**ASSIGNMENT**

1)Explain Different phases of Data Analytics Lifecycle?

The data analytics lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.

**Phase 1: Discovery –**

* The data science team learn and investigate the problem.
* Develop context and understanding.

# Phase 2: Data Preparation –

* Steps to explore, preprocess, and condition data prior to modeling and analysis.
* It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
* Data preparation tasks are likely to be performed multiple times and not in predefined order.
* Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

**Phase 3: Model Planning –**

* Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
* In this phase, data science team develop data sets for training, testing, and production purposes.
* Team builds and executes models based on the work done in the model planning phase.
* Several tools commonly used for this phase are – Matlab, STASTICA.

**Phase 4: Model Building –**

* Team develops datasets for testing, training, and production purposes.
* Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
* Free or open-source tools – Rand PL/R, Octave, WEKA.
* Commercial tools – Matlab , STASTICA.

**Phase 5: Communication Results –**

* After executing model team need to compare outcomes of modeling to criteria established for success and failure.
* Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
* Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

**Phase 6: Operationalize –**

* The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
* This approach enables team to learn about performance and related constraints of the model in production environment on small scale &nbsp, and make adjustments before full deployment.
* The team delivers final reports, briefings, codes.
* Free or open source tools – Octave, WEKA, SQL, MADlib.

2)Explain Hypothesis Testing and its Type.

Hypothesis testing is a statistical method that is used to make a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

**Defining Hypotheses :**

* **Null hypothesis (H0):** In statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.
* **Alternative hypothesis (H1):** The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

# One-Tailed Test

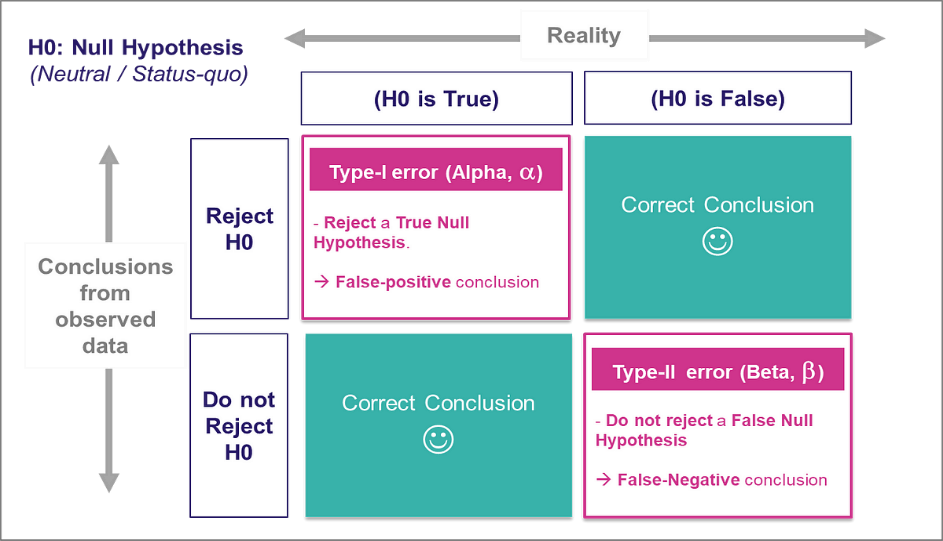
There are two types of one-tailed test:

* **Left-Tailed (Left-Sided) Test:** The alternative hypothesis asserts that the true parameter value is less than the null hypothesis.
* **Right-Tailed (Right-Sided) Test**: The alternative hypothesis asserts that the true parameter value is greater than the null hypothesis

# Two-Tailed Test

A two-tailed test considers both directions, greater than and less than a specified value.We use a two-tailed test when there is no specific directional expectation, and want to detect any significant difference.

1. **Non-parametric test**: Statistical tests that don't rely on specific distributional assumptions. They are used when data don't meet parametric assumptions. Examples include the Wilcoxon signed-rank test for paired samples and the Mann-Whitney U test for independent samples.
2. **T-test**: A parametric test used to compare means between two groups. It assumes normal distribution and equal variances. Commonly used for small sample sizes.
3. **Z-test**: A parametric test used to compare means or proportions when sample size is large. It assumes normal distribution and known population standard deviation. Typically used in hypothesis testing with large sample sizes.



3)Explain Apriori Algorithm.

The Apriori algorithm is a classical algorithm used in data mining and association rule learning. It's primarily employed for discovering frequent item sets within transactional databases. Here's a concise explanation:

1. **Frequent Item Sets**: The algorithm works by iteratively discovering frequent item sets, which are sets of items that frequently occur together in transactions above a specified minimum support threshold.
2. **Principle of Apriori**: The algorithm leverages the "apriori" property, which states that if an item set is frequent, then all of its subsets must also be frequent.

This principle helps in pruning the search space efficiently.

1. **Algorithm Steps**:
   * **Step 1**: Generate all frequent individual items (item sets of size 1) by scanning the database once and counting the occurrences of each item.
   * **Step 2**: Use the frequent items to generate candidate item sets of larger sizes, ensuring that all subsets of these candidate sets are also frequent, based on the apriori property. o **Step 3**: Count the occurrences of each candidate item set in the transactional database. o **Step 4**: Repeat steps 2 and 3 iteratively until no new frequent item sets can be generated. Terminate when no frequent item sets can be found or when the desired set size is reached.
2. **Association Rule Generation**: After finding frequent item sets, association rules can be derived. These rules indicate the likelihood of certain items co-occurring based on the discovered patterns. Association rules typically consist of an antecedent (if) and a consequent (then) part, with measures like confidence and support to evaluate their significance.
3. **Applications**: Apriori algorithm finds applications in market basket analysis, recommendation systems, web usage mining, and more, where discovering frequent item sets and association rules help understand patterns and correlations in large datasets.

