Summer Internship Project Report

On

Address Cleaning and Sanitization Program

Under the Guidance of

Mr. Sanjay Makhija, Head Analytics(General Manager) BSES Delhi

Submitted By

Rahul Kushwaha

ACKNOWLEDGEMENT

It gives me immense pleasure and satisfaction having completed the project successfully. I take this opportunity to express my sincere gratitude to the people who have been instrumental behind this success.

I express my sincere thanks to Mr. Sanjay Makhija , General Manager (Head Analytics) for giving me a great opportunity to work in such a dynamic organization and for guiding me in all stages of the project. I have a deep sense of gratitude and respect for the entire staff of BSES Delhi for sharing their knowledge and for assisting me. Their help has sparked my interest even more.

Rahul Kushwaha

INDEX

1.Introduction

The purpose of this project is to develop an address cleaning and sanitization program that converts a given raw address dataset into a usable dataframe. The company is facing challenges with addresses that contain special symbols, repeated words, extra useless information, and improper data formatting. These issues can impact the efficiency and accuracy of data analysis, as well as the retrieval of relevant information. The address cleaning and sanitization program aims to overcome these challenges by implementing a series of preprocessing steps to refine the address dataset. In this report, we provide an overview of the project process, along with detailed explanations for each step, and provide additional insights and information.

2. Objectives Of The Project

The objectives of the project are to develop an address cleaning and sanitization program that converts a raw address dataset into a usable dataframe. The program aims to address challenges such as special symbols, repeated words, useless information, and improper data formatting in the addresses. By implementing preprocessing steps, the program will refine the dataset to improve data analysis efficiency and accuracy. The specific objectives include manual removal of visible garbage values, handling and managing null values appropriately, inserting spaces between numerals and characters for better readability, replacing special characters with spaces for standardization, and creating address part markers for organized data storage. Additionally, the program will move Pincodes to separate columns for targeted analyses, remove unnecessary information from addresses to focus on pertinent details, and replace known words and abbreviations to ensure consistency. The project also involves creating a corpus of non-numeric words and fuzzy matching to enhance word data coverage. The final objective is to export the processed data to CSV files for seamless integration and utilization in various applications and projects.

3. Process of Address Cleaning and Sanitization

3.1 Reading Dataset and creating Dataframe:-

In this initial step, we read the address data from the system and create a dataframe to store the data in a structured format. This allows for efficient manipulation and analysis of the address dataset. By using a dataframe, the company can leverage the functionalities and benefits offered by dataframes, such as easy data access, filtering, and transformation.

	SDO_CD	CA_NO	ADDRESS_ORG
0	1211	101363856.0	#, 478 KALANDER COLONY, DILSHAD GARDEN, SHAHDA...
1	1211	101296041.0	#, 1438-A-45, GALI-4, #, BALBIR NAGAR EXTN SHA...
2	1211	101358487.0	#, 1/5484 , GALI NO-17, #, BALBIR NAGAR EXTN, ...
3	1211	101358755.0	#, 1/6221 GALI NO-4, #, EAST ROHTASH NAGAR SHA...
4	1211	101220086.0	T-10, #, #, NAVEEN SHAHDARA, #, 110032
5	1211	101209256.0	1/6421-A-1, #, #, EAST ROHTASH NAGAR SHAHDARA,...
6	1211	101209263.0	42-A, PKT R, #, DILSHAD GARDEN SHAHDARA, #, 11...
7	1211	101325114.0	E-23-F-1, #, #, DILSHAD COLONY SHAHDARA, #, 11...
8	1211	101157651.0	E-125-B, #, #, DILSHAD GARDEN SHAHDARA, #, 110095
9	1211	101229156.0	#, 1619/16 F/F, #, ULDHANPUR NAVEEN SHAHDARA, ...

3.2 Manually removing visible garbage values:-

Visible garbage values, such as random characters or nonsensical entries, can introduce noise and disrupt subsequent operations. The manual removal of these garbage values ensures the integrity and quality of the address data. By taking the time to inspect and remove these values, the program can produce a cleaner dataset, resulting in more accurate and reliable analysis outcomes.

