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

Data Mining	Data Profiling
Finding relevant information which has not been found before	Assess dataset for its uniqueness, consistency and logic
Turning raw data into valuable information	Unable to identify inaccurate data

2. What is Data Wrangling also commonly known as Data Preprocessing?

It is a process of discovering, structuring, cleaning, enriching, validating and analysing data.

3. How to deal with missing values from dataset?

Imputation with Mean/Median/Mode: use these values to fill in the missing values.

Interpolation, such as linear and spline interpolation are useful for continuous data. It is used to estimate missing values based on the relationship between neighbouring data points.

Listwise deletion (It involves removing entire observations (rows) from the dataset if any of the variables (columns) in that observation have missing values.)

4. What are the stages of data life cycle?

Plan: Decide what sort of data is needed for the specific project?

Capture: How do you get the data?

Manage: How should you store the data, to ensure it is safe and secure?

Analyse: Explore insights from data.

Archive: Securely store less frequently used data

Destroy: Remove data securely when it is no longer needed.

5. Best practices for data cleaning

Make a data cleaning plan by understand where the common error take place.

Identify and remove duplicates or outlier(subjective) before working with the data.

Focus on the accuracy of the data. Maintain the value types of the data, provide mandatory constraints (eg; mandatory email address or phone number) and set cross-field validation (Cross-field validation involves checking the relationships between multiple fields or columns in the dataset to ensure they are logically consistent.).

Standardise the data at the point of entry, thus leading to fewer errors.

6. What is normal distribution?

Normal distribution has a bell-shape curved with majority of the data clustered around the means. 68% of the data lies within 1 standard deviation (SD), 95% falls within 2 SD and 99.7% falls within 3 SD.

7. Overfitting vs underfitting

Overfitting	Underfitting
Model trains the data too well with training set	Model neither trains the data well nor can generalise to new data
Performance drops significantly with test set	Performs poorly both on train and test set
Overly complex model	Less data to build an accurate model or selecting wrong model for a certain data type

8. Non-relational vs relational database

Aspect	Relational Database (RDBMS)	Non-Relational Database (NoSQL)
Data Structure	Structured (Tables)	Flexible (Semi-Structured/Unstructured)
Schema	Enforced schema	Flexible or Schema-less
Scalability	Typically vertical scaling	Often designed for horizontal scaling
Examples	MySQL, PostgreSQL, Oracle, SQL Server	MongoDB (document-based), Cassandra (column-family), Redis (key-value), Neo4j (graph)

9. Exploratory data analysis (EDA)

EDA is primarily focused on gaining a deep understanding of the dataset, uncovering insights, and generating hypotheses. It aims to explore the data's characteristics, patterns, and relationships.

10. Descriptive vs Predictive vs Prescriptive

Descriptive

- Raw data into insights
- Uses data aggregation and data mining techniques

Predictive

- Use statistical model and forecasting techniques for future insights

Prescriptive

- combines predictive models with optimization algorithms and business rules to suggest the best course of action based on predicted outcomes.

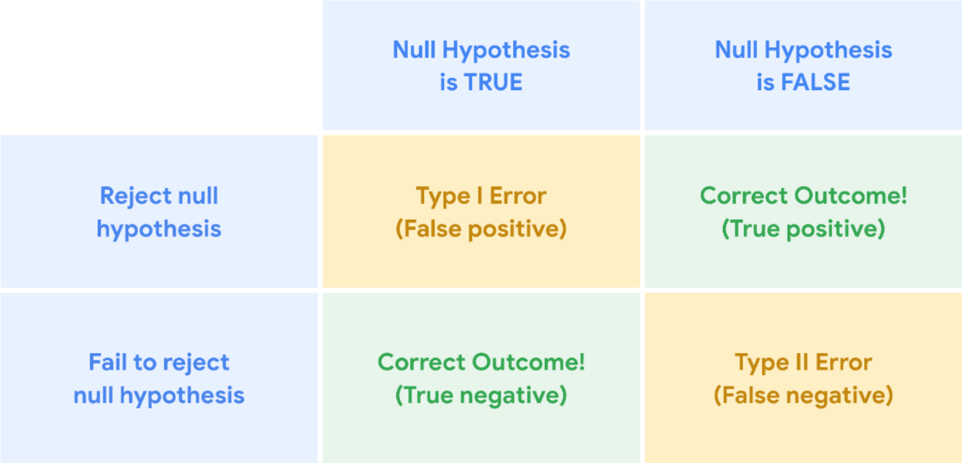
11. Sampling Techniques

Sampling is a statistical method to select a subset of data from an entire dataset (population) to estimate the characteristics of the whole population. Some examples are Cluster sampling, stratified sampling, simple random sampling.

12. Hypothesis Testing

Null Hypothesis: No relation between the predictor and outcome variable in the population.

Alternative Hypothesis: Some degree of relation between predictor and outcome variables



	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

P value: Measurement of the evidence against the null hypothesis. A low p-value indicates that the observed data are unlikely to occur if the null hypothesis is true. Common significance level, alpha, is 0.05. If p-value is lower than alpha, it is typically considered statistically significant.

13. **Supervised Learning**:

Definition: Supervised learning is a type of machine learning where the algorithm learns from labelled training data, which means the data includes both input features and corresponding target labels or outcomes. The goal is to learn a mapping from input to output, allowing the model to make predictions on new, unseen data.

- Linear Regression: Used for predicting a continuous numeric value (e.g., predicting house prices based on features like square footage and number of bedrooms).
- Logistic Regression: Applied for binary classification tasks (e.g., classifying whether an email is spam or not).
- Support Vector Machines (SVM): Useful for both classification and regression tasks, SVM aims to find a hyperplane that best separates data points into different classes.
- Classification is a specific type of supervised learning where the model's task is to categorize input data points into predefined classes or categories.
- Example Models:
    - Decision Trees: Constructs a tree-like model of decisions and their consequences (e.g., predicting whether a person will buy a product based on age, income, and other factors).
    - Random Forest: An ensemble of decision trees that can improve classification accuracy.
    - Naive Bayes: Based on Bayes' theorem, used for text classification tasks like spam
-

14. **Unsupervised Learning**:

Definition: Unsupervised learning involves working with unlabelled data, where the algorithm's goal is to find patterns, structure, or relationships within the data without guidance from predefined labels.

K-Means Clustering: Used for grouping similar data points into clusters (e.g., customer segmentation based on purchase behaviour).