

**ASSIGNMENTS: Five**

|  |  |
| --- | --- |
| **Student Name:** | **Peter Atem Anyuon** |
| **Course** | **Post Graduate Diploma in Monitoring and Evaluation** |
| **Student ID:** | **Aipms/290/2019** |
| **Lecturer:** | **Mr. Ratemo Fredrick** |
| **Module 5 Questions:** | **Diploma in Monitoring and Evaluation** |
| **Module Code:** | **D001** |
| **Assignment Number:** | **Five** |
| **Date Issued:** | **01/11/2019** |
| **Due Date:** | **30/11/2019** |
|  |  |
| **Assignment Brief:** | **Q1.** Explain the difference between data collection and data capture (10mrks)    **Q 2:** Explain the benefits of correctly interpreting data in an M&E process. (5 mrks)    **Q3.** Explain the main concerns for a data analyst while undertaking the task of data analysis. (10 mrks)    **Q4.** Describe key measures that are mandatory for data quality assurance at program level and explain the value of data quality assurance. (15 mrks).    **Q5:** In about 350 words, describe the main challenges to effective data interpretation and analysis. (10 mrsk) |

**Q1. Explain the difference between data collection and data capture (10mrks)**

Data collection is any process whose purpose is to acquire or assist in the acquisition of data. Collection is achieved by requesting and obtaining pertinent data from individuals or organizations via an appropriate. The data is either provided directly by the respondent (self-enumeration) or via an interviewer. Collection also includes the extraction of information from administrative sources which may require asking the respondent permission to link to administrative records.

Data capture refers to any process that converts the information provided by a respondent into electronic format. This conversion is either automated or involves staff keying the collected data (keyers). Data coding is any process that assigns a numerical value to a response. Coding is often automated, however, more complex decisions usually require human intervention (coders).

Often, survey operations involve a high degree of automation, which leads to the availability of paradata, information related to a survey process. Examples of paradata include an indicator of whether or not a unit is in the sample, history of calls and visits, trail of key strokes (audit trail), mode of collection, administrative information (e.g. interviewer profile) and cost information.

Data collection is not only the source of information, it is also the main contact a survey-taking agency has with the public who needs to be convinced to participate. Data capture and coding produce the formatted data used as input by all the subsequent survey processes. Data collection, data capture and coding operations often use a large portion of the survey budget and require considerable human and physical resources as well as time.

**Q 2: Explain the benefits of correctly interpreting data in an M&E process. (5 mrks)**

The process of Monitoring and Evaluation (M&E) can be multi- purpose.

The trick is to select the tools you need, combine them, and know how to use them when necessary. Information collection will be a critical aspect and function of all M&E. All purpose of data collection is to provide a factual basis for decision making.

Data management is the process of managing the entire process of data collection, capture, storage and its use in informing management and planning decisions across all levels of the organization and the stakeholder-community.

Data management entails the following:

Continually assessing whether project/program information needs are being met

Managing data collection process and ensuring that data is captured in time

Analyzing and using the information in time

Disseminating the information to ensure its timely use.

Proper data management confronts the following scenarios:

Information collected was not sufficient, detailed, inaccurate or incomplete

No backup information was stored

There were different perceptions in understanding and timelines within the project or program

Wrong people were interviewed

Information collection methods

All M&E work will involve collection of useful information (data) during the process of implementation.

This will include internal and external records to the program or project.

Data collection is the process of gathering and measuring raw information (data) on variables of interest, in an established systematic fashion that enables determination of stated project performance or research questions, test hypotheses, and or evaluate outcomes. It uses specific tools for collection purposes.

Data collection tools should have the following characteristics: Should be as simple and as clear as possible

Should be pre-tested to ensure they are user-friendly

Stick to a minimum set of questions needed to measure indicators.

Internal records: What it is - using internal records to track project activities, processes and output indicators such as numbers and demography of project members, supporters,

Participants:

This information is mainly quantitative in nature its benefits: It is useful for tracking activities and outputs

Limits – information about activities and outputs does not tell you what difference you are

Making

Tracking relevant secondary information

What it is - keeping records of relevant secondary information to track changes in outcomes and

Impacts and accompanying internal records e.g. Policy changes, media coverage, relevant surveys/data bases.

Type of information- mainly quantitative

The benefits: – useful for providing evidence of external changes and comparisons and is useful if you do not have the resources to carry out your own interviews or surveys.