	ADDRESS_ORG
112920	C-2, GROUND FLOOR KH NO-694 OLD NO-1412/80/7-K/C, GALI NO-10 MAIN MANDOLI ROAD, EAST NATHU COLONY, NEAR ,,, 110093
112921	36-B, MIG FLAT, #, MANSAROVER PARK SHAHDARA, #, 110032
112922	1/2303, #, #, RAM NAGAR SHAHDARA, #, 110032
112923	#, SR. SECTION ENGINEER. (POWER SUPPLY) KORIA PUL,, NORTHERN RAILWAY, KAURIA PUL, #, #, 110006
112924	10/9, BLOCK A, #, JHILMIL TAH INDSTL AREA, NEAR NEAR NEAR MCD DUSTBIN
112926	F-18/1 G/F, #, #, DILSHAD COLONY SHAHDARA, #, 110095
112927	#, B-43/10-A NEW NO E-115 GALI POLE-26, #, PREM EAST BABARPUR SHAHDARA, #, 110093
112928	#, E-505-B GALI-4, #, EAST BABARPUR SHAHDARA, #, 110093
112929	C-346 G/F, KHNO-16C-BLOCK,, GALI NO-7, PRATAP NAGAR SABOLI, SHIVLAYA INDIYAN PUBLICF SCHOOL, 110093
112930	E-416/2, G/F KH NO-24/20, MAIN 15 FT ROAD, WEST KARAWAL NAGAR, NEAR KANHAIYA PUBLIC SCHOOL, 110094

3.3 Checking and handling null in the Dataframe:-

Null values are a common challenge in datasets and can create problems during data processing. By checking for null values in the dataframe, the program identifies missing or empty entries in the address data. Handling these null values appropriately ensures that subsequent operations are not affected by missing or incomplete information. Techniques such as data imputation or data exclusion can be applied based on the specific requirements of the project.

3.4 Inserting space between numerals and characters:-

Numerals and characters are often written together in raw address data, which can make it challenging to parse and extract relevant information. By inserting spaces between numerals and characters, the program enhances the readability and comprehensibility of the address data. This step facilitates the accurate identification and separation of different address components, such as street numbers, building names, and apartment numbers, leading to more precise analysis and information retrieval.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS
0	1211	101363856.0	#, 478 KALANDER COLONY, DILSHAD GARDEN, SHAHDARA, NEAR RED CROSS HOSPITAL, 110095	#, 478 KALANDER COLONY, DILSHAD GARDEN, SHAHDARA, NEAR RED CROSS HOSPITAL, 110095
1	1211	101296041.0	#, 1438-A-45, GALI-4, #, BALBIR NAGAR EXTN SHAHDARA, #, 110032	#, 1438 -A- 45 , GALI- 4 , #, BALBIR NAGAR EXTN SHAHDARA, #, 110032
2	1211	101358487.0	#, 1/5484 , GALI NO-17, #, BALBIR NAGAR EXTN, #, 110032	#, 1 / 5484 , GALI NO- 17 , #, BALBIR NAGAR EXTN, #, 110032
3	1211	101358755.0	#, 1/6221 GALI NO-4, #, EAST ROHTASH NAGAR SHAHDARA, #, 110032	#, 1 / 6221 GALI NO- 4 , #, EAST ROHTASH NAGAR SHAHDARA, #, 110032
4	1211	101220086.0	T-10, #, #, NAVEEN SHAHDARA, #, 110032	T- 10 , #, #, NAVEEN SHAHDARA, #, 110032
5	1211	101209256.0	1/6421-A-1, #, #, EAST ROHTASH NAGAR SHAHDARA, #, 110032	1 / 6421 -A- 1 , #, #, EAST ROHTASH NAGAR SHAHDARA, #, 110032
6	1211	101209263.0	42-A, PKT R, #, DILSHAD GARDEN SHAHDARA, #, 110095	42 -A, PKT R, #, DILSHAD GARDEN SHAHDARA, #, 110095
7	1211	101325114.0	E-23-F-1, #, #, DILSHAD COLONY SHAHDARA, #, 110095	E- 23 -F- 1 , #, #, DILSHAD COLONY SHAHDARA, #, 110095
8	1211	101157651.0	E-125-B, #, #, DILSHAD GARDEN SHAHDARA, #, 110095	E- 125 -B, #, #, DILSHAD GARDEN SHAHDARA, #, 110095
9	1211	101229156.0	#, 1619/16 F/F, #, ULDHANPUR NAVEEN SHAHDARA, #, 110032	#, 1619 / 16 F/F, #, ULDHANPUR NAVEEN SHAHDARA, #, 110032

