DePaul Dataset Cleaning Documentation

Tool Used: Microsoft Excel

The following were the steps taken to clean the dataset:

1. Removal of Entirely Null Columns

Identified Issue:

Multiple columns contained 100% null values, providing no analytical or operational value.

Cleaning Step:

Deleted the following columns:

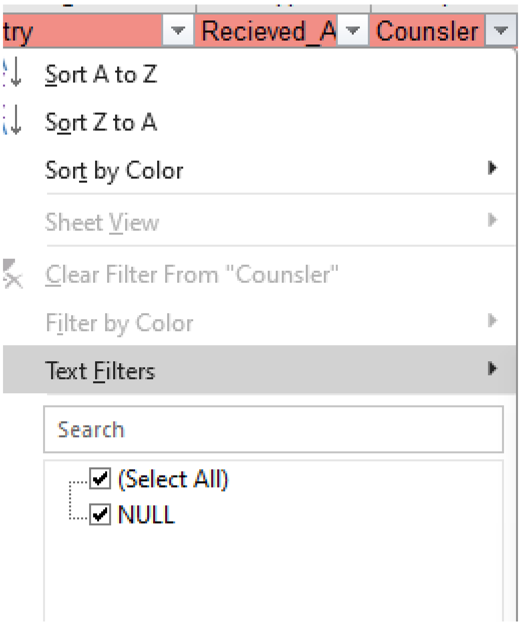
- counselor
- status
- major
- degree_type
- citizenship
- major_1st_choice
- street_3
- date_of_birth
- Most_Recent_Released_Decision
- RIT_Email_Created
- Housing_Contract
- Is_Global_Grad
- Is_Admitted
- SEVIS_ID
- Official_University_email_address

- Application_Agency_Code
- Outstanding_Checklist_Items
- Prior_I-20_Outreach_Detail
- Prior_Non_I-20_Outreach_Detail
- Most_Recent_Contact
- Most_Recent_User_and_Date
- Recent_GT_Form_Initiative
- Caller_Name_2
- Date_of_Contact_2
- Outcome_2
- Escalation_Required
- SLU_Start_Comment
- SLU_Start_City
- I_901_Status
- Intake-2
- Status-2
- City_and_Branch
- Header
- Template
- Region-2
- Final_Result
- Category
- Phone_Number_2

Rationale:

- Columns with no non-null entries add noise, increase file size, and complicate analysis.
- Retaining them offers zero utility for reporting, modeling, or operational use.

Example:



Before	**After**
Column counselor: all cells with nulls	Column removed entirely

2. Deduplication Based on reference_id

Identified Issue:

2,543 duplicate records identified using reference_id.

Cleaning Step:

- Removed duplicate rows where reference_id was identical.
- Kept the first occurrence (default Excel's behavior in "Remove Duplicates").

Rationale:

- reference_id is a unique identifier; duplicates likely stem from system sync errors or repeated imports.
- Keeping duplicates would inflate counts and distort analytics (e.g., enrollment stats).

Example:

Before	**After**
Row 10: ref_id = " 209623088", name = " Rohan Rameshbhai" Row 10: ref_id = " 209623088", name = " Rohan Rameshbhai"	Only one instance of ref_id "209623088" retained

3. Data Type Conversions

Identified Issue:

Incorrect or inconsistent data types impair sorting, filtering, and integration.

Cleaning Steps & Rationale:

Column	**Original Type**	**New Type**	**Rationale**	**Example**
reference_id	Integer (e.g. 45405320)	**Text**	IDs are identifiers, not numeric values. Leading zeros (if any) preserved; avoids scientific notation.	45405320 → "45405320"
given_name, last_name	General	**Text**	Ensures names aren’t interpreted as formulas or numbers.	"Deeksha Reddy" stays intact
phone_number	General	**Text**	Prevents truncation; ensures readability.	9.2E+11 → 920000000000
Street_2	General	**Text**	Prevents auto-conversion of street names	"Kotauratla Mandalam" remains as-is

Column	**Original Type**	**New Type**	**Rationale**	**Example**
Postal	General	**Text**	Postal codes may start with 0 (e.g., 02115); numeric format drops leading zeros.	02118-3096 → "02118-3096"
Admit_Date	Text or General	**Date**	Enables date-based filtering, sorting, and time-series analysis.	"02/20/2024" → Excel date serial

4. Column Merging: Name Fields

Identified Issue:

First and last names stored in separate columns (given_name, last_name), but downstream use may require full name.

