SSC/ N 9004: Provide data/information in standard formats

Session Overview

The Associate Analytics*Provide data/information in standard formats* module is designed to help participants understand the standard operating procedures in organizations pertaining to reporting data in a logical sequence and arriving at conclusive decisions models after analysis of data. This module is aimed at developing the sense of understanding in an individual when the individual works with data, of how to take the data and present it as relevant information in standardized formats.

Participants learn how to share information with other people inside or outside a specified work group and also how to arrive at decisions regarding certain problem types.

Session Goal

The primary goal of the session is for the participants to analyze data and present it in a suitable format, as is suitable for the given process or organization.

Successful candidates will be able to understand the process of standardized reporting and the nuances of a publishing a report with a specified end objective in mind.

Session Objectives

Upon completion of these parts of this course, the participants will be able to:

PC1. establish and agree with appropriate people the data/information you need to provide, the formats in which you need to provide it, and when you need to provide it

PC2. obtain the data/information from reliable sources

PC3. check that the data/information is accurate, complete and up-to-date

PC4. obtain advice or guidance from appropriate people where there are problems with the data/information

PC5. carry out rule-based analysis of the data/information, if required

PC6. insert the data/information into the agreed formats

PC7. check the accuracy of your work, involving colleagues where required

PC8. report any unresolved anomalies in the data/information to appropriate people

PC9. provide complete, accurate and up-to-date data/information to the appropriate people in the required formats on time

Reading Database using R

We can import Datasets from various sources having various files types for example,

- .csv format
- Big data tool – Impala
- **CSV File**

The sample data can also be in comma separated values (CSV) format. Each cell inside such data file is separated by a special character, which usually is a comma, although other characters can be used as well. The first row of the data file should contain the column names instead of the actual data. Here is a sample of the expected format.

```
Col1,Col2,Col3
100,a1,b1
200,a2,b2
300,a3,b3
```

After we copy and paste the data above in a file named "mydata.csv" with a text editor, we can read the data with the function read.csv.

[CODE]

```
> mydata = read.csv("mydata.csv")  # read csv file
> mydata
```

```
  Col1 Col2 Col3
1  100   a1   b1
2  200   a2   b2
3  300   a3   b3
```

In various European locales, as the comma character serves as the decimal point, the functionread.csv2 should be used instead. For further detail of the read.csv and read.csv2 functions, please consult the R documentation.

> help(read.csv)

- **Big data tool – Impala**

Cloudera 'Impala', which is a massively parallel processing (MPP) SQL query engine runs natively in Apache Hadoop.

R package, **RImpala**, connects Impala to R.

- **Creating New variables:-**

Use the assignment operator "<-" to create new variables.

For example,

mydata$sum <- mydata$x1 + mydata$x2

New variable is created using two already available variables.

- **Modifying existing variable:-**

We can rename the existing variable by *rename()* function.
For examples,
mydata<- rename(mydata, c(oldname="newname"))

We can also recode variables in R.
For example,
If we want to rename variable based on some criteria like below
mydata$agecat<- ifelse(mydata$age> 70, c("older"), c("younger"))

- Create a new variable Total_Sales using variables Sales_1 and Sales_2.
- Then, modify the above new variable name from Total_Sales to Sales_Total.

Outliers and Missing Data treatment

Inputting missing data using standard methods and algorithmic approaches (mice package R):

- In **R**, missing values are represented by the symbol **NA** (not available).
- Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number).
- Unlike SAS, **R** uses the same symbol for character and numeric data.

To test if there is any missing in the dataset we use *is.na ()* function.

For Example,
We have defined "y" and then checked if there is any missing value. T or True means that there is a missing value.
y <- c(1,2,3,NA)
is.na(y)
returns a vector (F FF T)

For example,

total <- rbind(data frameA, data frameB)

Note:-

If data frameA has variables that data frameB does not, then either:

1. Delete the extra variables in data frameA or
2. Create the additional variables in data frameB and set them to NA (missing)

before joining them with rbind( ).