Limits – secondary sources may not provide completely relevant or comparable information about the precise issue, neighborhood or people you are interested in.

In most cases, it needs to be supplemented by qualitative methods to assess your contribution to the observed outcomes i.e. by asking why and how change happened**.**

**Q3. Explain the main concerns for a data analyst while undertaking the task of data analysis.**

**(10 mrks)**

**What is a Data Analyst?**

Data Analysts deliver value to their companies by taking data, using it to answer questions, and communicating the results to help make business decisions. Common tasks done by data analysts include data cleaning, performing analysis and creating data visualizations.

Depending on the industry, the data analyst could go by a different title (e.g. Business Analyst, Business Intelligence Analyst, Operations Analyst, and Database Analyst). Regardless of title, the data analyst is a generalist who can fit into many roles and teams to help others make better data-driven decisions.

**The Data Analyst in Depth**

The data analyst has the potential to turn a traditional business into a data-driven one.  While often data analyst positions are "entry level" jobs in the wider field of data, not all analysts are junior level. As effective communicators with mastery over technical tools, data analysts are critical for companies that have segregated technical and business teams.

Their core responsibility is to help others track progress and optimize their focus. How can a marketer use analytics data to help launch their next campaign? How can a sales representative better identify which demographics to target? How can a CEO better understand the underlying reasons behind recent company growth? These are all questions that the data analyst provides the answer to by performing analysis and presenting the results.

They undertake the complex job of working with data to deliver value to their organization.

An effective data analyst will take the guesswork out of business decisions and help the entire organization thrive. The data analyst must be an effective bridge between different teams by analyzing new data, combining different reports, and translating the outcomes. In turn, this is what allows the organization to maintain an accurate pulse check on its growth.

The nature of the skills required will depend on the company's specific needs, but these are some common tasks:

Cleaning and organizing raw data.

Using descriptive statistics to get a big-picture view of their data.

Analyzing interesting trends found in the data.

Creating visualizations and dashboards to help the company interpret and make decisions with the data.

Presenting the results of a technical analysis to business clients or internal teams.

The data analyst brings significant value to both the technical and non-technical sides of an organization. Whether running exploratory analyses or explaining executive dashboards, the analyst fosters a greater connection between teams.

**What is a Data Scientist?**

A data scientist is a specialist who applies their expertise in statistics and building machine learning models to make predictions and answer key business questions. A data scientist still needs to be able to clean, analyze, and visualize data, just like a data analyst. However, a data scientist will have more depth and expertise in these skills, and will also be able to train and optimize machine learning models.

**The Data Scientist in Depth**

The data scientist is an individual who can provide immense value by tackling more open-ended questions and leveraging their knowledge of advanced statistics and algorithms. If the analyst focuses on understanding data from the past and present perspectives, then the scientist focuses on producing reliable predictions for the future.

The data scientist will uncover hidden insights by leveraging both supervised (e.g. classification, regression) and unsupervised learning (e.g. clustering, neural networks, anomaly detection) methods toward their machine learning models. They are essentially training mathematical models that will allow them to better identify patterns and derive accurate predictions.

**The following are examples of work performed by data scientists**:

* Evaluating statistical models to determine the validity of analyses.
* Using machine learning to build better predictive algorithms.
* Testing and continuously improving the accuracy of machine learning models.
* Building data visualizations to summarize the conclusion of an advanced analysis.

Data scientists bring an entirely new approach and perspective to understanding data. While an analyst may be able to describe trends and translate those results into business terms, the scientist will raise new questions and be able to build models to make predictions based on new data.

**What is a Data Engineer?**

Data engineers build and optimize the systems that allow data scientists and analysts to perform their work. Every company depends on its data to be accurate and accessible to individuals who need to work with it. The data engineer ensures that any data is properly received, transformed, stored, and made accessible to other users.

**The Data Engineer in Depth**

The data engineer establishes the foundation that the data analysts and scientists build upon. Data engineers are responsible for constructing data pipelines and often have to use complex tools and techniques to handle data at scale. Unlike the previous two career paths, data engineering leans a lot more toward a software development skill set.

At larger organizations, data engineers can have different focuses such as leveraging data tools, maintaining databases, and creating and managing data pipelines. Whatever the focus may be, a good data engineer allows a data scientist or analyst to focus on solving analytical problems, rather than having to move data from source to source.

The data engineer’s mindset is often more focused on building and optimization. The following are **examples of tasks that a data engineer might be working on:**

* Building APIs for data consumption.
* Integrating external or new datasets into existing data pipelines.
* Applying feature transformations for machine learning models on new data.
* Continuously monitoring and testing the system to ensure optimized performance.

