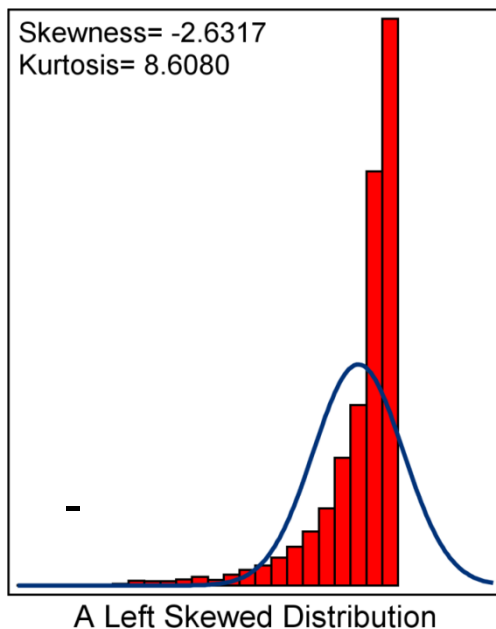
- ◆ A histogram of data from a sample drawn from a normal population generally show values of skewness and kurtosis near zero in SAS output.



Week 3 Topic:

Skewness

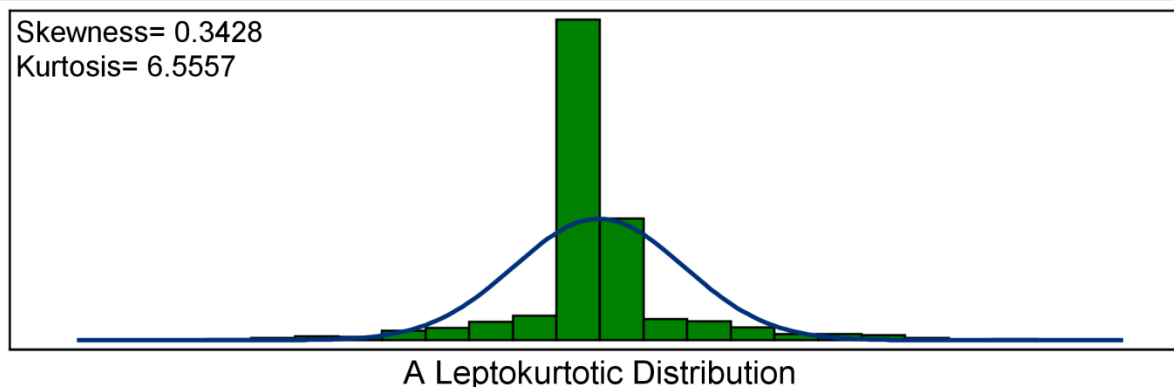
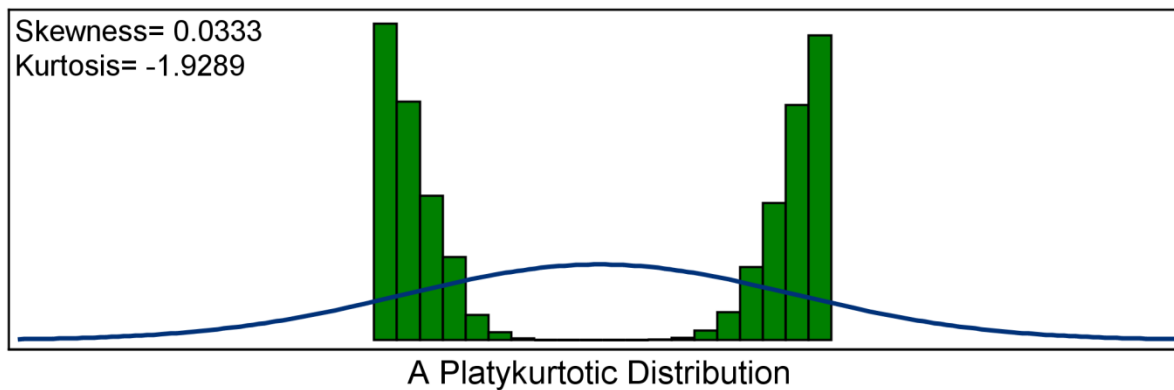
- ◆ One measure of the shape of a distribution is skewness. The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to zero.
- ◆ If your distribution is more spread out on the
- ◆ **left** side, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.
- ◆ **right** side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.



Week 3 Topic:

Kurtosis

- ◆ *Kurtosis* measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to zero in SAS. Kurtosis is often very difficult to assess visually.



Week 3 Topic:

Kurtosis (cont.)

- ◆ If the value of your kurtosis statistic is negative, the distribution is said to be *platykurtic*. If the distribution is both symmetric and platykurtic, then there tends to be a smaller-than-normal proportion of observations in the tails and/or a somewhat flat peak. Rectangular, bimodal, and multimodal distributions tend to have low (negative) values of kurtosis.
- ◆ If the value of the kurtosis statistic is positive, the distribution is said to be *leptokurtic*. If the distribution is both symmetric and leptokurtic, then there tends to be a larger-than-normal proportion of observations in the extreme tails and/or a taller peak than the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone distributions*.
- ◆ Distributions that are asymmetric also tend to have nonzero kurtosis. In these cases, understanding kurtosis is considerably more complex than in situations where the distribution is approximately symmetric.

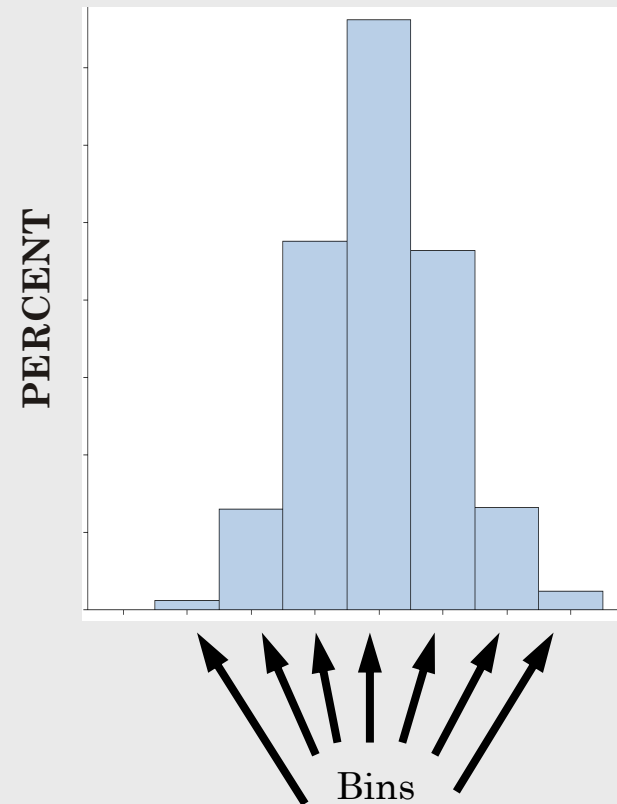
Week 3 Topic:

Graphical Displays of Distributions

- ◆ You can produce the following three types of plots for examining the **distribution** of your data values:
 - histograms
 - normal probability plots
 - box plots
 - scatter plots

Week 3 Topic: Histograms

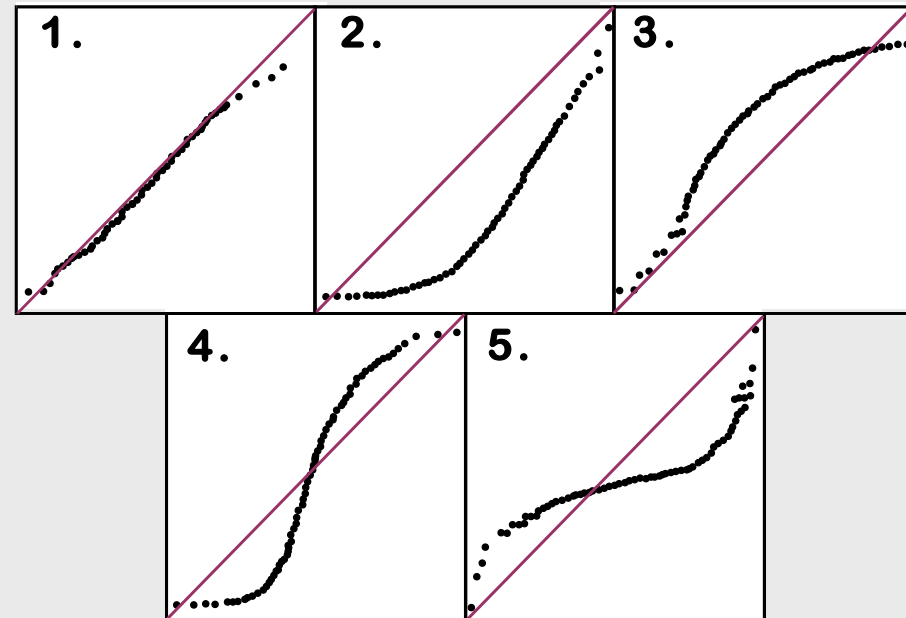
- ◆ Most elementary statistical procedures assume some underlying population probability distribution. It is a good idea to look at your data to see whether the distribution of your sample data can reasonably be assumed to come from a population with the assumed distribution.
- ◆ A histogram is a good way to determine how the probability distribution is shaped.



