**DATA SCIENCE STREAM ASSESSMENT**

**Question 1:**

**Data Cleansing**

The practice of checking, correcting or deleting inaccurate, damaged, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. There are many ways where data can be duplicated or incorrectly categorized when merging multiple data sources so data cleansing is helpful in removing that false information. Even if results and algorithms appear to be correct, they are unreliable if the data is inaccurate or not sufficient. Because the procedures will differ from dataset to dataset, there is no one definitive way to specify the correct phases in the data cleaning process. But it is very important to create a template for your data cleaning procedure so you can be sure you are carrying it out correctly each time.

**Ways of Data cleaning**

1. Develop a data quality strategy and Set expectations for our data.
2. Check data at the point of entry.
3. Validate accuracy for the data.
4. Manage data duplicity.
5. Change and restore missing data.
6. Promote the uses of clean data across the organisation.

**Question 2:**

**Data Mining**

* Data mining is the process which identifies patterns in a pre-built database. It carries out an analysis or knowledge discovery in the databases to analyse the existing database and large datasets to turn raw or unfinished data into useful information and find trends and patterns into it.
* It collects the patterns and knowledge from the available data, identify the valid, novel and potentially useful data and sets in data to solve problems through data analysis in otherwise scattered data.

**Data Profiling**

* Data Profiling, is also a process which analyses raw data from existing datasets, but it is instead used to collect or store the statistics, informative summaries about the data. Also called data archaeology. Data profiling is used to retrieve information about the data itself and assess the quality and improvement of the data. It also helps check data sets for consistency, uniqueness and logic while preparing it for subsequent cleansing, integration, and analysis.
* It primarily deals with the data quality, in areas such as enterprise data warehousing, to define anomalies in datasets. It displays and shows the wrong data at the initial stage of data i.e. at first place so that it can be corrected at the right time before entering.

What is the difference between data profiling and data mining?

|  |  |
| --- | --- |
| **Data Profiling** | **Data Mining** |
| Data profiling is a process where we analyse the data from existing dataset to develop the actual content, structure, and check quality of the data. | Data mining is process where we identify patterns and correlations within a large dataset to derive more useful information. |
| Involves analysis of raw data from existing datasets for purpose of statistics collection or informative summary about data | It involves computer-based methodology ad mathematical algorithm to extract useful information which is hidden inn data. |
| The goal is to create a knowledge based accurate info about the data which helps us to recognise the use and quality of data. | The purpose of data mining is to mine the data for further actions which indeed help to solve data related issues. |
| It employs an activity which includes discovery and analytical technique for collection of statistics. | The primary data mining tasks are classification, regression, clustering, estimation, and description. |

**Question 3:**

**Outliers**

Outliers are the instances which deviate original data from the distribution so that it is making typical set for the data which in turn makes data uncommon. How typical a given point is in relation to the distribution of the data is displayed by the distance from the centre of normal distribution. Each case can be differed based on how likely it is going to be normal or uncommon.

Different types of data mining can suffer from the existence of outliers. Before mining the data, anomaly detection can be performed to find outliers.

outliers are the values within a dataset which vary greatly from the others, they are either much larger, smaller.

Outliers may indicate variabilities in a measurement, experimental errors, or a novelty. When going through data analytics outliers can cause anomalies in the results obtained. This means that they require some special attention and, in some cases, will need to be removed in order to analyse data effectively.

The two main reasons for giving outliers a special attention because:

1. Outliers may have a negative effect on the result of an analysis

2. Their behaviour may be the information that a data analyst requires from the analysis.

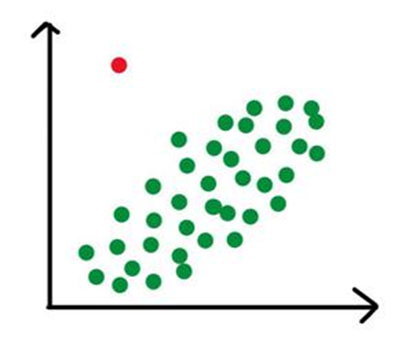
There are two kinds of outliers:

* Univariate outlier- it is an extreme value that relates to just one variable.
* Multivariate outlier- it is a combination of unusual or extreme values for at least two variables.

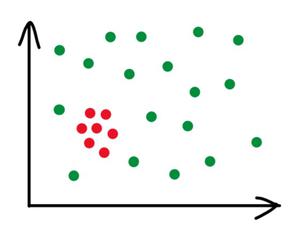
Example

1. In Intrusion Detection System, if a large number of packages are broadcasted in a very short span of time, then this is considered as a global outlier and we can say that that particular system has been potentially hacked.

