Dataset Overview & Documentation

DePaul University Students' Applications Dataset

1.1 Dataset Dimensions

Metric	Value
Total Records	7,543 rows.
Total Attributes	80 columns.
Unique Student Leads	Approximately 4,540 (based on Reference_ID).
Data Structure	Long-format event log (multiple rows per student representing different interactions or status updates).

2. Variable Classifications

2.1 Categorical Variables (60)

• Given_Name • College • Degree_Type • Counsler • Citizenship • Intake • Major_1st_Choice • Street_1 • Street_3 • Region • Comments • Most_Recent_Released_Decision • Housing_Contract • Is_Global_Grad • Official_University_email_address • University-2 • Prior_I-20_Outreach_Detail • Most_Recent_Contact • Recent_GT_Form_Initiative • Outcome_1 • Caller_Name_2 • Remark_2 • Escalation_Required • SLU_Start_City • Slate_Form_Filled • Category • University-3 • Status-2	• Last_Name • Major • Country • University • Status • College_1st_Choice • Phone Number • Street_2 • City • Email_ID • Phone_Number_2 • RIT_Email_Created • Application_Source • Is_Admitted • Application_Agency_Code • Outstanding_Checklist_Items • Prior_Non_I-20_Outreach_Detail • Most_Recent_User_and_Date • Caller_Name • Remark • Outcome_2 • Final_Result • SLU_Start_Comment • I_901_Status • Name • Intake-2 • Degree_Type-2 • City_and_Branch
• Type • Template	• Header • Region-2

2.2 Numerical and Date/Time Variables (21)

• Reference_ID • Received_At (epoch timestamp) • Campaign_Id • Modified_At (epoch timestamp) • Recieved_At-2 (epoch timestamp) • Attempts • Modified_At-2 (epoch timestamp) • Modified_At-3 (epoch timestamp) • Date_of_Birth • Date_of_Contact • Start_Date	• Postal • SEVIS_ID • Created_At (epoch timestamp) • Reference_ID-2 • ID • Created_At-2 (epoch timestamp) • Created_At-3 (epoch timestamp) • Date • Admit_Date • Date_of_Contact_2

3. General Dataset Themes

Variables are grouped into six high-level themes. The table below summarizes each theme, its purpose, and representative columns.

Theme	Description	Representative Variables	Count
Identity & Demographics	Columns related to personal identification and demographic information.	Reference_ID, Given_Name, Last_Name, Date_of_Birth, Citizenship, Country, Name	7
Academic Background	Columns capturing educational history, qualifications, and program interests.	College, Major, Degree_Type, University, College_1st_Choice, Major_1st_Choice, Intake, University-2, University-3, Degree_Type-2, Intake-2	11
Contact & Location Information	Columns containing addresses, phone numbers, and email addresses.	Phone_Number, Phone_Number_2, Street_1, Street_2, Street_3, City, Region, Postal, Email_ID, Official_University_email_address, City_and_Branch, Region-2, SLU_Start_City	13
Application & Admission Status	Columns tracking application progress, decisions, and admission details.	Status, Status-2, Admit_Date, Is_Admitted, Most_Recent_Released_Decision, Application_Source, Is_Global_Grad, SEVIS_ID, Housing_Contract, I_901_Status, Slate_Form_Filled, Outstanding_Checklist_Items, Final_Result, Escalation_Required	14
Theme	**Description**	**Representative Variables**	**Count**

Communication & Outreach History	Columns logging interactions, follow-ups, and communication records.	Counsler, Comments, Prior_I-20_Outreach_Detail, Prior_Non_I-20_Outreach_Detail, Most_Recent_Contact, Most_Recent_User_and_Date, Recent_GT_Form_Initiative, Date_of_Contact, Caller_Name, Outcome_1, Remark, Caller_Name_2, Date_of_Contact_2, Outcome_2, Remark_2, SLU_Start_Comment, Category, Type, Header, Template, Start_Date	21
System & Metadata	Columns used for internal tracking, system administration, and technical purposes.	Recieved_At, Created_At, Modified_At, Application_Agency_Code, Reference_ID-2, Recieved_At-2, Created_At-2, Modified_At-2, ID, Attempts, Created_At-3, Modified_At-3, Campaign_Id, RIT_Email_Created *Many of these columns are duplicates or system placeholders (like Created_At-2).*	14
Total			**80**

4. Initial Observations and Notable Characteristics

4.1 Data Layout Patterns

The dataset is structured in a Long Format. It was observed that the Reference_ID for individual students reappears across multiple rows (7,543 rows for approximately 4,540 unique students). This indicates the data is an event-based log where each row represents a specific touchpoint or status update rather than a single static student profile.

4.2 Data Quality Red Flags

- There are significant completeness issues.
- 47.5% of the columns (38 out of 80) are 100% null and contain no usable data. Additionally, system-generated timestamps (e.g., Created_At) are stored as Unix integers (e.g., 175749...), which will require mathematical conversion to standard date formats before time-series analysis (like "Leads per Month") can be performed.

4.3 Visual Patterns and Anomalies

- The data shows a clear concentration of leads in the South Asian region (specifically India and Pakistan).
- An anomaly was noted in the outreach volume: while the vast majority of students have 1–3 'Attempts,' a small group of outliers has over 10 attempts. These outliers warrant further investigation to determine if they represent high-priority leads or automated system errors.

