## Source of Dataset

## Pakistan General Elections Dataset 1970-2018

* <https://www.kaggle.com/datasets/tahminashoaib86/pakistan-general-elections-dataset-1970-2018?resource=download>

## Task

* Data Pre Processing

## Dataset Description:

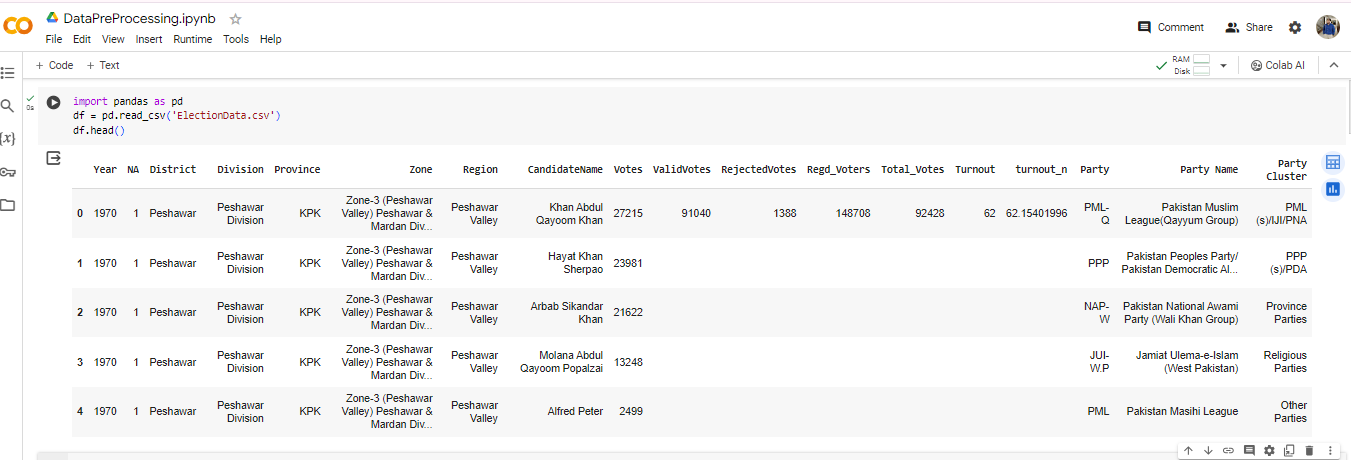
* This dataset comprises the electoral results of the last 10 elections, providing year-wise and national constituency-wise data. It includes the number of registered voters, total votes polled, rejected votes, valid votes, and candidate-wise vote counts. The dataset also contains information on the candidate's affiliated party and the geographical identification of the national constituency, such as province, division, district, zone, and region.

## Data Preprocessing Steps

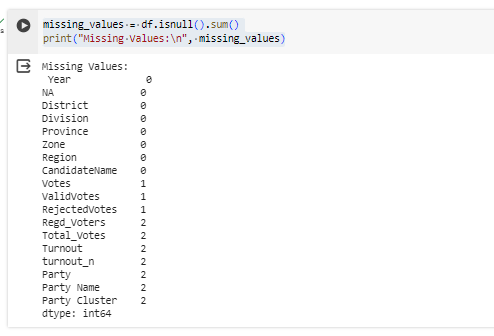
Data Pre-processing includes the following steps which are described further with code.

1. [Data quality assessment](https://monkeylearn.com/blog/data-preprocessing/#quality-assesment)
2. [Data cleaning](https://monkeylearn.com/blog/data-preprocessing/#cleaning)
3. [Data transformation](https://monkeylearn.com/blog/data-preprocessing/#transformation)
4. [Data reduction](https://monkeylearn.com/blog/data-preprocessing/#reduction)

