<u>Checkpoint 1</u>

BFI-Sustainability Team

Team Members - Akash Kesani, Ravindu Tharanga Perera, Samuel Kannan

Summary of Data

The dataset appears to be a collection of records, potentially from a customer service or issue tracking system, with a wide range of attributes. Here's a preliminary analysis based on dataset:

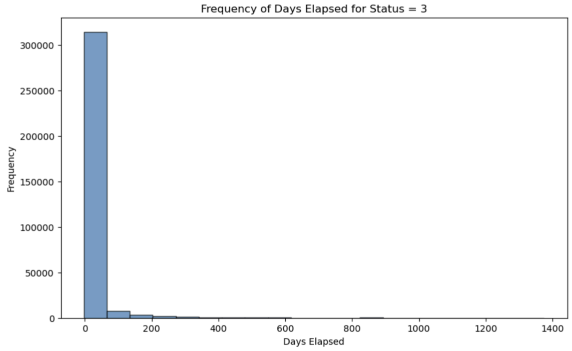
-Unit of Analysis: Each record represents a unique issue or case.

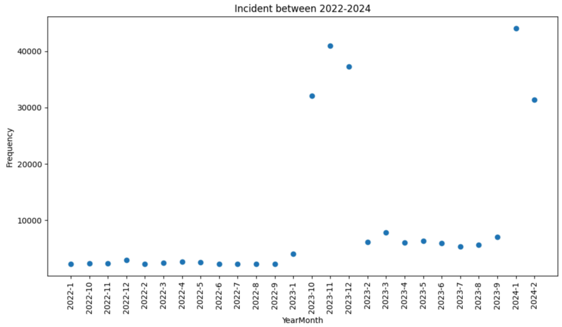
-Total Observations: The dataset contains 377,060 observations.

-Unique Observations: There are 377,060 unique IDs, indicating each observation is unique.

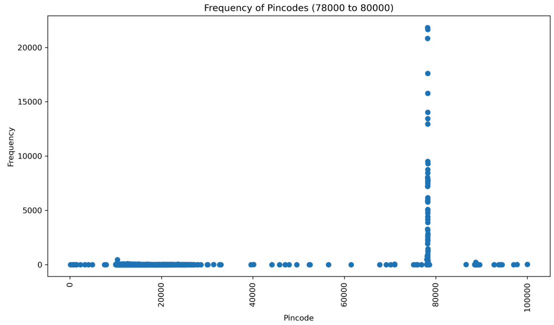
-Time Period Covered: Time Period Covered: The dataset covers issue or cases from April 18, 2015, to February 24, 2024

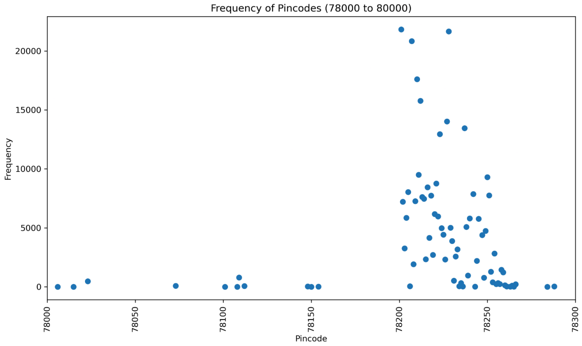
-Some critical extracts from the dataset





Majority of the cases are closed within the first 10 days, with some spilling over the next 100 days. The incidents recorded in the dataset have been between 2014 to 2024, but majority of the cases are from the years 2023 and 2024.





Majority of the data collected has been in the San Antonio area whose pincodes fall in the region of 78200 to 78300.

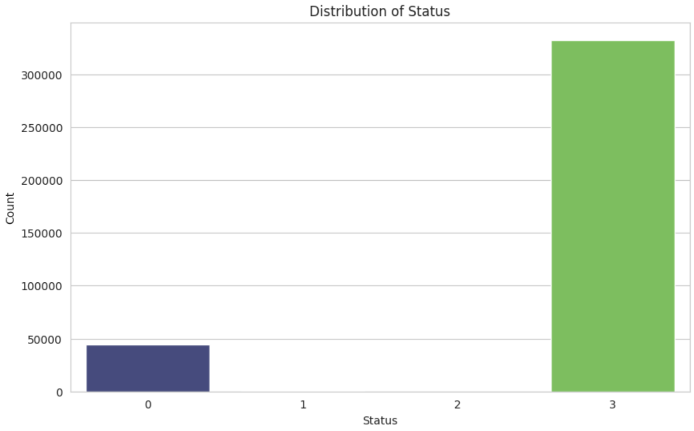
Data Cleaning

Identified several columns with more than 50% missing values, we removed irrelevant columns, especially those with non-informative values (like empty lists or largely NaN values).

Handling missing values or NaNs in critical columns such “Flag Category ID”, “Flag Subcategory ID”, Parsed date-related columns in our dataset like ‘Date Created’ and ‘Date Closed’ into appropriate datetime formats for analysis.

Description of Outcome

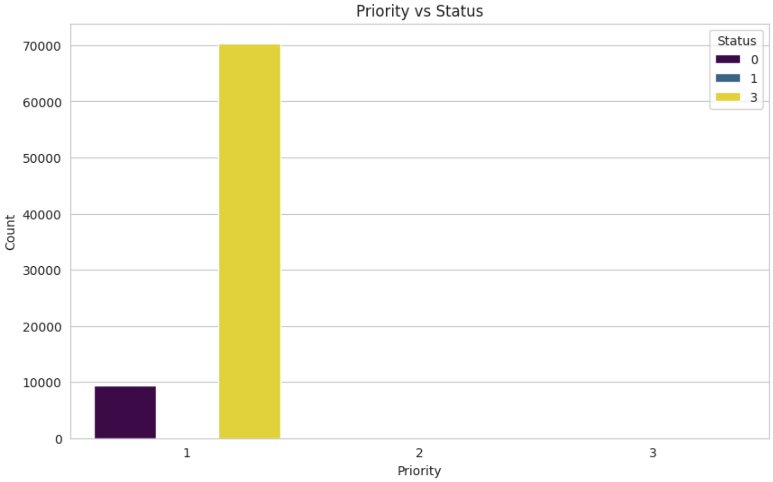
The outcome variable identified for analysis is 'Status', which represents the current state of each issue or case.



Where ‘0’ indicates pending cases, ‘1’ indicates accepted/open cases, ‘2’ indicates rejected cases and ‘3’ indicates solved cases. The reason for ‘Status’ as the outcome variable is because it shows the overall effectiveness of the ‘LAGAN’ system in resolving reported cases which was based on our domain knowledge. In the dataset most of the cases are solved followed by pending cases.

Description of Key Predictors

Based on evaluating all predictors we have selected 'Priority' as one of the key predictors . Priority is a continuous variable with values from 0-3. 'Priority' directly indicates the urgency or importance assigned to each case and would be critical in influencing the Status of each case



Based on the above plot, we can infer completed cases have higher priority than incomplete cases adding cause to ‘Priority’ being one of the key predictors for our response variable ‘Status’.