Wadzanayi Kuweta Theory Questions Assignment 2

1. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

There are various techniques and practices to keep a nice and clean database and below are some best practices for cleansing data:

a. Develop a data quality strategy

- Set expectations for your data.
- Create data quality key performance indicators (KPIs) - What are they, and how will you meet them? How will you track the health of your data? How will you maintain data hygiene on an ongoing basis?
- Find out where most data quality errors occur.
- Identify incorrect data.
- Understand the root cause of the data problem.
- Develop a plan for ensuring the health of your data.

b. Correct data at the point of entry

To keep a clean database, it is important to have clean and standardized data to ensure all important attributes are free of issues and mistakes at the point of entry. This can help save time and effort for the team before going any further.

A standard operating procedure for entering data should be created and enforced by everyone in the team. This will ensure that only high-quality data can be entered into the system.

c. Validate the accuracy of your data

In this step, we need to validate the data to make sure it meets all the requirements, which can be done manually with a small data set. However, with larger and more complex data sets, the manual method is extremely time-consuming, labour-intensive, and ineffective as people are prone to make mistakes. Therefore, data quality control tools are made to help with this issue.

d. Manage duplicates

Duplicates are harmful and are a waste of time and effort. They interfere with various functions of the company and slow down the firm operating process.

Companies must avoid them as best as they can. And after removing all duplicate data at the entrance, it is important to consider the following:

- Standardising: Converting data to one single format to process and analyze.
- Normalising: Ensuring that all data is recorded consistently.
- Merging: When data is scattered across multiple datasets, merging is the act of combining relevant parts of those datasets to create a new file.
- Aggregating: Sorting data and expressing it in a summary form.
- Filtering: Narrowing down a dataset to only include the information that users want.
- Scaling: Transforming data so that it fits within a specific scale such as 0-100 or 0-1.
- Removing duplicate and outlier data points to prevent a bad fit in linear regression.

e. Append missing data

Append is a process of filling in missing information in the required field of the records, such as phone number, email address, last and first name, home address, etc. But finding the missing information can be tricky. To do this step effectively, it is recommended that firms should use a reliable third-party data source to help fill in the gaps.

f. Promote the use of clean data across the organization

After everything is done, you need to communicate with everybody across the organization about the importance of clean data. Ensure that employees, regardless of their functions, understand and maintain the practice of clean data.

2.

Difference between Data Profiling and Data Mining:

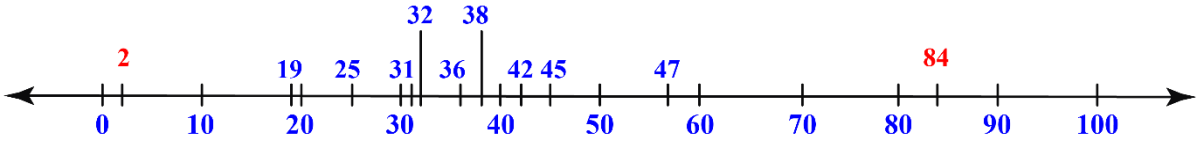
DATA MINING	**DATA PROFILING**
Data mining is the process of identifying the patterns in a pre-built database.	1. Data profiling is a process of analyzing data from the existing one.

It is also called KDD (Knowledge Discovery in Databases).	It is also known as data archaeology.
The purpose of data mining is to build machine learning techniques for real-time needs.	The purpose of data profiling is to provide us with accuracy, consistency, uniqueness, and error-free within a dataset.
It extracts data by applying some computer-based methodologies and some algorithm.	It extracts from the existing raw dataset.
The point of data mining is to dig out the data from the sources to resolve some issues through data analysis.	The purpose is to collect accurate data for recognizing the use and quality of that data.
It is usually executed on structured data.	It is executed on structured as well as unstructured data.
This involves Classification, Clustering, Regression, Association rule, and neural networks to perform tasks.	This involves the discovery and Analytical Techniques to collect informative summaries related to the data.
The applications of data mining involve customer behaviour, credit analysis, fraud detection, business intelligence, etc.	The applications of data profiling involve targeted advertising, fraud, and risk detection, image recognition, delivery logistics, etc.
Tools used for data mining are Weka, RapidMiner, Orange, KNIME, Sisense, SPSS, SPSS Modeler, Rattle, Data Melt, etc.	Tools used for data profiling are Atlan, Aggregate Profiler, IBM Infosphere Information Analyzer, Informatica Data Explorer, Melissa Data Profiler, Microsoft Docs, etc.

3.

A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.

Example: For a data set containing 2, 19, 25, 32, 36, 38, 31, 42, 57, 45, and 84



In the above number line, we can observe the numbers 2 and 84 are at the extremes and are thus outliers.

