Q1: Why is it important to define the objectives for any Business problem?

Defining objectives helps in:

- Clarity: Ensures all stakeholders have a common understanding.
- Measuring success: Provides clear success metrics (e.g., operational efficiency or revenue increases).
- Resource allocation: Guides time and budget allocation.
- Focus: Keeps efforts aligned with business needs.

Q2: How to maintain the quality of the Machine Learning model developed for the Business problem?

- Data quality checks: Ensure data consistency, handle missing values, and remove duplicates.
- Cross-validation: Regularly assess model performance using k-fold cross-validation.
- Hyperparameter tuning: Optimize model parameters using techniques like grid search or random search.
- Model monitoring: Post-deployment, monitor for performance drift or bias.

Q3: What is the first document created/drafted for any ML project?

- The Problem Statement Document (PSD) is typically the first. It outlines business goals, success metrics, data sources, potential challenges, and expected outcomes.

Q4: How to load data with multiple sheets?

In Python using pandas, you can load multiple sheets from an Excel file as follows:

```
import pandas as pd

excel_file = pd.ExcelFile("filename.xlsx")

sheets = {sheet_name: pd.read_excel(excel_file, sheet_name) for sheet_name in excel_file.sheet_names}
```

Q5: What are the Auto EDA techniques?

- Pandas Profiling: Generates a report with data types, distributions, correlations, and missing values.
- Sweetviz: Similar to pandas profiling but with additional visual insights.
- Dtale: An interactive tool to explore datasets.
- Autoviz: Auto-generates visualizations for different data types and relationships.

Q6: What are four business moments, and what insights can we draw from them?

The four business moments (statistical moments) are:

- Mean (1st moment): Average value.
- Variance (2nd moment): Measure of data spread.
- Skewness (3rd moment): Indicates asymmetry.
- Kurtosis (4th moment): Measures the "tailedness" or outlier propensity. These moments provide insights into central tendency, variability, asymmetry, and the presence of outliers in the data.

Q7: Write the techniques in data Pre-Processing.

- Handling missing values: Imputation, removal.
- Outlier detection/removal: Z-score, IQR method.
- Encoding categorical variables: Label encoding, one-hot encoding.
- Scaling: Normalization (Min-Max scaling), Standardization (Z-score scaling).
- Feature engineering: Create new features from existing ones (e.g., date splits).
- Data transformation: Log transformation, square root, etc.

Q8: When do we use label encoding and one-hot encoding?

- Label encoding is used for ordinal categorical variables where there is a meaningful order (e.g., low, medium, high).
- One-hot encoding is used for nominal categorical variables where the categories are not ordered (e.g., color: red, green, blue).

Q9: What is the technique to remove outliers?

- Z-score: Data points with Z-scores beyond ±3 are considered outliers.
- IQR method: Values below Q1 - 1.5IQR or above Q3 + 1.5IQR are considered outliers.
- Isolation Forest and DBSCAN (in ML contexts) can also detect outliers.

Q10: What are the techniques to check whether the data is normally distributed or not?

- Histogram: Visual inspection for a bell-shaped curve.

Q-Q plot: Data is compared against a normal distribution.

- Shapiro-Wilk test and Kolmogorov-Smirnov test: Statistical tests for normality.
- Skewness and Kurtosis values: If skewness ≈ 0 and kurtosis ≈ 3, data is likely normal.

Q12: How to make data scale-free?

- Min-Max scaling: Scales values between 0 and 1.
- Standardization (Z-score scaling): Transforms data to have a mean of 0 and a standard deviation of 1.
- RobustScaler: Handles outliers by using the median and interquartile range.

Q13: What types of graphs are used to depict bivariate analysis?

- Scatter plots: For continuous variables.
- Box plots: For categorical vs continuous variables.
- Heatmaps: To show correlations between multiple variables.
- Pair plots: Depict relationships between multiple variables.

Q14: What do you mean by bivariate frequency distribution?

- It refers to the distribution of two variables, examining the frequency of their combinations (e.g., contingency tables for categorical data).

Q15: Which libraries are used in Hierarchical clustering?

- SciPy: Provides tools for agglomerative clustering and dendrogram plotting.
- Scikit-learn: Has AgglomerativeClustering for hierarchical clustering.

Q16: What is the difference between Agglomerative clustering and Divisive Clustering?

- Agglomerative clustering: A bottom-up approach where clusters are merged.
- Divisive clustering: A top-down approach where clusters are split.

Q17: Which metric is used to find distance/similarities between two data points and between a record and a cluster?

- Euclidean distance, Manhattan distance, Cosine similarity, and Minkowski distance are common metrics.

Q18: What are the parameters needed to plot the Dendrogram?

- Linkage matrix: Defines how clusters are merged.
- Distance threshold: Sets where to cut the dendrogram to form clusters.
- Color threshold: Helps visualize clusters.

Q19: How to perform cluster evaluation? Which are the techniques used for cluster evaluation?

- Silhouette score: Measures how similar a point is to its own cluster compared to other clusters.
- Davies-Bouldin Index: Measures the average similarity ratio of each cluster with its most similar one.
- Calinski-Harabasz Index: Ratio of cluster dispersion to cluster compactness.

Q20: What do the Silhouette coefficient, Calinski-Harbasz, and Davies-Bouldin Index indicate in hierarchical clustering?

- Silhouette coefficient: Ranges from -1 to 1, with higher values indicating better clustering.
- Calinski-Harabasz Index: Higher values indicate more distinct and compact clusters.
- Davies-Bouldin Index: Lower values indicate better clustering.