We use **cbind()** function to combine data by column the syntax is same as **rbind().**

Plyr package: Tools for Splitting, Applying and Combining Data.

We use rbind.fill() in plyr package in R. It binds or combines a list of data frames filling missing columns with NA.
For example,
rbind.fill(mtcars[c("mpg", "wt")], mtcars[c("wt", "cyl")])

In this all the missing value will be filled with NA.

Discuss Function and Loops

Using for and ifelse in R :

"FOR" Loop:-

To repeat an action for every value in a vector by using a "for" loop.
We construct a "for" loop in R as follows:

for(i in values){

... do something ...

}

This for loop consists of the following parts:

- The keyword for, followed by parentheses.
- An identifier between the parentheses. In this example, we use i, but that can be any object name you like.
- The keyword in, which follows the identifier.
- A vector with values to loop over. In this example code, we use the object values, but that again can be any vector you have available.
- A code block between braces that has to be carried out for every value in the object values.

In the code block, you can use the identifier. Each time R loops through the code, R assigns the next value in the vector with values to the identifier.

19. Workplace gossip
20. Exercise
21. Needless interruptions
22. Defining contribution
23. Aimless Internet surfing
24. Irrelevant phone calls

Suggested Answers:

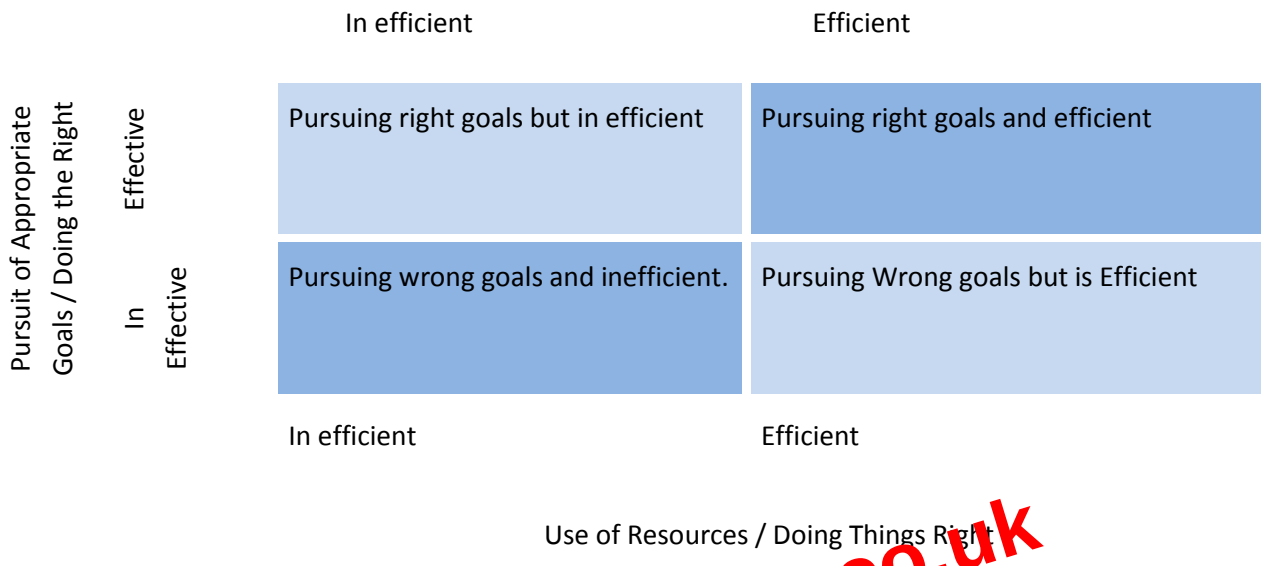
Depends on rationale shared

1. Wildly important goal – Q1
2. Last minute assignments from boss – Q1
3. Busy work – Q4 – Consumes time however not pressing
4. Personal health – Q4 – requires planning and care not pressing
5. Pressing problems – Q1 – has to be solved immediately
6. Crises – Q1 – have to tended to immediately
7. Planning – Q2 – Important but not urgent; should be done before crisis
8. Time wasters – Q4
9. Professional development – Q2
10. Win-win performance agreement – Q2 – Expectation setting part of planning
11. Too many objectives – Q3 – Prioritize further to establish which are important and pressing
12. Vital customer call – Q1 – Customer centricity
13. Major Deadlines – Q1
14. Unimportant pre scheduled meetings – Q3
15. Meaningless management reports – Q3 – Prioritize further to establish which are important and pressing
16. Coaching and mentoring team – Q2
17. Low priority email – Q3 – Prioritize further to establish which are important and pressing
18. Other people's minor issues – Q3 – May not be urgent but important for team building
19. Workplace gossip – Q4 – Non value add; occasionally creates negativity
20. Exercise – Q4 – Important for health and personal well being. To be done in spare and leisure time. Cannot be ignored.
21. Needless interruptions – Q3
22. Defining contribution – Q2
23. Aimless Internet surfing – Q4
24. Irrelevant phone calls – Q4 – Reserve and avoid

Summary

- It is important to manage time.
- To manage time one must:
  - Prioritize
  - Define Urgency
  - Define Importance

Efficiency vs Effectiveness



Service Level Agreements

Service Level Agreement (SLA) is a contract between a service provider and its internal or external customers that documents what services the provider will furnish

SLA measures the service provider's performance and quality in a number of ways.

