Milestone 3 – Data preprocessing and Feature engineering stage

Data preprocessing refers to the steps and techniques applied to raw data before it can be used for analysis or machine learning tasks. The quality and accuracy of the preprocessing stage significantly impact the outcomes and effectiveness of subsequent data analysis.

Here are some common steps involved in data preprocessing:

**Data Cleaning:** This involves handling missing data, duplicate records, and correcting inconsistencies or errors in the data. Missing data can be imputed or deleted based on the nature and quantity of missing values. Duplicate records can be removed to avoid bias and duplication during analysis.

Following columns have null values:

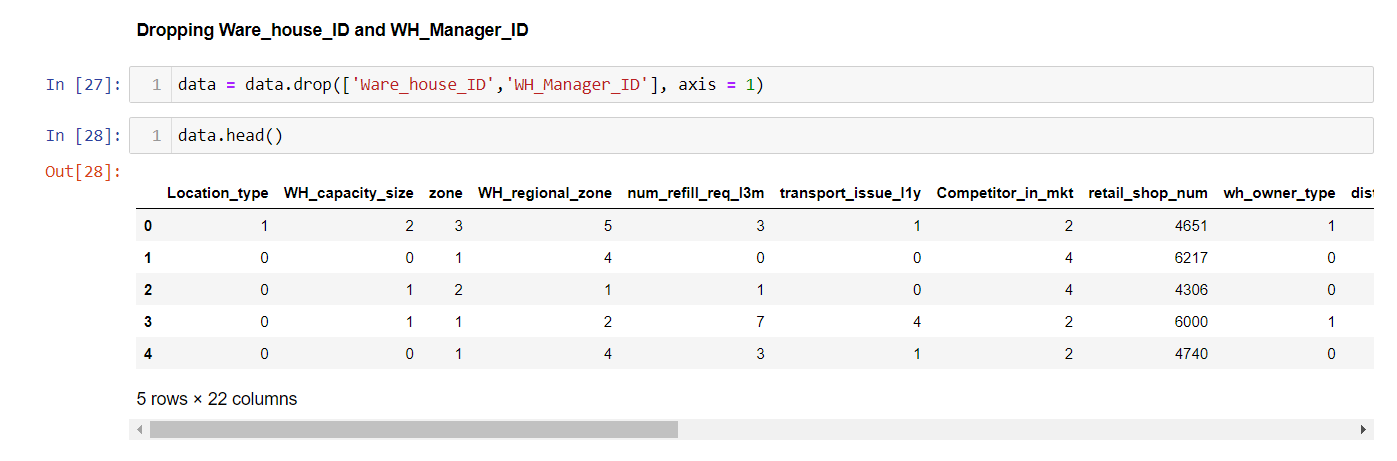
**workers\_num:** Missing values were replaced by Median value

