ML worksheet answers are as follows:

1. b) 4
2. d) 1,2 and 4
3. d) formulating the clustering problem
4. a) Euclidean distance
5. b) divisive clustering
6. d) all answers are correct
7. a) divide the data points in groups
8. b) unsupervised learning
9. d) all of the above
10. a) k means clustering algorithm
11. d) all of the above
12. a) Labeled data
13. To make the clusters, we start by measuring the distance from each data point to each of the 3 centroids. And we assign the points to the cluster closest to it.
    The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters. First, we have to select the variables upon which we base our clusters.
14. If all the data objects in the cluster are highly similar then the cluster has high quality.
15. Cluster analysis is a multivariate data mining technique whose goal is to group objects (e.g., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.
    Types of Cluster Analysis are 1) Hierarchical Cluster Analysis 2) K means 3) DB scan (density based)

SQL worksheet answers are as follows:

1. a & d) create and alter
2. a,b,c) update, delete and select
3. b) structured query language
4. b) data definition language
5. a) data manipulation language
6. c) Create Table A (B int,C float)
7. b) Alter Table A ADD COLUMN D float
8. b) Alter Table A Drop Column D
9. b) Alter Table A Alter Column D int
10. c) Alter Table A Add Primary key B
11. Data-Warehouse: Data-Warehousing integrates data and information collected from various sources into one comprehensive database. It is used for easily extracting, transforming and loading(ETL) data for various analytical use.
12. OLTP: OLTP stands for On-Line Transactional processing. OLTP captures, stores, and processes data from transactions in real time.

    OLAP: OLAP stands for On-Line Analytical Processing. OLAP uses complex queries to analyse aggregated historical data from OLTP systems.
13. Various characteristics of data-warehouse:

    Subject Oriented: A data warehouse provides information on a topic such as sales inventory or supply chain rather than company operations

    Time Variant: Time variant keys example date, month, time are typically present

    Integrated: A data warehouse combines data from various sources. The sources are combined in such a manner that a consistent, relatable and ideally certifiable, providing a business with confidence in the data's quality

    Persistent and non-volatile: Prior data is not deleted when new data is added. Historical data is preserved for comparisons, trends and analytics
14. Star-Schema: A star schema is a multi-dimensional data model used to organize data in a database so that it is easy to understand and analyze. Star schemas can be applied to data warehouses, databases, data marts, and other tools. The star schema design is optimized for querying large data sets.
15. SETL: SETL (set language) is very high level programming language based on the mathematical theory of sets. It is developed by Jack Schwartz.

Statistics worksheet answers are as follows:

1. a) true
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship
10. Normal distribution is a continuous probability distribution wherein values lie in a symmetrical fashion mostly situated around the mean. The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the "bell curve."
11. When dealing with missing data, we can use two primary methods to solve the error: imputation or the removal of data. The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model. The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias.
12. A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better.
13. Mean imputation is not considered a good practice as it distorts relationships between variables. It distorts multivariate relationships and affects statistics such as correlation.
14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
15. The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions. Descriptive statistics include mean (average), variance, skewness, and kurtosis.