Some sample metrics SLAs may specify or include

- Availability and uptime – the percentage of the time services will be available
- The no of users being served, the bandwidth or volume being addressed or the quantum f work being performed in work units
- Specify performance benchmarks to which actual performance will be periodically compared
- Turnaround time

In addition to establishing performance metrics, an SLA may include a plan for addressing downtime and documentation for how the service provider will compensate customers in the event of a contract breach. SLAs, once established, should be periodically reviewed and updated to reflect changes in technology and the impact of any new regulatory directives

its data in their studies because of its poor overall performance.

Historically, the third and fourth standardized moments (skewness and kurtosis) were some of the earliest tests for normality. The Jarque–Bera test is itself derived fromskewness and kurtosis estimates. Mardia's multivariate skewness and kurtosis tests generalize the moment tests to the multivariate case. Other early test statistics include the ratio of the mean absolute deviation to the standard deviation and of the range to the standard deviation.

More recent tests of normality include the energy test (Székely and Rizzo) and the tests based on the empirical characteristic function (ecf) (e.g. Epps and Pulley, Henze–Zirkler, BHEP test). The energy and the ecf tests are powerful tests that apply for testing univariate or multivariate normality and are statistically consistent against general alternatives.

The normal distribution has the highest entropy of any distribution for a given standard deviation. There are a number of normality tests based on this property, the first attributable to Vasicek.

3. **Bayesian tests:**

Kullback–Leibler divergences between the whole posterior distributions of the slope and variance do not indicate non-normality. However, the ratio of expectations of these posteriors and the expectation of the ratios give similar results to the Shapiro–Wilk statistic except for very small samples, when non-informative priors are used.

Spiegelhalter suggests using a Bayes factor to compare normality with a different class of distributional alternatives. This approach has been extended by Farrell and Rogers-Stewart.

Facilitator Preparation

Responsibilities
✓ **Review examples provided: reflect on your own experiences and determine when to share them.** ✓ **Review all material – Facilitator Guide, Presentation, Guides and Handouts (if any)** ✓ **Make sure you have copies of all the handouts.** ✓ **Make sure the learning resources are loaded on your computer.** ✓ **Conduct a run through of the content. Conduct a dress rehearsal of the session as you move through the content. Make sure you are comfortable with the tools and interactions recommended in the facilitator guide.** ✓ **Note that all examples are in italics to emphasize key learning points; however, you may use your own professional experience to enhance the learning.** ✓ **Make sure you create folders for all breakout activities.**

Topic: Team Work

Team Work

Ask participants to share their thoughts on:

- What is team work?
- How is it more advantageous?

What is a Team?

A team comprises a group of people linked in a common purpose.