**wh\_est\_year:** Missing values were replaced by Median value

**approved\_wh\_govt\_certificate:** Missing values were replaced by Mode value.

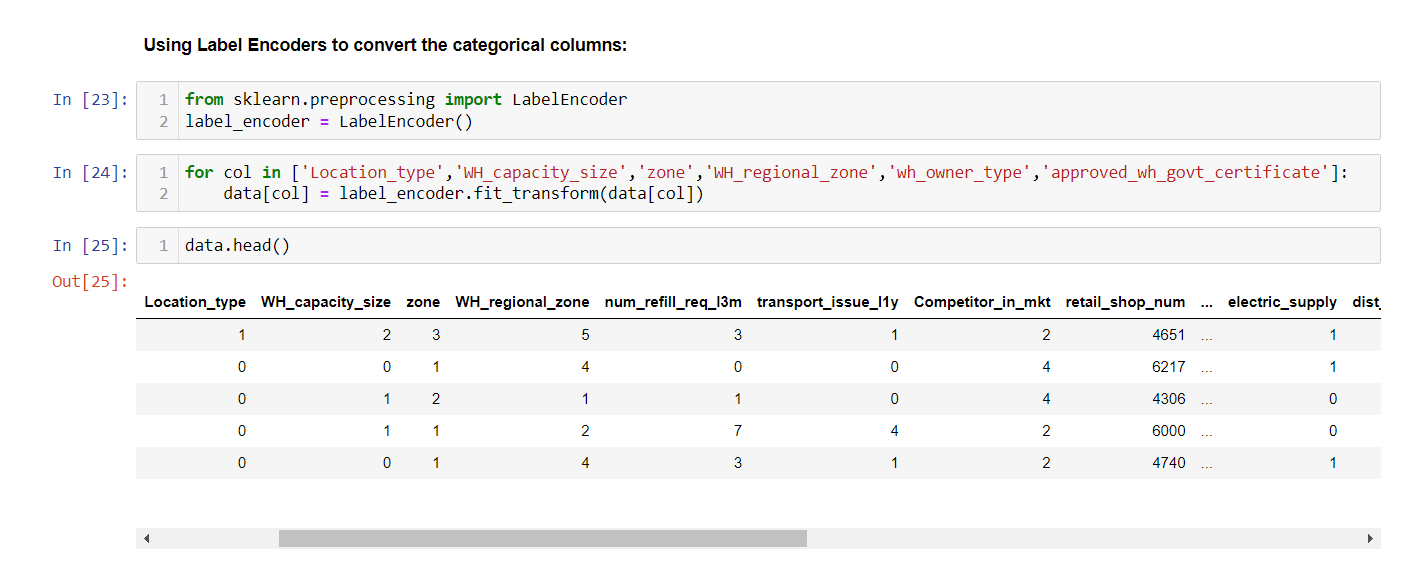
. 

Since Warehouse\_ID and Manager\_ID are distinct ID, we will be dropping it since these two features does not contribute in predicting the target.

****

**2. Data Transformation:** Data transformation involves converting the data from one format to another, often to meet the requirements of the analysis or modelling techniques. Common transformations include scaling, normalization, logarithmic transformations, and categorical variable encoding (e.g., one-hot encoding).

Following columns are of categorical type, so here we will be converting them to numerical type with the help of label encoders:



**3. Feature Selection/Extraction:** Feature selection aims to identify the most relevant features that contribute the most to the analysis or modelling task while reducing dimensionality. Feature extraction techniques like principal component analysis (PCA) or singular value decomposition (SVD) can be used to derive new features from existing ones.

Following tests were carried out to check for feature importance:

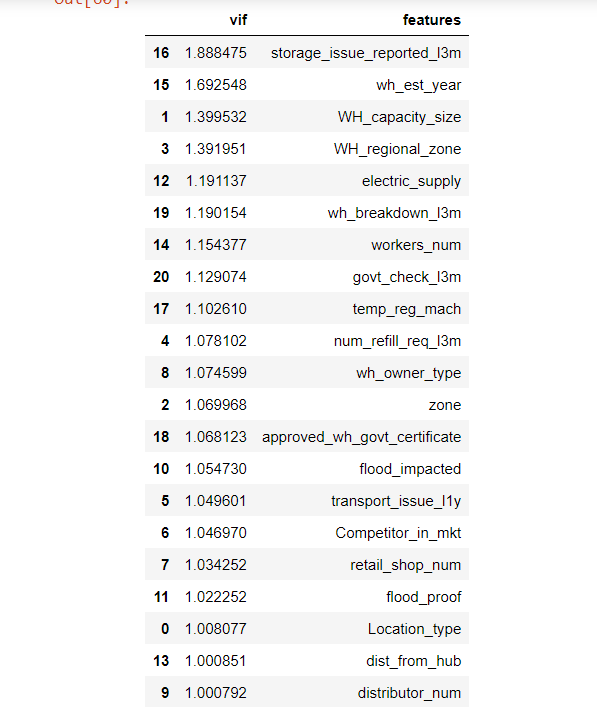
**Variance Inflation Factor (VIF):**

It is a measure of multicollinearity, which is the presence of high correlation among predictor variables in a regression model. VIF assesses how much the variance of the estimated regression coefficient is inflated due to multicollinearity.

