THEORY QUESTIONS ASSIGNMENT

Data Science Stream - Ruvimbo Hungwe

1. **What does "Data Cleansing" mean? What are the best ways to practice this?**

Data cleansing, sometimes referred to as "Data Cleaning" is the process of identifying, removing or replacing incomplete, corrupted, duplicate or incorrectly formatted data within a dataset. This is an important technique which is carried out prior to any data analysis in order to maintain the integrity and quality of the data so that accurate, consistent and reliable insights can be produced from the dataset. Failure to cleanse the data will result in inaccurate conclusions and could lead to poor business decisions. The best way to practice this is by following the ETL (Extract, Transform, Load) process which is illustrated in the various stages in figure 1 below.

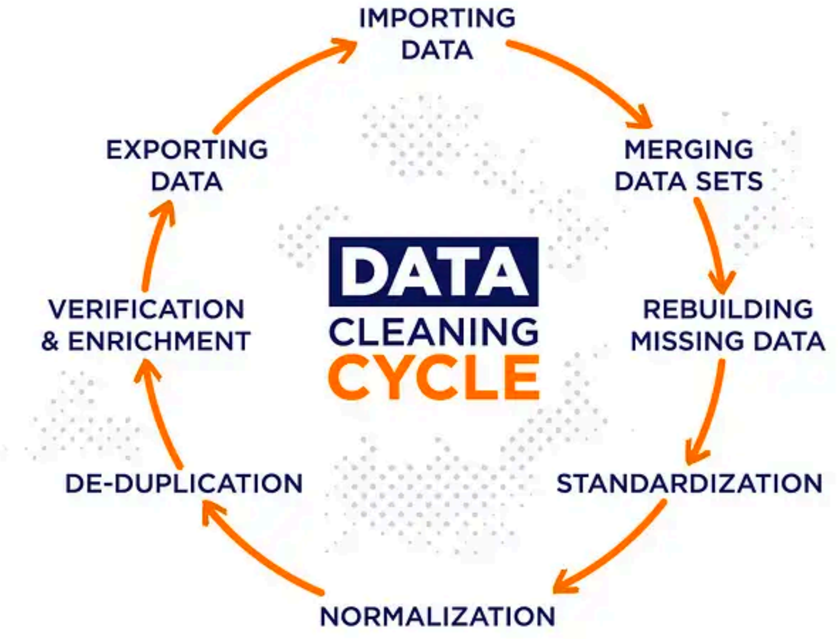


Figure 1. Data cleaning cycle

Following the steps in this continuous cycle will ensure the quality of the data. The standard practices for data cleaning can be broken down into 5 different stages;

- Data Audit
- Workflow specification
- Workflow Execution
- Validation
- Reporting

Data auditing is the process which is used to determine what kind of errors are in the data set and where they are located. Workflow execution is the stage which the data is cleaned using specific operations. This sequence is referred to as the “workflow’. Workflow execution is the data cleaning stage and the main objective is to remove or correct the data. This could be duplicates, errors and outliers. The validation stage of the process is for auditing the data and ensuring that the relevant processes, rules and constraints have been executed. The final stage is reporting which involves creating summaries of the data and then insights can then be generated.

2. What is the difference between data profiling and data mining?

Data profiling is a process of evaluating data from existing datasets in order to collect statistics or informative summaries about the data. It is also known as data archaeology. The main goal of data profiling is to assess the quality of the data by identifying anomalies, incorrect values and missing values during the initial stage of data analysis. Data mining, on the other hand is the process of identifying patterns in a pre-built database. It involves evaluating the existing database and large datasets through analysis or knowledge discovery to turn the raw data into useful information and find new trends and patterns. It is also known as knowledge discovery in databases (KDD).

The purpose of data profiling is to obtain information about the data and assess the quality in order to find anomalies in the dataset. The aim is to provide accurate and consistent information about the data and ensure that it is error free. Whereas the purpose of data mining is to mine for actionable information through the use of mathematical algorithms in order to segment the data and produce future trends so it can be used in general areas of Business Intelligence. Another difference between the two processes is that data mining is usually executed on structured data whereas data profiling is executed on both structured and unstructured data.