In summary, the Apriori algorithm is a fundamental tool in data mining for discovering frequent item sets in transactional databases, leveraging the apriori property to efficiently prune the search space and generate meaningful association rules. 4)Explain Association Rules?

Association rules are a fundamental concept in data mining and data science, particularly in the context of market basket analysis and recommendation systems.

These rules identify relationships or associations between different items in a dataset. Here's an explanation:

1. **Definition**: Association rules describe the co-occurrence patterns of items in transactional datasets. They are typically represented as "if-then" statements, indicating that if certain items are present in a transaction (the antecedent), then other items are likely to be present as well (the consequent).
2. **Metrics**:
   * **Support**: The frequency of occurrence of a particular item set in the dataset. It measures the proportion of transactions that contain the item set.
   * **Confidence**: The conditional probability that a transaction containing the antecedent also contains the consequent. It quantifies how often the items in the consequent appear in transactions that contain the items in the antecedent. o **Lift**: Indicates the ratio of the observed support of the item set to the expected support if the antecedent and consequent were independent. Lift > 1 indicates a positive association, meaning that the presence of the antecedent increases the likelihood of the consequent.
3. **Example**:
   * An association rule might be: {Diapers} => {Beer}, with support of 0.05 and confidence of 0.6.
   * Interpretation: This rule indicates that diapers and beer are frequently purchased together (support of 0.05), and when diapers are purchased, there's a 60% chance that beer is also bought in the same transaction (confidence of 0.6).
4. **Applications**:
   * **Market Basket Analysis**: Understanding which products are frequently purchased together in retail transactions.
   * **Recommendation Systems**: Suggesting additional products or items based on the items already in a user's cart or purchase history.
   * **Cross-Selling**: Identifying opportunities to promote complementary products to customers.
5. **Algorithmic Approaches**:
   * Association rule mining algorithms, such as Apriori and FP-Growth, are used to discover frequent item sets and generate association rules from transactional data. o These algorithms identify patterns of co-occurrence by examining the support and confidence of item sets.

In summary, association rules provide valuable insights into patterns of co-occurrence in transactional data, enabling businesses to make informed decisions about product placement, marketing strategies, and personalized recommendations. They are essential tools in data science for understanding and leveraging the relationships between different items or variables in datasets.

5)What Is ANOVA? ,

It seems like there might be a typo in your question. If you meant to ask about ANOVA (Analysis of Variance), I can provide an explanation.

ANOVA (Analysis of Variance) is a statistical technique used to compare means across two or more groups. It assesses whether there are statistically significant differences between the means of these groups. ANOVA works by partitioning the total variance observed in the data into different sources, such as variation within groups and variation between groups. It then compares the variability between groups to the variability within groups to determine if the differences in means are likely due to random chance or if there is a significant effect of the independent variable (the variable defining the groups).

ANOVA can be used in various scenarios, such as:

1. **One-way ANOVA:** Used when comparing means across two or more independent groups.
2. **Two-way ANOVA:** Used when there are two independent variables influencing the dependent variable.
3. **ANCOVA (Analysis of Covariance):** Extends ANOVA by incorporating one or more continuous covariates into the analysis.

ANOVA provides statistical significance testing through p-values, indicating whether the observed differences between group means are statistically significant or likely to have occurred by chance. If the p-value is below a predetermined significance level (usually 0.05), it suggests that there are significant differences between at least two groups.

ANOVA is widely used in various fields, including psychology, biology, economics, and engineering, to compare means across multiple groups and to determine the effects of different treatments, interventions, or factors on a dependent variable

6)What is Frequent Itemset?

In data science, a frequent itemset refers to a collection of items that frequently appear together in a dataset, particularly in transactional databases. Frequent itemsets are crucial in association rule mining, where the goal is to identify patterns of co-occurrence among items.

Here's a more detailed explanation:

1. **Definition**:
   * A frequent itemset is a subset of items from a larger set of items, such as products in a retail store or words in a document, that occur together frequently within transactions or instances in a dataset. o The frequency of an itemset is measured by its support, which represents the proportion of transactions or instances in which the itemset occurs.
2. **Example**:
   * Consider a retail transaction dataset where each transaction represents a customer purchase. o If items {A, B, C} appear together in 50 out of 100 transactions, then the support of the itemset {A, B, C} is 50%.
3. **Importance**:
   * Frequent itemsets are important because they reveal common associations or patterns in data, such as frequently purchased items in market basket analysis or frequently co-occurring words in text mining. o These patterns can be leveraged for various purposes, including product recommendations, cross-selling strategies, and targeted marketing campaigns.
4. **Mining Frequent Itemsets**:
   * Mining frequent itemsets typically involves using algorithms such as the Apriori algorithm or FP-Growth algorithm.
   * These algorithms systematically search the dataset to identify itemsets that meet a specified minimum support threshold.
   * By pruning the search space efficiently, these algorithms can handle large datasets and discover frequent itemsets effectively.
5. **Support Threshold**:
   * The support threshold is a user-defined parameter that determines the minimum level of support required for an itemset to be considered frequent.
   * Adjusting the support threshold allows analysts to control the number and specificity of the frequent itemsets discovered.

In summary, frequent itemsets represent sets of items that occur together frequently in a dataset, and they play a central role in uncovering meaningful associations and patterns in data, particularly in transactional databases. These patterns can be valuable for making data-driven decisions in various domains, including retail, marketing, and ecommerce.