EDA REPORT

Team 1

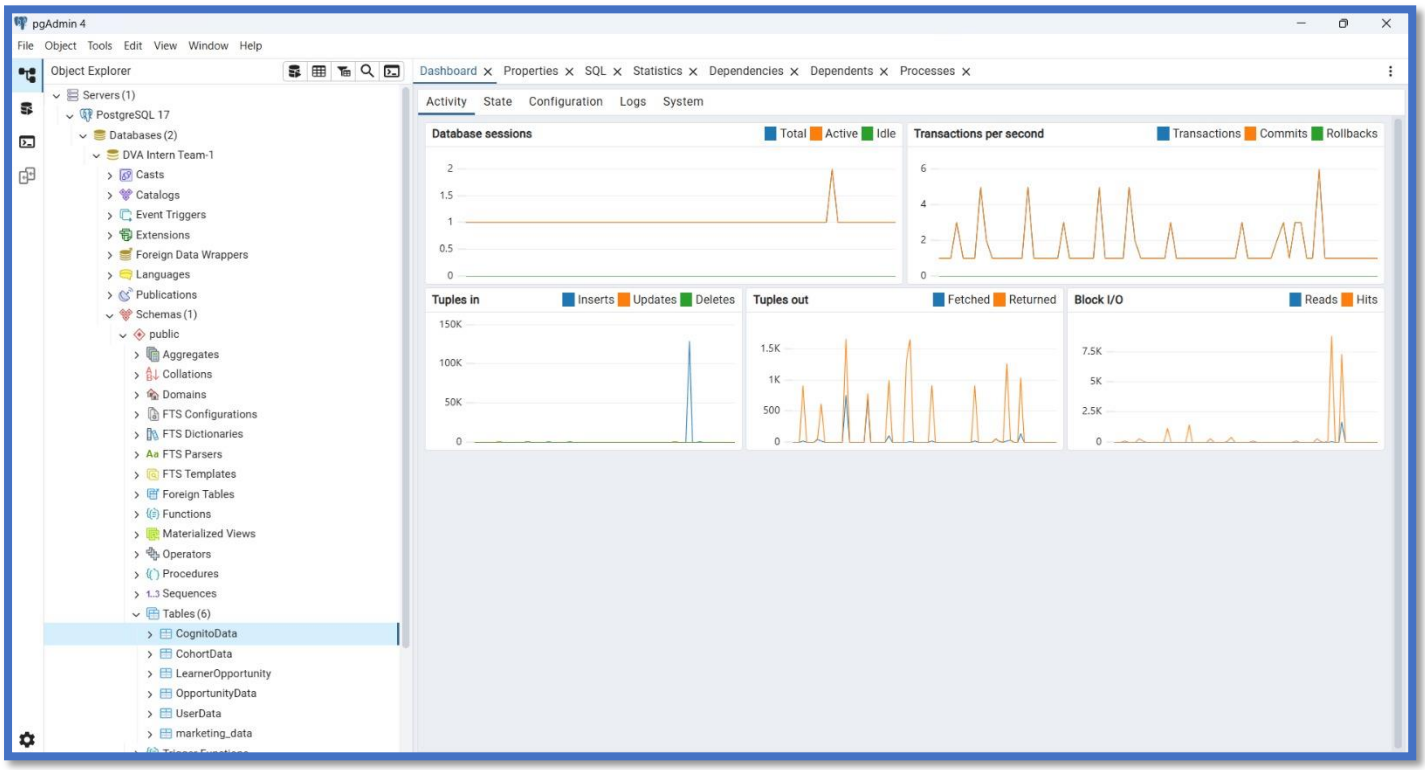
Team Member Details:

NAME	EMAIL ID
TARINEE	tarinee@s.amity.edu
SWARA	birjeswara2005@gmail.com
NIKHIL AWASTHI	nikhilawasthi944@gmail.com
DARSHAN DHOLA	darshandhola2002@gmail.com
MH. ASIF HOSSAIN	asifhossain8612@gmail.com
PRASHANT KASHYAP	prashantkashyap5579@gmail.com
VARUN JOSE MADANU	varunjosemadanu@gmail.com
MUSTAFA OUN	wnvyvcgc@gmail.com
ANKIT RAJ	226301029@gkv.ac.in
MOLEBOGENG SEROBATSE	molebogeng.serobatse@gmail.com
OMAR IMAM	omaruw3@gmail.com

INDEX

Sr. No.	Data Set	Page No.
1.	PostgreSQL Setup	2
2.	User Data	3-5
3.	Opportunity Data	6-9
4.	Marketing Data	10-13
5.	Learning Opportunity Data	14-17
6.	Cohort Data	18-20
7.	Cognito Data	21-25

PostgreSQL Loaded with Datasets



User Data

This dataset provides demographic and educational background data of learners from different countries. The primary goal of this exploratory data analysis (EDA) is to understand the data structure, detect missing or inconsistent entries, and generate insights that aid in further transformation, modeling, or decision-making.