3. Define Outlier with an example.

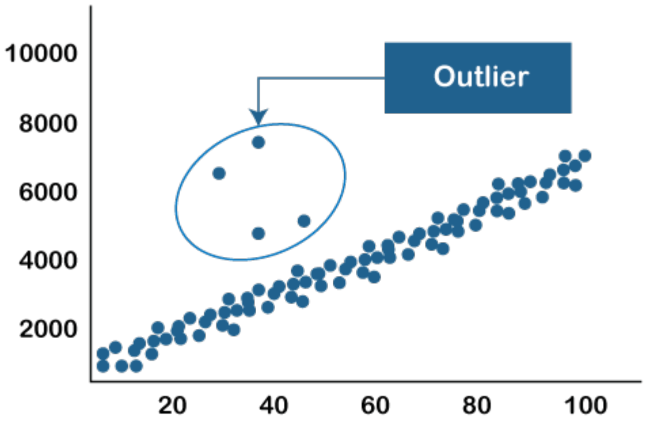


Figure 2. Example of a dataset containing outliers

An outlier is a data point that is a numerically abnormal distance to other data points . It is data that lies outside the other values in the set (figure 2). Outliers are usually much larger or significantly smaller than the rest of the data. When carrying out data analysis outliers can have a negative effect on the result of the analysis and thus need special attention to decide whether they need to be removed or not in order to analyse the data effectively. Outliers may be errors that can be excluded from the analysis however other times they can reveal insights into special cases in the data that would have otherwise remained unknown. Let’s take the example in figure 3 below. This shows the amount of money spent on food monthly over a 12 month period. This shows that in the 6th and 12th month the amount of money spent on food more than doubled so these are classed as outliers. One of the reasons could be that the person could have hosted parties in those months therefore they spent more money. A decision would have to be made whether the outliers will be kept in the dataset because if we were to calculate the average of the amount spent over the 12 month period including the outliers it would be £287 whereas if the outliers were removed the average would be £220. This demonstrates that the outliers can skew the results of an analysis.

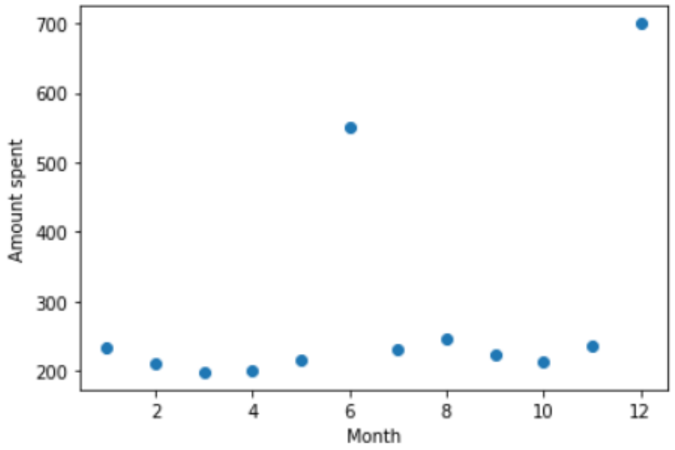
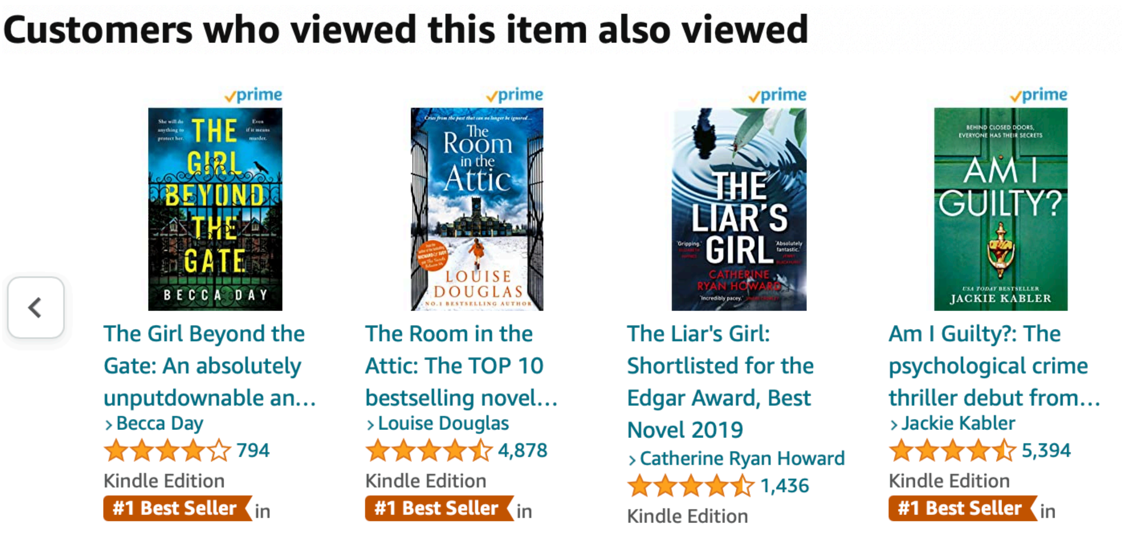


