1A.

The five V's of Big data is as follows:

- Volume – It indicates the amount of data that is growing at a high rate i.e. data volume in Petabytes
- Velocity – Velocity of data means the rate at which data grows. Social media contributes a major role in the velocity of growing data
- Variety – Term Variety in Big Data refers to the different data types i.e. various data formats like text, audios, videos, etc.
- Veracity – It indicates the uncertainty of available data. The main reason for arising uncertainty is the high volume of data that brings incompleteness and inconsistency
- Value –It refers to turning data into value. By turning accessed big data into values, businesses may generate revenue

2A.

Very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers.

It is a term which is commonly used by data analysts while referring to a value that appears to be far removed and divergent from a set pattern in a sample.There are two kinds of outliers – Univariate and Multivariate.

3A.

The process of clustering involves the grouping of similar objects into a set known as a cluster. In Clustering objects in one cluster are likely to be different when compared to objects grouped under another cluster. It is one of the main tasks in data mining and is also a technique used in statistical data analysis. Hierarchical, partitioning, density-based, and model-based. These are some of the popular clustering methods.

4A.

Linear Regression:

- It requires independent variables to be continuous
- It is based on least squares estimation
- It requires 5 cases per independent variable
- It is aimed at finding the best fitting straight line where the distance between the points and the regression lines are the error

Logistic Regression :

- It can have dependent variables with more than two categories
- It is based on maximum likelihood estimation
- It required at least 10 events per independent variable
- It is used to predict a binary outcome, the resultant graph is an S-curved one