IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

February 9,11 2016

Week 5 Presentation

Week 5 Topic: Agenda

- Assignment Q&A
- Cluster Analysis
- Data Smart Example on Wine

Week 5 Topic: Pothole Repair Analysis

We review pothole repair budget and service request metrics data from the past 4 years (2011 ~ 2014) to determine whether enough budget was appropriated to meet the Target Response Days of 7.

- Pothole Repair (PR) data was acquired from data.gov. Category information did not match year to year. Some assumptions were made to gather the budget amounts which are documented in the Excel workbook.
- Weather data (average temperatures and # of snow days) was gathered from wunderground.com for the 4 years.
- Main assumptions for analysis include:
  - Only Asphalt related budget amounts are considered for the analysis
  - Only labor costs are considered as majority of the cost for pothole repairs is assumed to be labor related

Best use of Friday TA sessions and Discussion Topics

Fridays:

- Be ready to show questions/problems on laptop or other means
- TA will have example analysis on hand. If you need to schedule a call/google hangout for a review of this, contact the TA for scheduling.
- Good question/answer session from Fridays will be posted for all in Discussion Topics.

Discussion Topics:

- It is best to show examples of the help you are seeking. It is difficult to recreate your problem especially if it has to do with debugging of software.
- It should be more for long term discussions on general how-to tips and tricks. What it is not meant for is last minute, immediate answer needed situations an hour before your assignment is due.

Week 5 Topic:
Feedback on Week 3 Zip Code

Point deduction pointers - the following items incur deductions:

1) Unsorted data, especially when showing ranking.
2) Unformatted numbers – text, charts, graphs, and tables.
3) More than a few grammar issues. Always do one last grammar check.
4) Unorganized/Unformatted slides. Watch for titles, bullets, font sizing, inserts, and spacing of all of the above. Unorganized information is negative information.
5) Using inappropriate statistics. There should be a reason for Sums and Counts and inclusion of such calculations. Do not include grand totals in charts & graphs. Default pivot calculation is SUM, don't leave it as is unless it makes sense to do so. E.g., sum of zip code means nothing
6) Excel copy/paste that looks like Excel with tool bars, grid lines, etc.
7) Any list/rows of data, whether it be bulleted on a slide or in a table, that are quoted as having a certain number of items or rows, any list that represents a ranking, etc. should be numbered.
8) Always label charts, graphs, axis, and tables with appropriate words related to the analysis. No generic terms without context e.g., histogram, counts, frequency, totals, etc.

Week 5 Topic: Feedback on Week 3 Zip Code (cont.)

The following are also important for keeping points:

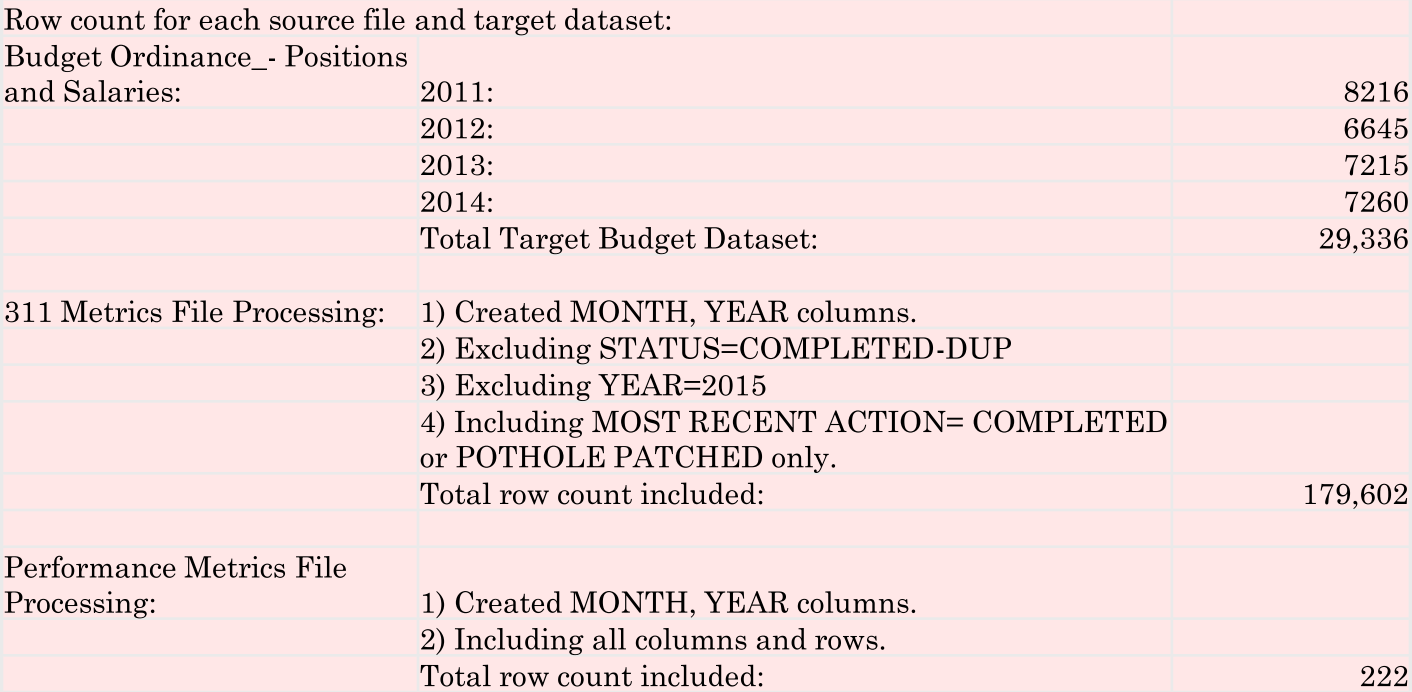
- ◆ Watch for phrasing and tone of the presentation:
  - ▪ “Maximum of” should be “Highest number of …” or “Most number of…”
  - ▪ Try to avoid using “We” or “I” or “They” in sentences. Less use of pronouns the better
  - ▪ Use factual statements. No conversations or slang. Remember you are presenting not trying to have a discussion with the reader.
- ◆ Proper use of Histograms (metric comparisons, rankings), Pie Charts (percentage breakouts), and Frequency Distribution Charts (counting of categorical data or binning of continuous numeric variables)

Administrative:

- ◆ If asked to post PPT and XLS files, always post PPT first followed by the XLS file(s).
- ◆ Amendment to the zipping rule, you can zip XLS file(s) – ONLY. For those that are wondering, XLS previews are generally not available in Blackboad.
- ◆ Still, do not zip PPT files so previews can be used in Blackboard.
- ◆ Clear previous submissions if you submit multiple times. (?)

Week 5 Topic:
Feedback on Week 4 Pothole Repair I

- As mentioned in the announcement on (2/8), the Excel file submitted did NOT follow directions as stated in class. Only submit data that is used for the analysis.
- The following states how many rows to be expected for unfiltered data. In general, if you perform the same tasks:



Row count for each source file and target dataset:		
Budget Ordinance_- Positions and Salaries:	2011:	8216
	2012:	6645
	2013:	7215
	2014:	7260
	Total Target Budget Dataset:	29,336
311 Metrics File Processing:	1) Created MONTH, YEAR columns.	
	2) Excluding STATUS=COMPLETED-DUP	
	3) Excluding YEAR=2015	
	4) Including MOST RECENT ACTION= COMPLETED or POTHOLE PATCHED only.	
	Total row count included:	179,602
Performance Metrics File Processing:	1) Created MONTH, YEAR columns.	
	2) Including all columns and rows.	
	Total row count included:	222

Week 5 Topic: Presentation Pointers (*revisited*)

- **Actionable** – analyst does the analysis, not the reader. Information, not data, needs to be presented in a way that is easy to understand. Decisions should be made from presentation of data.
  - Listed data – is a presentation of just data
  - Sorted or Ranked data – is information that can be acted upon

- **Anticipate** – what the reader will ask, think, and need. Don't just stop with what you have, think of what might be asked and do the next ten steps of analysis.
  - Derive data if further analysis needs it
  - Get additional data if the next question to be asked needs it
  - Be creative

- **Purposeful** – any graph, chart, dataset that you present needs to have purpose. Don't waste the reader's time if there is no decision to be made from the analysis. Ask the "so what" in the analysis.