Week 3 Topic:

Normal Probability Plots

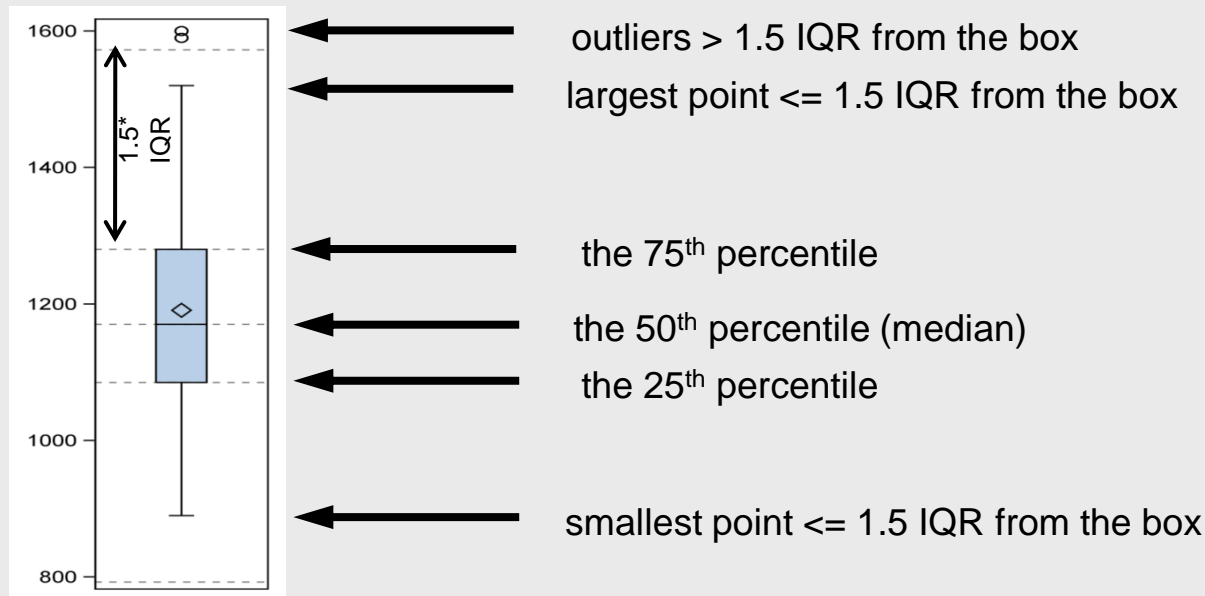
- ◆ A *normal probability plot* is a visual method for determining whether your data comes from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.
- ◆ The above diagrams illustrate some possible normal probability plots for data from the following:
- ◆ normal distribution (The observed data follow the reference line.)
- ◆ skewed-to-the-right distribution
- ◆ skewed-to-the-left distribution
- ◆ light-tailed distribution
- ◆ heavy-tailed distribution



Week 3 Topic:

Box Plots

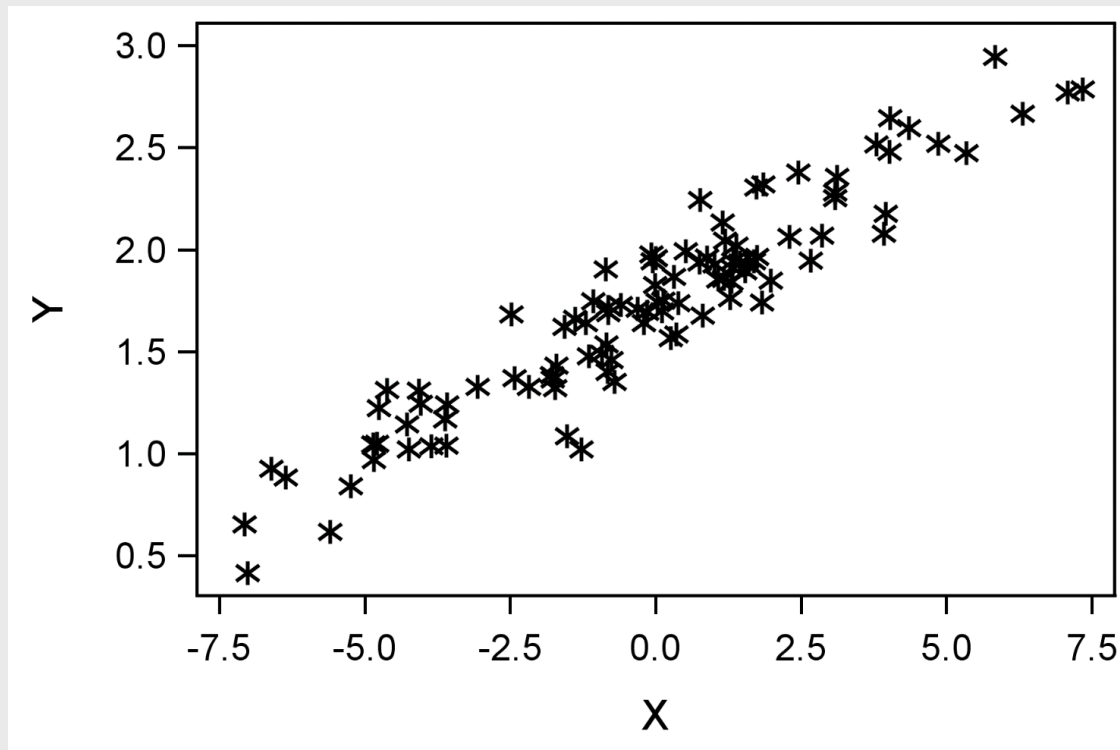
- ◆ *Box plots* (Tukey 1977) (sometimes referred to as *box-and-whisker plots*) provide information about the variability of data and the extreme data values. The box represents the middle 50% of your data (between the 25th and 75th percentile values). You get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as by assessing the symmetry of the box and whiskers around the median line. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile range (IQR) units. If any values lay more than 1.5 IQR units from either end of the box, they are represented in SAS by individual plot symbols.
- ◆ The plot above shows that the data are approximately symmetric.



The mean is denoted by a \diamond .

Week 3 Topic: Scatter Plots

- ◆ *Scatter plots* are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.



Week 3 Topic: Scatter Plots (cont.)

- ◆ Scatter plots are useful to accomplish the following:
 - explore the relationships between two variables
 - locate outlying or unusual values
 - identify possible trends
 - identify a basic range of Y and X values
 - communicate data analysis results
- ◆ The predicted value can be thought of as the best estimate of the value of the response at a given value of the predictor variable. Scatter plots show graphically the relationship between predictor variables and response variables.
- ◆ Traditionally, predictor variables are plotted on the x axis and response variables are plotted on the y-axis. A preliminary analysis of associations involves discovery of the presence of associations and their nature.

Week 3 Topic:

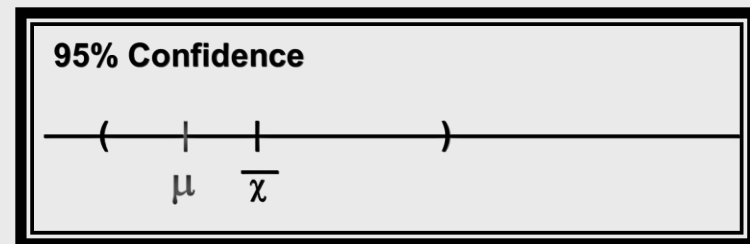
Standard Error of a Mean

- ◆ A statistic that measures the variability of your estimate is the *standard error of the mean*. In statistics, assumptions are often made about distributions of parameters. A common one is that the sampling distribution of parameters is normal. This does not necessarily mean that the units of the population are normally distributed. It is often assumed that the parameter itself is normally distributed. Even though most statisticians only take one sample and get one point estimate for the population parameters, it is useful if they can assume normality of the parameter. The variability of a parameter is measured by its standard error.
- ◆ It differs from the sample standard deviation in that:
 - the sample standard deviation is a measure of the variability of data;
 - the standard error of the mean is a measure of the variability of the sample mean.
 - Standard error of the mean = $\frac{s}{\sqrt{n}} = s_{\bar{x}}$
 - where
 - s is the sample standard deviation.
 - n is the sample size.
- ◆ The standard error of the mean is a measure of precision of the parameter estimate. The smaller the standard error, the more precise your estimate.

Week 3 Topic:

Confidence Interval

- ◆ A *confidence interval* is a range of values that you believe is likely to contain the population parameter of interest is defined by an upper and lower bound around a parameter estimate.
- ◆ To construct a confidence interval, a significance level must be chosen.
- ◆ A 95% confidence interval is commonly used to assess the variability of the sample mean. In the Ames housing sales example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean sale price for your population of home sales.
- ◆ You want to be as confident as possible, but remember that if you increase the confidence level too much, the width of your interval increases beyond the point where it is informative. For example, a 100% confidence interval would have confidence bounds of negative and positive infinity.
- ◆ A 95% confidence interval represents a range of values within which you are 95% certain that the true population mean exists.
 - One interpretation is that if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.



Week 4 Topic:


Defining Cluster Analysis

“*Cluster analysis* is a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual.”

Everitt (1998), *The Cambridge Dictionary of Statistics*


Week 4 Topic:

It's about Pattern Discovery

- ◆ **Cluster** is a group of similar objects (observations, customers, patients, buyers, locations, etc.)
- ◆ **Cluster Analysis** is a set of data-driven partitioning techniques designed to group a collection of objects into clusters. Its about data explorations, searching for patterns in complex data, that is conducted in repetitive  fashion. Finding these patterns can lead to business decisions.