Figure 3. Scatter diagram showing the amount of money spent on food per month

4. What is “Collaborative Filtering”?

Collaborative filtering is a recommender systems technique which is used to filter items that a user might like based on data and interactions by other users. The premise of it is that people who agree on their evaluation of certain items are likely to agree in the future. It works by finding a smaller set of users within a large group of people that have similar tastes to a particular user. Collaborative filtering systems look at the relationship between the users and items to determine the similarity of the rating of the items both users have rated. An example of collaborative filtering is Amazon’s “customers who have views this item also viewed” seen in figure 4 below. This measures the similarity in the user’s search behaviour with other users and provides a recommendation.



Customers who viewed this item also viewed

prime
The Girl Beyond the Gate: An absolutely unputdownable an…
› Becca Day
★★★★☆ 794
Kindle Edition
#1 Best Seller in

prime
The Room in the Attic: The TOP 10 bestselling novel…
› Louise Douglas
★★★★☆ 4,878
Kindle Edition
#1 Best Seller in

prime
The Liar's Girl: Shortlisted for the Edgar Award, Best Novel 2019
› Catherine Ryan Howard
★★★★☆ 1,436
Kindle Edition

prime
Am I Guilty?: The psychological crime thriller debut from…
› Jackie Kabler
★★★★☆ 5,394
Kindle Edition
#1 Best Seller in

Figure 4. Amazon collaborative filtering

5. What is “Time Series Analysis”?

Time series analysis is method of analysing a sequence of data points that have been collected over an interval of time. In time series analysis the data points are recorded at consistent intervals over a set period of time. This period of time could be daily, weekly, monthly, quarterly etc. This type of analysis isn’t only just about collecting the data over time, it also shows how variables can change over time. Time series analysis generally requires a large number of data points to ensure the results are consistent and reliable. It also ensures that any trends or patterns discovered that are seasonal are not counted as outliers. Time series data can be used for forecasting based on historical data. Time series analysis allow businesses to better understand trends, patterns and their causes over time. Time series analysis is used for non-stationary data which can fluctuate over time for example stock prices, weather data, census etc.

6. Explain the core steps of a Data Analysis project?

There are 6 fundamental steps of a Data Analysis project.

The first step is to understand the business issues. This is done by defining the business objectives, gathering information from the stakeholders and determining the appropriate analysis methods for the business needs. This step also helps in clarifying the scope of works which then identifies the deliverables. These elements need to be clearly defined before any analysis can take place in order to provide the best deliverable possible.

Once the objective of the project has been clearly understood, the next step is to collect the initial data, identify the data requirements and get an understanding of the dataset. Key variables should then be identified to help categorise the data. It is also important to look for errors, missing variables or duplicate data in preparation for data cleaning.

When the data has been identified and organised, the data cleaning can commence. This involves verifying if the data types are compatible or not, identifying outliers and missing values and modifying them if necessary but in such a way that the data is not skewed. The next step is to perform exploratory analysis and modelling. During this step, models will be built to test the data and seek the answers to the given objectives. This is done through the use of statistical modelling methods by selecting the one most appropriate for the project. Once the models have been crafted, the data needs to be assessed in order to determine if the correct information for the deliverable is present. This is the opportunity to determine if the data needs more cleaning, if the model works properly and finally if the data found the outcome the client was looking for. If this is not the case previous steps will need to be repeated in order to achieve the necessary outcome.