**Q4. Describe key measures that are mandatory for data quality assurance at program level and explain the value of data quality assurance. (15 mrks).**

 Below lists 5 main criteria used to measure data quality:

* **Accuracy:**for whatever data described, it needs to be accurate.
* **Relevancy:**the data should meet the requirements for the intended use.
* **Completeness:** the data should not have missing values or miss data records.
* **Timeliness:**the data should be up to date.
* **Consistency:** the data should have the data format as expected and can be cross reference-able with the same results.
* **However, if a company can manage the data quality of each dataset at the time when it is received or created, the data quality is naturally guaranteed. There are 7 essential steps to making that happen:**
* **1. Rigorous data profiling and control of incoming data**
* In most cases, bad data comes from data receiving. In an organization, the data usually comes from other sources outside the control of the company or department. It could be the data sent from another organization, or, in many cases, collected by third-party software. Therefore, its data quality cannot be guaranteed, and a rigorous data quality control of incoming data is perhaps the most important aspect among all data quality control tasks. A good data profiling tool then comes in handy; such a tool should be capable of examining the following aspects of the data:
* Data format and data patterns
* Data consistency on each record
* Data value distributions and abnormalies
* Completeness of the data

It is also essential to automate the data profiling and data quality alerts so that the quality of incoming data is consistently controlled and managed whenever it is received — never assume an incoming data is as good as expected without profiling and checks. Lastly, each piece of incoming data should be managed using the same standards and best practices, and a centralized catalog and KPI dashboard should be established to accurately record and monitor the quality of the data.

**2. Careful data pipeline design to avoid duplicate data**

Duplicate data refers to when the whole or part of data is created from the same data source, using the same logic, but by different people or teams likely for different downstream purposes. When a duplicate data is created, it is very likely out of sync and leads to different results, with cascading effects throughout multiple systems or databases. At the end, when a data issue arises, it becomes difficult or time-consuming to trace the root cause, not to mention fixing it.

In order for an organization to prevent this from happening, a data pipeline needs to be clearly defined and carefully designed in areas including data assets, data modeling, business rules, and architecture. Effective communication is also needed to promote and enforce data sharing within the organization, which will improve overall efficiency and reduce any potential data quality issues caused by data duplications. This gets into the core of data management, the details of which are beyond the scope of this article. On a high level, there are 3 areas that need to be established to prevent duplicate data from being created:

1. A data governance program, which clearly defines the ownership of a dataset and effectively communicates and promotes dataset sharing to avoid any department silos.
2. Centralized data assets management and data modeling, which are reviewed and audited regularly.
3. Clear logical design of data pipelines at the enterprise level, which is shared across the organization.

With today’s rapid changes in technology platforms, solid data management and enterprise-level data governance are essential for future successful platform migrations.

**Accurate gathering of data requirements**

An important aspect of having good data quality is to satisfy the requirements and deliver the data to clients and users for what the data is intended for. It is not as simple as it first sounds, because:

* It is not easy to properly present the data. Truly understanding what a client is looking for requires thorough data discoveries, data analysis, and clear communications, often via data examples and visualizations.
* The requirement should capture all data conditions and scenarios — it is considered incomplete if all the dependencies or conditions are not reviewed and documented.
* Clear documentation of the requirements, with easy access and sharing, is another important aspect, which should be enforced by the Data Governance Committee.

The role of Business Analyst is essential in requirement gathering. Their understanding of the clients, as well as current systems, allows them to speak both sides’ languages. After gathering the requirements, business analysts also perform impact analysis and help to come up with test plans to make sure the data produced meets the requirements.

**4. Enforcement of data integrity**

An important feature of the relational database is the ability to enforce data Integrity using techniques such as foreign keys, check constraints, and triggers. When the data volume grows, along with more and more data sources and deliverables, not all datasets can live in a single database system. The referential integrity of the data, therefore, needs to be enforced by applications and processes, which need to be defined by best practices of data governance and included in the design for implementation.

In today’s big data world, referential enforcement has become more and more difficult. Without the mindset of enforcing integrity in the first place, the referenced data could become out of date, incomplete or delayed, which then leads to serious data quality issues.