Week 4 Topic:

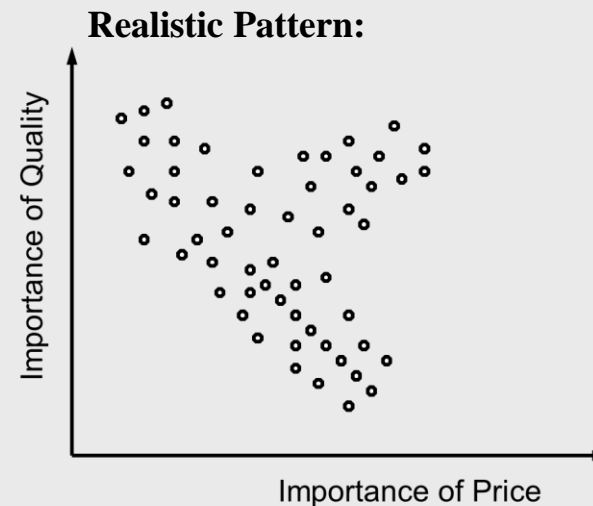
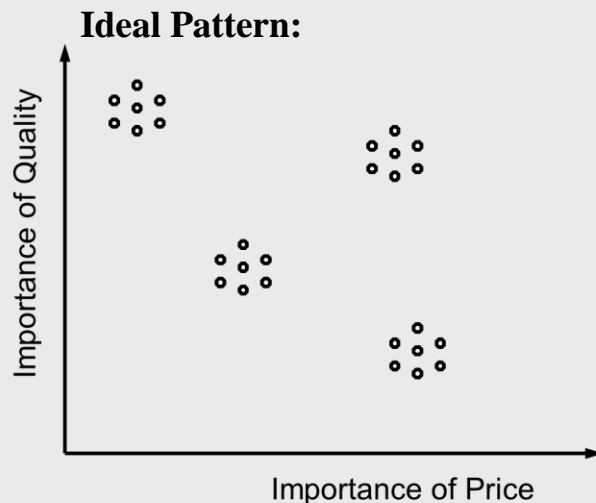
Further definitions - Cluster Analysis

- ◆ *Cluster analysis* is the generic name for a wide variety of procedures that can be used to create a classification of entities/objects 
- ◆ *Cluster analysis* is a **convenient** method commonly used in many disciplines to categorize entities (individuals, objects, and so on) into groups that are **homogenous** along a range of observed characteristics (variables).
- ◆ The goal of cluster analysis is to partition/classify data into groups/objects (in our case, individuals, households or families) so that each object in a cluster is similar to the other objects in the same cluster; however objects in different clusters are dissimilar to each other.
- ◆ Therefore, if classification is successful, the objects within the cluster will be close together when plotted geometrically and different clusters will be far apart. A plot (showing clearly separated clusters) in two dimensions is shown on the next slide.

Week 4 Topic:

Cluster Analysis – Grouping

- ◆ Realistically speaking, it is highly unlikely to find natural groups that are very well separated, even in the two-dimensional space. It is more likely that a two-dimensional plot of customer importance variables will produce a messy, not well-separated plot that could indicate the possibility of different numbers of groups depending on how we view and choose to partition the data.



Week 4 Topic:

Cluster Analysis – Grouping Options

- ◆ Many other cluster (grouping) solutions could possibly be seen in this simple data set. Clearly, the real-world data will be much more complicated (greater number of variables/dimensions, each variable may be measured with different measurement scales, possibility of measurement error, possibility of random error, missing values, extreme values, etc.) than this simple two-dimensional example.



Week 4 Topic:

Clustering Guidelines - 1


Which variables should be used for clustering?

- ◆ The variables for clustering should primarily be chosen based on business objectives for segmentation.
- ◆ In an ideal world, the variables should be relatively small in number, with low correlations among each other, have good (interval) measurement properties, have approximately normal distribution (kurtosis and skewness values close to zero).
- ◆ In the real world, the conditions mentioned above almost never exist.
- ◆ Fortunately, we have tools to handle some of these issues

Week 4 Topic:

Clustering Guidelines - 2

How should similarity between observations be operationalized? Choose among:

- ◆ distance metrics - when you have only numerical variables
- ◆ similarity coefficients - when you have only non-numerical variables only 
- ◆ distance metrics - when you have both numerical and non-numerical variables).

Week 4 Topic:

Clustering Guidelines – 3

How to form clusters?

- ◆ Choose among different linkages or variance
- ◆ Typically, variables with higher variances tend to have more impact in determining the cluster solution than variables with lower variances. In some cases, data may need to be standardized to measure linkages. Range standardization is preferred because it tends to preserve the differences among groups better than standardizing to 0 means and unit variance (Milligan and Cooper, 1988). We will revisit this later using our case study data set.

Week 4 Topic:

Clustering Guidelines – 4

How many clusters?

- ◆ Practical (managerial) considerations in segmentation studies often dictate a small number of clusters (somewhere in the range of 2-10).
- ◆ Use relative sizes of each cluster in making this decision (in most applications, the preference is for somewhat balanced sizes or number of observations in each cluster).
- ◆ Use the relative change in distances at which clusters are combined (or, relative changes in the overall heterogeneity measure) to help with this decision. There are some statistics (such as pseudo- F , pseudo t^2 and CCC) that can be used judiciously to guide your decisions as well.

Week 4 Topic:

Cluster Analysis – Methods of execution

- ◆ **Unsupervised learning** -- learn from raw data (no examples of correct classification). In other words, class label (e.g., income bands, purchase power, etc.) information is unavailable. Unsupervised methods set the model's parameters without prior knowledge about the classification of samples.
- ◆ **Supervised learning** -- algorithms use class variables to generate its solution. An example is market segmentation on the next slide.
- ◆ Plotting the data can help to see if there is any evidence of cluster structure at all. It can also give you an idea of how many clusters there are, as well as helping you to identify potentially problematic non-spherical (irregular) clusters.
- ◆ It might be necessary to preprocess the data to optimize them for clustering. Common preparation steps include creating a distance matrix, standardizing the variables, or other transformations.

Week 4 Topic:

What is Market Segmentation?

“Market segmentation is grouping people (with the willingness, purchasing power, and the authority to buy) according to their similarity in several dimensions related to a product under consideration.”

Market Segmentation is supervised learning -- learn from data where the correct classification of examples is given (class label information is available) e.g., Naïve Bayes Classifier.

These can be results of a query or simple segmentations using dimensions:

- ◆ Demographics: Age, Gender, Education, Income, Home ownership, etc.
- ◆ Psychographics: Lifestyle, Attitude, Beliefs, Personality, Buying motives, etc.
- ◆ Brand Loyalty
- ◆ Geography: State, ZIP, City size, Rural vs. Urban, etc.

Week 4 Topic:

Applications – Retail Example

| Application | Business Decision Support |
|----------------------------|---|
| Profiling and Segmentation | Understand customer behaviors and needs by segment. Direct efforts to like customer groups. |
| Cross-sell and Up-sell | Determine what customers are likely to buy. Better target/recommend product/ service offerings. |
| Acquisition and Retention | Understand customer preferences and purchase patterns. Determine how to grow and maintain valuable customers. |
| Campaign Management | Execute better customer communications. Determine right offer to the right person at the right time. Determine which customers to invest in and how to best appeal to them. |

Week 4 Topic:

Types of Clustering

- ◆ **Partitional (k-means/optimization) clustering**
 - A division of objects into non-overlapping subsets (clusters) such that each object is in exactly one cluster
 - SAS offers PROC FASTCLUS (k-means clustering)
- ◆ **Hierarchical clustering**
 - A set of nested clusters organized as a hierarchical tree. Hierarchical clustering creates clusters that are hierarchically nested within clusters at earlier iterations, similar to the identification of species taxonomy in biology.

Week 4 Topic:

k-means clustering

- ◆ *K-means clustering* is, perhaps, the most popular partitive clustering algorithm. One reason for its popularity is that the time required to reach convergence on a solution is proportional to the number of observations being clustered, which means it can be used to cluster larger data sets.
- ◆ In fact, *k*-means clustering is inappropriate for small data sets (< 100 cases); the solution becomes sensitive to the order in which the observations appear. Changing the observation ordering, neither adding nor deleting observations, produces vastly different cluster solutions. This is known as the **order effect**.
- ◆ In SAS, PROC FASTCLUS implements the *k*-means algorithm. As its name suggests, the PROC FASTCLUS finds clusters in only a few (default=1) passes through the data. It also produces a description of the typical member of each cluster, which is useful both as a summary of its members, and as the basis for scoring new cases.

Week 4 Topic:

Limitations of K-means

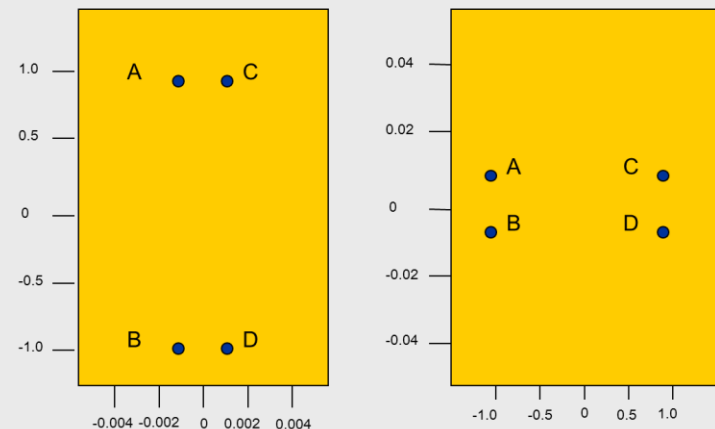
- ◆ Need to specify K (number of clusters) in advance
- ◆ Applicable only for numeric data
- ◆ Has problems when clusters are of differing sizes or densities
- ◆ Unable to handle noisy data and outliers
- ◆ May be indeterminate

Week 5 Topic:

Cluster Analysis - Scaling

- ◆ As these plots indicate, **grouping (and hence clustering) is heavily dependent** on the measurement scales of clustering variables.
- ◆ Euclidean distance metric assume equal weight to each input variables. But, in reality, the input variables with a wider scale of measurement (more variance) get weighted more in determining distances.
- ◆ So, the solution is standardization. 1) Many different ways of doing this in SAS (STDIZE procedure). 2) Range standardization (where each input variable is scaled by first subtracting the minimum value and then dividing by the range) is often preferred to standardizing to 0 mean, 1 standard deviation.

What's the natural grouping in these plots?