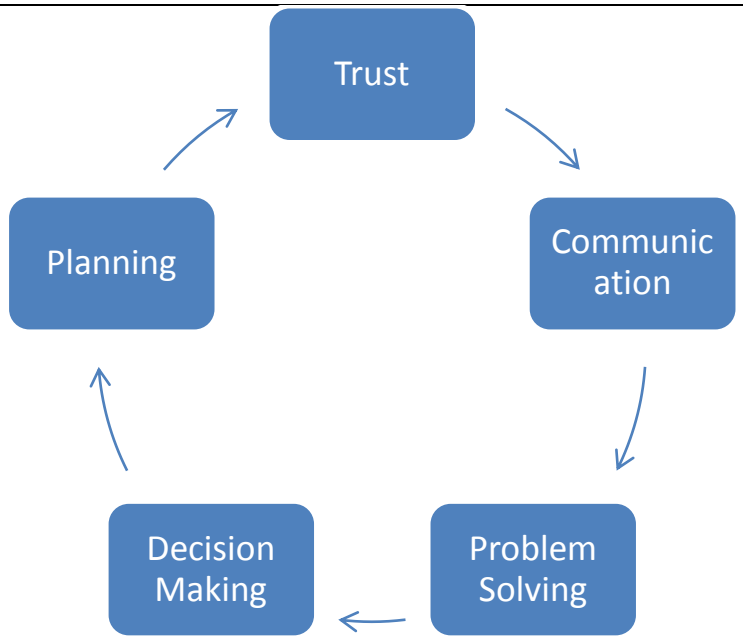
Teams are especially appropriate for conducting tasks that are high in complexity and have many interdependent subtasks.

Team Work

Coming together is a beginning, keeping together is progress and working together is success. A team is a number of people associated together in work or activity. In a good team members create an environment that allows everyone to go beyond their limitation.

Why do we need teamwork – The overriding need of all people working together for the same organization is to make the organization profitable.

Team Building



Team work vs. Individual work

- Team Work

  - Agree on goals/ milestones
  - Establish tasks to be completed
  - Communicate / monitor progress
  - Solve Problem
  - Interpret Results
  - Agree completion of projects

- Individual work

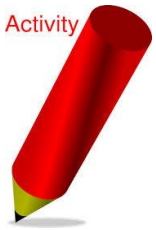
  - Work on tasks
  - Work on new / revised tasks

Team Development

Key Points

Importance of Professionalism

Provide a brief overview of the session. Discuss the importance of professional behavior in the organization.



Activity Description:

Ask the candidates "What does professionalism mean to you?"

Take a few minutes and write down your thoughts… as a definition or description.

Summarize

Who is professional?

A person who has achieved an acclaimed level of proficiency in any trade and whose competencies can be measured against fixed set of standards or guidelines. The following are the key characteristics in a person that make him stand out as a professional.

- **Positively proactive.** Professionals demonstrate behaviours that are positive, proactive instead of negative, and reactive.
- **Respect.** Through this ethic and value of respect, professionals are known and trusted within and without their respective organizations.
- **Opportunities to help others.** Those who avow before understand they have a responsibility to help others whether it is to grow self-leadership skills or provide some expert advice.
- **Follow-up.** No one likes to wait for un-returned phone calls or emails. Professionals make it a habit to follow-up on everything and accept responsibility when they fail to engage in that behavior.
- **Empathy.** Professionals know how to be empathetic. This characteristic is a one of the signs of high emotional intelligence and a predictor for leadership success.
- **Self-confident.** When individuals are self confident, they do not have to put others down at their own expense. These individuals have a high sense of balanced self-esteem and role awareness.
- **Sustainable.** Professionals are truly sustainable in that they can continue forward when times become difficult. Their ethics and beliefs keep them focused.
- **Integrity.** Integrity is putting your values into action; doing the right thing when no one else is looking without personal gain or benefit; and accepting a potential personal cost.
- **Optimize all interactions.** This is critical because professionals do not negate the value of people. They look to see how one interaction can benefit someone else even before himself or herself.