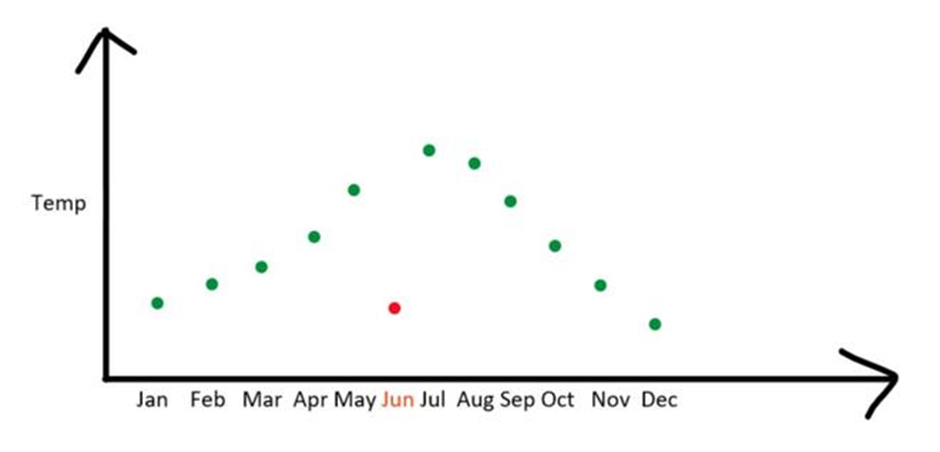
A temperature reading of 50°C may behave as an outlier in the context of a “winter season” but will behave like a normal data point in the context of a “summer season”.



2. In an Intrusion Detection System, a DOS (denial-of-service) package from one computer to another may be considered as normal behaviour. However, if this happens if there are several computers running at the same time, then this may be considered as abnormal behaviour and as a whole they can be termed as collective outliers.



3. A temperature reading of 40°C may behave as an outlier in the context of a “winter season” but will behave like a normal data point in the context of a “summer season”.



**Question 4:**

**Collaborative filtering -** Collaborative filtering (CF) is a technique which is used by recommender systems. Collaborative filtering has two senses, a narrow one and a general one.

In the newer, narrower sense, collaborative filtering is a method which makes automatic predictions (filtering) about the interests or choices of a user by collecting preferences or taste information from many users (collaborating). The approach of collaborative is that if a person *A* has the same opinion as a person *B* on a problem, A is more likely to have B's opinion on a different issue than that of a randomly chosen person. For example, amazon helps people recommend different products to buy each and every time they try to search a product.

Collaborative filtering is the process of filtering information or patterns using techniques involving collaboration among multiple viewpoints, data sources, etc.Applications of collaborative filtering are basically involving very large data sets. Collaborative filtering methods have been applied to many different kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data, etc.

Collaborative filtering algorithms often require

(1) Users active participation

(2) an easy way to represent users' interests

(3) algorithms that are able to match people with similar interests.

A key problem of collaborative filtering is how it combines and weights the preferences or choices of user neighbours. Sometimes, users can immediately rate the recommended items. As a result, the system gains an increasingly accurate representation of user preferences over time.

**Question 5:**

**Time Series Analysis**

A particular method of analysing a set of data points gathered over a period of time is called a "time series analysis." Instead of just capturing the data points intermittently or arbitrarily, time series analysers record the data points at regular intervals over a predetermined length of time.

To maintain consistency and reliability, time series analysis often needs a lot of data. A large data collection guarantees that your analysis can sift through erratic data and that your sample size is representative. Additionally, it guarantees that any trends or patterns are not outliers and can take seasonal variation into consideration. Time series data can also be utilised for forecasting, which is the process of making predictions about the future based on the past.

WHY?

Organizations can better comprehend systemic patterns across time by using time series analysis. Business users can examine seasonal trends and learn more about their causes using data visualisations. These visualisations can do much more than just display line graphs with today's analytics solutions.

Organizations can use time series forecasting to estimate the likelihood of future events when they analyse data at regular intervals. Predictive analytics includes the predicting of time series data. It can indicate expected data changes, such as seasonality or cyclical behaviour, which improves forecasting and gives a better understanding of data factors.