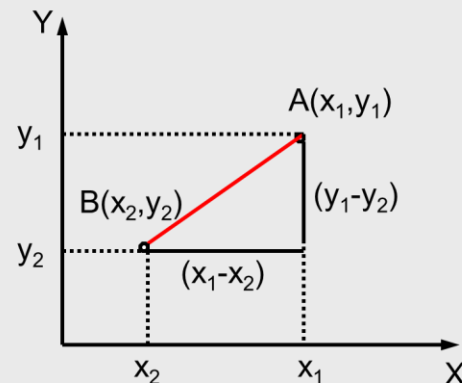
The Data points (A, B, C, and D) are the same but X and Y axes are scaled differently in the two plots!

Week 5 Topic:

Cluster Analysis – Distance Measure

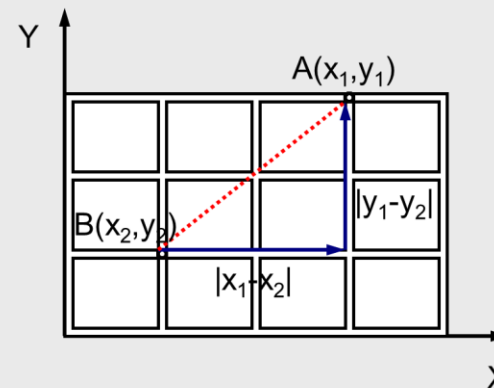
In most practical applications of segmentations, Euclidean distance metric is used. This has perhaps happened because of many people's familiarity with this distance metric and the wide availability of computer programs that used this metric for clustering by default. The city-block metric has also been used in some applications.

Assume two observations, A and B, in a two-dimensional space have coordinates (x_1, y_1) and (x_2, y_2) .



Squared Euclidean Distance between A and B,
 $(D_{AB})^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$ or, $D_{AB} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Assume two observations, A and B, in a two-dimensional space have coordinates (x_1, y_1) and (x_2, y_2) .



City Block or Manhattan Distance between A and B,
 $D_{AB} = |x_1 - x_2| + |y_1 - y_2|$

Week 5 Topic:

Cluster Analysis – Example

Assume we have the following data from three customers about what they consider important in buying a product (rated on a scale with 1=not at all important and 11=extremely important).

| Customer | Price | Quality |
|----------|-------|---------|
| A | 9 | 8 |
| B | 5 | 9 |
| C | 6 | 8 |

Euclidean distance between A and B is $\sqrt{(9-5)^2 + (8-9)^2} \approx 4.123$

City Block distance between A and B is $|9-5| + |8-9| = 5$

Euclidean distance between B and C is $\sqrt{(5-6)^2 + (9-8)^2} = 1.414$

City Block distance between B and C is $|5-6| + |9-8| = 2$

Euclidean distance between A and C is $\sqrt{(9-6)^2 + (8-8)^2} = 3$

City Block distance between A and C is $|9-6| + |8-8| = 3$

Week 5 Topic:

Cluster Analysis – Example (cont.)

What if we also had data on a third variable, say importance of service, from the same three customers?

| Customer | Price | Quality | Service |
|----------|-------|---------|---------|
| A | 9 | 8 | 7 |
| B | 5 | 9 | 8 |
| C | 6 | 8 | 9 |

Euclidean distance between A and B is $\sqrt{(9-5)^2 + (8-9)^2 + (7-8)^2} = 4.243$

Euclidean distance between B and C is $\sqrt{(5-6)^2 + (9-8)^2 + (8-9)^2} = 1.732$

Euclidean distance between A and C is $\sqrt{(9-6)^2 + (8-8)^2 + (7-9)^2} = 3.606$

Week 5 Topic:

Cluster Analysis – Example (cont.)

This is a likely situation in many practical segmentation problems. Unfortunately, in such situations (involving a mix of numerical and categorical variables), complications arise because theoretically it is unclear what “distance” really means! Often the only practical solution is to convert the categorical variables into binary (1/0) variables and then use the binary variables along with the other numeric variables in calculating distance metrics.

Suppose we also know customers’ marital status, and we would like to use that in our distance calculation.

| Customer | Price | Quality | Service | Marital Status |
|----------|-------|---------|---------|----------------|
| A | 9 | 9 | 7 | Single |
| B | 5 | 9 | 8 | Married |
| C | 6 | 8 | 9 | Divorced |

Convert marital status to dummy variables and use those in distance calculations.

| Customer | Price | Quality | Service | Single | Married | Divorced |
|----------|-------|---------|---------|--------|---------|----------|
| A | 9 | 9 | 7 | 1 | 0 | 0 |
| B | 5 | 9 | 8 | 0 | 1 | 0 |
| C | 6 | 8 | 9 | 0 | 0 | 1 |

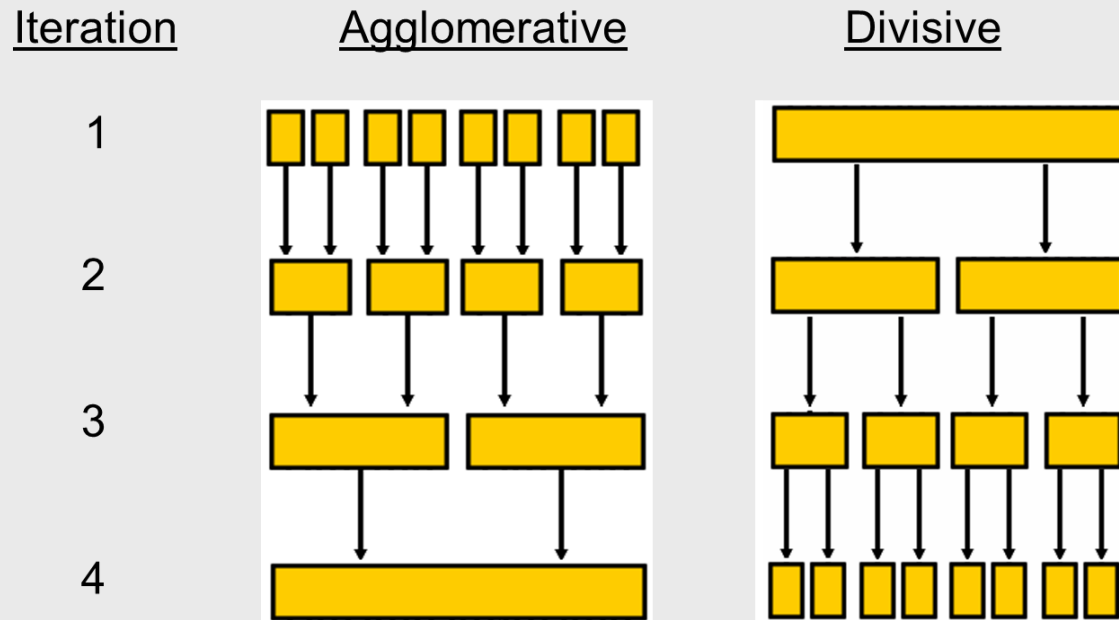
Euclidean distance between A and B is

$$\sqrt{(9-5)^2 + (9-9)^2 + (7-8)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2} = 4.359$$

Week 5 Topic:

Hierarchical Clustering - Definition

In hierarchical methods, clusters at each stage are hierarchically nested within clusters at earlier stages. Although these methods are commonly used in marketing practice, it is difficult to theoretically justify the use of these methods unless we expect a natural hierarchical structure in grouping of observations. There are two types of hierarchical clustering, *agglomerative* and *divisive*. Agglomerative methods are more commonly used.



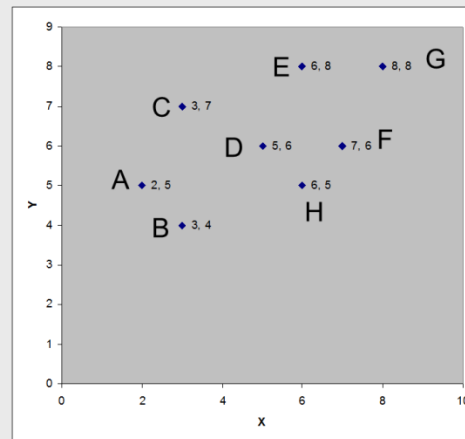
Week 5 Topic:

Hierarchical Clustering - Example

- ◆ Using the Euclidean distance formula, the distance matrix among these eight subjects is calculated as follows (below is the output from PROC DISTANCE).
- ◆ Note that the distance matrix is symmetric (across the diagonal) and hence the top half of the matrix is redundant. Also, the diagonal elements are all zeros as these reflect distance between each ID and itself. The smallest distance (1.41421) is between (A, B), (H, D), and (H, F). The next smallest distance (2.00000) is between (F, D) and (G, E). These distances will be used in the agglomerative clustering methods to assign IDs to clusters.

Data from eight subjects (A, B, C, D, E, F, G, H) on two variables, X and Y.

| ID | X | Y |
|----|---|---|
| A | 2 | 5 |
| B | 3 | 4 |
| C | 3 | 7 |
| D | 5 | 6 |
| E | 6 | 8 |
| F | 7 | 6 |
| G | 8 | 8 |
| H | 6 | 5 |



| ID | A | B | C | D | E | F | G | H |
|----|---------|---------|---------|---------|---------|---------|---------|---|
| A | 0.00000 | . | . | . | . | . | . | . |
| B | 1.41421 | 0.00000 | . | . | . | . | . | . |
| C | 2.23607 | 3.00000 | 0.00000 | . | . | . | . | . |
| D | 3.16228 | 2.82843 | 2.23607 | 0.00000 | . | . | . | . |
| E | 5.00000 | 5.00000 | 3.16228 | 2.23607 | 0.00000 | . | . | . |
| F | 5.09902 | 4.47214 | 4.12311 | 2.00000 | 2.23607 | 0.00000 | . | . |
| G | 6.70820 | 6.40312 | 5.09902 | 3.60555 | 2.00000 | 2.23607 | 0.00000 | . |
| H | 4.00000 | 3.16228 | 3.60555 | 1.41421 | 3.00000 | 1.41421 | 3.60555 | 0 |

Week 5 Topic:

Hierarchical Clustering – Example (cont.)