- Also, trying to read something into every movement others make can get in the way of effective interactions.

Forms of non verbal communication

1. **Ambulation** is the way one walks. Whether the person switches, stomps, or swaggers can indicate how that person experiences the environment.
2. **Touching** is possibly the most powerful nonverbal communication form. People communicate trust, compassion, tenderness, warmth, and other feelings through touch. Also, people differ in their willingness to touch and to be touched. Some people are "touchers" and others emit signals not to touch them.
3. **Eye contact** is used to size up the trustworthiness of another. Counselors use this communication method as a very powerful way to gain understanding and acceptance. Speakers use eye contact to keep the audience interested.
4. **Posturing** can constitute a set of potential signals that communicate how a person is experiencing the environment. It is often said that a person who sits with his/her arms folded and legs crossed is defensive or resistant. On the other hand, the person may just be cold.
5. **Tics** are involuntary nervous spasms that can be a key to indicate one is being threatened. For example, some people stammer or jerk when they are threatened. But these mannerisms can easily be misinterpreted.
6. **Sub-vocals** are the non-words one says, such as "ugh" or "um." They are used when one is trying to find the right word. People use a lot of non-words trying to convey a message to another person. Another example is the use of "you know." It is used in place of the "ugh" and other grunts and groans commonly used.
7. **Distancing** is a person's psychological space. If this space is invaded, one can become somewhat tense, alert, or "jammed up." People may try to move back to reestablish their personal space. The kind of relationship and the motives toward one another determines this personal space.
8. **Gesturing** carries a great deal of meaning between people, but different gestures can mean different things to the sender and the receiver. This is especially true between cultures. Still,

gestures are used to emphasize our words and to attempt to clarify our meaning.

9. **Vocalism** is the way a message is packaged and determines the signal that is given to another person. For example, the message, "I trust you," can have many meanings. "I trust you" could imply that someone else does not. "I trust you" could imply strong sincerity. "I trust you" could imply that the sender does not trust others.

Written Communication

Written communication involves any type of message that makes use of the written word. Written communication is the most important and the most effective of any mode of business communication.

Examples of written communications generally used with clients or other businesses include email, Internet websites, letters, proposals, telegrams, faxes, postcards, contracts, advertisements, brochures, and news releases.

Advantages and disadvantages of written communication:

Advantages

- Creates permanent record
- Allows to store information for future reference
- Easily distributed
- All recipients receive the same information
- Written communication helps in laying down apparent principles, policies and rules for running on an organization.
- It is a permanent means of communication. Thus, it is useful where record maintenance is required.
- Written communication is more precise and explicit
- Effective written communication develops and enhances organization's image
- It provides ready records and references
- Written communication is more precise and explicit.
- Effective written communication develops and enhances an organization's image
- Necessary for legal and binding documents

Disadvantages of Written Communication

- Written communication does not save upon the costs. It costs huge in terms of stationery and the manpower employed in writing/typing and delivering letters.

1. **The use of jargons.** Over Over-complicated, unfamiliar and/or technical terms.
2. **Emotional barriers and taboos.** Some people may find it difficult to express their emotions and some topics may be completely 'off-limits' or taboo.
3. **Lack of attention, interest, distractions, or irrelevance to the receiver.**
4. **Differences in perception and viewpoint.**
5. **Physical disabilities** such as hearing problems or speech difficulties.
6. **Physical barriers to non verbal communication.** Not being able to see the non-verbal cues, gestures, posture and general body language can make communication less effective. Accents.
7. **Language differences and the difficulty in understanding unfamiliar accents.**
8. **Expectations and prejudices which may lead to false assumptions or stereotyping.** People often hear what they expect to hear rather than what is actually said and jump to incorrect conclusions.
9. **Cultural differences.** The norms of social interaction vary greatly in different cultures, as do the way in which emotions are expressed. For example, the concept of personal space varies between cultures and between different social settings.