3.5 Replacing all special characters with space:-

Special characters, such as punctuation marks or symbols, do not provide valuable information for address analysis and can create issues during data processing. By replacing all special characters with spaces, the program eliminates potential conflicts or errors that may arise from the presence of

these characters. This step also helps standardize the address data, making it more consistent and easier to work with.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS
0	1211	101363856.0	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 GALI 4 BALBIR NAGAR EXTN SHAHDARA 110032
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 GALI NO 17 BALBIR NAGAR EXTN 110032
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 GALI NO 4 EAST ROHTASH NAGAR SHAHDARA 110032
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA 110032
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA 110032
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A PKT R DILSHAD GARDEN SHAHDARA 110095
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA 110095
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA 110095
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 F F ULDHANPUR NAVEEN SHAHDARA 110032

3.6 Creating address part markers:-

Address data is typically composed of multiple components, including street names, city names, state names, postal codes, and more. By creating separate columns as markers for each address part, the program enables efficient data organization and retrieval. This segmentation provides clear boundaries between different address components, simplifying subsequent operations such as filtering or aggregating specific parts of the address data.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS	PLOT	FLOOR	BLOCK	STREET	POCKET
0	1211	101363856.0	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	None	None	None	None	None
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 GALI 4 BALBIR NAGAR EXTN SHAHDARA 110032	None	None	None	None	None
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 GALI NO 17 BALBIR NAGAR EXTN 110032	None	None	None	None	None
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 GALI NO 4 EAST ROHTASH NAGAR SHAHDARA 110032	None	None	None	None	None
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA 110032	None	None	None	None	None
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA 110032	None	None	None	None	None
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A PKT R DILSHAD GARDEN SHAHDARA 110095	None	None	None	None	None
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA 110095	None	None	None	None	None
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA 110095	None	None	None	None	None
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 F F ULDHANPUR NAVEEN SHAHDARA 110032	None	None	None	None	None

3.7 Moving pincodes to another column:-

Postal codes, or pincodes, are essential for location-based analysis and are often included in addresses. By extracting and moving pincodes to a separate column, the program enhances the accessibility and usability of this

information. Isolating pincodes allows the company to perform targeted analyses, such as identifying clusters of customers in specific geographic areas or optimizing logistics routes based on postal code zones.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS	PLOT	FLOOR	BLOCK	STREET	POCKET	PINCODE
0	1211	101363856.0	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL	None	None	None	None	None	110095
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 GALI 4 BALBIR NAGAR EXTN SHAHDARA	None	None	None	None	None	110032
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 GALI NO 17 BALBIR NAGAR EXTN	None	None	None	None	None	110032
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 GALI NO 4 EAST ROHTASH NAGAR SHAHDARA	None	None	None	None	None	110032
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA	None	None	None	None	None	110032
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA	None	None	None	None	None	110032
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A PKT R DILSHAD GARDEN SHAHDARA	None	None	None	None	None	110095
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA	None	None	None	None	None	110095
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA	None	None	None	None	None	110095
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 F F ULDHANPUR NAVEEN SHAHDARA	None	None	None	None	None	110032

3.8 Removing unnecessary info from the address:-

Raw address data may contain extraneous information that is not relevant to the analysis or objective of the project. By removing unnecessary information, the program improves the accuracy and focus of the address dataset. This step ensures that subsequent analyses are based on the most pertinent and meaningful address details, reducing noise and increasing the precision of results.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS	PLOT	FLOOR	BLOCK	STREET	POCKET	PINCODE
0	1211	101363856.0	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA RED CROSS HOSPITAL	None	None	None	None	None	110095
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 STREET 4 BALBIR NAGAR SHAHDARA	None	None	None	None	None	110032
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 STREET 17 BALBIR NAGAR	None	None	None	None	None	110032
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 STREET 4 EAST ROHTASH NAGAR SHAHDARA	None	None	None	None	None	110032
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA	None	None	None	None	None	110032
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA	None	None	None	None	None	110032
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A POCKET R DILSHAD GARDEN SHAHDARA	None	None	None	None	None	110095
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA	None	None	None	None	None	110095
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA	None	None	None	None	None	110095
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 FIRST FLOOR ULDHANPUR NAVEEN	None	None	None	None	None	110032