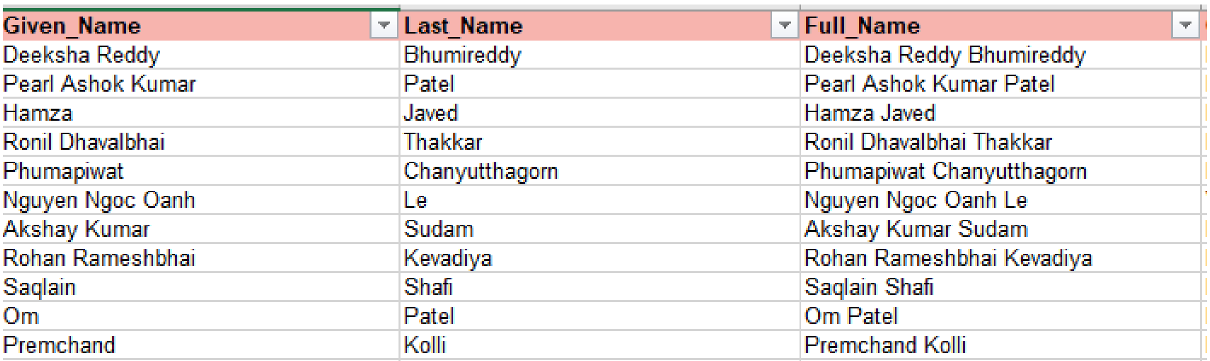
Cleaning Step:

- Created a new column full_name using formula: =TRIM(given_name & " " & last_name)

  Rationale:

- Simplifies display, mailing lists, and reporting.
- TRIM() removes extra spaces from concatenation.

Example:



Given_Name	Last_Name	Full_Name
Deeksha Reddy	Bhumireddy	Deeksha Reddy Bhumireddy
Pearl Ashok Kumar	Patel	Pearl Ashok Kumar Patel
Hamza	Javed	Hamza Javed
Ronil Dhavalbhai	Thakkar	Ronil Dhavalbhai Thakkar
Phumapiwat	Chanyutthagorn	Phumapiwat Chanyutthagorn
Nguyen Ngoc Oanh	Le	Nguyen Ngoc Oanh Le
Akshay Kumar	Sudam	Akshay Kumar Sudam
Rohan Rameshbhai	Kevadiya	Rohan Rameshbhai Kevadiya
Saqlain	Shafi	Saqlain Shafi
Om	Patel	Om Patel
Premchand	Kolli	Premchand Kolli

5. Removal of Redundant/System-Generated Columns

Identified Issue:

Columns are auto-generated by system and provide no user-facing value.

Cleaning Step:

Dropped the following:

- Received_At
- Created_At
- Modified_At
- Reference_ID-2
- Recieved_At_2
- University-2, University-3
- Degree_Type_2
- Created_At-2, Created_At-3
- Modified_At-2, Modified_At-3
- ID
- Attempts
- Start_Date
- Date_of_Contact

Rationale:

- These appear to be audit/log fields from CRM or database replication.
- Multiple versions (University-2, -3) suggest failed merges or sync artifacts.
- Not useful for business analysis or student outreach.

6. City Name Standardization

Identified Issue:

Inconsistent, misspelled, or overly verbose city names reduce geospatial accuracy and grouping.

Cleaning Steps:

Used manual find-and-replace and standardized naming conventions:

Before	**After**	**Reason**
accea	Accra	Typo
acra	Accra	Typo
Ado-Ekiti	Ado Ekiti	Remove hyphen for consistency
Agege Lagos	Agege	Remove redundant state/country
Ago Palace way	Ago Palace	Truncate for uniformity
Ahmedabad variants	Ahmedabad	Correct spelling
Accra Ghana	Accra	Remove country (assumed known)
Addis Abeba	Addis Ababa	Standard English spelling
Ajah Lagos	Ajah	Remove state
ALEXANDRIA	Alexandria	Fix capitalization
ambala city ambala	Ambala City	Deduplicate and standardize
Ananthapur	Anantapur	Correct spelling
Bengaluru, Karnataka → bengaluru	Bengaluru	Normalize casing and remove state
Benin city	Benin	Remove “city”
Chang^ (with special char)	Chang	Remove trailing symbol

Before	After	Reason
Chennai, Tamil Nadu	Chennai	Keep only city
Chickmagluru	Chikkamagaluru	Correct spelling
Chitoor	Chittoor	Spelling fix
Coimbatore Rural	Coimbatore	Simplify
Dublin:1	Dublin	Remove artifact
Fct Abuja	Fct	Standard abbreviation
Gautam Budh Nagar → Gautam Buddha Nagar	Correct spelling	
Hanumakonda, Warangal → Hanumakonda	Remove district	
Hsinchu City → Hsinchu	Remove “City”	
Ifako-Ijaiye → Ifako	Shorten for consistency	
Ikorodu/Lagos → Ikorodu	Remove slash + state	
Islmabad → Islamabad	Spelling fix	

Rationale:

- Enables accurate grouping by city (e.g., for regional reports).
- Reduces false uniqueness (e.g., “Chennai” vs “Chennai, Tamil Nadu” counted as two cities).
- Improves data quality for mapping or logistics.

Recommendations for Future

1. Store phone numbers as Text to preserve formatting.

2. Validate city names against a master list (e.g., ISO cities) during intake.
3. Avoid importing system audit columns unless needed for debugging.
4. Use data validation rules in Excel forms to prevent typos at entry.