**. Integration of data lineage traceability into the data pipelines**

For a well-designed data pipeline, the time to troubleshoot a data issue should not increase with the complexity of the system or the volume of the data. Without the data lineage traceability built into the pipeline, when a data issue happens, it could take hours or days to track down the cause. Sometimes it could go through multiple teams and require data engineers to look into the code to investigate.

Data Lineage traceability has 2 aspects:

* **Meta-data:**the ability to trace through the relationships between datasets, data fields and the transformation logic in between.
* **Data itself:** the ability to trace a data issue quickly to the individual record(s) in an upstream data source.

Meta-data traceability is an essential part of effective data governance. This is enabled by clear documentation and modeling of each dataset from the beginning, including its fields and structure. When a data pipeline is designed and enforced by the data governance, the meta-data traceability should be established at the same time. Today, meta-data lineage tracking is a must-have capability for any data governance tool on the market, which makes it easier to store and trace through datasets and fields by a few clicks, instead of having data experts go through documents, databases, and even programs.

Data traceability is more difficult than meta-data traceability. Below lists some common techniques to enable this ability:

1. Trace by unique keys of each dataset: This first requires each dataset has one or a group of unique keys, which is then carried down to the downstream dataset through the pipeline. However, not every dataset can be traced by unique keys. For example, when a dataset is aggregated, the keys from the source get lost in the aggregated data.
2. Build a unique sequence number, such as transaction identifier or record identifier when there are no obvious unique keys in the data itself.
3. Build link tables when there are many-to-many relationships, but not 1-to-1or 1-to-many.
4. Add timestamp (or version) to each data record, to indicate when it is added or changed.
5. Log data change in a log table with the value before a change and the timestamp when the change happens

Data traceability takes time to design and implement. It is, however, strategically critical for data architects and engineers to build it into the pipeline from the beginning; it is definitely worth the effort considering it will save a tremendous amount of time when a data quality issue does happen. Furthermore, data traceability lays the foundation for further improving data quality reports and dashboards that enables one to find out data issues earlier before the data is delivered to clients or internal users.

1. **Automated regression testing as part of change management**

Obviously, data quality issues often occur when a new dataset is introduced or an existing dataset is modified. For effective change management, test plans should be built with 2 themes: 1) confirming the change meets the requirement; 2) ensuring the change does not have an unintentional impact on the data in the pipelines that should not be changed. For mission critical datasets, when a change happens, regular regression testing should be implemented for every deliverable and comparisons should be done for every field and every row of a dataset. With the rapid progress of technologies in big data, system migration constantly happens in a few years. Automated regression test with thorough data comparisons is a must to make sure good data quality is maintained consistently.

**7. Capable data quality control teams**

Lastly, 2 types of teams play critical roles to ensure high data quality for an organization:

**Quality Assurance:** This team checks the quality of software and programs whenever changes happen. Rigorous change management performed by his team is essential to ensure data quality in an organization that undergoes fast transformations and changes with data-intensive applications.

**Production Quality Control:**Depending on an organization, this team does not have to be a separate team by itself. Sometime it can be a function of the Quality Assurance or Business Analyst team. The team needs to have a good understanding of the business rules and business requirements, and be equipped by the tools and dashboards to detect abnormalities, outliers, broken trends and any other unusual scenarios that happen on Production. The objective of this team is to identify any data quality issue and have it fixed before users and clients do. This team also needs to partner with customer service teams and can get direct feedback from customers and address their concerns quickly. With the advances of modern AI technologies, efficiency can be potentially improved drastically. However, as stated at the beginning of this article, quality control at the end is necessary but not sufficient to ensure a company creates and sustains good data quality. The 6 steps stated above are also required.

**Summary**

In conclusion, good data quality requires disciplined data governance, rigorous management of incoming data, accurate requirement gathering, thorough regression testing for change management and careful design of data pipelines, in addition to data quality control programs for the data delivered both externally and internally. For all quality problems, it is much easier and less costly to prevent the data issue from happening in the first place, rather than relying on defending systems and ad hoc fixes to deal with data quality problems. Finally, by following the 7 steps in this article, good data quality can not only be guaranteed and but also sustained.

**Q5: In about 350 words, describe the main challenges to effective data interpretation and analysis. (10 mrsk)**

**What Is Data Interpretation?**

Data interpretation refers to the implementation of processes through which data is reviewed for the purpose of arriving at an informed conclusion. The interpretation of data assigns a meaning to the information analyzed and determines its signification and implications.

