Data Wrangling Report – HR Dataset

This report outlines the data wrangling and cleaning procedures performed on the HR Dataset obtained from Kaggle. The purpose of this process was to prepare the dataset for subsequent analysis and Visualization by ensuring its accuracy, consistency, and completeness.

1. Data Collection

The dataset was imported directly from Kaggle using the 'kagglehub' library. Multiple Excel files were loaded, including 'SatisfiedLevel.xlsx' and 'Employee(E.xlsx)'. Each dataset was explored using functions such as head(), info(), and isnull() to understand its structure, column types, and missing data patterns.

2. Data Assessment

- ✓ The data quality assessment involved checking for missing values and duplicate records across all datasets. Using the pandas functions duplicated() and isnull(), issues were identified that required cleaning actions.

  Duplicate entries were found in the Employee dataset, totaling 19 records. These duplicates were removed using the 'drop_duplicates()' method to ensure that each employee entry is unique.
- ✓ Missing data was identified in key categorical columns such as 'Gender', 'Department', and 'State'. To address these gaps, missing values were imputed using the most frequent value (mode) for each column. This approach preserves categorical distribution while minimizing bias.

3. Data Cleaning

- ✓ For numerical features, the cleaning process primarily involved **examining data ranges** to detect potential outliers rather than modifying them.
  The column *"DistanceFromHome (KM)"* was reviewed by checking its minimum and maximum values to identify any extreme or unrealistic entries.
  However, since no severe anomalies were detected, no transformations or capping were applied at this stage.
- ✓ No missing numerical values were found that required mean or median imputation.
  Therefore, the focus of this stage was limited to verification and ensuring that numerical data distributions were reasonable and consistent with expected business logic.

4. Data Verification

After data cleaning, the dataset was rechecked for remaining null values and duplicates to confirm the effectiveness of the cleaning process. The data structure was verified using the 'info()' and 'isnull().sum()' functions to ensure consistency.

5. Summary

The HR dataset has been successfully cleaned and prepared for further analysis. The main issues resolved include removal of duplicates and imputation of missing categorical values. The resulting dataset is now consistent, complete, and ready for use in exploratory data analysis or modeling.