Week 5 Topic: Presentation Pointers (*revisited*)

- **Simple** – thoughtful presentation of information is always simple. Analysis can be complicated but this does not mean the presentation has to be. Spend the time to simplify the presentation. Don't mask the decisions in complexity
  - One chart, graph, pivot, or table per slide usually works better
  - Organize the information so that there is a *flow*

- **Annotate** – always explain what the reader is reading. The reader shouldn't have to guess.
  - Add an intro/summary sentence to a chart, graph, or table
  - Titles, labels, legends, etc. all should be labeled
  - Freeze columns and rows as needed to make sure the reader knows when scrolling

Week 5 Topic: Pothole Spending Analysis examples

- What position incurs the most salary and how does it compare?
- How does the weather affect budget appropriations? Does it?
- Is the city able to meet it's goal of responding within 7 days?
- Is there a seasonality to the repair response times?
- What percentage of calls result in pothole repairs? How are the calls addressed if not addressed by pothole repairs?
- What percentage of calls are completed? Are there any that are not and why?
- Can you guess, applying some metric, what the projected 2016 budget would be?
- Looking at the whole Transportation budget, what percentage is attributed to pothole repair?
- Is there a location/geographic dependency/pattern to pothole repairs?

Week 5 Topic: Assignment 3

1) Read Chapter 2 of Data Smart – we will review the wine example on Thursday.
2) Case Study Part 1:
   a) Collect datasets (no need to submit raw data)
   b) Create master workbook with merged, 2011 ~ 2014 data, into one worksheet. Then, add transportation metrics, 311, and weather information, as necessary, into the same workbook.
   c) Document data validations, manipulations, derivations, assumptions, & observations made while collecting and integrating data.
   d) Document metadata.

Week 5 Topic: Assignment 4

Pothole Assignment Part 2

Presentation portion of your pothole repair spending assignment:

a) Intro/Background (1-2 slides)
b) Data Processing/Metadata (1-2 slides)
c) Analysis/Findings (3-5 slides)
d) Conclusion (1-2 slides)

Use template already provided.

Week 4 Topic Review: 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 Review: 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.

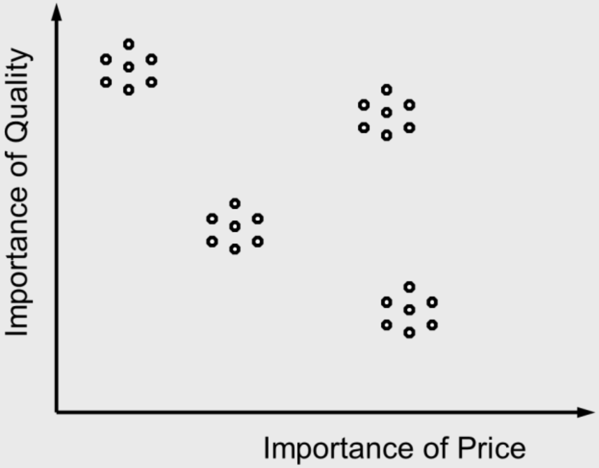
Week 4 Topic Review:
Further definitions - Cluster Analysis

- *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.

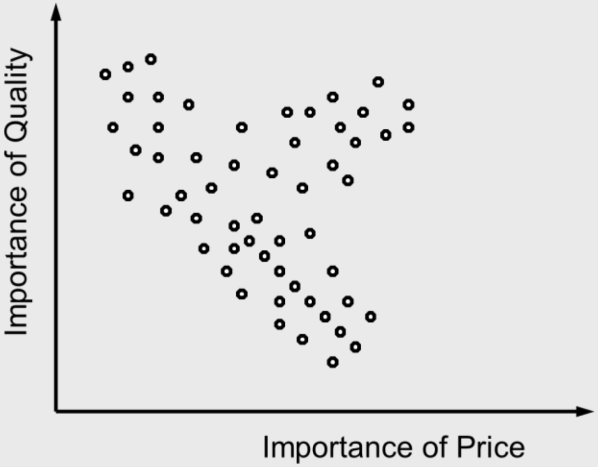
Week 4 Topic Review: Cluster Analysis – Grouping

- 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.

Ideal Pattern:



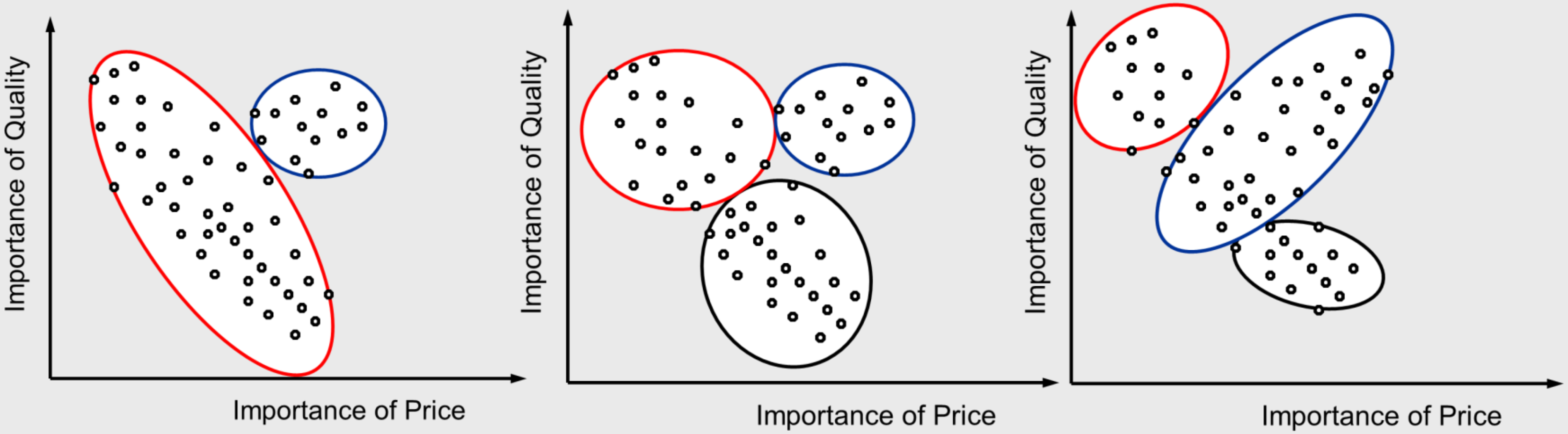
Realistic Pattern:



ITM - 527

Week 4 Topic Review: 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 Review: 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 Review: 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 Review: 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 Review: 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 Review: 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 Review: 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 Review: 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 5 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 5 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.

