**WORKSHEET-1**

1. B
2. D
3. D
4. A
5. B
6. D
7. A
8. B
9. D
10. A
11. D
12. A
13. . Calculation(say K-means)

K denotes the number of clusters. The algorithm assigns each datapoint to one of the k groups

Step 1- select the number of clusters to be identified(k value).

Step2-randomly select k no. of datapoints

Step3-measure the distance between 1st point and selected k clusters(It’s the Euclidean distance in case of 2-D)

Step4-now assign the 1st point to the nearest cluster

Step5- calculate the mean value including the new point for the 1st cluster formed.

Step6- pick up the second point and calculate the distance between this point and selected k clusters(where the newly formed cluster replaces the old one)

Step7-again assign this point and calculate the mean and repeat the process

1. Cluster quality

By adding up the variance within each cluster. So the algorithm iterates over again and again unless and until the datapoints within each cluster stop to change. For the second iteration, different initial k points would be chosen. By following this process, an iteration would be arrived at where the sum of variation within each cluster is minimum

1. Clustering is an unsupervised learning technique whereby a dataset is divided into groups consisting of similar data-points. These points in the same are as similar as possible whereas the points in different groups are as dissimilar as possible. The types of Clustering is as follows:-

Exclusive clustering- A type of hard clustering where a datapoint exclusively belongs to one cluster. Ex: K- means

Overlapping clustering- A type of soft clustering where a datapoint belongs to multiple clusters.

Ex: C-means

Hierarchical clustering- Ex: Dendogram whereby n number of datapoints culminate towards a single cluster.

**WORKSHEET-2**

1. e
2. e
3. a
4. a
5. b
6. b
7. a
8. d
9. –
10. A
11. F
12. E
13. The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center
14. Advantages of k-means

Relatively simple to implement.

Scales to large data sets.

Guarantees convergence.

Can warm-start the positions of centroids.

Easily adapts to new examples.

Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

1. The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results.

**WORKSHEET SQL-1**

1. A,D
2. A,B
3. B
4. B
5. A
6. C
7. B
8. B
9. B
10. A
11. A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

It is a blend of technologies and components which aids the strategic use of data. It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

1. OLTP and OLAP both are the online processing systems. OLTP is a transactional processing while OLAP is an analytical processing system. OLTP is a system that manages transaction-oriented applications on the internet for example, ATM. OLAP is an online system that reports to multidimensional analytical queries like financial reporting, forecasting, etc.

The basic difference between OLTP and OLAP is that OLTP is an online database modifying system, whereas, OLAP is an online database query answering system.

1. Subject-oriented –

A data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operations. It can be achieved on specific theme. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc.

A data warehouse never put emphasis only current operations. Instead, it focuses on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

Integrated –

It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various data warehouse in shared and generally granted manner.

A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benefits in effective analysis of data. Reliability in naming conventions, column scaling, encoding structure etc. should be confirmed. Integration of data warehouse handles various subject related warehouse.

Time-Variant –

In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated.

Non-Volatile –

As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantity on logical business. It evaluates the analysis within the technologies of warehouse.

In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:

Data Loading

Data Access

1. Star Schema in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets.
2. Short for Set Theory as a Language (or Set Language), SETL is a high-level programming language that’s based on the mathematical theory of sets. It was developed in the early 1970’s by mathematician Professor J. Schwartz. SETL is an interpreted language with a syntax that is resembles C and in many cases similar to Perl. In SETL every statement is terminated by a semicolon. Variable names are case-insensitive and are automatically determined by their last assignment.

**WORKSHEET SQL-2**

1. D
2. D
3. A
4. A
5. B
6. C
7. A
8. –
9. B
10. D
11. B
12. C
13. A
14. B.c
15. B,c

**WORKSHEET STATISTICS-1**

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C
10. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.
11. First, determine the pattern of your missing data. There are three types of missing data:

Missing Completely at Random: There is no pattern in the missing data on any variables. This is the best you can hope for.

Missing at Random: There is a pattern in the missing data but not on your primary dependent variables such as likelihood to recommend or SUS Scores.

Missing Not at Random: There is a pattern in the missing data that affect your primary dependent variables. For example, lower-income participants are less likely to respond and thus affect your conclusions about income and likelihood to recommend. Missing not at random is your worst-case scenario. Proceed with caution.

And here are seven things you can do about that missing data:

Listwise Deletion: Delete all data from any participant with missing values. If your sample is large enough, then you likely can drop data without substantial loss of statistical power. Be sure that the values are missing at random and that you are not inadvertently removing a class of participants.

Recover the Values: You can sometimes contact the participants and ask them to fill out the missing values. For in-person studies, we’ve found having an additional check for missing values before the participant leaves helps.

Imputation

Imputation is replacing missing values with substitute values. The following methods use some form of imputation.

Educated Guessing: It sounds arbitrary and isn’t your preferred course of action, but you can often infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all “4s”, assume that the missing value is a 4.

Average Imputation: Use the average value of the responses from the other participants to fill in the missing value. If the average of the 30 responses on the question is a 4.1, use a 4.1 as the imputed value. This choice is not always recommended because it can artificially reduce the variability of your data but in some cases makes sense.

Common-Point Imputation: For a rating scale, using the middle point or most commonly chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it’s still among the more risky options. Use caution unless you have good reason and data to support using the substitute value.

Regression Substitution: You can use multiple-regression analysis to estimate a missing value. We use this technique to deal with missing SUS scores. Regression substitution predicts the missing value from the other values. In the case of missing SUS data, we had enough data to create stable regression equations and predict the missing values automatically in the calculator.

Multiple Imputation: The most sophisticated and, currently, most popular approach is to take the regression idea further and take advantage of correlations between responses. In multiple imputation [pdf], software creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions. It is one of a number of examples where computers continue to change the statistical landscape. Most statistical packages like SPSS come with a multiple-imputation feature. More on multiple imputation.

Missing data is like a medical concern: ignoring it doesn’t make it go away. Ideally your data is missing at random and one of these seven approaches will help you make the most of the data you have.

1. A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. [1] A/B tests consist of a randomized experiment with two variants, A and B.[2][3] It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.
2. Key fact to note is that the drawbacks of using a mean apply when the missing data is MAR (Missing At Random)

Mean and mode ignore feature correlations

Mean reduces a variance of the data . a smaller variance leads to the narrower confidence interval in the probability distribution[3]. This leads to nothing else than introducing a bias to our model.

1. . Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.
2. The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics. descriptive statistics deals with the presentation and collection of data. inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics

**WORKSHEET STATISTICS-2**

1. C
2. C
3. D
4. B
5. B
6. B
7. D
8. B
9. D
10. A
11. C
12. D
13. D
14. A
15. D