The formula for calculating VIF for a particular feature is as follows:

VIF (feature) = 1 / (1 - R^2)

Where R^2 is the coefficient of determination obtained by regressing the feature against all other predictor variables.



Results for VIF test shows that there is no multicollinearity between features.

**Mutual Information Gain:**

Mutual Information Gain (MIG) is a measure used in feature selection or feature ranking for regression tasks. It quantifies the amount of information that a feature provides about the target variable in a dataset. In the context of regression, MIG assesses the relationship between a feature and the continuous target variable.



MIG scores as plotted above shows that,

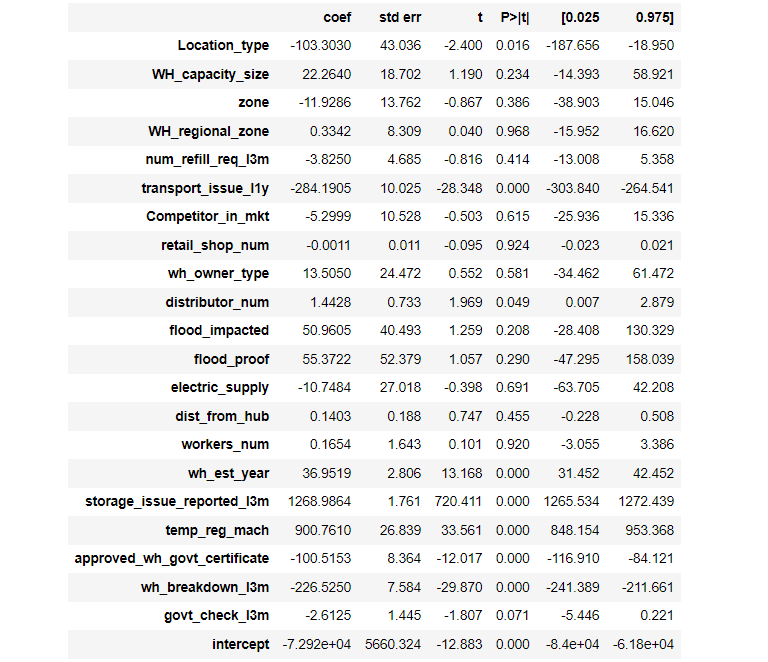
* storage\_issue\_reported\_l3m,
* wh\_breakdown\_l3m
* wh\_est\_year
* transport\_issue\_l1y
* approved\_wh\_govt\_certificate

These features can explain most of the variance with regards to the target variable.

**OLS Regression Method:**

OLS (Ordinary Least Squares) is a widely used regression technique that estimates the coefficients of a linear regression model by minimizing the sum of squared differences between the observed and predicted values of the target variable. OLS assumes a linear relationship between the predictors and the target variable, and it aims to find the best-fit line that minimizes the residual errors.

OLS can be used for feature selection through the examination of the estimated coefficients or p-values associated with each predictor variable.



Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 2.63e+06. This might indicate that there are