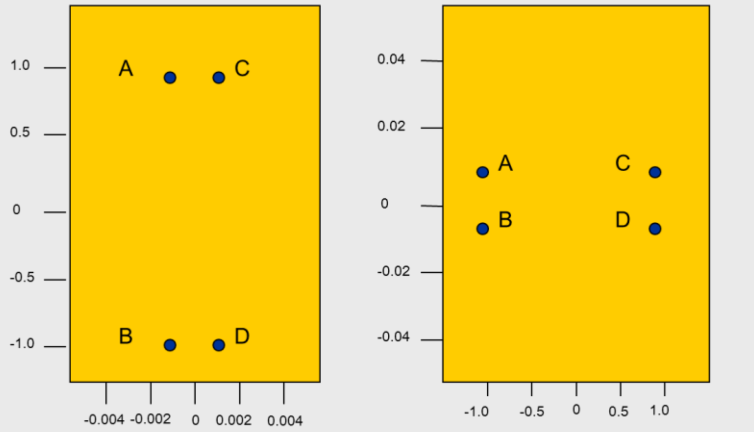
Week 5 Topic: Limitations of K-means

- 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?

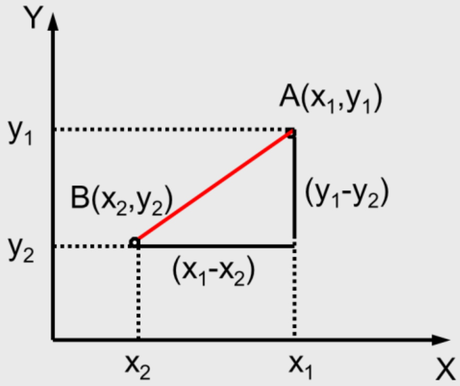


The Data points (A, B, C, and D) are the same but X and Y axes are scaled differently in the two plots!

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.

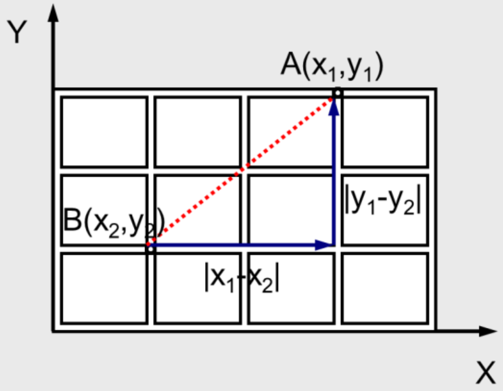
Assume two observations, A and B, in a two-dimensional space have coordinates (x_1,y_1) and (x_2,y_2).



Squared Euclidean Distance between A and B,

$(D_{AB})^2 = (x_1\text{-}x_2)^2 + (y_1\text{-}y_2)^2$ 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,

$D_{AB} = |x_1\text{-}x_2| + |y_1\text{-}y_2|$

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 $\sqrt{(9-5)^2+(8-9)^2}$= 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.

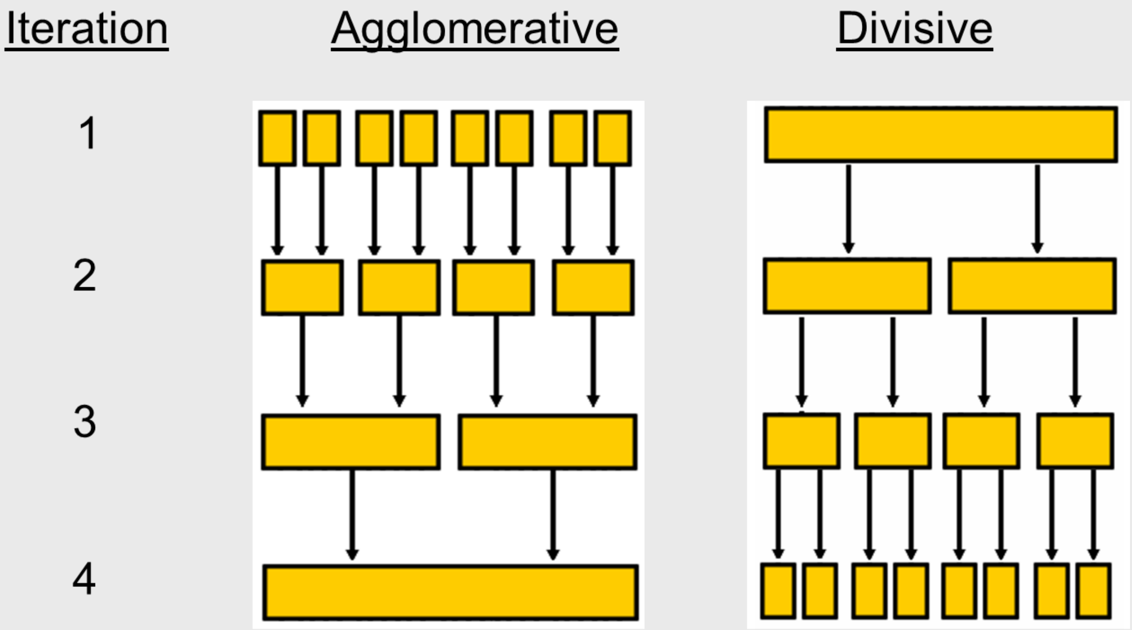
Customer	Price	Quality	Service	Single	Married	Divorced
A	9	9	7	1	0	0
B	5	9	8	0	1	0
C	6	8	9	0	0	1

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.

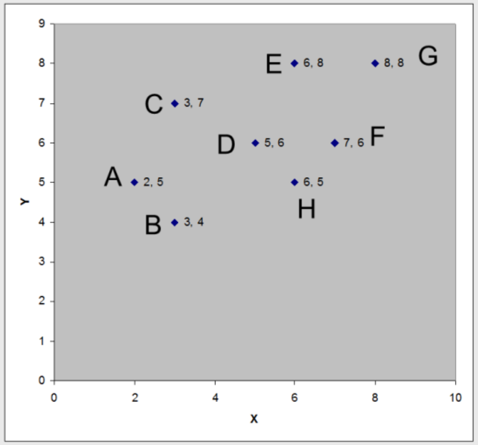


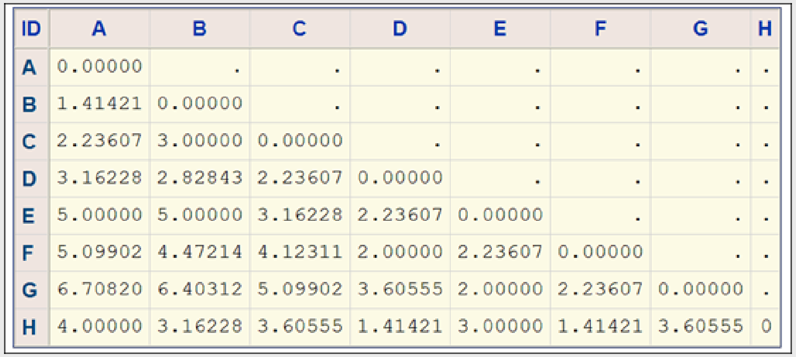
Week 5 Topic: Hierarchical Clustering - Example

- 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 (2.0000) 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

