Data Cleaning and Preprocessing Report: Glassdoor Job Listings

Objective

The objective of this project is to clean and preprocess raw Glassdoor job listings data using Python. This involves removing irrelevant or inconsistent entries, handling missing values, and standardizing key fields. The goal is to prepare the dataset for reliable analysis, such as comparing average salaries across job titles.

Steps Performed in Data Cleaning

1. **Data Import and Initial Inspection**
    - Loaded the dataset using **pandas**.
    - Checked dimensions (shape), data types (info()), missing values (isnull().sum()), duplicates (duplicated().sum()), and performed summary statistics (describe()).
2. **Dropping Unnecessary Columns**
    - Removed the Job Description and index columns to reduce noise and memory usage.
3. **Reset Index**
    - Index was reset to start from 1 for better readability.
4. **Cleaning Company Names**
    - Removed newline characters and extra text like \n from the Company Name column using : .apply(lambda x: x.split('\n')[0]).
5. **Salary Cleaning**
    - Created a temporary column SALARY to clean the Salary Estimate.
    - Removed garbage text like Glassdoor est., $, K, and /yr using regular expressions.
    - Split salary into **Minimum** and **Maximum Salary** columns.
    - Converted these fields to numeric types using pd.to_numeric().

6. **Average Salary Calculation**
    - Calculated **Average Salary** as the mean of Min and Max Salary columns.
7. **Standardizing Job Titles**
    - Converted job titles into Title Case format to maintain consistency.
8. **Placeholder Replacement**
    - Replaced all placeholder values like -1 with NaN for accurate missing value handling.
    - If the **column's data type** is **object or string**, replaced all values with np.nan
    - If the **column's data type** is **numeric** (int, float), replaced all values with 1
9. **Duplicate Removal**
    - Dropped any duplicate entries to avoid data skew.
10. **Final Overview**
    - This project focused on cleaning and preparing job listings data from Glassdoor to enable accurate analysis. Key steps included removing unnecessary columns, handling missing values, standardizing formats for salary and job titles, and eliminating duplicates. After preprocessing, the dataset became suitable for extracting insights