Examples

Non-stationary data—things that change over time or are impacted by time—are studied using time series analysis. Time series analysis is commonly used in sectors like banking, retail, and economics because currency and sales are always fluctuating. When automated trading algorithms are used, stock market analysis is a fantastic illustration of time series analysis in action. Time series analysis is also excellent for predicting weather variations, assisting meteorologists in foreseeing everything from tomorrow's weather report to upcoming years of climate change. Following are some instances of time series analysis in action:

* Weather data
* Rainfall measurements
* Temperature readings
* Heart rate monitoring (EKG)
* Brain monitoring (EEG)
* Quarterly sales
* Stock prices
* Automated stock trading
* Industry forecasts
* Interest rates

**Question 6:**

**Core steps of Data Analysis Project**

The CRISP-DM methodology's six main steps make up the data analytics lifecycle, which explains the process of carrying out a data analytics project. Understanding the business problem, comprehending the data set, preparing the data, exploratory analysis, validation, and visualisation and presentation are some of these phases.

6 steps for data analysis are:

1. Understand the business issues:

You will be provided with a brief summary of the requirements when you are offered a data project. You should be able to deduce the main goals the company is pursuing from that outline. You should consider the project's general scope, business objectives, the information stakeholders are looking for, the sort of analysis they prefer, and the deliverables (the project's results) they require.

In order to produce the finest deliverable possible, you must have these components established precisely before starting your data project. Furthermore, it's crucial to get all the information you need up front because you frequently won't have another chance to do so until the project is finished.

1. Understand your data set:

You can organise your data using a variety of techniques. Excel can be used to investigate and prepare a small dataset, but for heavier projects, you'll probably want to use more rigorous tools. Muoz proposes using Tableau Prep, Tableau Desktop, R, Python, Apteryx, or Python to assist get your data ready for cleaning.

You should locate important factors within these programmes to aid in classifying the data. Look for data errors as you go through the data sets. These may include missing data, information that doesn't make logical sense, redundant information, or even grammatical faults. You must include these missing variables before you can fully clean your data.

1. Prepare the data:

Once we start cleaning your dataset once all the variables have been arranged and recognised. You will fill in any missing variables in this stage, add new broad categories to assist organise data that doesn't have a home, and eliminate any duplicates from your data. The data will be handled more quickly and accurately without being skewed if average data scores are imputed for categories where there are missing values.

1. Perform exploratory analysis and model:

You will start creating models in this step to test your data and look for solutions to the stated objectives. You can choose the statistical modelling approach that works best for your data by comparing various options. Typical models include, among others, decision trees, random forest modelling, and linear regressions.

1. Validate your data:

Once your models are created, you must evaluate the data to see if you have the right information for your deliverable. Did the models operate correctly? Do the data still need to be cleaned up? Did you discover the resolution the client was looking for? If not, you might have to go over the prior actions once more. A great deal of trial and error is to be expected!

1. Visualise and present your findings:

You can start your data visualisation once all of your deliverables have been met. Data visualisation will frequently be essential in conveying your findings to the client. Because not all clients are adept with data, interactive visualisation tools like Tableau are quite helpful in explaining your findings to clients. Your data must be able to convey a narrative. A tale will help the client understand the significance of your findings.

You need to explicitly define your objectives, just as with any project. You can acquire the best deliverables for your clients if you outline your task. Even though each of these phases is crucial, you can end up having to go back if you begin the job without all the necessary information.

Data analysts require a wide range of abilities to be successful in their jobs, from hard abilities like statistical modelling to soft abilities like communication and presentation. A solid mix of non-technical abilities can help you advance your career, even while technical skills are essential to developing a successful career in analytics. As an illustration, the ability to arrange your big data projects in accordance with the data analytics lifecycle is a crucial soft talent that enables you to effectively manage your projects from start to finish.

**Question 7:**

**Characteristics of a good data model**

A data model can link several teams inside an organisation, including IT, business analysts, and management, to provide an information system or applications that are completely functioning. A formal, structured data model is allocated to each application or information system in order to gather and store data, process it, and disseminate useful information.

Making a data model is the process of data modelling. Developers will collaborate extensively with stakeholders at every stage (i.e., potential users of the to-be-created systems). Data modelling starts with gathering information from stakeholders regarding business requirements and standards. In order to construct a specific database architecture, these concepts are subsequently translated into data structures. Data modelling from the bottom up Bottom-up data modelling concentrates on identifying and resolving the data requirements of small groups inside a firm and then combining these separate findings together to meet the requirements of the entire company.