The importance of data interpretation is evident and this is why it needs to be done properly. Data is very likely to arrive from multiple sources and has a tendency to enter the analysis process with haphazard ordering. Data analysis tends to be extremely subjective. That is to say, the nature and goal of interpretation will vary from business to business, likely correlating to the type of data being analyzed. While there are several different types of processes that are implemented based on individual data nature, the two broadest and most common categories are “quantitative analysis” and “qualitative analysis”.

Yet, before any serious data interpretation inquiry can begin, it should be understood that visual presentations of data findings are irrelevant unless a sound decision is made regarding scales of measurement. Before any serious data analysis can begin, the scale of measurement must be decided for the data as this will have a long-term impact on data interpretation ROI. The varying scales include:

* **Nominal Scale:**non-numeric categories that cannot be ranked or compared quantitatively. Variables are exclusive and exhaustive.
* **Ordinal Scale:**exclusive categories that are exclusive and exhaustive but with a logical order. Quality ratings and agreement ratings are examples of ordinal scales (i.e., good, very good, fair, etc., OR agree, strongly agree, disagree, etc.).
* **Interval:**a measurement scale where data is grouped into categories with orderly and equal distances between the categories. There is always an arbitrary zero point.
* **Ratio:** contains features of all three.
* When interpreting data, an analyst must try to discern the differences between correlation, causation and coincidences, as well as many other bias – but he also has to consider all the factors involved that may have led to a result. There are various data interpretation methods one can use.
* The interpretation of data is designed to help people make sense of numerical data that has been collected, analyzed and presented. Having a baseline method (or methods) for interpreting data will provide your analyst teams a structure and consistent foundation. Indeed, if several departments have different approaches to interpret the same data, while sharing the same goals, some mismatched objectives can result. Disparate methods will lead to duplicated efforts, inconsistent solutions, wasted energy and inevitably – time and money. In this part, we will look at the two main methods of interpretation of data: with a qualitative and a quantitative analysis.

### Qualitative Data Interpretation

Qualitative data analysis can be summed up in one word – categorical. With qualitative analysis, data is not described through numerical values or patterns, but through the use of descriptive context (i.e., text). Typically, narrative data is gathered by employing a wide variety of person-to-person techniques. These techniques include:

* **Observations:** detailing behavioral patterns that occur within an observation group. These patterns could be the amount of time spent in an activity, the type of activity and the method of communication employed.
* **Documents:** much like how patterns of behavior can be observed, different types of documentation resources can be coded and divided based on the type of material they contain.
* **Interviews:**one of the best collection methods for narrative data. Enquiry responses can be grouped by theme, topic or category. The interview approach allows for highly-focused data segmentation.

A key difference between qualitative and quantitative analysis is clearly noticeable in the interpretation stage. Qualitative data, as it is widely open to interpretation, must be “coded” so as to facilitate the grouping and labeling of data into identifiable themes. As person-to-person data collection techniques can often result in disputes pertaining to proper analysis, qualitative data analysis is often summarized through.

**Quantitative Data Interpretation**

If quantitative data interpretation could be summed up in one word (and it really can’t) that word would be “numerical.” There are few certainties when it comes to data analysis, but you can be sure that if the research you are engaging in has no numbers involved, it is not quantitative research. Quantitative analysis refers to a set of processes by which numerical data is analyzed. More often than not, it involves the use of statistical modeling such as standard deviation, mean and median. Let’s quickly review the most common statistical terms:

* **Mean:**a mean represents a numerical average for a set of responses. When dealing with a data set (or multiple data sets), a mean will represent a central value of a specific set of numbers. It is the sum of the values divided by the number of values within the data set. Other terms that can be used to describe the concept are arithmetic mean, average and mathematical expectation.
* **Standard deviation:**this is another statistical term commonly appearing in quantitative analysis. Standard deviation reveals the distribution of the responses around the mean. It describes the degree of consistency within the responses; together with the mean, it provides insight into data sets.
* **Frequency distribution:**this is a measurement gauging the rate of a response appearance within a data set. When using a survey, for example, frequency distribution has the capability of determining the number of times a specific ordinal scale response appears (i.e., agree, strongly agree, disagree, etc.). Frequency distribution is extremely keen in determining the degree of consensus among data points.

# References

* (n.d.). http://www.businessdictionary.com/definition/data-analysis.html.
* (n.d.). https//blog.taskque.com/reasons-project-fail/.
* http://www.businessdictionary.com/definition/data-analysis.html. (n.d.).
* <http://www.un.cv/files/Template%20for%20M&E%20plan.pdf>