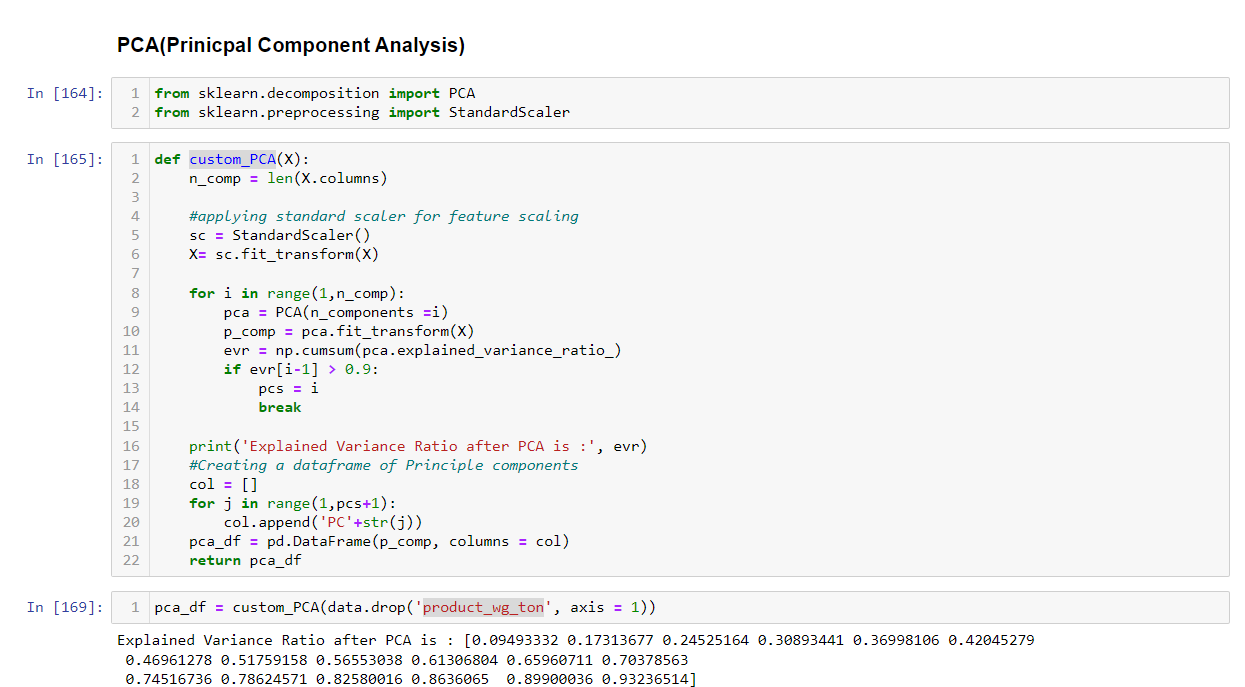
strong multicollinearity or other numerical problems.

Since the condition number is large, the dependent variable or target variable will have high impact with small change in independent variable. This means that the correct solution/answer to the equation will become hard to find.

**Summary:**

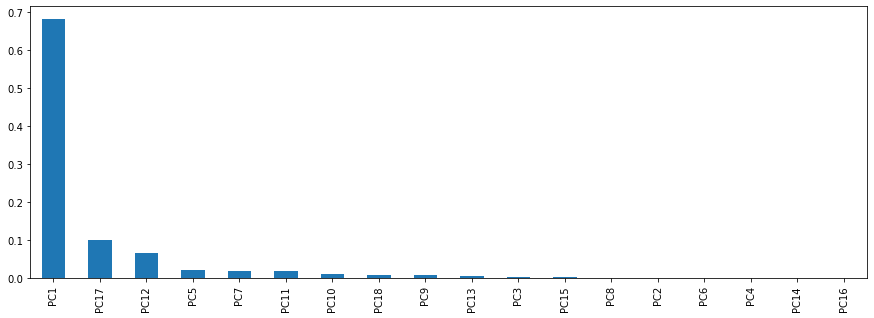
* Only few features in the dataset contribute to the prediction of target variable, and few features are dominant and change in them have higher impact in the variance of the target variable.
* Performing PCA (Principal component Analysis) might reduce the dependency of target variable on only few features.

**Principal component Analysis:**

****

After performing PCA we can see that the features got reduced to 18 and these are able to explain 93% of variance in the target.

MIG score post PCA:



From the graph we can see that MIG score is slightly well distributed showing that PCA has reduced the dependency of target on single variable.

THANK YOU