# [Data quality assessment](https://monkeylearn.com/blog/data-preprocessing/#quality-assesment)

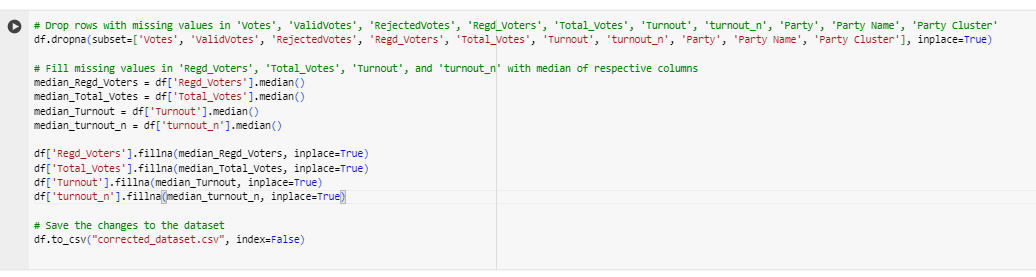


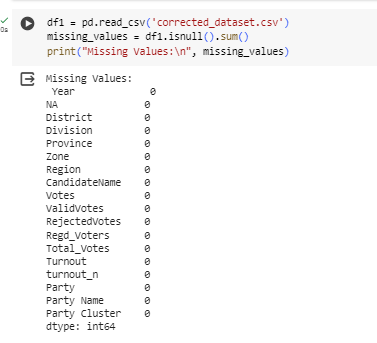
Checking how many missing values are there in each column of a table and shows the result.



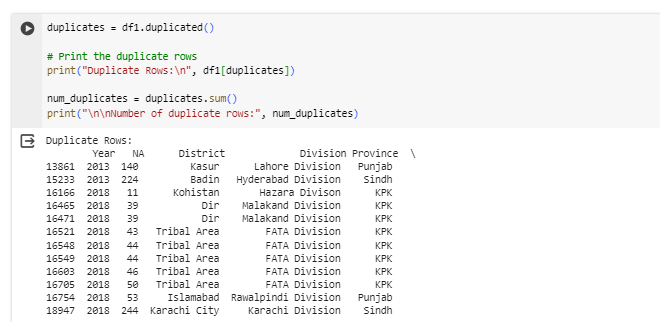
# [Data cleaning](https://monkeylearn.com/blog/data-preprocessing/#cleaning)

Removing rows with missing values in specific columns, fills missing values in other columns with their median values, and saves the changes to a new CSV file.



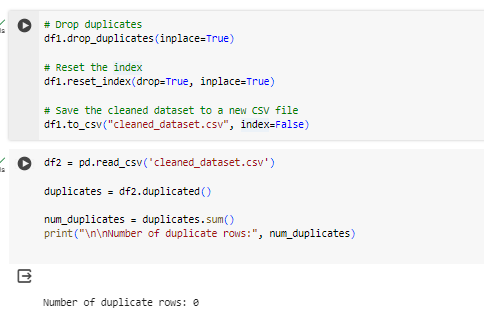


Finds and prints duplicate rows in the Data Frame `df1`, and then calculates and prints the total number of duplicate rows.