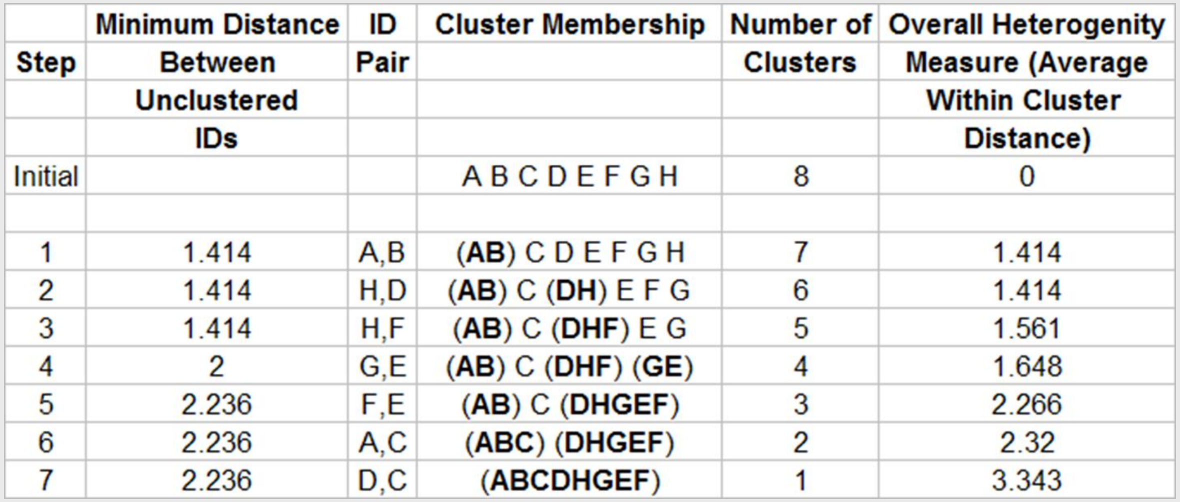




ID	A	B	C	D	E	F	G	H
A	0.00000	.	.	.	.	.	.	.
B	1.41421	0.00000	.	.	.	.	.	.
C	2.23607	3.00000	0.00000	.	.	.	.	.
D	3.16228	2.82843	2.23607	0.00000	.	.	.	.
E	5.00000	5.00000	3.16228	2.23607	0.00000	.	.	.
F	5.09902	4.47214	4.12311	2.00000	2.23607	0.00000	.	.
G	6.70820	6.40312	5.09902	3.60555	2.00000	2.23607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

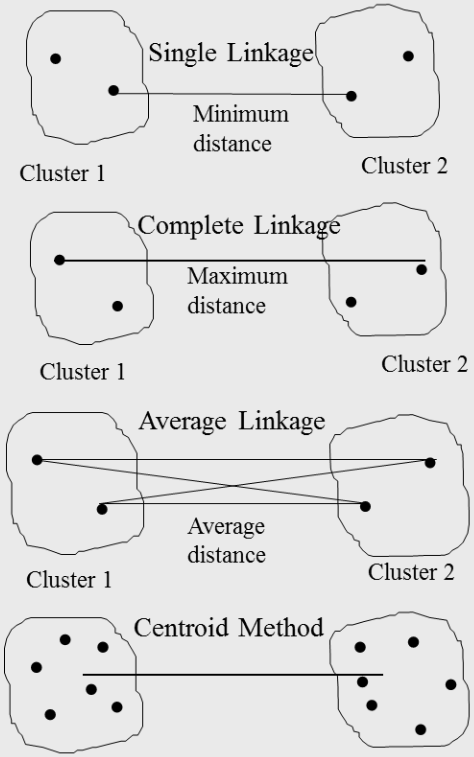


Step	Minimum Distance Between Unclustered IDs	ID Pair	Cluster Membership	Number of Clusters	Overall Heterogenity Measure (Average Within Cluster Distance)
Initial			A B C D E F G H	8	0
1	1.414	A,B	**(AB)** C D E F G H	7	1.414
2	1.414	H,D	**(AB)** C **(DH)** E F G	6	1.414
3	1.414	H,F	**(AB)** C **(DHF)** 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

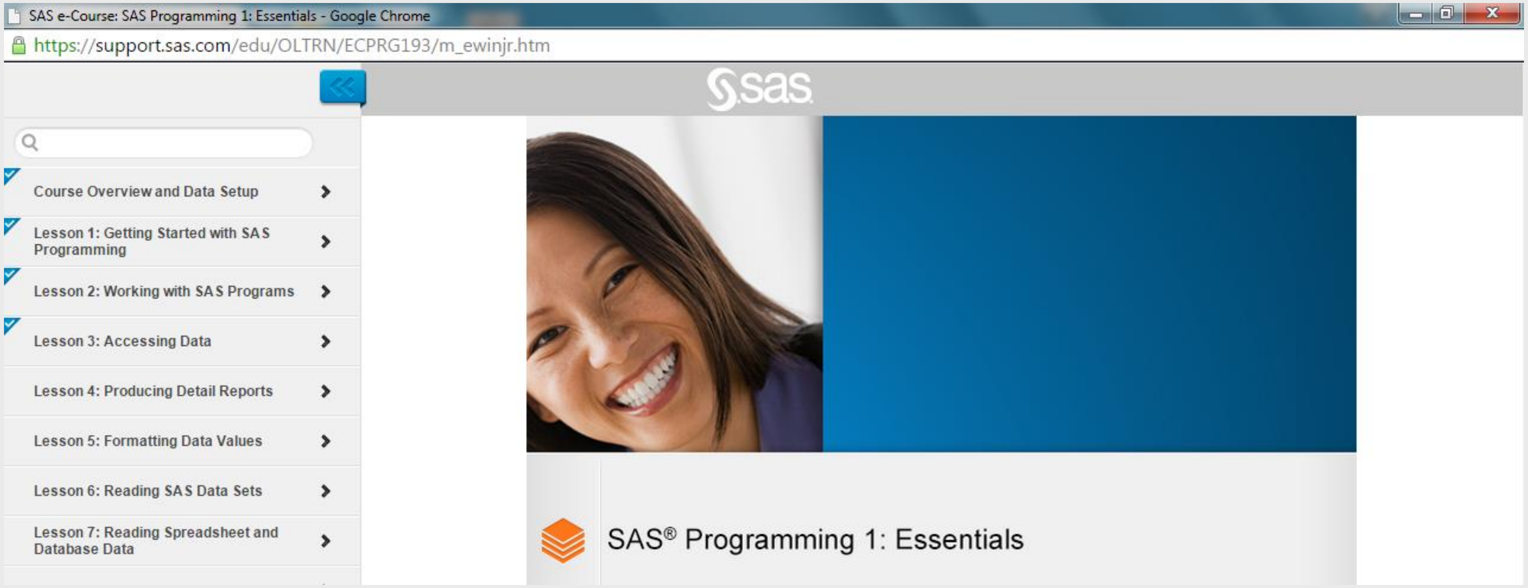
There are many approaches within agglomerative methods in measuring similarity between clusters when one or both clusters have multiple members.