The final step is the deployment and visualisation. This is done once all the deliverables have been met. This is a very important step in communicating the findings to the client and it must be presented in such a way that it is easily explainable to the client. Data visualisation is done through the use of graphical representation of the information and data by using charts, graphs and maps.

7. What are the characteristics of a good data model?

There are 4 characteristics of a good data model. Data in a good model can be easily consumed for actionable results. Any large changes to the data should be scalable. A good model also provides predictive performance and lastly a good model should be able to adapt to any changes in requirements but not at the expense of the aforementioned characteristics.

8. Explain and provide examples of univariate, bivariate, and multivariate analysis?

Univariate, bivariate and multivariate analysis are all methods of exploratory analysis. Univariate data is the type of data that only consists of one dependable variable. As it is a single variable it does not deal with causes or relationships. The main purpose of univariate data analysis is to derive the data, define it and summarise it to analyse the pattern that exists in it. Patterns derived from univariate analysis can be made by drawing conclusions using central tendency measures which are mean, median mode, frequency distribution tables such as histograms, pie charts and bar charts and also dispersion or spread of data such as range, minimum, maximum quartiles etc. An example of univariate data analysis is the chosen subjects of a number of students which is illustrated in figure 5 below.

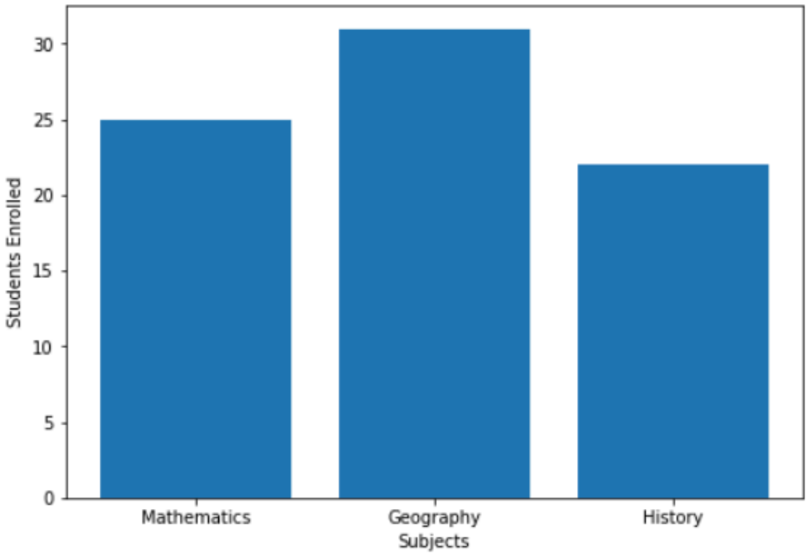


Figure 5. Bar chart showing number of students enrolled to specific subjects

Bivariate analysis is related to the comparison of two variables and studying their relationship. The variables can be dependent or independent of each other. An example of bivariate data could be the temperature and the number of barbecue stands sold. This is represented in figure 6 below.

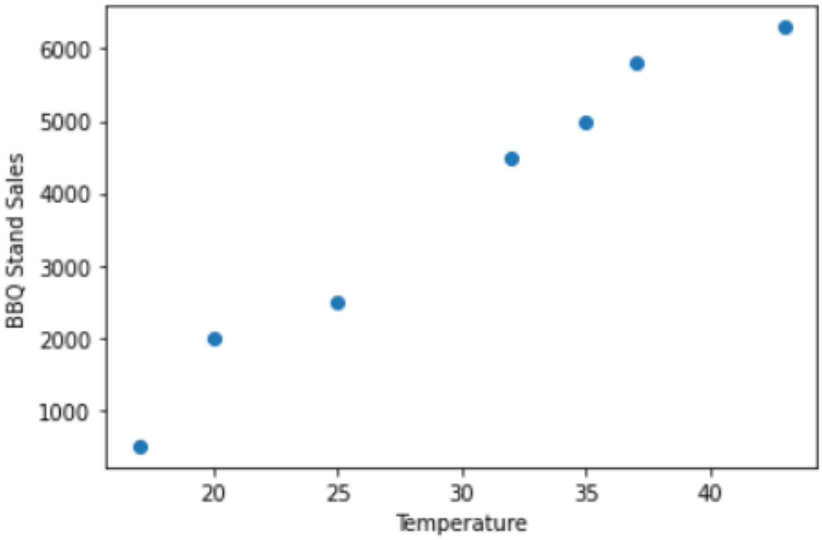


Figure 6. Scatter diagram showing the temperature in relation to number of barbecue stand sales