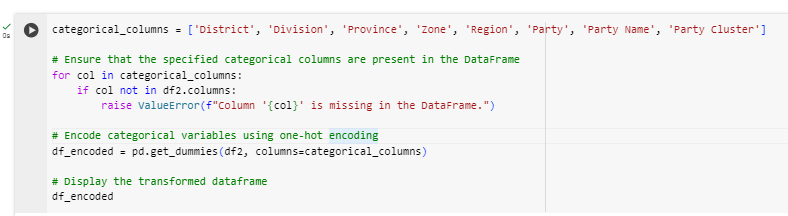


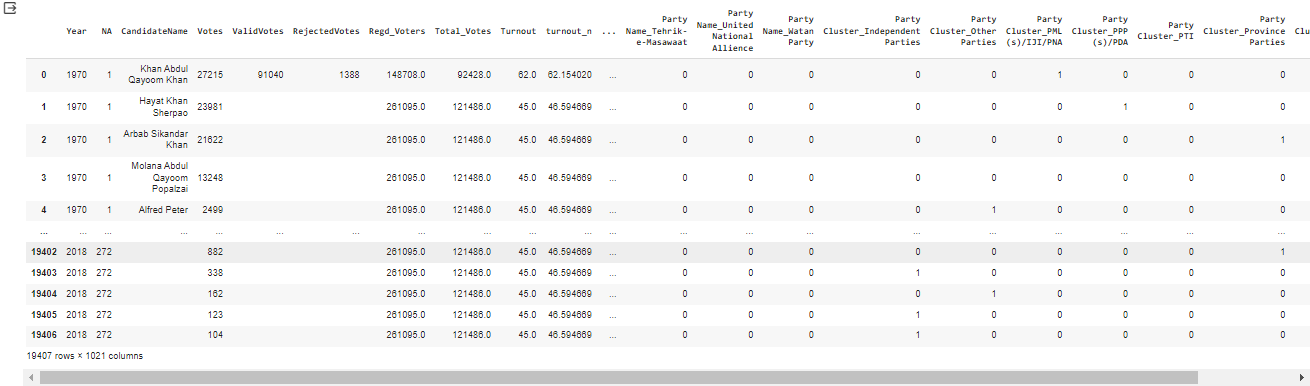
This code drops duplicate rows from the Data Frame `df1`, resets the index, saves the cleaned dataset to a new CSV file named "cleaned\_dataset.csv", reads the cleaned dataset back into a new Data Frame `df2`, and finally calculates and prints the total number of duplicate rows in `df2`.



# [Data transformation](https://monkeylearn.com/blog/data-preprocessing/#transformation)

This code performs one-hot encoding on the specified categorical columns in the Data Frame `df2`, ensuring that these columns exist in the Data Frame before encoding. The resulting transformed Data Frame, `df\_encoded`, contains binary indicators for each category in the original categorical columns. Then code is saved after one-hot encoding to a new CSV file named "encoded\_dataset.csv". The parameter `index=False` ensures that the index of the Data Frame is not included in the CSV file.

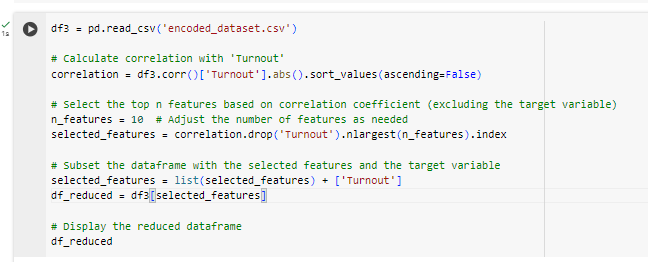


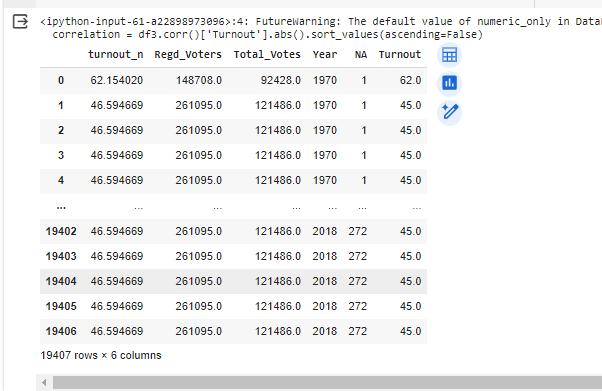


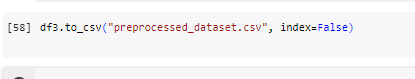


# [Data reduction](https://monkeylearn.com/blog/data-preprocessing/#reduction)

This code reads the encoded dataset from the file "encoded\_dataset.csv" into a Data Frame `df3`. Then, it calculates the correlation of each feature with the target variable 'Turnout', selects the top n features based on their correlation coefficients, and subsets the Data Frame to include only these selected features along with the target variable 'Turnout'. Finally, it displays the reduced Data Frame `df\_reduced`. Then this code is saved by the Data Frame `df3` (which represents the preprocessed dataset after encoding and reduction) to a new CSV file named "preprocessed\_dataset.csv". The parameter `index=False` ensures that the index of the Data Frame is not included in the CSV file.







After all the preprocessing steps are completed and the final preprocessed dataset is saved to the "preprocessed\_dataset.csv" file.