The modelling of top-down data The ideas and issues facing the entire business are the first focus of top-down data modelling. After then, the model is divided into smaller parts to accommodate the unique needs of each group inside a corporation.

There are three primary categories of data models they are:

• Conceptual data model: When high-level information is needed at the beginning of a project, a conceptual data model describes business concepts that are usually used. It alludes to an organised picture of the data that underpins business operations across the entire organisation. Instead of focusing on the unique traits of each entity, its main goal is to determine the classes of entities.

• Logical data model: A logical data model is more in-depth than a conceptual model. Along with entities, it also describes the data structure, the characteristics of entities, and the connections between them.

• Physical data model: A physical data model provides a thorough description of the data model. With main and foreign keys, data types, validation rules, triggers, and procedures, it builds atop a logical data model.

Benefits of Data Models include

• Greater calibre

• Disclosure of data sources

• Transparent data movement within organisations

• Cost cutting

• Improved performance

Which Qualities Make Up a Strong Data Model?

(1) A solid data model makes it simple to consume data.

(2) A solid model can adapt to significant data changes.

(3) A good model's performance can be predicted.

(4) A strong model can change to meet new needs.

**Question 8: Explain and provide examples of univariate, bivariate, and multivariate analysis?**

Univariate- These studies, which fall under the category of "univariate," are descriptive statistical analysis procedures that can be distinguished according to the amount of variables they include at any given time. For instance, the pie charts of sales by territory involve just one variable, making this type of analysis known as a univariate analysis. Due to the fact that the information only pertains to one variable that varies, univariate data analysis is the simplest type of analysis. The major goal of the analysis is to explain the data and identify patterns that exist within it; it does not deal with causes or relationships.

Bivariate: This study seeks to comprehend the distinction between two variables simultaneously, like a scatterplot. Bivariate analysis is an example that may be used to illustrate how to analyse spending and sales volume. It is done to determine the relationship between the two variables during the examination of this sort of data, which deals with causes and relationships. Ice cream sales and temperature during the summer are two examples of bivariate data.

Multivariate: This analysis looks at more than two variables to better understand how they affect the replies. It is comparable to bivariate but has more dependent variables than that. The methods used to analyse this data depend on the objectives to be met. Regression analysis, path analysis, factor analysis, and multivariate analysis of variance are a few of the methodologies.

Example of univariate analysis:

* + Height can be taken as an example as univariate example because height cannot be dependent on any relation or cause. We can take mean, median and mode for height values.
  + Height of a students in a class can be represented in bar, histograms, pie charts and distribution frequency tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Height | 162 | 170.2 | 176 | 156 | 180 | 180.1 |

Example of bivariate analysis:

* + Sales of beverages throughout the summer are a good example of a bivariate analysis because there will be more than one variable and they are all directly proportional to one another.
  + These kinds of independent and dependent bivariate examples are shown using the x and y axes, respectively.

|  |  |
| --- | --- |
| **Temperature** | **Sales of drinks** |
| 20 | 1000 |
| 22 | 1500 |
| 30 | 1700 |
| 32 | 2500 |

Multivariate analysis example:

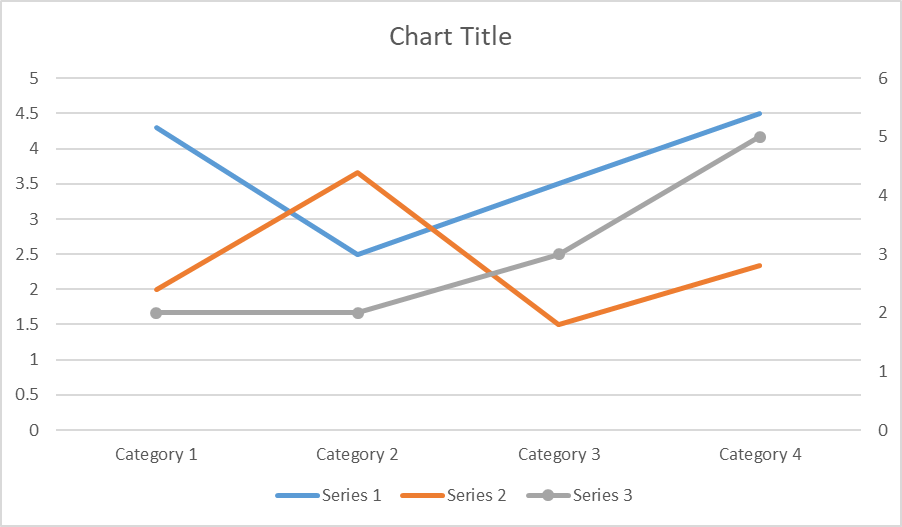
* + As there are more than one or two variables that are reliant on one another to provide an analysis of the patient's condition, a doctor collecting blood levels, cholesterol, height, weight, and gender is a good example of a multivariate analysis.

