# ASSIGNMENT 1 answers:

1-option B

2-option D

3- option B

4-option A

5-option C

6-Option D

7-option D

8\_option B

9-option A

10-option A

11-option D

12-option D

13. a)Clusters the data into *k* groups where *k*  is predefined.

b)Select *k* points at random as cluster centers.

c)Assign objects to their closest cluster center according to the *Euclidean distance* function.

d)Calculate the centroid or mean of all objects in each cluster.

e)Repeat steps b, c and d until the same points are assigned to each cluster in consecutive rounds.

14. There are majorly two types of measures to assess the clustering performance. (i) Extrinsic Measures which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

15. Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

TYPES

**Centroid Clustering**-This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify.

**Density Clustering-**Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

**Distribution Clustering-**Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.

**Connectivity Clustering**-Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.

# Assignment 2

1-option A and D

2-option A and C

3- option B

4-option B

5-option A

6-Option C

7-option B

8\_option B

9-option B

10-option A

11. **What is Data Warehousing?**

**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.

12. **What is OLTP?**

An OLTP system captures and maintains transaction data in a database. Each transaction involves individual database records made up of multiple fields or columns. Examples include banking and credit card activity or retail checkout scanning.

In OLTP, the emphasis is on fast processing, because OLTP databases are read, written, and updated frequently. If a transaction fails, built-in system logic ensures data integrity.

**What is OLAP?**

OLAP applies complex queries to large amounts of historical data, aggregated from OLTP databases and other sources, for data mining, analytics, and BUISNESS INTELLIGENCE projects. In OLAP, the emphasis is on response time to these complex queries. Each query involves one or more columns of data aggregated from many rows. Examples include year-over-year financial performance or marketing lead generation trends. OLAP databases and datawarehouses  give analysts and decision-makers the ability to use custom reporting tools to turn data into information. Query failure in OLAP does not interrupt or delay transaction processing for customers, but it can delay or impact the accuracy of business intelligence insights.

13. The key characteristics of a data warehouse are as follows:

* Some data is denormalized for simplification and to improve performance
* Large amounts of historical data are used
* Queries often retrieve large amounts of data
* Both planned and ad hoc queries are common
* The data load is controlled

14. A star schema is a data warehousing architecture model where one fact table references multiple dimension tables, which, when viewed as a diagram, looks like a star with the fact table in the center and the dimension tables radiating from it. It is the simplest among the data warehousing schemas and is currently in wide use.

The star schema is the simplest form of a dimensional model used in business intelligence and data warehousing wherein data is arranged in dimensions and facts. In the star schema, there is a single fact table, which is usually expressed in the third normal form (3NF), and multiple de-normalized dimension tables connected to it, radiating out like the points of a star. The star schema has been optimized for querying large data sets and is generally used in data marts and warehouses in order to support OLAP cubes, ad hoc queries, analytic applications and business intelligence.

The fact tables in a star schema usually have two columns: the first is for the foreign keys pointing to the dimension tables, and the second is for the measures that contain numeric facts, hence, the name fact table. The dimension tables are actually structures that are usually composed of multiple hierarchies that categorize data.

15. **SETL** (SET Language) is a very high-level programming language based on the mathematical theory of sets. It was originally developed by (Jack) Jacob T. Schwartz at the New York University (NYU) Courant Institute of Mathematical Sciences in the late 1960s.

# ASSIGNMENT 3

1-option B

2-option B

3- option B

4-option D

5-option C

6-Option A

7-option B

8\_option A

9-option b

10-A **normal distribution** is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

Here’s an example of a normal distribution curve:

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same.

11. One of the most common problems I have faced in Data Cleaning/Exploratory Analysis is handling the missing values. Firstly, understand that there is NO good way to deal with missing data. I have come across different solutions for data imputation depending on the kind of problem — Time series Analysis, ML, Regression etc. and it is difficult to provide a general solution. In this blog, I am attempting to summarize the most commonly used methods and trying to find a structural solution.

Imputation vs Removing Data

Before jumping to the methods of data imputation, we have to understand the reason why data goes missing.

1. **Missing at Random (MAR):**Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
2. **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
3. **Missing not at Random (MNAR):**Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable’s value (e.g. Let’s assume that females generally don’t want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

# 

* **Imputation Using (Mean/Median) Values** -This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

. Imputation Using (Most Frequent) or (Zero/Constant) Values:

* **Most Frequent-**isanother statistical strategy to impute missing values and YES!! It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.

# Imputation Using k-N- The k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses ‘****feature similarity****’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the k’s closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood. Let’s see some example code using Impute library which provides a simple and easy way to use KNN for imputation:

12. like any type of scientific testing, A/B testing is basically statistical hypothesis testing, or, in other words, [**statistical inference**](https://en.wikipedia.org/wiki/Statistical_inference). It is an analytical method for making decisions that estimates population parameters based on sample statistics.

The population refers to all the visitors coming to your website (or specific group of pages), while the sample refers to the number of visitors that participated in the test.

Let’s say, you make a decision to implement some change on your product pages based on A/B test results that tested a “sample” of the visitors to your website. Ultimately, only a percentage of the visitors saw the challenger, so that  of course means not all the visitors. However, with A/B testing, you assume if the challenger (i.e. variation) in the test increased conversions for a group of visitors on product pages, it will thus have the same result for all the visitors of your product pages (we will delve into the accuracy of a variation’s validity later).

To recap, the A/B testing process can be simplified as follows:

* You start the A/B testing process by making a claim (hypothesis).
* You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
* The final data shows you whether your hypothesis was correct, incorrect or inconclusive,.

13. Bad practice in general

* If just estimating means: mean imputation preserves the mean of the observed data
* Leads to an underestimate of the standard deviation
* Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables.  There are many names for a regression’s dependent variable.  It may be called an outcome variable, criterion variable, endogenous variable, or regressand.  The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable.  Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes.  That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.  A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values.  The regression analysis can be used to get point estimates.  A typical question is, “what will the price of gold be in 6 months?”

Types of Linear Regression

**Simple linear regression**  
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

**Multiple linear regression**  
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

**Logistic regression**  
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

**Ordinal regression**  
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

**Multinomial regression**  
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

**Discriminant analysis**  
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting.  Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R²).  However, overfitting can occur by adding too many variables to the model, which reduces model generalizability.  Occam’s razor describes the problem extremely well – a simple model is usually preferable to a more complex model.  Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

15.Branches of statistics

**Descriptive Statistics**

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

**Inferential Statistics**

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.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.