- ◆ There are many different ways agglomerative clustering can be performed on the distance-matrix data. In this example, use a simple rule: identify two most similar (smallest distance) IDs not already in the same cluster and join their clusters. Using this procedure, at the initial step, each of the eight IDs is considered to be in eight separate clusters.
- ◆ In Step 1, A and B are joined into one cluster as these two have one of the shortest distances (1.414); that is, these two IDs are the most similar. The number of clusters at this stage is seven, with AB in one cluster and each of the other IDs in six separate clusters. The average heterogeneity measure is calculated here as the average distance between observations within clusters (there are many other ways of operationalizing this measure). This measure generally increases as more observations are combined into clusters. This is expected because more observations getting clustered makes it likely that more dissimilar observations are being combined into clusters.

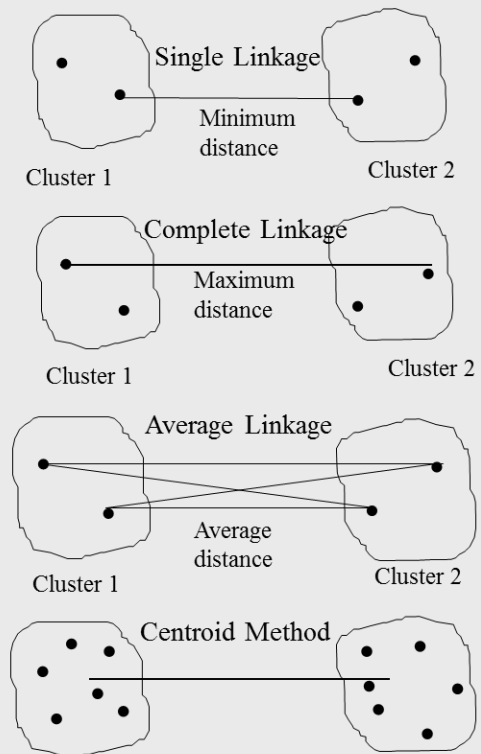
| | Minimum Distance | ID | Cluster Membership | Number of | Overall Heterogeneity |
|---------|------------------|------|--------------------|-----------|-----------------------|
| Step | Between | Pair | | Clusters | Measure (Average |
| | Unclustered | | | | Within Cluster |
| | IDs | | | | Distance) |
| Initial | | | A B C D E F G H | 8 | 0 |
| 1 | 1.414 | A,B | (AB) C D E F G H | 7 | 1.414 |
| 2 | 1.414 | H,D | (AB) C (DH) E F G | 6 | 1.414 |
| 3 | 1.414 | H,F | (AB) C (DHF) E G | 5 | 1.561 |
| 4 | 2 | G,E | (AB) C (DHF) (GE) | 4 | 1.648 |
| 5 | 2.236 | F,E | (AB) C (DHGEF) | 3 | 2.266 |
| 6 | 2.236 | A,C | (ABC) (DHGEF) | 2 | 2.32 |
| 7 | 2.236 | D,C | (ABCDHGEF) | 1 | 3.343 |

Week 5 Topic:

Hierarchical Clustering – Linkage

There are many approaches within agglomerative methods in measuring similarity between clusters when one or both clusters have multiple members.

- ◆ In **single-linkage**, the similarity between clusters is defined as the shortest distance from any member in one cluster to any member in the other cluster. Single-linkage is probably the most versatile algorithm, but poorly delineated cluster structures within the data produce unacceptable snakelike “chains” for clusters.
- ◆ In **complete-linkage**, the similarity between clusters is defined as the maximum distance from any member in one cluster to any member in the other cluster. Complete linkage eliminates the chaining problem, but only considers the outermost observations in a cluster, thus affected more by outliers.
- ◆ In **average-linkage**, the similarity between clusters is defined as the average similarity from all members in one cluster to all members in the other cluster. Average linkage is based on the average similarity of all individuals in a cluster and tends to generate clusters with small within-cluster variation and is less affected by outliers.
- ◆ In **Centroid method**, the similarity between clusters is defined as the distance between the two-cluster centroids. A cluster centroid is the mean values of all members in a cluster on the variables used in the analysis. Centroid linkage measures distance between cluster centroids and like average linkage is less affected by outliers.



Appendix:

SAS Framework and File Types

SAS framework:

- ◆ **Access data:** Using SAS, you can read any kind of data.
- ◆ **Manage data:** SAS gives you excellent data management capabilities
- ◆ **Analyze data:** For statistical analysis, SAS is the gold standard.
- ◆ **Present data:** You can use SAS to present your data meaningfully.

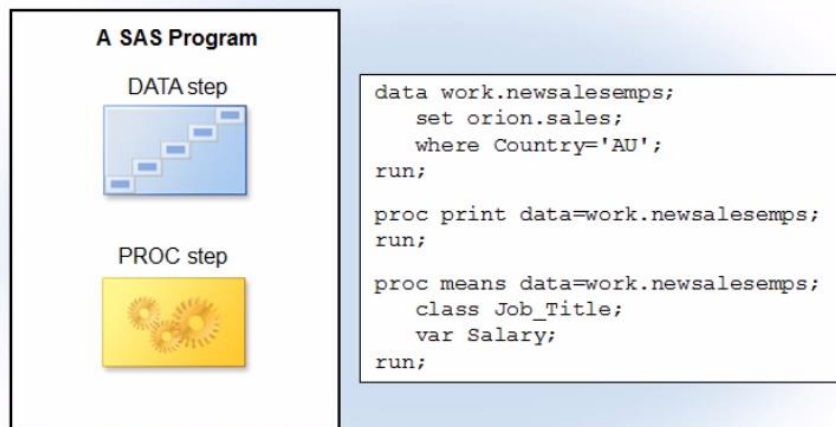
Three major file types:

- ◆ **Raw data files** contain data that has not been processed by any other computer program. They are text files that contain one record per line, and the record typically contains multiple fields. Raw data files aren't reports; they are unformatted text.
- ◆ **SAS data sets** are specific to SAS. A SAS data set is data in a form that SAS can understand. Like raw data files, SAS data sets contain data. But in SAS data sets, the data is created only by SAS and can be read only by SAS.
- ◆ **SAS program files** contain SAS programming code. These instructions tell SAS how to process your data and what output to create. You can save and reuse SAS program files.

Appendix: SAS Steps

- ◆ A SAS program consists of DATA steps and PROC steps. A SAS programming step is comprised of a sequence of statements. Every step has a beginning and ending step boundary. SAS compiles and executes each step independently, based on the step boundaries.
- ◆ A SAS program can also contain global statements, which are outside DATA and PROC steps, and typically **affect the SAS session**. A TITLE statement is a global statement. After it is defined, a title is displayed on every report, unless the title is cleared or canceled.
- ◆ SAS statements usually begin with an identifying keyword, and always end with a semicolon. SAS statements are free format and can begin and end in any column. A single statement can span multiple lines, and there can be more than one statement per line. Unquoted values can be lowercase, uppercase, or mixed case. This flexibility can result in programs that are difficult to read.

SAS Programming Steps



Appendix:

SAS Comments

- Comments are used to document a program and to mark SAS code as non-executing text. There are two types of comments: *block comments* and *comment statements*.

/ comment */*
** comment statement;*

```
/* create a temporary data
set, newsalesemps, from
the data set orion.sales */

data work.newsalesemps;
  set orion.sales;
  where Country='AU';
run;

proc print data=work.newsalesemps;
run;

proc means data=work.newsalesemps;
  class Job_Title;
  var Salary;
run;
```

/ comment */*

- any length
- internal semicolons
- X** nested

```
*create a temporary data set,
newsalesemps, from the data set
orion.sales;

data work.newsalesemps;
  set orion.sales;
  *where Country='AU';
run;

proc print data=work.newsalesemps;
run;

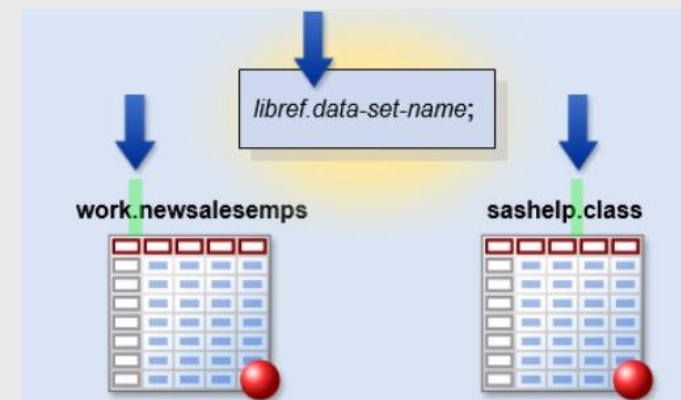
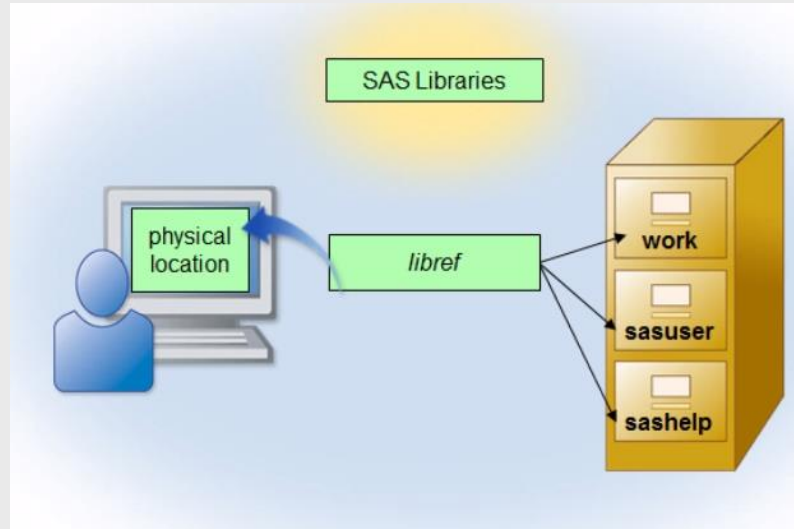
proc means data=work.newsalesemps;
  class Job_Title;
  var Salary;
run;
```

** comment statement;*

- complete statements
- X** internal semicolons

Appendix: SAS Libraries

- ◆ SAS data sets are stored in SAS libraries. A SAS library is a collection of one or more SAS files that are recognized by SAS. SAS automatically provides one temporary and at least one permanent SAS library in every SAS session.
- ◆ **Work** is a temporary library that is used to store and access SAS data sets for the duration of the session. **Sasuser** and **sashelp** are permanent libraries that are available in every SAS session.
- ◆ You refer to a SAS library by a library reference name, or libref. A libref is a shortcut to the physical location of the SAS files.
- ◆ All SAS data sets have a two-level name that consists of the libref and the data set name, separated by a period. Data sets in the **work** library can be referenced with a one-level name, consisting of only the data set name, because **work** is the default library. Data sets in permanent libraries must be referenced with a two-level name.