3.9 Replacing known words and abbreviation in address:-

Known words and abbreviations can introduce inconsistencies or variations in the address dataset. By replacing them with standardized terms, the program promotes consistency and enhances the usability of the address data. This standardization improves data quality and reduces the potential for errors or misinterpretations during analysis.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS	PLOT	FLOOR	BLOCK	STREET	POCKET	PINCODE
0	1211	101363856.0	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA RED CROSS HOSPITAL	None	None	None	None	None	110095
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 STREET 4 BALBIR NAGAR SHAHDARA	None	None	None	None	None	110032
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 STREET 17 BALBIR NAGAR	None	None	None	None	None	110032
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 STREET 4 EAST ROHTASH NAGAR SHAHDARA	None	None	None	None	None	110032
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA	None	None	None	None	None	110032
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA	None	None	None	None	None	110032
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A POCKET R DILSHAD GARDEN SHAHDARA	None	None	None	None	None	110095
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA	None	None	None	None	None	110095
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA	None	None	None	None	None	110095
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 FIRST FLOOR ULDHANPUR NAVEEN	None	None	None	None	None	110032

3.10 Segregating address markers:-

The segregation of address markers into different columns builds upon the earlier creation of address part markers. By storing data in separate columns based on these markers, the program further enhances data organization and accessibility. This enables more focused analyses of specific address components and facilitates quick data retrieval for various use cases, such as customer segmentation by city or state.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS	PLOT	FLOOR	BLOCK	STREET	POCKET	PINCODE
0	1211	101363856.0	EXTN 478 KALANDER COLONY CLONI KOLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY CLONI COLONY DILSHAD GARDEN SHAHDARA RED CROSS HOSPITAL	478	None	None	None	None	110095
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 BALBIR NAGAR SHAHDARA	1438	None	None	4	None	110032
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 BALBIR NAGAR	1 5484	None	None	17	None	110032
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 EAST ROHTASH NAGAR SHAHDARA	1 6221	None	None	4	None	110032
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA	10	None	None	None	None	110032
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA	1 6421	None	None	None	None	110032
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A POCKET R DILSHAD GARDEN SHAHDARA	42	None	None	None	R	110095
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA	23	None	None	None	None	110095
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA	125	None	None	None	None	110095
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 ULDHANPUR NAVEEN SHAHDARA	1619 16	FIRST	None	None	None	110032

3.11 Creating corpus that are NonNumeric and bigger/equal to 3 :-

The creation of a corpus specifically containing non-numeric words with a length of three or more characters is a valuable step in address analysis. This corpus serves as a reference for identifying significant words within the address dataset. By focusing on non-numeric words of a certain length, the program excludes common short words and numeric values that might not provide meaningful insights or contribute to the analysis objectives.

	WORD	NO. OF MATCHES	BEST_MATCHES
0	OPPOSITESENIORSECOUNDARYSCHOOL	0	
1	AMBEYENCLAVECHAUHANPATTISABHA	0	
2	SCHOOLJAVASCRIPTDATESELECTED	0	
3	MUKANDVIHARCHOWKCYCLEFACTORY	0	
4	BHAGATSINGHMOHLLANEWUSMANPUR	0	

3.12 Finding similar words in corpus through fuzzy and adding to the word only dataframe:-

Fuzzy matching techniques are employed to identify similar words within the created corpus. This process expands the coverage and accuracy of the word data by capturing variations or misspellings that may exist within the dataset. By adding these similar words to a dedicated dataframe, the program enables more comprehensive analysis and enhances the completeness of the word data for further exploration or information retrieval tasks.

