Q1.) Explain the methods to detect an Outlier.

ANS. Detecting Outliers is crucial for ensuring data quality in statistical Analysis. There are several methods to identify Outliers.

1. Visual Methods :-

- Box plot : A Box plot displays the distribution of data and highlights outlier as points outside the "whiskers" (usually 1.5 times the interquartile range from the lower and upper quartiles.)
- Histogram : A Histogram can help visualize the distribution of the data, making it easier to spot any unusual peaks or gaps that may indicate outliers.
- Scatter plot : In Multivariate data, a scatter plot can be used to detect points that are far removed from the general cluster of data.

2. Statistical Methods:-

- Z score : The Z score measures how many standard deviations a data point is from the mean. A Z score greater than 3 or less than -3 is often considered an outlier.

$$\boxed{Z = \frac{(x-\mu)}{\sigma}}$$

, where x → Data point, μ → mean, σ → standard Deviation.

- IQR (Interquartile Range) Method : IQR is the range between the 25th percentile (Q1) and the 75th percentile (Q3). Data points outside the range defined by $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ are considered outliers.

3. Distance Based Methods:-

- K Nearest Neighbours (KNN) : KNN identifies outliers as datapoints whose K nearest neighbours are far away from them.
- Local Outlier Factor (LOF) : This method calculates the local density of data points and identifies outliers as those with significantly lower density compared to their neighbours

4. Clustering Based Methods:-

- Density based Spatial Clustering of Application with Noise (DBSCAN): In this, cluster data points based on their density and identifies outliers as points not belonging to any cluster
- Hierarchical clustering: Hierarchical clustering involves building a hierarchy of clusters by iteratively merging or spliting clusters based on their similarity.

5. Isolation Forest:-

Isolation Forest randomly isolates data points by splitting Features and identifies outliers as those isolated quickly and easily.

6. One Class Support Vector Machines (OCSVM):-

- One Class SVM learns a boundary around the normal data and identifies outliers as points falling outside the boundary

Q2) Explain Data Transformation Methods.

ANS. Data transformation methods modify data to improve its usability and meet the assumptions of statistical models or machine Learning Algorithms.

1. Normalization (Min-max Scaling)

Rescales data to a fixed range (e.g., [0,1]) to ensure all features are on the same scale.

2. Standardization (Z-score)

- Centers data around mean of 0 with a standard deviation of 1, useful for algorithms that assume normal distribution

3. Log Transformation

Applies the Logarithmic function to reduce skewness often used for data with wide ranges.

4. Square root Transformation.

Used to reduce right skewness by applying the square root of data, commonly used to count data.

5. Box-cox Transformation:
~~Used to reduce~~ A power transformation that stabilizes variance and normalizes data, useful when data is skewed.

6. Power Transformation: Applies a power function (e.g., squaring) to data to reduce skewness.

7. Categorical Encoding.
Converts categorical data into numeric formats (e.g. one-hot Encoding, Label Encuding) for use in machine learning models.

8. Binning (Discretization)
Converts continuous data into discrete categories or bins, simplifying data for certain algorithms.

9. Clipping: Limits extreme values by setting them to a specific threshold, often used to handle Outliers.

10. Rank Transformation: Converts data into ranks instead of raw values, useful in non-parametric Analysis.

Q3.) Write an Algorithm to display the statistics of Null values present in the Dataset.

ANS. Algorithm:-

STEP 1: Load the Dataset into a pandas DataFrame.

STEP 2: Check for the Null values in the Dataset.

STEP 3: Calculate the number of null values ~~of~~ for each column.

STEP 4: Calculate the percentage of Null values for each column.

STEP 5: Display the statistics (both count and percentage of Null values.)

Sample Python code:-

```
import pandas as pd
df = pd.read_csv('dataset.csv')
// To check null values.
df.isnull()
null_count = df.isnull().sum()
null_percentage = (null_count / len(df)) * 100
null_stats = pd.DataFrame ({
    'Null Count' : null_count,
    'Null Percentage' : null_percentage
})
print (null_stats)
```

Q4.) Write an Algorithm to replace the Outlier value with the mean of the variable.

ANS. Algorithm:-

1. Load the dataset into a pandas DataFrame.
2. Calculate the IQR for each column (numeric type)
3. Identify outliers using the IQR Method.
   - Outliers are defined as values below Q1 − 1.5 × IQR or above Q3 + 1.5 × IQR, where Q1 is the 25th percentile and Q3 is the 75th percentile
4. Replace Outliers with the mean of the respective column.
5. Display the Modified Dataset.