Appendix:

SAS Libraries (cont.)

- ◆ You can create and access your own SAS libraries. User-defined libraries are permanent but are not automatically available in a SAS session. You must assign a libref to a user-created library to make it available. You use a LIBNAME statement to associate the libref with the physical location of the library, that is, the physical location of your data. You can submit the LIBNAME statement alone at the start of a SAS session, or you can store it in a SAS program so that the SAS library is defined each time the program runs. If your program needs to reference data sets in multiple locations, you can use multiple LIBNAME statements.
- ◆ In an interactive SAS session, a libref remains in effect until you cancel it, change it, or end your SAS session. To cancel a libref, you submit a LIBNAME statement with the CLEAR option. This clears or disassociates a libref that was previously assigned. To specify a different physical location, you submit a LIBNAME statement with the same libref name but with a different filepath.
- ◆ When a SAS session ends, everything in the **work** library is deleted. The librefs are also deleted. Remember that the contents of permanent libraries still exist in the operating environment, but each time you start a new SAS session, you must resubmit the LIBNAME statement to redefine a libref for each user-created library that you want to access.

```
LIBNAME libref 'SAS-library' <options>;
```

```
LIBNAME libref CLEAR;
```

Appendix:

SAS PROC CONTENTS AND PRINT

- ◆ Use PROC CONTENTS with *libref._ALL_* to display the contents of a SAS library. The report will list all the SAS files contained in the library, as well as the descriptor portion of each data set in the library. Use the NODS option in the PROC CONTENTS statement to suppress the descriptor information for each data set.

```
PROC CONTENTS DATA=libref._ALL_ NODS;  
RUN;
```

```
PROC CONTENTS DATA=libref.SAS-data-set;  
RUN;
```

- ◆ After associating a libref with a permanent library, you can write a PROC PRINT step to display a SAS data set within the library.

```
PROC PRINT DATA=libref.SAS-data-set;  
RUN;
```

Appendix:

SAS Data Sets

- ◆ SAS data sets are specially structured data files that SAS creates and that only SAS can read. A SAS data set is displayed as a table composed of variables and observations. A SAS data set contains a descriptor portion and a data portion.
- ◆ The descriptor portion contains general information about the data set (such as the data set name and the number of observations) and information about the variable attributes (such as name, type, and length). There are two types of variables: **character** and **numeric**. A character variable can store any value and can be up to 32,767 characters long. Numeric variables store numeric values in floating point or binary representation in 8 bytes of storage by default. Other attributes include formats, informats, and labels. You can use PROC CONTENTS to browse the descriptor portion of a data set.
- ◆ The data portion contains the data values. Data values are either **character** or **numeric**. A valid value must exist for every variable in every observation in a SAS data set. **A missing value is a valid value in SAS.** A missing character value is displayed as a **blank**, and a missing numeric value is displayed as a **period**. You can specify an alternate character to print for missing numeric values using the MISSING= SAS system option. You can use PROC PRINT to display the data portion of a SAS data set.
- ◆ SAS variable and data set names must be 1 to 32 characters in length and **start with a letter or underscore**, followed by letters, underscores, and numbers. Variable names are **not case sensitive**.

/salesemps

| Last_Name | Job_Title | Salary |
|-----------|---------------|--------|
| Benny | Sales Rep. II | 26780 |
| Betschkus | Sales Rep. IV | . |
| Brown | | 26955 |
| Boltau | Sales Rep. II | 27440 |

missing values

Week 6 Topic:

VAR and SUM Statements

- ◆ You can use the VAR statement in a PROC PRINT step to subset the variables in a report. You specify the variables to include and list them in the order in which they are to be displayed.
- ◆ You can use the SUM statement in a PROC PRINT step to calculate and display report totals for the requested numeric variables.

```
PROC PRINT DATA=SAS-data-set;  
    VAR variable(s);  
    SUM variable(s);  
RUN;
```


Week 6 Topic:

WHERE Statement

- ◆ The WHERE statement in a PROC PRINT step subsets the observations in a report. When you use a WHERE statement, the output contains only the observations that meet the conditions specified in the WHERE expression. This expression is a sequence of operands and operators that form a set of instructions that define the condition. The operands can be constants or variables. Remember that variable operands must be defined in the input data set. Operators include comparison, arithmetic, logical, and special WHERE operators.

Comparison:

| Symbol(s) | Mnemonic | Definition |
|-----------|----------|--------------------------|
| = | EQ | equal to |
| ^= ^= ~= | NE | not equal to |
| > | GT | greater than |
| < | LT | less than |
| >= | GE | greater than or equal to |
| <= | LE | less than or equal to |
| | IN | equal to one of a list |

Examples

```
where Gender='M';
where Gender eq 'M';
where Salary ne .;
where Salary>50000;
where Salary lt 50000;
where Salary<=60000;
where Country in ('AU','US');
```

Week 6 Topic:

WHERE Statement (Cont.)

Arithmetic:

| Symbol | Definition |
|--------|----------------|
| ** | exponentiation |
| * | multiplication |
| / | division |
| + | addition |
| - | subtraction |

Example

```
where Salary+Bonus<=10000;
```

Logical:

WHERE *where-expression-1* AND | OR
where-expression-n;

| Symbol(s) | Mnemonic | Definition |
|-----------|----------|--------------------|
| & | AND | logical <i>and</i> |
| | OR | logical <i>or</i> |
| ^ ~ | NOT | logical <i>not</i> |

Examples

```
where Country ne 'AU' and Salary>=50000;
where Gender eq 'M' or Salary ge 50000;
where Country='AU' | Country='US';
where Country in ('AU' 'US');
where Country not in ('AU', 'US');
```

Week 6 Topic: WHERE Statement (Cont.)

Contains:

WHERE *where-expression*;

| Symbol | Mnemonic | Definition |
|--------|----------|----------------------|
| ? | CONTAINS | includes a substring |

Examples

```
where Country='AU' and
      Job_Title contains 'Rep';
```

```
where Country='AU' and
      Job_Title ? 'Rep';
```

case sensitive

Mnemonic:

| Mnemonic | Definition |
|----------------|----------------------------|
| BETWEEN-AND | an inclusive range |
| WHERE SAME AND | augment a where expression |
| IS NULL | a missing value |
| IS MISSING | a missing value |
| LIKE | matches a pattern |

| Symbol | Replaces |
|--------|--------------------------|
| % | any number of characters |
| - | one character |

- 1) Use of ~not~ to exclude
- 2) Use of ~ same and~ to augment
- 3) IS NULL and IS MISSING can be used for both numeric and character variables

Week 6 Topic:

Sorting and Grouping

- ◆ The SORT procedure sorts the observations in a data set. You can sort on one variable or multiple variables, sort on character or numeric variables, and sort in ascending or descending order. By default, SAS replaces the original SAS data set unless you use the OUT= option to specify an output data set. PROC SORT does not generate printed output.
- ◆ Every PROC SORT step must include a BY statement to specify one or more BY variables. These are variables in the input data set whose values are used to sort the data. By default, SAS sorts in ascending order, but you can use the keyword DESCENDING to specify that the values of a variable are to be sorted in descending order. When your SORT step has multiple BY variables, some variables can be in ascending and others in descending order.
- ◆ You can also use a BY statement in PROC PRINT to display observations grouped by a particular variable or variables. The groups are referred to as BY groups. Remember that the input data set must be sorted on the variables specified in the BY statement.

```
PROC SORT DATA=input-SAS-data-set  
             <OUT=ouput-SAS-data-set>;  
             BY <DESCENDING> by-variable(s);  
RUN;
```

Week 6 Topic:

ID, TITLE, FOOTNOTE Statement

- ◆ You can use the ID statement in a PROC PRINT step to specify a variable to print at the beginning of the row instead of an observation number. The variable that you specify replaces the Obs column.

```
ID variable(s);
```

- ◆ You can enhance a report by adding titles, footnotes, and column labels. Use the global TITLE statement to define up to 10 lines of titles to be displayed at the top of the output from each procedure. Use the global FOOTNOTE statement to define up to 10 lines of footnotes to be displayed at the bottom of the output from each procedure.

```
TITLEn 'text';  
FOOTNOTEn 'text';
```

- ◆ Titles and footnotes remain in effect until you change or cancel them, or until you end your SAS session. Use a null TITLE statement to cancel all titles, and a null FOOTNOTE statement to cancel all footnotes.