- 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: Reading Assignments

#	Reading
1	Read Chapter 2 in Data Smart
2	Take Lessons 1 ~ 3 in SAS Programming Essentials



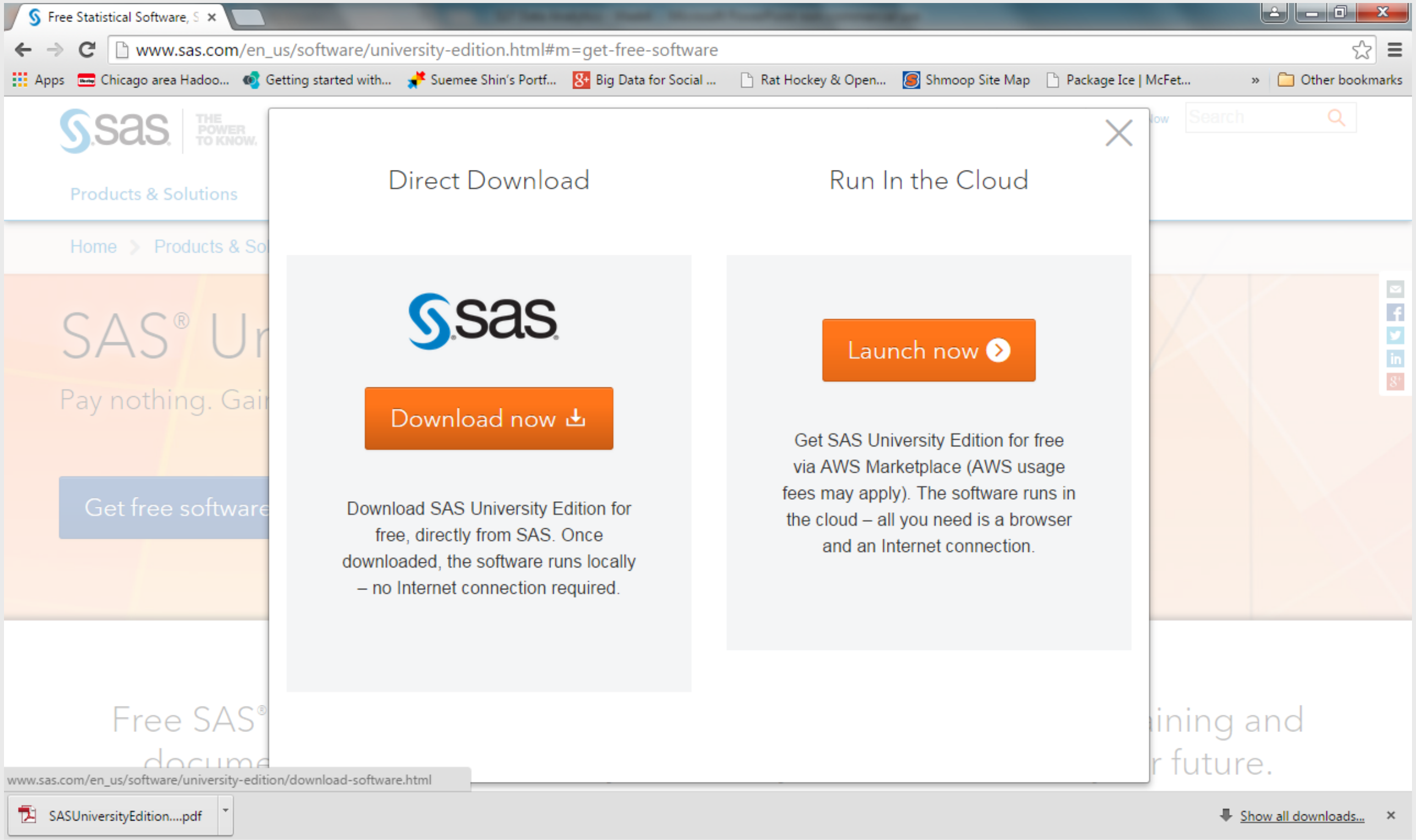
SAS e-Course: SAS Programming 1: Essentials - Google Chrome

https://support.sas.com/edu/OLTRN/ECPRG193/m_ewinjr.htm

sas
Course Overview and Data Setup
Lesson 1: Getting Started with SAS Programming
Lesson 2: Working with SAS Programs
Lesson 3: Accessing Data
Lesson 4: Producing Detail Reports
Lesson 5: Formatting Data Values
Lesson 6: Reading SAS Data Sets
Lesson 7: Reading Spreadsheet and Database Data
SAS® Programming 1: Essentials

Appendix:
Installing SAS University Edition

- Go to: http://www.sas.com/en_us/software/university-edition.html
- Create a student account. You will get free software and e-learning for one year.

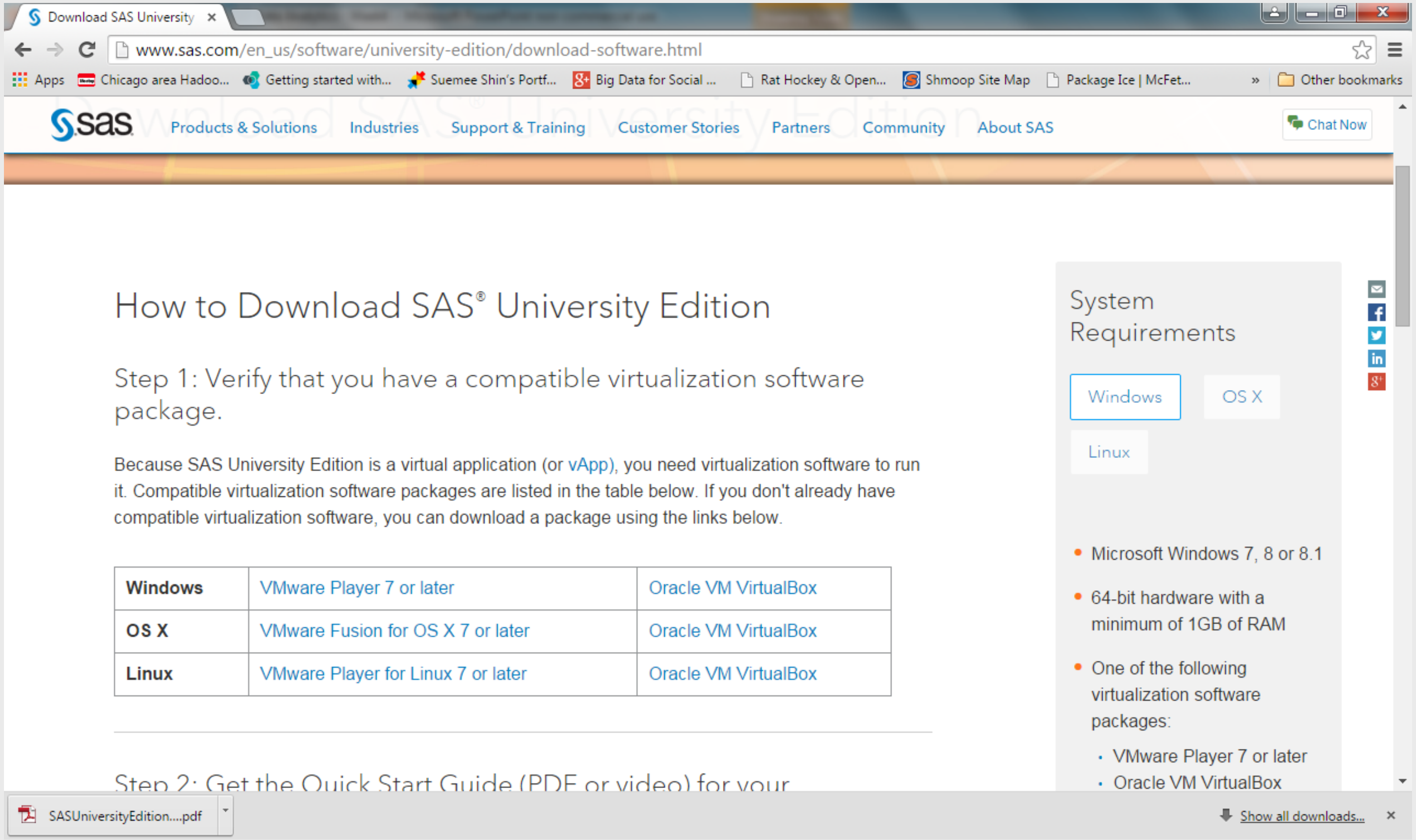

Free Statistical Software, S ×
www.sas.com/en_us/software/university-edition.html#m=get-free-software
Apps | Chicago area Hadoo... | Getting started with... | Suemee Shin's Portf... | Big Data for Social ... | Rat Hockey & Open... | Shmoop Site Map | Package Ice | McFet... | Other bookmarks
sas THE POWER TO KNOW.
Products & Solutions
Home > Products & So
SAS® Ur
Pay nothing. Gai
Get free software
Free SAS® docume
aining and r future.
Direct Download
Run In the Cloud
Download now
Launch now
www.sas.com/en_us/software/university-edition/download-software.html
SASUniversityEdition....pdf
Show all downloads...

Download SAS University Edition for free, directly from SAS. Once downloaded, the software runs locally – no Internet connection required.

Get SAS University Edition for free via AWS Marketplace (AWS usage fees may apply). The software runs in the cloud – all you need is a browser and an Internet connection.