Key Columns in the Dataset:

- Learner ID
- Country
- Degree
- Institution
- Major

This exploratory data analysis (EDA) aims to:

- Understand the structure and quality of the dataset
- Detect missing or inconsistent data
- Examine the diversity of learners by country and academic background
- Provide insights through summaries and visual representations
- Recommend data cleaning steps for further analysis or modeling

The insights drawn will be instrumental in cleaning and transforming the data for modeling and in making data- driven decisions.

Dataset Overview:

The table below tells the number of categorical and numerical fields in the dataset.

Column Category	Number of fields
Numerical Fields	0
Categorical Fields	5

The table below tells the datatype of each field/ column, no. of blanks and recommendation to handle.

Column No.	Column Name	Column Category	No. of NULLs	How to Handle? (Recommendations)
1.	Learner ID	Categorical	0	-
2.	Country	Categorical	2275	Replace with 'Undeclared'
3.	Degree	Categorical	52693	Replace with 'Undeclared'
4.	Institution	Categorical	52693	-Replace with 'Undeclared'

5.	Major	Numerical	52694	- Replace with ‘Undeclared’

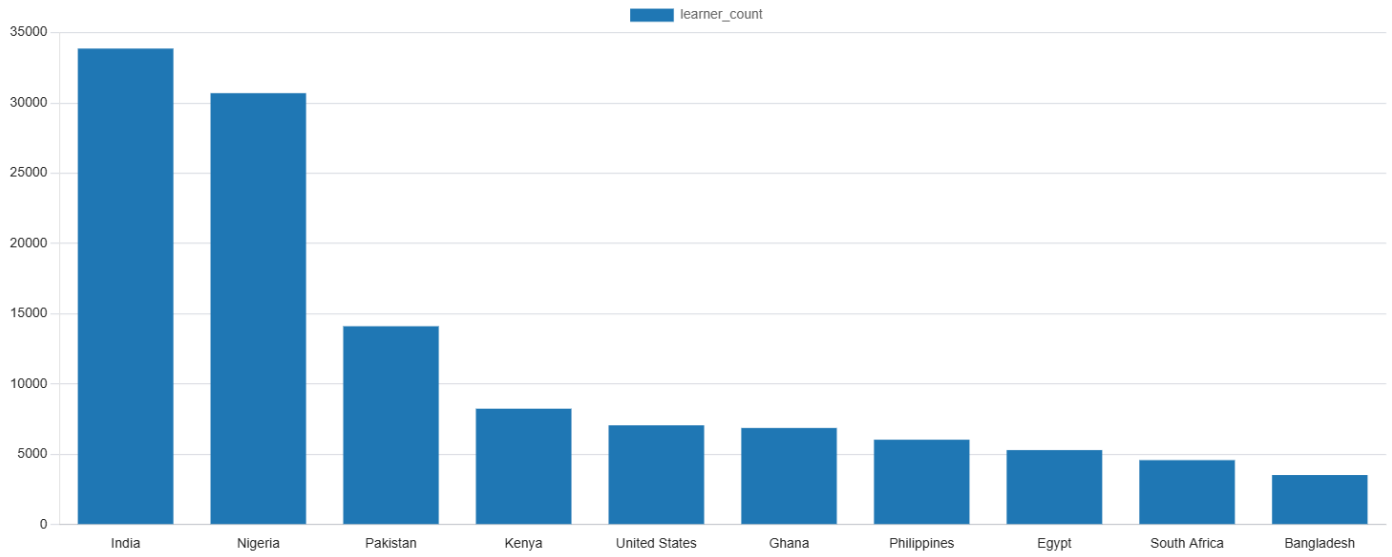
No. of Duplicate Rows in dataset: 0

Statistical Summary (Categorical Counts)

Degree	Count
Graduate Student	31806
Undergraduate Student	30709
Not in Education	6319
High School Student	4109
Other Professional	2997
Teacher/Educator	562
Parent of Student	64
NULL/Blank	52693