Week 6 Topic:

LABEL Statement

- ◆ Use the LABEL statement in a PROC PRINT step to define temporary labels to display in the report instead of variable names. Labels can be up to 256 characters in length. Most procedures use labels automatically, but PROC PRINT does not. Use the LABEL option in the PROC PRINT statement to tell SAS to display the labels. Alternatively, the SPLIT= option tells PROC PRINT to use the labels and also specifies a split character to control line breaks in column headings.

```
PROC PRINT DATA=SAS-data-set LABEL;  
    LABEL variable='label'  
          variable='label'  
          ... ;  
RUN;
```

```
SPLIT='split-character';
```

Week 6 Topic:

FORMAT Statement

- ◆ A format is an instruction that tells SAS how to display data values in output reports. You can add a **FORMAT** statement to a PROC PRINT step to specify temporary SAS formats that control how values appear in the report. There are many existing SAS formats that you can use. Character formats begin with a dollar sign, but numeric formats do not.

FORMAT *variable(s) format;*

- ◆ SAS stores date values as the number of days between January 1, 1960, and a specific date. To make the dates in your report recognizable and meaningful, you must apply a SAS date format to the SAS date values.

Week 6 Topic:

FORMAT Statement Examples

| Format | Stored Value | Displayed Value |
|-----------|--------------|-----------------|
| MMDDYY6. | 0 | 010160 |
| MMDDYY8. | 0 | 01/01/60 |
| MMDDYY10. | 0 | 01/01/1960 |
| DDMMYY6. | 365 | 311260 |
| DDMMYY8. | 365 | 31/12/60 |
| DDMMYY10. | 365 | 31/12/1960 |

| Format | Stored Value | Displayed Value |
|------------|--------------|-----------------|
| \$4. | Programming | Prog |
| 12. | 27134.5864 | 27135 |
| 12.2 | 27134.5864 | 27134.59 |
| COMMA12.2 | 27134.5864 | 27,134.59 |
| DOLLAR12.2 | 27134.5864 | \$27,134.59 |
| COMMAX12.2 | 27134.5864 | 27.134,59 |
| EUROX12.2 | 27134.5864 | €27.134,59 |

| Format | Definition |
|-----------|--|
| \$w. | writes standard character data. |
| w.d | writes standard numeric data. |
| COMMAw.d | writes numeric values with a comma that separates every three digits and a period that separates the decimal fraction. |
| DOLLARw.d | writes numeric values with a leading dollar sign, a comma that separates every three digits, and a period that separates the decimal fraction. |
| COMMAXw.d | writes numeric values with a period that separates every three digits and a comma that separates the decimal fraction. |
| EUROXw.d | writes numeric values with a leading euro symbol (€), a period that separates every three digits, and a comma that separates the decimal fraction. |

Week 6 Topic:

User Defined FORMAT Statement

- ◆ You can create your own user-defined formats. When you create a user-defined format, you don't associate it with a particular variable or data set. Instead, you create it based on values that you want to display differently. The formats will be available for the remainder of your SAS session. You can apply user-defined formats to a specific variable in a PROC PRINT step.
- ◆ You use the FORMAT procedure to create a format. You assign a format name that can have up to 32 characters. The name of a character format must begin with a dollar sign, followed by a letter or underscore, followed by letters, numbers, and underscores. Names for numeric formats must begin with a letter or underscore, followed by letters, numbers, and underscores. A format name cannot end in a number and cannot be the name of a SAS format.
- ◆ You use a VALUE statement in a PROC FORMAT step to specify the way that you want the data values to appear in your output. You define value-range sets to specify the values to be formatted and the formatted values to display instead of the stored value or values. The value portion of a value-range set can include an individual value, a range of values, a list of values, or a keyword. The keyword OTHER is used to define a value to display if the stored data value does not match any of the defined value-ranges.
- ◆ When you define a numeric format, it is often convenient to use numeric ranges in the value-range sets. Ranges are inclusive by default. To exclude the endpoints, use a less-than symbol after the low end of the range or before the high end.
- ◆ The LOW and HIGH keywords are used to define a continuous range when the lowest and highest values are not known. Remember that for character values, the LOW keyword treats missing values as the lowest possible values. However, for numeric values, LOW does not include missing values.

```
PROC FORMAT;  
  VALUE format-name value-or-range1='formatted-value1'  
                                value-or-range2='formatted-value2'  
                                ...;  
RUN;
```

Week 6 Topic:

Using VALUE Statement

Using the VALUE Statement

```
PROC FORMAT;  
  VALUE format-name value-or-range1= 'formatted-value1 '  
                                value-or-range2= 'formatted-value2 '  
                                ...;  
RUN;
```

value-range sets

| | <i>value-or-range</i> | = | <i>formatted-value</i> |
|---------|-----------------------|---|------------------------|
| value → | 'AU' 1 | = | 'Australia' |
| range → | 'B'-'D' 0-50000 | = | 'Tier 1' |
| list → | 'U','V' 1,2,3 | = | 'Below 49.9' |

Week 6 Topic:

Creating new datasets

- ◆ You use a DATA step to create a new SAS data set from an existing SAS data set. The DATA step begins with a DATA statement, which provides the name of the SAS data set to create. Include a SET statement to name the existing SAS data set to be read in as input.
- ◆ You use the WHERE statement to subset the input data set by selecting only the observations that meet a particular condition. To subset based on a SAS date value, you can use a SAS date constant in the WHERE expression. SAS automatically converts a date constant to a SAS date value.

```
DATA output-SAS-data-set;  
  SET input-SAS-data-set;  
  WHERE where-expression;  
RUN;
```

- ◆ By default, the SET statement reads all of the observations and variables from the input data set and writes them to the output data set. You can customize the new data set by selecting only the observations and variables that you want to include. You can use a WHERE statement to select the observations, as long as the variables included in the condition come from the input data set. You can use a DROP statement to list the variables to exclude from the new data set, or use a KEEP statement to list the variables to include. If you use a KEEP statement, you must include every variable to be written, including any new variables.

```
DROP variable-list;  
KEEP variable-list;
```

Week 6 Topic:

Creating new datasets (cont.)

- ◆ You can subset the original data set with a WHERE statement for variables that are defined in the input data set, and a subsetting IF statement for new variables that are created in the DATA step. Remember that, although IF expressions are similar to WHERE expressions, you cannot use special WHERE operators in IF expressions.

IF expression;

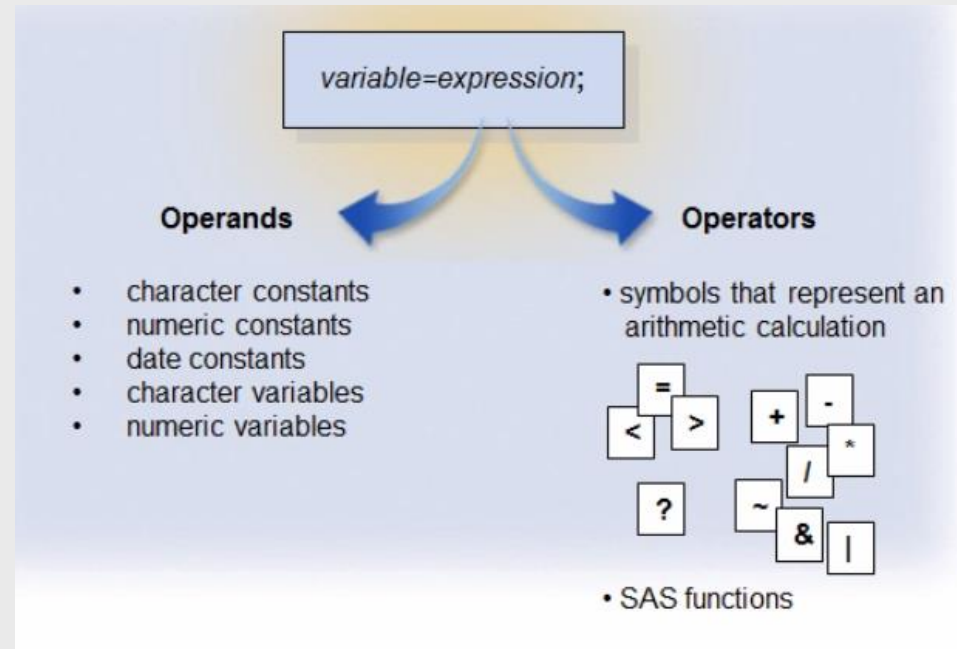
- ◆ To subset observations in a PROC step, you must use a WHERE statement. You cannot use a subsetting IF statement in a PROC step. To subset observations in a DATA step, you can always use a subsetting IF statement. However, a WHERE statement can make your DATA step more efficient because it subsets on input.
- ◆ When you use the LABEL statement in a DATA step, SAS permanently associates the labels to the variables by storing the labels in the descriptor portion of the data set. Using a FORMAT statement in a DATA step permanently associates formats with variables. The format information is also stored in the descriptor portion of the data set. You can use PROC CONTENTS to view the label and format information. PROC PRINT does not display permanent labels unless you use the LABEL or SPLIT= option.

Week 6 Topic:

Assignment Statement

- ◆ You use an assignment statement to create a new variable. The assignment statement evaluates an expression and assigns the resulting value to a new or existing variable. The expression is a sequence of operands and operators. If the expression includes arithmetic operators, SAS performs the numeric operations based on priority, as in math equations. You can use parentheses to clarify or alter the order of operations.

variable=expression;



Week 6 Topic:

Assignment Statement Examples

- ◆ Watch for order of execution:

| Symbol | Definition | Priority |
|--------|----------------|----------|
| ** | exponentiation | I |
| * | multiplication | II |
| / | division | II |
| + | addition | III |
| - | subtraction | III |

| Example | Type |
|------------------------------|-----------------------|
| Salary=26960; | numeric constant |
| Gender='F'; | character constant |
| Hire_Date='21JAN1995'd; | date constant |
| Bonus=Salary*.10; | arithmetic expression |
| BonusMonth=month(Hire_Date); | SAS function |

Week 6 Topic:

PROC CLUSTER – general format

General form of the CLUSTER procedure:

```
PROC CLUSTER DATA=SAS-data-set  
  METHOD=method <options>;  
  VAR variables;  
  FREQ variable;  
  RMSSTD variable;  
RUN;
```

The required METHOD= option specifies the hierarchical technique to be used to cluster the observations.