The outliers are a part of the group but are far away from the other members of the group.

4.

Collaborative Filtering is a Machine Learning technique used to identify relationships between pieces of data. This technique is frequently used in recommender systems to identify similarities between user data and items. This means that if Users A and B both like Product A, and User B also likes Product B, then Product B could be recommended to User A by the system.

The model keeps track of what products users like and their characteristics to see what users, who like products with similar characteristics, enjoyed. The model then makes its recommendations accordingly.

5. Time series analysis is a technique in statistics that deals with time series data and trend analysis. Time series data follows periodic time intervals that have been measured in regular time intervals or have been collected at time intervals. In other words, a time series is simply a series of data points ordered in time, and time series analysis is the process of making sense of this data.

The data is considered in three types:

- Time series data: A set of observations on the values that a variable takes at different times.
- Cross-sectional data: Data of one or more variables, collected at the same point in time.
- Pooled data: A combination of time series data and cross-sectional data.

6. Core Steps of a Data analysis project

- **Define the question**—What business problem are you trying to solve? Frame it as a question to help you focus on finding a clear answer.
- **Collect data**—Create a strategy for collecting data. Which data sources are most likely to help you solve your business problem?
- **Clean the data**—Explore, scrub, tidy, de-dupe, and structure your data as needed. Do whatever you have to! But don't rush…take your time!
- **Analyze the data**—Carry out various analyses to obtain insights. Focus on the four types of data analysis: descriptive, diagnostic, predictive, and prescriptive.
- **Validate the data**— assess the data and determine if you have the correct information for your deliverable. Did the models work properly? Does the data need more cleaning?
- **Share your results**—How best can you share your insights and recommendations? A combination of visualization tools and communication is key.
- **Embrace your mistakes**—Mistakes happen. Learn from them. This is what transforms a good data analyst into a great one.

7. Characteristics of a good data Model

a. Data in a good model can be easily consumed.

b. large data changes in a good model are scalable.

c. A good model provides predictable performance.

d. A good model can adapt to changes in requirements, but not at the expense of a-c.

8.

1. Univariate data –
This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. An example of univariate data can be height.

Heights (in cm)	164	167.3	170	174.2	178	180	186

2. Bivariate data –
This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables. An example of bivariate data can be temperature and ice cream sales in summer season.

TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

3. Multivariate data –
When the data involves **three or more variables**, it is categorized under multivariate. It is like bivariate but contains more than one dependent variable. The ways to perform analysis on this data depend on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis, and multivariate analysis of variance (MANOVA).

9.

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

(1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Three major uses for regression analysis are

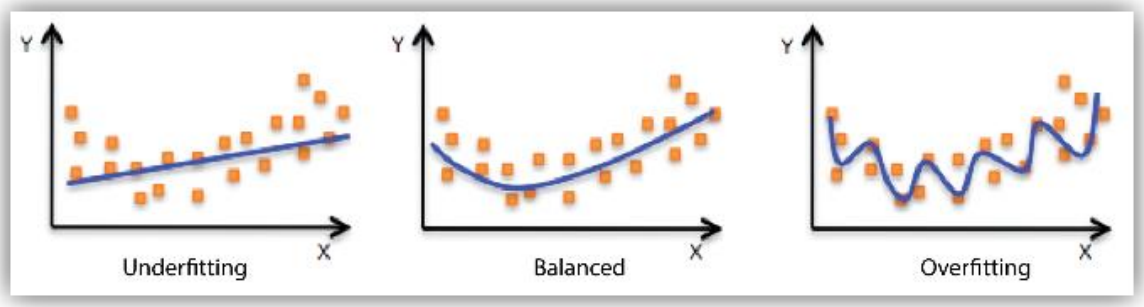
(1) determining the strength of predictors,

(2) forecasting an effect, and

(3) trend forecasting.

10.

Understanding model fit is important for understanding the root cause of poor model accuracy. This understanding guides us to take corrective steps. We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data.



A model is underfitting the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples and the target values. A model is overfitting the training data when the model performs well on the training data but does not perform well on the evaluation data. This is because the model is memorizing the data it has seen and is unable to generalize to unseen examples.

References

https://blog.trginternational.com/data-cleansing-best-practices

https://www.geeksforgeeks.org/difference-between-data-profiling-and-data-mining/

https://www.northeastern.edu/graduate/blog/data-analysis-project-lifecycle/

https://www.modernanalyst.com/Careers/InterviewQuestions/tabid/128/ID/4904/Describe-the-difference-between-univariate-bivariate-and-multivariate-analysis.aspx

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html