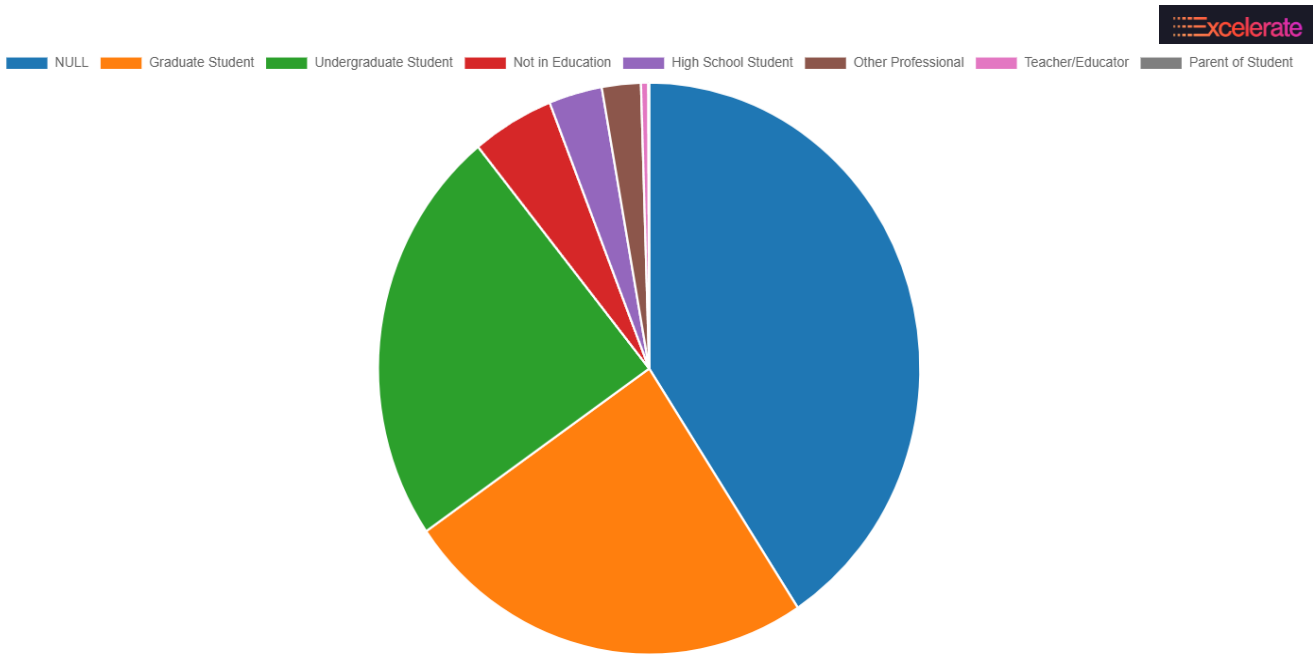
Visualizations:

- **No of learners from different country**



- **Pie Chart** of Degree Distribution:

- Majority of learners are Graduate and Undergraduate students.
- A significant portion of records contains NULL values in the "degree" column.
- Professional, High School, and other categories form a minority.



Data Dictionary

Column Name	Description
Learner ID	Unique identifier for each learner
Country	Country of the learner
Degree	Educational qualification of learner
Institution	Educational institution name
Major	Area of academic focus or study

Conclusion

This EDA helped in profiling the dataset, identifying key data issues (like NULL values and skewed degree distribution), and preparing actionable insights for cleaning and transformation. These insights will guide the next steps in the project including dashboard development and model integration.

Opportunity Data

The dataset under analysis contains detailed information about various opportunities, also it classifies, structures, plus makes them available. Key performance indicators (KPI's) that it includes are such things as:

- Opportunity ID
- Opportunity Name
- Opportunity Code
- Category
- Tracking Question

The aim in this exploratory data analysis (EDA) is:

- Understand dataset completeness and consistency as well as the structure
- Identify trends as well as patterns in opportunity distribution across categories
- Detect any of the missing, inconsistent, and duplicate values and handle all of them
- Evaluate names with unique identifiers like opportunity_id
- Visual and statistical summaries aid in future data preparation and modeling.

A cleaner more reliable dataset will enable data-driven decision-making for opportunity management as well as calculated planning, built from perceptions gained from this EDA.

Dataset Overview:

The table below tells the number of categorical ,alpha-numerical,text fields in the dataset.

Column Category	Number of fields
Alpha-Numerical Fields	2
Categorical Fields	2
Text	1

The table below tells the datatype of each field/ column, no. of blanks and recommendation to handle.