Multivariate analysis is used when comparing more than two variables. Cluster analysis, factor analysis, multiple regression analysis, principal component analysis are all types of multivariate analysis that are used to study more complex sets of data. As there are more than 20 ways to person multivariate analysis, it is important that the type of data and the end goal is considered before choosing a type of analysis. An example of multivariate analysis is a streaming service collecting the following data on its users; minutes watched per day, total viewing sessions per week, number of unique shows viewed per month. This data can then be presented in a cluster analysis to identify which customers they should target when advertising.

9. What is a Linear Regression?

Linear regression analysis used to predict the value of a variable based on the value of another variable by modeling the relationship between two variables and fitting a linear equation to the observed data. The variable that the model tries to predict is called the independent variable whereas the variable being used to predict the independent variable is called the dependent variable. Figure 7 illustrates the relationship between the independent and dependant variables.

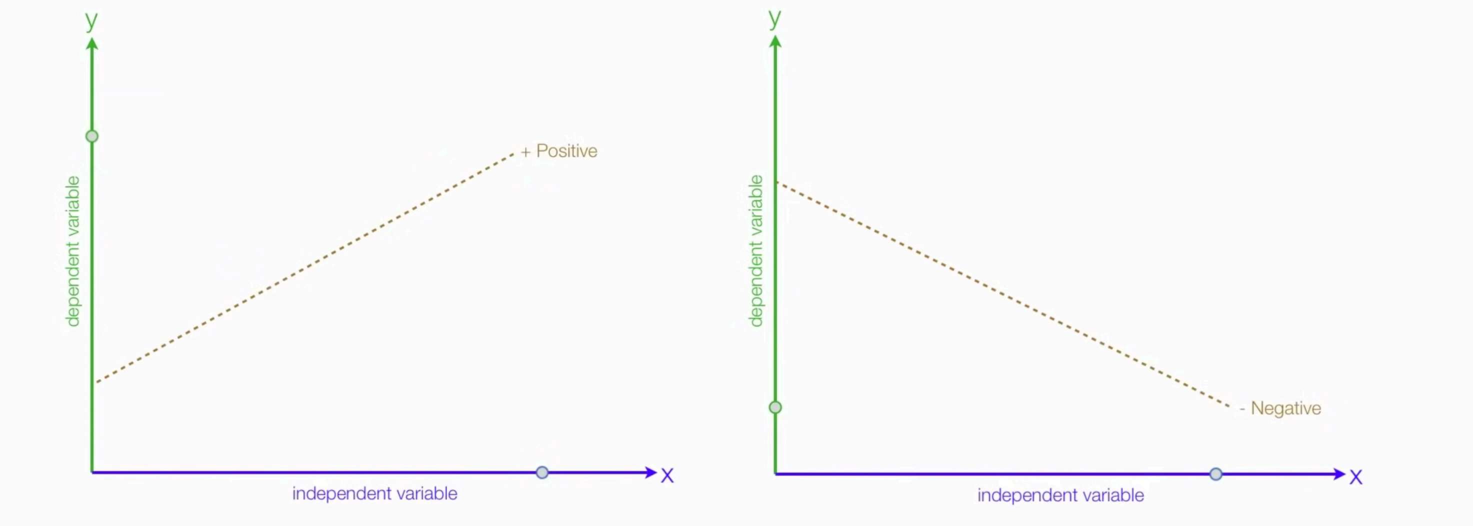


Figure 7. Relationship between independent and dependant variables

If both the independent and dependent variables increase this is a positive relationship. If the independent variable increases and the dependent variable decreases this is a negative relationship.

