ApplicantData Performance Analysis: Week 1

Executive Summary :-

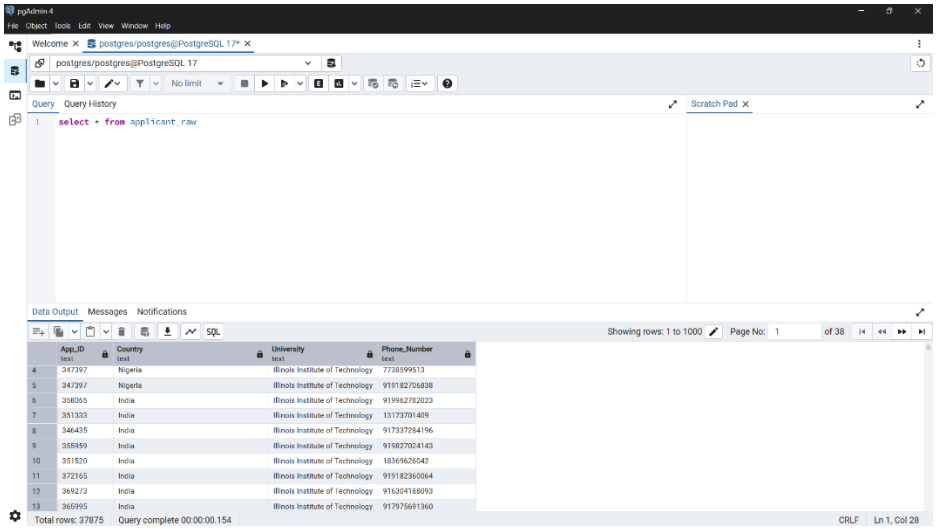
The dataset contains 37,882 rows and 4 columns: App_ID, Country, University, and Phone_Number. All columns are currently of the object data type, which may require conversion to other data types depending on your analysis needs.

Here are some specific observations to consider for data cleaning:

- The App_ID column has one missing value. Additionally, this column contains non-numeric characters, such as -, which may need to be addressed before any numerical analysis.
- The Phone_Number column also consists of numerical values, but it is currently stored as an object data type. Depending on the analysis, this column may need to be converted to a numerical type.
- The Country and University columns appear to be suitable for categorical analysis.

Database Setup and Data Retrieval for EDA :-

- For this project, I set up a PostgreSQL database to manage the applicant data. Using the `psycopg2` library in Python, I established a connection and created a dedicated table called `applicant_data`. The raw data was then efficiently loaded into this table using a bulk `COPY` command.
- After successfully ingesting the data, I used SQL queries via `psycopg2` to perform initial exploratory data analysis (EDA). This process allowed me to quickly profile the dataset, confirming the total number of records, counting unique values in key columns, and identifying the most frequent entries for further analysis.



Data Profile and Quality Assessment :-

A data quality assessment was performed on the applicant dataset to prepare it for analysis. The raw data contained four columns: App_ID, Country, University, and Phone_Number.

The **App_ID** column contained inconsistencies, including a small number of missing values and a mix of numeric and non-numeric characters. The **Country** and **University** columns were mostly clean, though the University column showed a single value (Illinois Institute of Technology) across all records. The **Phone_Number** column varied in format and length, requiring standardization for consistent analysis.

Overall, the assessment identified key data cleaning priorities to ensure the dataset is reliable for further reporting and analysis.

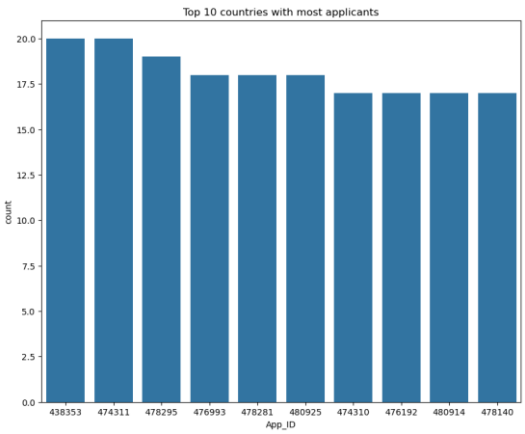
Data Schema:

```
RangeIndex: 37875 entries, 0 to 37874
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   App_ID        37875 non-null  object
 1   Country       37875 non-null  object
 2   University    37875 non-null  object
 3   Phone_Number  37875 non-null  object
```

Key Analytical Findings :-

1. Distribution of Applicants by Country

An analysis of the Country column shows a high concentration of applicants from a few key countries. India accounts for a significant portion of the applications, followed by Nigeria. This suggests a strong focus on these regions for outreach or a high level of interest from these countries.

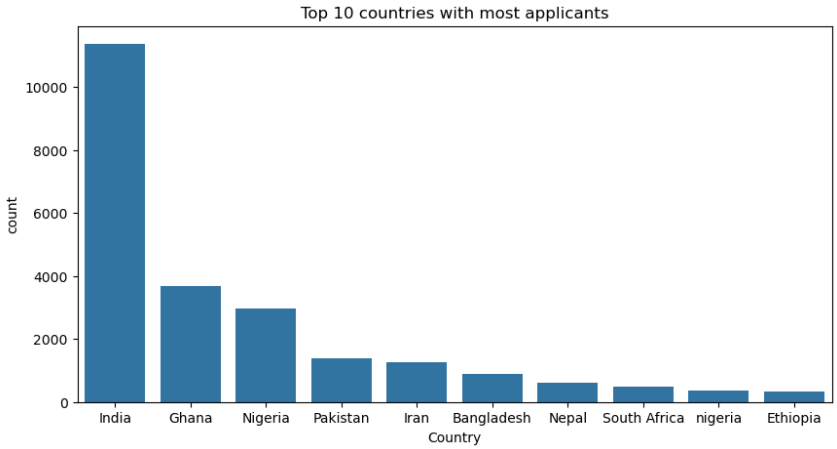


2. University Application Trends

The dataset shows a distinct trend in university applications, with a vast majority of applicants applying to a single institution. This indicates that the dataset is highly skewed towards a specific university, which is important to consider for any further analysis or interpretation of the data.

3. Top Countries by Applicant Count

A closer look at the data reveals that India has the highest number of applicants, far surpassing all other countries. This is followed by Nigeria, Ghana, Kenya, and Vietnam. The top 10 countries account for the majority of the total applications.



4. Unique Identifiers

The App_ID column contains a large number of unique values, confirming its role as a primary identifier for applicants. However, the presence of the single missing value and inconsistent formatting will need to be resolved to ensure data integrity.

Next Steps:-

Based on the findings from this initial analysis, the following actions are recommended to prepare the data for more detailed analysis:

- **Data Cleaning:** Address the missing value in App_ID and standardize the formatting of the App_ID and Phone_Number columns.
- **Feature Engineering:** Extract country codes or other relevant information from the Phone_Number to enrich the dataset for geographical analysis.
- **Deep Dive Analysis:** Conduct a more in-depth analysis of application trends over time (if date data is available) and explore the relationship between the top countries and university applications