Topics

**VBA Macros -** <https://youtube.com/shorts/CHpkncNdIFk?si=JLrbVF2HuAqyNn_u>

**Data Engineering**

**Data Acquisition**

**Data Storing**

**Data Pre-processing**

**Missing Data**

1. **Getting Extra data**
2. **Dropping Rows/Columns:**
   * Remove rows or columns with excessive missing values if they provide limited utility.
3. **Imputation:**
   * Use statistical methods like **mean**, **median**, or **mode** for numerical data.
   * Apply **forward-fill** or **backward-fill** for time-series data.
4. **Model-Based Imputation:**
   * Use predictive models like regression or k-Nearest Neighbours (KNN) to estimate missing values.
5. **Category-Specific Imputation:**
   * For categorical data, replace missing values with the most frequent category or a placeholder (e.g., "Unknown").
6. **Flagging Missing Data:**
   * Add a binary indicator column to capture the presence of missing values.
7. **Assess impact and validate** – checking for potential biases or distortions
8. **Document and communicate**

**Data Types**

**Duplicates**

**Detecting and handling duplicate rows in a table**

**1. Identifying duplicates**

Define duplicate criteria

e.g. Player name, Country combination

NO JERSEY #

**2. Resolving Duplicates**

Identify records to keep based on competency / recency Runs / Wickets / Career Span

Removing the duplicate rows or archiving them

**3. Preventing future duplicates**

Adding unique constraints on keys

Automatic quality checks

3 tables: ODI, test, T20

One for Runs, one for wickets

Career start year

**How would you ensure Automatic quality checks in data?**

1. Using like ‘%string’ grammar error
2. Terminal <spaces>
3. Concatenate 2 columns as key
4. Distance Functions

**Outliers**

System error / Human error

1. **Detection Methods:**
   * Use statistical techniques like the **IQR rule** or **Z-scores** to identify outliers.
   * Visualize data using box plots or scatter plots for manual inspection.
2. **Trimming or Winsorizing:**
   * Trim outliers or cap them at a specific percentile (e.g., 5th and 95th).
3. **Transformation:**
   * Apply log, square root, or normalization transformations to reduce the impact of outliers.
4. **Robust Models:**
   * Use algorithms like Random Forest or median-based regression that are less sensitive to outliers.
5. **Separate Analysis:**
   * Analyze outliers independently to determine their relevance or potential errors.

**Standard Scalar / Normalization**

1. **Distance-Based Models**
   * K-Nearest Neighbours (KNN): Relies on Euclidean or other distance metrics; unscaled features can dominate the distances.
   * Support Vector Machines (SVM): With kernels (especially RBF or polynomial), the performance depends on the scale of the input features.
   * K-Means Clustering: Distance metrics are used to assign clusters, so normalization ensures fair treatment of all features.
   * Principal Component Analysis (PCA): Sensitive to feature scales because it maximizes variance in the data.
2. **Gradient-Based Optimization Models**
   * Logistic Regression: Normalization ensures faster convergence when using gradient descent.
   * Neural Networks: Scaled features help in stable weight updates and faster training.

**Data Inconsistencies**

variations in data formats and recording practices.

1. comprehensive data mapping and cleansing strategy.

2. establish standardized data formats

Created primary key of Name, Team

3. Removed duplicates

4. Used data dictionaries- For Team name

4. cross verified with online records

**Feature Engineering**

**Test- Train split**

**Large Data Sets**

Anime Recommendation Project