Week 6 Topic:

PROC CLUSTER – example options

Our example PROC CLUSTER considered is:

```
PROC CLUSTER <options>;  
ID ;  
VAR var1 var2 var3 ... varn;  
RUN;
```

Here the options control the printing, computational, and output of the procedures:

- ◆ ID variable identify observations. If the ID statement is omitted, each observation is denoted by OBn , where n is the observation number.
- ◆ NOPRINT - suppresses any printed output,
- ◆ NOEIGEN - suppresses printing of eigenvalues, specifying the NOEIGEN option saves time if the number of variables is large,
- ◆ SIMPLE - produces simple descriptive statistics for each variable,
- ◆ METHOD = - controls the clustering method used (required option), we will use CENTROID
- ◆ STANDARD - Uses the correlation matrix for computation,
- ◆ RMSSTD - displays the root-mean-square standard deviation of each cluster.
- ◆ RSQUARE - displays the R^2 and semipartial R^2 to evaluate cluster solution.
- ◆ NONORM - prevents the distances from being normalized to unit mean or unit root mean square with most methods.
- ◆ OUTTREE = -create an output dataset for cluster diagrams.

The VAR statement lists the variables to be considered as responses.

Week 6 Topic:

PROC FASTCLUS - overview

- ◆ By default, the FASTCLUS procedure uses Euclidean distances, so the cluster centers are based on least squares estimation. This kind of clustering method is often called a k-means model, since the cluster centers are the means of the observations assigned to each cluster when the algorithm is run to complete convergence. Each iteration reduces the least squares criterion until convergence is achieved.
- ◆ Often there is no need to run the FASTCLUS procedure to convergence. PROC FASTCLUS is designed to find good clusters (but not necessarily the best possible clusters) with only two or three passes through the data set.
- ◆ The initialization method of PROC FASTCLUS guarantees that, if there exist clusters such that all distances between observations in the same cluster are less than all distances between observations in different clusters, and if you tell PROC FASTCLUS the correct number of clusters to find, it can always find such a clustering without iterating. Even with clusters that are not as well separated, PROC FASTCLUS usually finds initial seeds that are sufficiently good that few iterations are required. Hence, by default, PROC FASTCLUS performs only one iteration.
- ◆ The initialization method used by the FASTCLUS procedure makes it sensitive to outliers. PROC FASTCLUS can be an effective procedure for detecting outliers because outliers often appear as clusters with only one member.

Week 6 Topic:

PROC FASTCLUS – general format

- ◆ General form of the FASTCLUS procedure:

```
PROC FASTCLUS DATA=SAS-data-set  
                <MAXC=>|<RADIUS=><options>;  
    VAR variables;  
RUN;
```

- ◆ Because PROC FASTCLUS produces relatively little output, it is often a good idea to create an output data set, and then use other procedures such as PROC MEANS, PROC SGPLOT, PROC DISCRIM, or PROC CANDISC to study the clusters.

Week 6 Topic:

PROC FASTCLUS – example options

Our example for PROC FASTCLUS considered is:

```
PROC FASTCLUS MAXCLUSTERS=n RADIUS=t <options>;  
VAR var1 var2 var3 ... varn;  
RUN;
```

Here the options control the printing, computational, and output of the procedures:

- ◆ ID variable identify observations. If the ID statement is omitted, each observation is denoted by OBn , where n is the observation number.
- ◆ MAXCLUSTERS= n specifies the maximum number of clusters permitted. If you omit the MAXCLUSTERS= option, a value of 100 is assumed
- ◆ RADIUS= t establishes the minimum distance criterion for selecting new seeds. No observation is considered as a new seed unless its minimum distance to previous seeds exceeds the value given by the RADIUS= option. The default value is 0.
- ◆ REPLACE= specifies seed replacement method. If you specify the REPLACE=RANDOM option, the RADIUS= option is ignored.
- ◆ MAXITER= specifies maximum number of iterations
- ◆ DISTANCE displays distances between cluster centers
- ◆ LIST displays cluster assignments for all observations

The VAR statement lists the variables to be considered as responses.

Week 8 Topic:

FASTCLUS vs CLUSTER

| | # of Clusters | # of Observations | Takes Matrix as input | Model | Sensitivity |
|---------------|----------------|-----------------------|-----------------------|-------------------------|-----------------------------------|
| PROC CLUSTER | <i>small k</i> | <i>small data set</i> | Yes | Hierarchical Clustering | # of Observations, # of Variables |
| PROC FASTCLUS | 20~100 | >100 | No | K-means Clustering | Outliers, Order of Observations |

Notes:

- The time required by PROC FASTCLUS is roughly proportional to the number of observations, whereas the time required by PROC CLUSTER with most methods varies with the square or cube of the number of observations. Therefore, you can use PROC FASTCLUS with much larger data sets than PROC CLUSTER.
- If you want to hierarchically cluster a data set that is too large to use with PROC CLUSTER directly, you can have PROC FASTCLUS produce, for example, 50 clusters, and let PROC CLUSTER analyze these 50 clusters instead of the entire data set.

Week 8 Topic:

FASTCLUS vs CLUSTER (cont.)

From <http://analytics.ncsu.edu/sesug/2010/SDA10.Reiss.pdf>

Table 2 summarizes some of the advantages, disadvantages, and overall differences of both clustering methods as applied to this project.

Table 2: PROC FASTCLUS vs. PROC CLUSTER

| | PROC FASTCLUS | PROC CLUSTER |
|----------------------|--|---|
| Method | distance-based disjoint | hierarchical |
| Steps | <ul style="list-style-type: none">• Arbitrarily choose k observations as the "seeds"• Assign all other observations to their closest "seed"• Update the cluster mean and re-define the center | <ul style="list-style-type: none">• Each observation begins in a cluster by itself• The two closest clusters are merged to form a new cluster• Merging continues until only one cluster is left |
| Advantages | <ul style="list-style-type: none">• Faster• Can handle large datasets | <ul style="list-style-type: none">• Produces a tree visualization• Provides more process details |
| Disadvantages | <ul style="list-style-type: none">• Must specify number of clusters• Different seeds = different results | <ul style="list-style-type: none">• Slower• Not suitable for larger datasets• Missing value imputation necessary |

Week 8 Topic:

Analysis Considerations I

- ◆ **# of Observations:** The FASTCLUS procedure is intended for use with large data sets, with 100 or more observations. With small data sets, the results can be highly sensitive to the order of the observations in the data set.
- ◆ **Outliers:** Most cluster solutions are affected heavily by presence of outliers and/or observations that are just too different from the others. These observations can also indicate potential business opportunities. You could subset the data using the WHERE statement in your DATA step. The initialization method used by the FASTCLUS procedure makes it sensitive to outliers. PROC FASTCLUS can be an effective procedure for detecting outliers because outliers often appear as clusters with only one member.
- ◆ **Standardization:** Before using PROC FASTCLUS, decide whether your variables should be standardized in some way, since variables with large variances tend to have more effect on the resulting clusters than those with small variances. If all variables are measured in the same units, standardization might not be necessary. Otherwise, some form of standardization is strongly recommended. The STANDARD procedure can standardize all variables to mean zero and variance one. PROC FASTCLUS uses algorithms that place a larger influence on variables with larger variance, so it might be necessary to standardize the variables before performing the cluster analysis.

Week 8 Topic:

Analysis Considerations II

- ◆ **Convergence:** If PROC FASTCLUS runs to complete convergence, the final cluster seeds will equal the cluster means or cluster centers. If PROC FASTCLUS terminates before complete convergence, which often happens with the default settings, the final cluster seeds might not equal the cluster means or cluster centers. If you want complete convergence, specify CONVERGE=0 and a large value for the MAXITER= option.
- ◆ PROC FASTCLUS always selects the first complete (no missing values) observation as the first seed. The next complete observation that is separated from the first seed by at least the distance specified in the RADIUS= option becomes the second seed. Later observations are selected as new seeds if they are separated from all previous seeds by at least the radius, as long as the maximum number of seeds is not exceeded.

Week 8 Topic:

Analysis Considerations III - CCC

- ◆ **CCC:** The best way to use the CCC is to plot its value against the number of clusters, ranging from one cluster up to about one-tenth the number of observations. The CCC may not behave well if the average number of observations per cluster is less than ten. The following guidelines should be used for interpreting the CCC:
 - Peaks on the plot with the CCC greater than 2 or 3 indicate good clusterings.
 - Peaks with the CCC between 0 and 2 indicate possible clusters but should be interpreted cautiously.
 - There may be several peaks if the data has a hierarchical structure.
 - Very distinct nonhierarchical spherical clusters usually show a sharp rise before the peak followed by a gradual decline.
 - Very distinct nonhierarchical elliptical clusters often show a sharp rise to the correct number of clusters followed by a further gradual increase and eventually a gradual decline.
 - If all values of the CCC are negative and decreasing for two or more clusters, the distribution is probably unimodal or long-tailed.
 - Very negative values of the CCC, say, -30, may be due to outliers. Outliers generally should be removed before clustering and their removal documented.
- ◆ If the CCC increases continually as the number of clusters increases, the distribution may be grainy or the data may have been excessively rounded or recorded with just a few digits. A final and very important warning: neither the CCC nor R^2 is an appropriate criterion for clusters that are highly elongated or irregularly shaped. If you do not have prior substantive reasons for expecting compact clusters, use a nonparametric clustering method such as Wong and Lane's (1983) rather than Ward's method or k-means clustering.