1. **Keep in mind**
   - Understand the context of the data and its domain before applying anything.
   - Avoid data leakage, for example scaling the data before splitting.
   - Try to be as simple as possible.
   - Visualize the data at stages of preprocessing to check for anomalies and trends.
   - Understand the models you want to use.
     - Many models require different techniques, some a lot and some none.

2. **Data Cleaning**
   - **Handle Missing Values:**
     - **Remove missing values:** Drop rows or columns with missing data if the percentage is small.
     - **Impute missing values:** Fill in missing values
       - Mean
       - Median
       - Mode
       - KNN or regression imputation.
   - **Detect and Handle Outliers:**
     - **Z-score:** Identify outliers using statistical methods and either remove or cap them.
     - **Winsorizing:** Limit extreme values in the dataset to reduce the impact of outliers. Especially if extreme values make no sense (i.e. age = 200).
   - **Correct Data Errors:**
     - Fix typos, inconsistent formatting, and incorrect entries.

3. **Data Transformation**
   - **Normalization/Standardization:**
     - **Normalization:** Rescale data to a range, typically [0, 1].
     - **Standardization:** Same as normalization.
     - **Log or power transformations:** Useful for skewed data.
   - **Encoding Categorical Variables:**
     - **One-hot encoding:** Binary columns for each category.
     - **Label encoding:** Numerical labels to categorical values.
     - **Target encoding:** Categorical features using the mean of the target variable.

- **Binning:**
  - Transform continuous data into discrete bins.
  - i.e. states to regions, prices to low, medium, high etc..

4. **Feature Engineering**
   - **Feature Creation:**
     - Combine features into new ones. i.e. extract year from a date.
   - **Feature Scaling:**
     - Scaling features helps reduce bias.
   - **Polynomial Features:**
     - Higher-order terms for non-linear relationships.
   - **Handling Text Data:**
     - Tokenization or stemming.
5. **Dimensionality Reduction**
   - **Principal Component Analysis (PCA):**
     - Reduce the dimensionality of data while retaining the most important information.
   - **Singular Value Decomposition (SVD):**
     - Especially for sparse data like text.
   - **Feature Selection:**
     - Recursive feature elimination (RFE), Lasso, stepwise, PCA.
   - **Variance Threshold:**
     - Remove features with low variance.

6. **Handling Class Imbalance**
   - **Oversampling:**
     - **Random Oversampling:** Duplicate examples from the minority class.
     - **SMOTE (Synthetic Minority Over-sampling Technique):** Synthetic samples based on existing data.
   - **Undersampling:**
     - **Random Undersampling:** Remove examples from the majority class to balance the dataset.

    - **Tomek Links/ENN (Edited Nearest Neighbors):** Good undersampling techniques.
- **Adjust Class Weights:**
    - Penalize misclassification of the minority class more heavily.

7. **Dealing with Multicollinearity**

- **Correlation Matrix:**
    - Calculate the correlation between features and remove highly correlated features (typically above 0.9).
- **Variance Inflation Factor (VIF):**
    - Identify features that are highly collinear and remove them to avoid issues in regression models.

8. **Data Splitting**

- **Train-Test Split:**
    - Split the data into training and testing sets (usually 80/20 or 70/30).
- **Stratified Sampling:**
    - When splitting data, use stratified sampling to maintain the same proportion of each class in both the training and testing sets.

9. **Imbalance within features**

- **Transform Skewed Data:**
    - Use log, square root, or box-cox transformation.