	WORD	NO. OF MATCHES	BEST_MATCHES
0	OPPOSITESENIORSECOUNDARYSCHOOL	0	
1	AMBEYENCLAVECHAUHANPATTISABHA	0	
2	SCHOOLJAVASCRIPTDATESELECTED	0	
3	MUKANDVIHARCHOWKCYCLEFACTORY	0	
4	BHAGATSINGHMOHLLANEWUSMANPUR	0	
5	RAJDHANIPUBLICSCHOOLWALIGALI	0	
6	AMBEDKARBASTIGHONDA	0	
7	MASJIDWALIGALITIRPALFECTORY	0	
8	VIDHYALAYASEELAMPURSHAHDARA	0	
9	IMRATKIRANASTIRESHIVMANDIR	0	
10	GAUTAMPURISHAHDARA	0	
11	BHAJANPURASHAHDARA	2	BHAJANPURISHAHDARANULLNULL BHAJANPURSHAHDARANULLNULL
12	BHAJANPURISHAHDARA	2	BHAJANPURASHAHDARANULLNULL BHAJANPURSHAHDARANULLNULL
13	BEHINDALOKPUNJPUBLICSCHOOL	0	
14	ZAFARABADSHAHDARA	4	JAFFARABASHAHDARANULLNULL ZAFRABADSHAHDARANULLNULL ZAFRABADSHADHARANULLNULL ZAFRABASHAHDARANULLNULL

3.13 Removing NULL from end of words:-

In the preprocessing stage, some words may end up with the string "NULL" appended to them. This string is likely an artifact of the data processing and does not carry any useful information. By removing "NULL" from the end of words, the program ensures data accuracy and prevents potential discrepancies or misinterpretations caused by this artifact.

	WORD	NO. OF MATCHES	BEST_MATCHES
0	OPPOSITESENIORSECOUNDARYSCHOOL	0	nan
1	AMBEYENCLAVECHAUHANPATTISABHA	0	nan
2	SCHOOLJAVASCRIPTDATESELECTED	0	nan
3	MUKANDVIHARCHOWKCYCLEFACTORY	0	nan
4	BHAGATSINGHMOHLLANEWUSMANPUR	0	nan
5	RAJDHANIPUBLICSCHOOLWALIGALI	0	nan
6	AMBEDKARBASTIGHONDA	0	nan
7	MASJIDWALIGALITIRPALFECTORY	0	nan
8	VIDHYALAYASEELAMPURSHAHDARA	0	nan
9	IMRATKIRANASTIRESHIVMANDIR	0	nan
10	GAUTAMPURISHAHDARA	0	nan
11	BHAJANPURASHAHDARA	2	BHAJANPURISHAHDARA BHAJANPURSHAHDARA
12	BHAJANPURISHAHDARA	2	BHAJANPURASHAHDARA BHAJANPURSHAHDARA
13	BEHINDALOKPUNJPUBLICSCHOOL	0	nan
14	ZAFARABADSHAHDARA	4	JAFFARABASHAHDARA ZAFRABADSHAHDARA ZAFRABADSHADHARA ZAFRABASHAHDARA

3.14 Exporting to CSVs:-

The program exports the processed word data to CSV files to ensure data persistence and easy sharing with other stakeholders. By storing the processed data in a structured and standardized format, the program enables seamless integration with other systems or analysis tools. This step facilitates collaboration and the utilization of the cleaned address data across various applications and projects.