4.4 Data Content Observations

The Intake column is a compound field, meaning it merges the Academic Program, Degree Level, and Start Term (e.g., 'Human Resources-MS: 2024 Fall'). To allow for separate reporting on popular programs versus enrollment semesters, a data-splitting process (Text-to-Columns) will be required during the cleaning phase.

Data Quality Report

This notebook documents an assessment of data quality issues in the dataset, including missing values, duplicate records, and unusual or problematic entries that may affect analysis.

In [13]:
```
#importing libraries
import pandas as pd
import numpy as np
```

In [15]:
```
#Loading dataset
df = pd.read_csv(r"C:\Users\user\Desktop\DePaul_Data.csv")
df.head()
```

Out[15]:

	Reference_ID	Given_Name	Last_Name	College	Major	Degree_Type	Country	R
0	45405320	Deeksha Reddy	Bhumireddy	INDIA - Osmania University - Bachelor's Degree	NaN	NaN	India	1757
1	858032003	Pearl Ashok Kumar	Patel	INDIA - Manipal University Jaipur Bachelor's...	NaN	NaN	India	1757
2	902518555	Hamza	Javed	PAKISTAN - Shaheed Zulfikar Ali Bhutto Institu...	NaN	NaN	Pakistan	1757
3	902518555	Hamza	Javed	PAKISTAN - Shaheed Zulfikar Ali Bhutto Institu...	NaN	NaN	Pakistan	1757
4	218755608	Ronil Dhavalbhai	Thakkar	INDIA - Swarrnim Institute of Technology - Swa...	NaN	NaN	India	1757

In [17]:
```
#Checking shape
df.shape
```

Out[17]: (7543, 80)

5 rows × 80 columns

3.1 Missing Value Analysis(count and percentage)

This section evaluates the extent of missing data across all variables in the dataset.

In [42]:
```
#Creating DataFrame to show the missing counts and percentages for each
variable. missing_summary = pd.DataFrame({
    "Missing Count": df.isnull().sum(),
    "Missing Percentage": (df.isnull().mean() * 100).round(2)
})
#Showing the first 50 records
missing_summary.head(50)
```

Out[42]:

	Missing Count	Missing Percentage
Reference_ID	0	0.00
Given_Name	2	0.03
Last_Name	0	0.00
College	71	0.94
Major	7543	100.00
Degree_Type	7543	100.00
Country	11	0.15
Recieved_At	0	0.00
Counsler	7543	100.00
University	0	0.00
Citizenship	7543	100.00
Status	7543	100.00
Intake	3	0.04
College_1st_Choice	7	0.09
Major_1st_Choice	7543	100.00
Phone_Number	587	7.78
Street_1	15	0.20
Street_2	4638	61.49
Street_3	7543	100.00
City	33	0.44
Region	914	12.12
Postal	226	3.00
Email_ID	1	0.01
Date_of_Birth	7543	100.00
Comments	1625	21.54
Phone_Number_2	7543	100.00
Admit_Date	1754	23.25
Most_Recent_Released_Decision	7543	100.00
RIT_Email_Created	7543	100.00

	Missing Count	Missing Percentage
Housing_Contract	7543	100.00
Application_Source	0	0.00
Is_Global_Grad	7543	100.00
Is_Admitted	7543	100.00
SEVIS_ID	7543	100.00
Official_University_email_address	7543	100.00
Application_Agency_Code	7543	100.00
Created_At	0	0.00
Modified_At	0	0.00
Reference_ID-2	3194	42.34
Recieved_At-2	3194	42.34
University-2	3194	42.34
Outstanding_Checklist_Items	7543	100.00
Prior_I-20_Outreach_Detail	7543	100.00
Prior_Non_I-20_Outreach_Detail	7543	100.00
Most_Recent_Contact	7543	100.00
Most_Recent_User_and_Date	7543	100.00
Recent_GT_Form_Initiative	7543	100.00
Date_of_Contact	3194	42.34
Caller_Name	3194	42.34
Outcome_1	3194	42.34

Observation: This dataset is a mix of several variables that have high percentages of missing data, with some columns showing even 100% missing values thus offering little to no value in analysis and several variables that have little percentages of missing data which is good for analysis.

3.2 Duplicate Record Check

This section identifies whether duplicate rows exist within the dataset.

In [48]:
```
#Checking duplicate count.
duplicate_count = df.duplicated().sum()
duplicate_count
```

Out[48]: 0

Observation: The assessment showed that there are no exact duplicate rows within the dataset.

3.3 Outliers and Unusual Entries

This section examines unusual data characteristics such as non-standard formats

In [56]:
```
#Checking data types
df.dtypes
```

Out[56]:
```
Reference_ID       int64
Given_Name        object
Last_Name         object
College           object
Major            float64
...
Template         float64
Start_Date        object
Region-2         float64
Created_At-3     float64
Modified_At-3    float64
Length: 80, dtype: object
```

Observation: The assessment showed that there are some inconsistencies with variable datatypes for example, date is marked as float instead of datetime and for phone number it is marked as float which has led to observation of unusual entries in the data for it. This affects the integrity of the data.

3.4 Summary of Data Quality Issues

The following data quality issues were identified:

- High percentages of missing data in multiple variables.
- Inconsistent variable datatypes with the data they hold.

In [ ]:
```
```