Appendix: Getting the right download file

◆ Select your system specific download files. Recommend using VirtualBox for Mac users.



How to Download SAS® University Edition

Step 1: Verify that you have a compatible virtualization software package.

Because SAS University Edition is a virtual application (or vApp), you need virtualization software to run it. Compatible virtualization software packages are listed in the table below. If you don't already have compatible virtualization software, you can download a package using the links below.

Windows	VMware Player 7 or later	Oracle VM VirtualBox
OS X	VMware Fusion for OS X 7 or later	Oracle VM VirtualBox
Linux	VMware Player for Linux 7 or later	Oracle VM VirtualBox

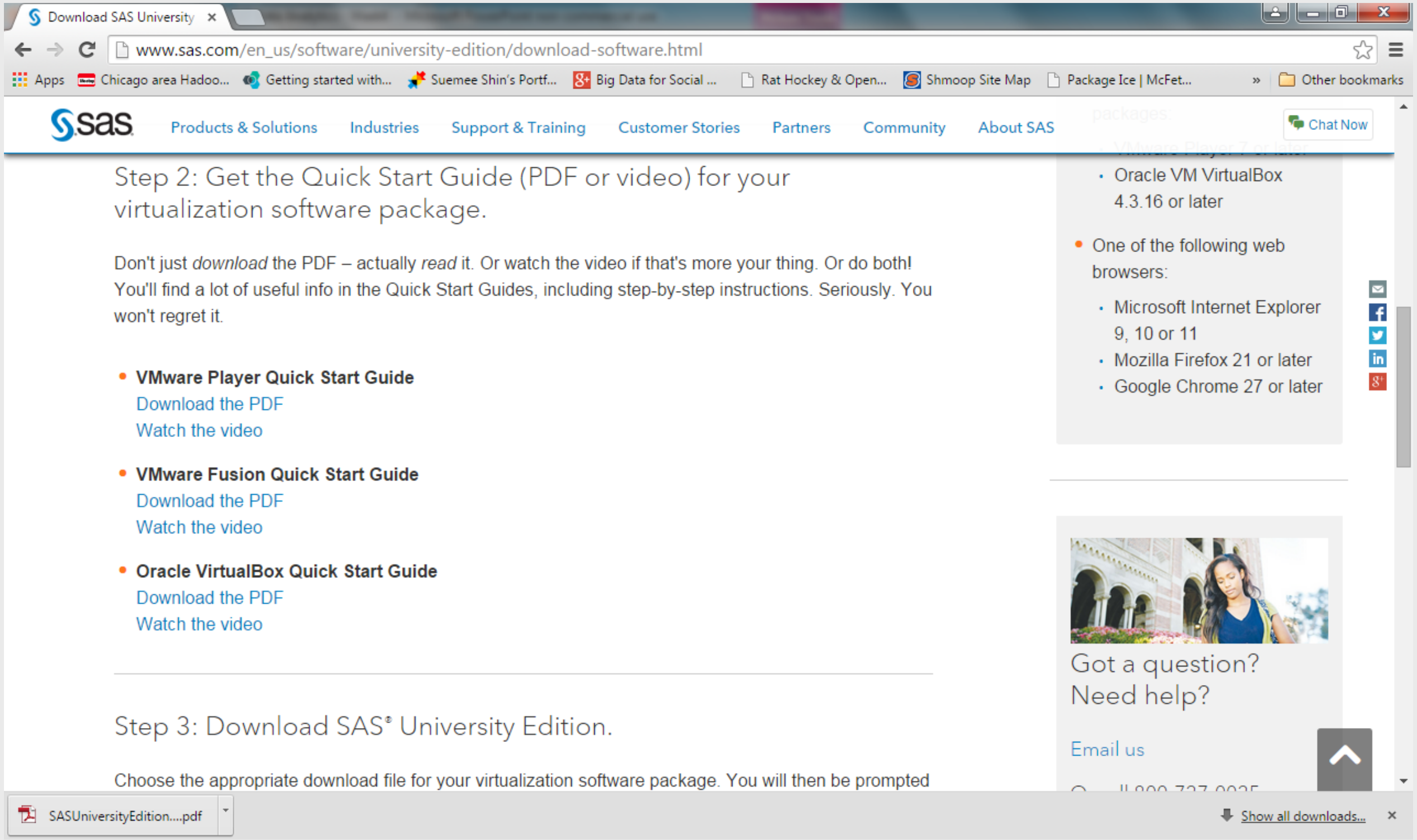
Step 2: Get the Quick Start Guide (PDF or video) for your

System Requirements

- Microsoft Windows 7, 8 or 8.1
- 64-bit hardware with a minimum of 1GB of RAM
- One of the following virtualization software packages:
  - VMware Player 7 or later
  - Oracle VM VirtualBox

Appendix: Follow installation directions

- Follow directions in the Quick Start Guide PDF



Step 2: Get the Quick Start Guide (PDF or video) for your virtualization software package.

Don't just *download* the PDF – actually *read* it. Or watch the video if that's more your thing. Or do both! You'll find a lot of useful info in the Quick Start Guides, including step-by-step instructions. Seriously. You won't regret it.

- **VMware Player Quick Start Guide**
  Download the PDF
  Watch the video
- **VMware Fusion Quick Start Guide**
  Download the PDF
  Watch the video
- **Oracle VirtualBox Quick Start Guide**
  Download the PDF
  Watch the video

Step 3: Download SAS® University Edition.

Choose the appropriate download file for your virtualization software package. You will then be prompted

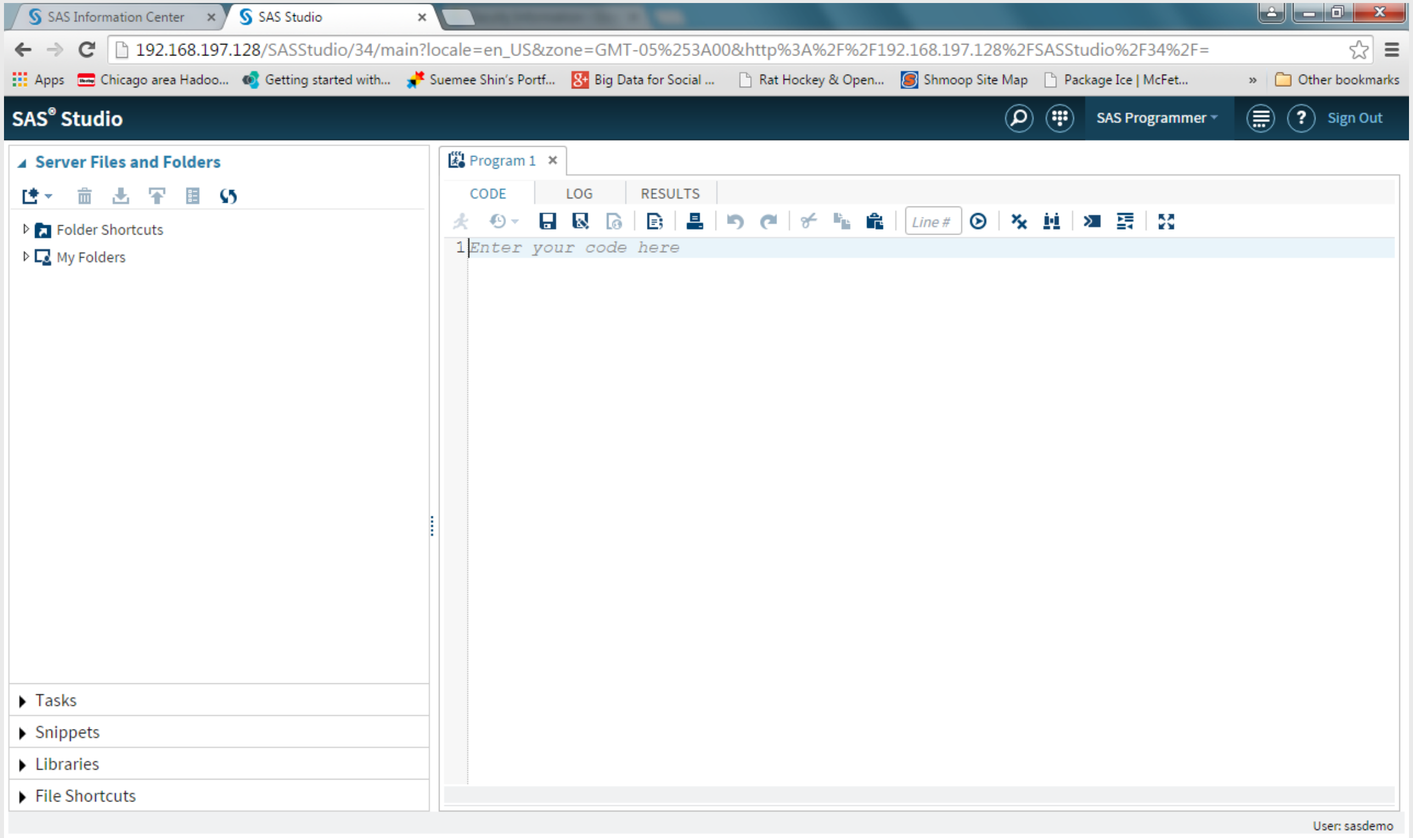
- Oracle VM VirtualBox 4.3.16 or later
- One of the following web browsers:
  - Microsoft Internet Explorer 9, 10 or 11
  - Mozilla Firefox 21 or later
  - Google Chrome 27 or later