Check Your Understanding

1. True or False? A good definition of communication is the sending of information from

one person to another.

a. True

b. False

Suggested Responses:

False

2. True or False? Good working relationships between people form an important foundation for effective communication

a. True

b. False

10. True or False? In order to have an effective communication program, top management

must take an active part

a. True

b. False

Suggested Responses:

True

11. True or False? Where persuasion is needed, oral communication is better than written

Communication.

a. True

b. False

Suggested Responses:

True

12. True or False? In keeping others informed, it is better to under-communicate than over communicate

a. True

b. False

Suggested Responses:

False

13. True or False? The best way to be sure we understand a communication is to repeat it

back to the communicator.

a. True

b. False

Suggested Responses:

True

Check your understanding

1. How do you read excel dataset in R?
2. What are the types of No SQL tools based on Data Models?
3. Why do we use No SQL?
4. Is No SQL a query language like SQL?

Summary

- For integration of SQL and R we use SQLDF package.
- A NoSQL (originally referring to "non SQL" or "non-relational") database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.
- NoSQL support SQL like query language.
- NoSQL is used primarily in compliment to Big Data tools.
- Excel can be integrated with R using R-Connector.
- How to execute VBA code in R tool?

1. **Divide the class into groups of 4-5 participants**
2. **Give the Dataset to the participants.**
3. **Give 10 minutes to the class for each group to discuss the various change points between SQL and NO SQL along with a discussion on the methods type which they would like to use**
4. **Each group presents their examples with justification. (5 min each)**

Fixes:-

There are four common corrections for heteroscedasticity. They are:

- View logarithmized data. Non-logarithmized series that are growing exponentially often appear to have increasing variability as the series rises over time. The variability in percentage terms may, however, be rather stable.

- Use a different specification for the model (different X variables, or perhaps non-linear transformations of the X variables).

- Apply a weighted least squares estimation method, in which OLS is applied to transformed or weighted values of X and Y. The weights vary over observations, usually depending on the changing error variances. In one variation the weights are directly related to the magnitude of the dependent variable, and this corresponds to least squares percentage regression.

- Heteroscedasticity-consistent standard errors (HCSE), while still biased, improve upon OLS estimates. HCSE is a consistent estimator of standard errors in regression models with heteroscedasticity. This method corrects for heteroscedasticity without altering the values of the coefficients. This method may be superior to regular OLS because if heteroscedasticity is present it corrects for it, however, if the data is homoscedastic, the standard errors are equivalent to conventional standard errors estimated by OLS. Several modifications of the White method of computing heteroscedasticity-consistent standard errors have been proposed as corrections with superior finite sample properties.

Dummy Variables

In regression analysis, a dummy variable (also known as an indicator variable, design variable, Boolean indicator, categorical variable, binary variable, or qualitative variable) is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into mutually exclusive categories (such as smoker/non-smoker, etc.).

In other words, Dummy variables are "proxy" variables or numeric stand-ins for qualitative facts in a regression model. In regression analysis, the dependent variables may be influenced not only by quantitative variables (income, output, prices, etc.), but also by qualitative variables (gender, religion, geographic region, etc.). A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept.

For example,

Suppose Gender is one of the qualitative variables relevant to a regression. Then, female and male would be the categories included under the Gender variable. If female is arbitrarily assigned the value of 1, then male would get the value 0. Then the intercept (the value of the dependent variable if all other explanatory variables hypothetically took on the value zero) would be the constant term for males but would be the constant term plus the coefficient of the gender dummy in the case of females.

S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology; often written as SMART) is a monitoring system included in computer hard disk drives (HDDs) and solid-state drives (SSDs) that detects and reports on various indicators of drive reliability, with the intent of enabling the anticipation of hardware failures.

When S.M.A.R.T. data indicates a possible imminent drive failure, software running on the host system may notify the user so stored data can be copied to another storage device, preventing data loss, and the failing drive can be replaced.

Understand the business problem related to engineering, Identify the critical issues. Set business objectives.

The BA process can solve problems and identify opportunities to improve business performance. In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks. The latter, identifying opportunities can be viewed as a problem of strategy choice requiring a solution.