	WORD	NO. OF MATCHES	BEST_MATCHES
13	BHAJANPURASHAHDARA	2	BHAJANPURISHAHDARA BHAJANPURSHAHDARA
14	BHAJANPURISHAHDARA	2	BHAJANPURASHAHDARA BHAJANPURSHAHDARA
15	BEHINDALOKPUNJPUBLICSCHOOL	0	
16	ZAFARABADSHAHDARA	4	JAFFARABASHAHDARA ZAFRABADSHAHDARA ZAFRABADSHADHARA ZAFRABASHAHDARA
17	INTERNATIONALPUBLICSCHOOL	0	
18	KAITHWARAUSMANPUR	0	
19	BRAHAMPURSHAHDARA	2	BRAHMPURISHAHDARA BRAHMPURSHAHDARA
20	MASJIDSUBHASHVIHARNORTHGH	0	
21	JAFFARABASHAHDARA	2	ZAFARABADSHAHDARA ZAFRABASHAHDARA
22	BHAJANPURSHAHDARA	2	BHAJANPURASHAHDARA BHAJANPURISHAHDARA
23	SOMBAZARPRAGATIVIHARGAMRI	0	
24	SEELAMPURSHAHDARA	2	SILAMPURSHAHDARA SILAMPUSHAHDARA
25	ADRDNEARCHANDMASZIDCHUHAN	0	
26	BHAJANPURABHAJANPURADELHI	0	
27	BRAHMPURISHAHDARA	2	BRAHAMPURSHAHDARA BRAHMPURSHAHDARA
28	MAINROADPRATAPNAGARSABOLI	0	
29	BRAHMPURSHAHDARA	2	BRAHAMPURSHAHDARA BRAHMPURISHAHDARA
30	MADINAMASZIDVIJAYPARKMAU	0	
31	MANDUBAHAJANPURA	0	

3.15 Keeping necessary/useful word only:-

In the final step, the program filters the word-only dataframe to retain only the necessary and useful words for the specific objectives of the project. By focusing on essential information, the program improves the efficiency of subsequent analyses and reduces the complexity associated with working with a large number of words. This streamlined dataset provides valuable insights and facilitates accurate and targeted information retrieval.

	SDO_CD	CA_NO	ADDRESS_ORG	ADDRESS	PLOT	FLOOR	BLOCK	STREET	POCKET	PINCODE
0	1211	101363856.0	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA NEAR RED CROSS HOSPITAL 110095	478 KALANDER COLONY DILSHAD GARDEN SHAHDARA RED CROSS HOSPITAL	478	None	None	None	None	110095
1	1211	101296041.0	1438-A-45 GALI-4 BALBIR NAGAR EXTN SHAHDARA 110032	1438 A 45 BALBIR NAGAR SHAHDARA	1438	None	None	4	None	110032
2	1211	101358487.0	1/5484 GALI NO-17 BALBIR NAGAR EXTN 110032	1 5484 BALBIR NAGAR	1 5484	None	None	17	None	110032
3	1211	101358755.0	1/6221 GALI NO-4 EAST ROHTASH NAGAR SHAHDARA 110032	1 6221 EAST ROHTASH NAGAR SHAHDARA	1 6221	None	None	4	None	110032
4	1211	101220086.0	T-10 NAVEEN SHAHDARA 110032	T 10 NAVEEN SHAHDARA	10	None	None	None	None	110032
5	1211	101209256.0	1/6421-A-1 EAST ROHTASH NAGAR SHAHDARA 110032	1 6421 A 1 EAST ROHTASH NAGAR SHAHDARA	1 6421	None	None	None	None	110032
6	1211	101209263.0	42-A PKT R DILSHAD GARDEN SHAHDARA 110095	42 A POCKET R DILSHAD GARDEN SHAHDARA	42	None	None	None	R	110095
7	1211	101325114.0	E-23-F-1 DILSHAD COLONY SHAHDARA 110095	E 23 F 1 DILSHAD COLONY SHAHDARA	23	None	None	None	None	110095
8	1211	101157651.0	E-125-B DILSHAD GARDEN SHAHDARA 110095	E 125 B DILSHAD GARDEN SHAHDARA	125	None	None	None	None	110095
9	1211	101229156.0	1619/16 F/F ULDHANPUR NAVEEN SHAHDARA 110032	1619 16 ULDHANPUR NAVEEN SHAHDARA	1619 16	FIRST	None	None	None	110032

Conclusion

In conclusion, the address cleaning and sanitization program successfully addresses the challenges faced by the company with their raw address dataset.By implementing a comprehensive series of preprocessing steps, the program cleans and transforms the address data into a usable dataframe. This allows for efficient data analysis, accurate information retrieval, and improved decision-making. The program's methodologies, such as removing garbage values, handling null values, and replacing known words, enhance the data quality and consistency, reducing errors and increasing the reliability of analysis outcomes. The program provides valuable insights and information from the address dataset, enabling the company to optimize various processes, enhance customer analytics, and make informed business decisions based on clean and accurate address data.