Got a question? Need help?

Email us

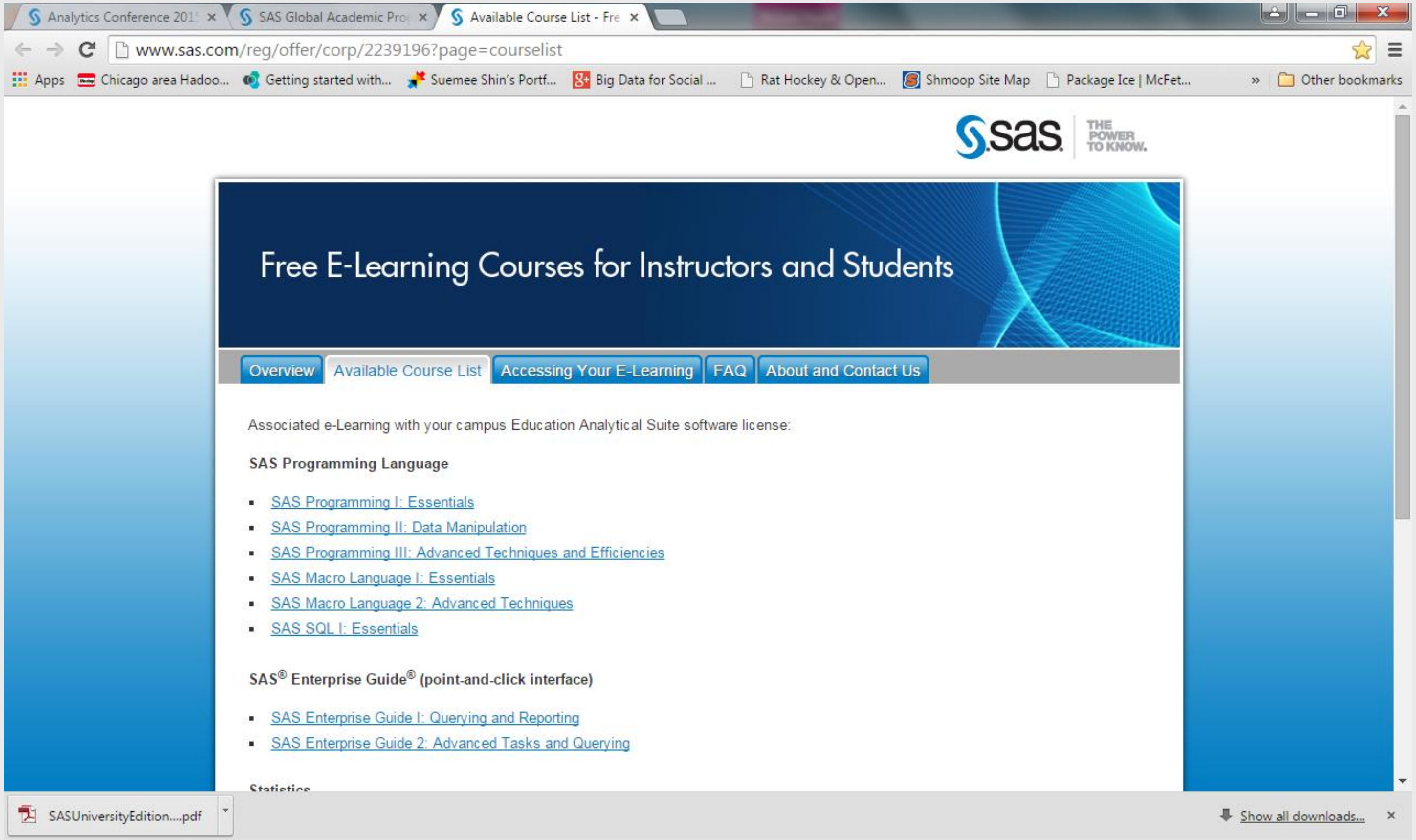
Appendix: SAS Studio – start window

- Get to the point of when you can open a start window:



Appendix: SAS Programming I Essentials

- Start on SAS Programming I Essentials e-learning course as time allows. Ideally, we'll want to finish this course before Week 8:



Free E-Learning Courses for Instructors and Students

Associated e-Learning with your campus Education Analytical Suite software license:

SAS Programming Language

- SAS Programming I: Essentials
- SAS Programming II: Data Manipulation
- SAS Programming III: Advanced Techniques and Efficiencies
- SAS Macro Language I: Essentials
- SAS Macro Language 2: Advanced Techniques
- SAS SQL I: Essentials

SAS® Enterprise Guide® (point-and-click interface)

- SAS Enterprise Guide I: Querying and Reporting
- SAS Enterprise Guide 2: Advanced Tasks and Querying

Statistics

Appendix:
SAS Installation of University Edition

Available Tools and Functions:

- ◆ Site name: 'UNIVERSITY EDITION 2.2 9.4M3 WITH ETS FOR PLAYER'.
- ◆ Expiration: 15JUN2016.
- ◆ Product expiration dates:
- ◆ ---Base SAS Software 15JUN2016 (CPU A)
- ◆ ---SAS/STAT 15JUN2016 (CPU A)
- ◆ ---SAS/ETS 15JUN2016 (CPU A)
- ◆ ---SAS/IML 15JUN2016 (CPU A)
- ◆ ---SAS/ACCESS Interface to PC Files 15JUN2016 (CPU A)
- ◆ ---SAS/IML Studio 15JUN2016 (CPU A)
- ◆ ---SAS Workspace Server for Local Access 15JUN2016 (CPU A)
- ◆ ---SAS Workspace Server for Enterprise Access 15JUN2016 (CPU A)
- ◆ ---High Performance Suite 15JUN2016 (CPU A)

Appendix: SAS Framework and File Types

SAS framework:

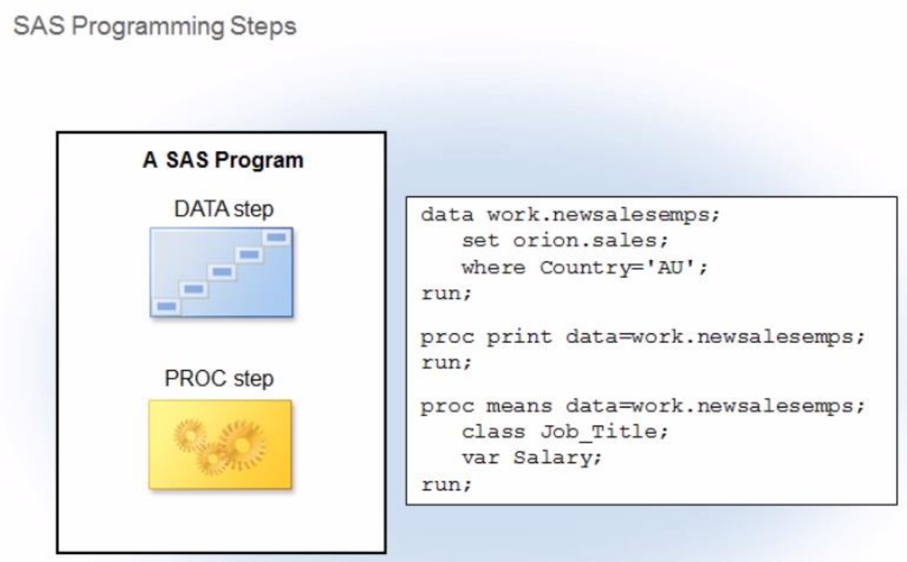
- **Access data**: Using SAS, you can read any kind of data.
- **Manage data**: SAS gives you excellent data management capabilities
- **Analyze data**: For statistical analysis, SAS is the gold standard.
- **Present data**: You can use SAS to present your data meaningfully.

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.

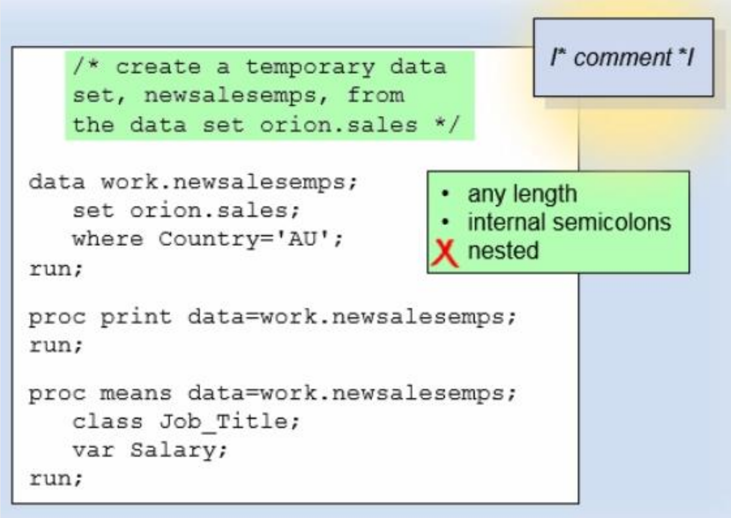


```
data work.newsalesemps;
   set orion.sales;
   where Country='AU';
run;

proc print data=work.newsalesemps;
run;

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

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*;**

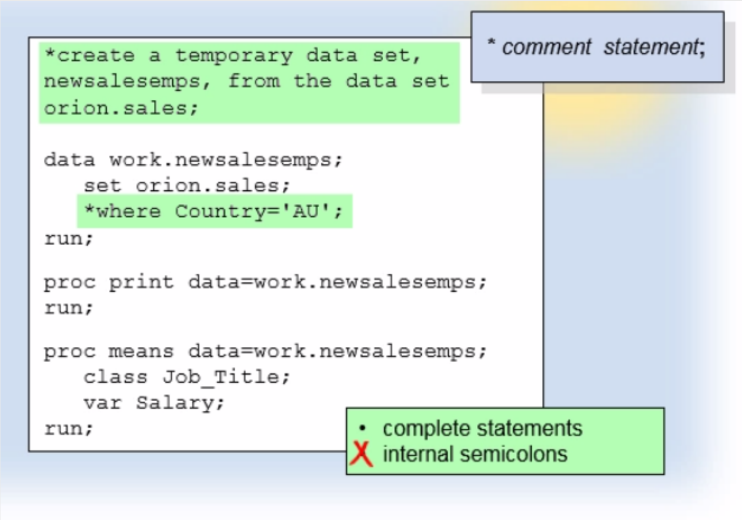


```
/* create a temporary data
set, newsalesemps, from
the data set orion.sales */

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

✗ 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;
```

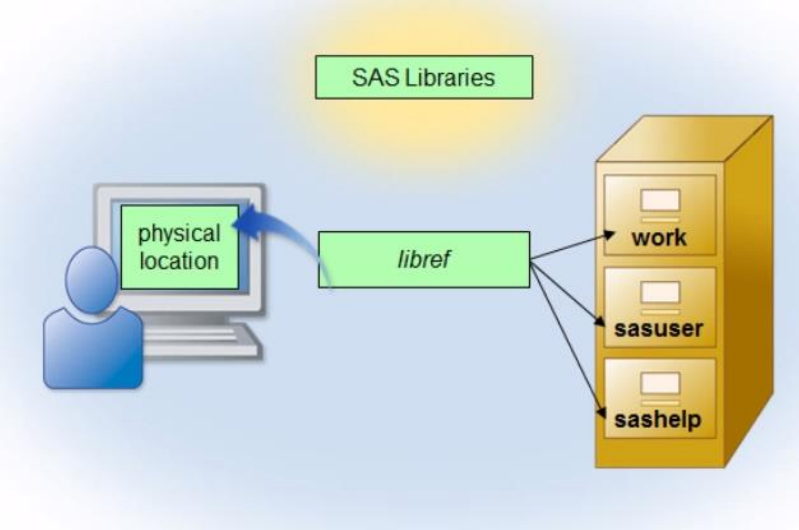
** comment statement*;

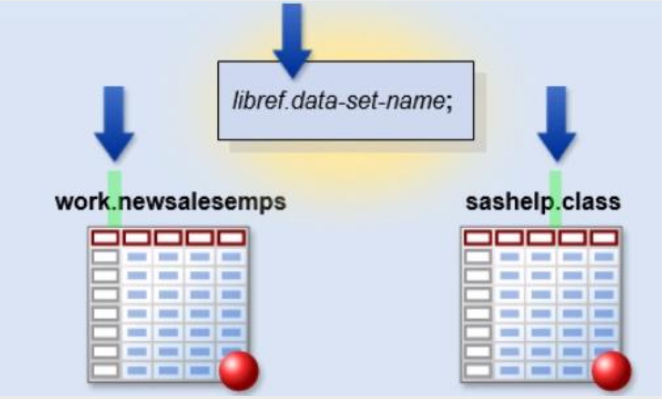
- complete statements

✗ 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 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;
```

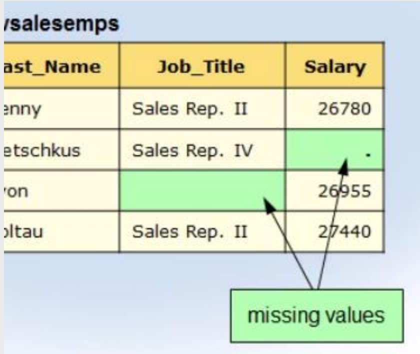
```
PROC CONTENTS DATA=libref.SAS-data-set;
RUN;
```

- 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

ast_Name	Job_Title	Salary
enny	Sales Rep. II	26780
etschkus	Sales Rep. IV	.
on		26955
oltau	Sales Rep. II	27440

missing values