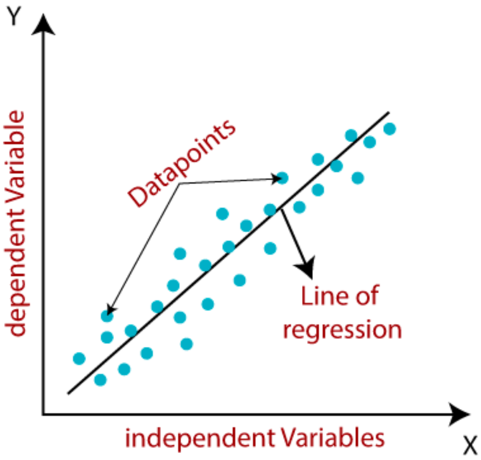


Figure 8. Line of regression

The line of regression is a straight line or surface that fits through the data points (figure 8). This is based upon the least squares method. An example of this would be looking at the relationship between study time and grades. The study time is the independent variable and the grades are the dependent variable. As the study time increases it is expected that the grades will also increase this is a positive relationship. Whereas if the time spent watching Netflix is considered the independent variable, the grades are expected to decrease which is a negative relationship. This type of analysis is used for trend forecasting, forecasting an effect and determining the strength of predictors.

10. In terms of modelling data, what do we mean by Over-fitting and Under-fitting?

A model is under-fitting when the model has not trained enough or hasn't been given enough input variables to determine a meaningful relationship between the input and output variables. The model performs so poorly on the training data that it cannot reflect the complexity of the data. This means it will generalise poorly to unseen data and have a high bias and less variance. Conversely, the results of overtraining the model will be overfitting. Overfitting is when a model performs well on the training data but fails on the evaluation data. This is because it memorises all the specific details of the data that it fails to perform accurately against unseen data. These are illustrated in figure 9. The optimum would be to have a model that has a low training error and low test error. This can be achieved by increasing the model flexibility if the model is under-fit by adding more specific features to the training data or increasing the amount of training examples. Or in order to overcome overfitting, the model needs to be trained for a shorter period of time or simplify the training data.

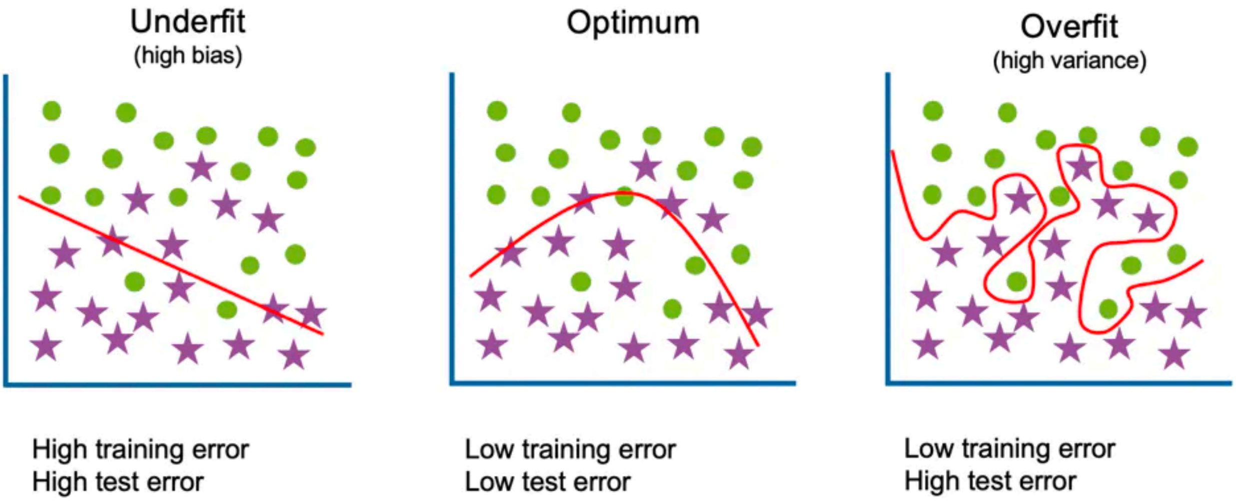


Figure 9. Examples of Underfit, Optimum and Overfit models.