**Question 9:**

**Linear Regression**

The supervised machine learning approach known as "linear regression" identifies the linear relationship between the dependent and independent variables by finding the best-fitting linear line between them. The least squares approach is used in linear regression.

Drawing a line through each of the shown data points is the idea. The line is set up to be as close to all of the data points as possible. The distance is referred to as "errors" or "residuals." The distance between data points and the mathematical function illustrated is represented by dotted symbols on lines.



import pandas as pd

import matplotlib.pyplot as plt

from scipy import stats

full\_data = pd.read\_csv("entry.csv", header=0, sep=",")

x = full\_data["A"]

y = full\_data ["B"]

slope, intercept, r, p, std\_err = stats.linregress(x, y)

def myfunc(x):

return slope \* x + intercept

mymodel = list(map(myfunc, x))

plt.scatter(x, y)

plt.plot(x, slope \* x + intercept)

plt.ylim(ymin=0, ymax=1000)

plt.xlim(xmin=0, xmax=300)

plt.xlabel("A")

plt.ylabel ("B")

plt.show()

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

The two main issues that affect machine learning and lower the effectiveness of the machine learning models are overfitting and underfitting. Each machine learning model's primary objective is to generalise effectively.

The ability of an ML model to adjust the provided set of unknown input to produce an acceptable output is defined here as generalisation. It indicates that it can generate trustworthy and accurate output after receiving training on the dataset. As a result, underfitting and overfitting are the two words that need to be examined in order to determine how effectively the model performs and generalises.

Let's first learn some fundamental terms that will aid in a thorough understanding of overfitting and underfitting:

o Signal: The genuine underlying pattern of the data that enables the machine learning model to derive knowledge from the data is referred to as the signal.

o Noise: Noise is unneeded and irrelevant data that degrades the model's performance.

o Bias: When machine learning algorithms are oversimplified, bias is a prediction inaccuracy that enters the model. Alternately, it could be the discrepancy between the expected and actual values.

o Variance: This arises when a machine learning model performs well with the training dataset but poorly with the test dataset.

Overfitting

When our machine learning model tries to include all the data points—or more—that are present in the dataset, this is known as overfitting. As a result, the model begins to cache erroneous values and noise from the dataset, which lowers the model's efficiency and accuracy. High variance and low bias characterise the overfitted model.

The more training we give our model; the more likely overfitting is to occur. This indicates that the likelihood of an overfitted model increasing as we train our model. The fundamental issue with supervised learning is overfitting.

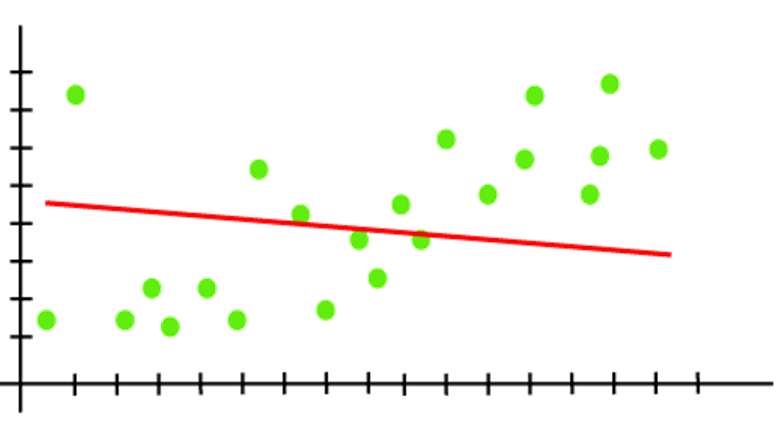
Example: The graph of the results from linear regression shown below helps explain the idea of overfitting:



How to prevent models from being overfitted?

The performance of the machine learning model is deteriorated by both overfitting and underfitting. However, overfitting is the primary culprit, thus there are a few ways we may lessen its incidence in our model.

Cross-validation, training with more data, removing features, ending training early, regularisation, and assembling are some examples of these steps.



As we see above, the model is unable to capture the data points present in the plot.

How to prevent underfitting:

- By lengthening the model's training period.

- By expanding the selection of features.