Column No.	Column Name	Column Category	No. of Blanks	How to Handle? (Recommendations)
1.	Opportunity id	Alpha-Numerical	0	-
2.	Opportunity name	Categorical	0	-
3.	Category	Categorical	0	-
4.	Opportunity code	Alpha-Numerical	0	-

5.	Tracking questions	Text	0	-

No. of Duplicate Rows in dataset:

There are so many duplicate values in category ,opportunity_ name and tracking_ question columns but there opportunity_ id and opportunity_ code are different that's why I didn't delete those values.

To Check Dataset Importation:

To check the data is imported successfully or not.There are some steps that we have to follow :

1. Click on the new database which you have created on POSTGRE SQL and select Query Tool.
2. Write a query

CREATE TABLE opportunity (opportunity_id varchar(150),opportunity_name varchar(150),category varchar(50),opportunity_code varchar(20),tracking_questions text);

3. To check table is created or not ,write a query

SELECT * FROM opportunity;

OR

SELECT opportunity_id,opportunity_name,category,opportunity_code,tracking_questions FROM opportunity;

If you see the table,which means you created the table successfully but it is empty because we don't insert any data yet.

4. After importing the data, you have to check that our data is imported successfully or not .For that we use query

SELECT opportunity_id,opportunity_name,category,opportunity_code,tracking_questions FROM opportunity;

If you see the data in the form of rows and columns ,it means the data is imported successfully.

To Handle Duplicates and Null Values:

To check there is no null and duplicate values we use a query

SELECT DISTINCT opportunity_id FROM opportunity;

If the No. of Columns = No. of Columns in Output ,it means there is no duplicate and null values.

To check all the columns we use a query SELECT DISTINCT column_name FROM opportunity;

If the No. of Columns ≠ No. of Columns in Output ,it means there is duplicate and null values.

In this dataset, there are so many duplicate values in category ,opportunity_ name and tracking_questions columns but there opportunity_ id and opportunity_ code are different so that's why I didn't delete those values.

I delete tracking_questions column using a query

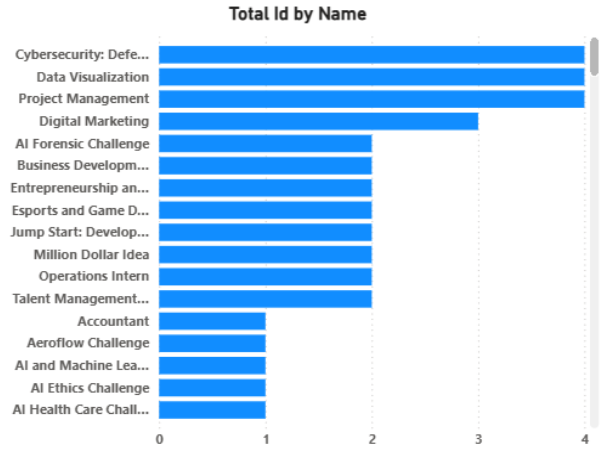
ALTER TABLE opportunity

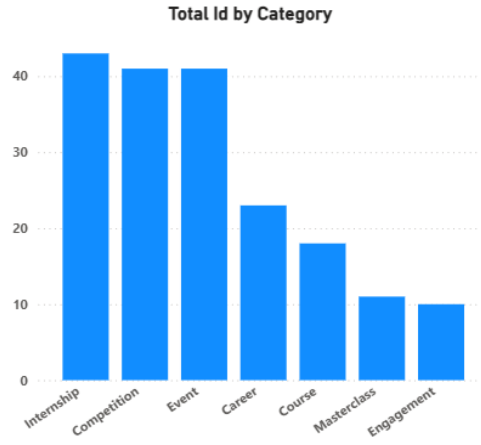
DROP COLUMN tracking_questions;

Because it contains so many Null values and Metadata.

Visualizations:

Opportunity Data Overview





These charts indicate that the dataset is dominated with a few names also categories of opportunity. Opportunities can appear most frequently in areas such as “Cybersecurity: Defence”, “Data Visualization”, and “Project Management” because these areas demand or focus so strongly. Regarding category, Internship, Competition, and Event make up most entries, and this highlights a strategy geared towards practical experience, skill-building, and engagement. Masterclass and Engagement are smaller categories, maybe showing specialized areas. Further exploration or further development could unlock some untapped potential from within those categories.

Recommendations

Here are some key takeaways from our analysis to improve future marketing performance:

1. **Leverage High-Interest Areas:**
   Continue promoting popular opportunity themes such as Cybersecurity, Data Visualization, and Project Management, as they show high engagement and alignment with learner or participant interests.

2. **Balance Opportunity Categories:**
   While Internships, Competitions, and Events dominate the dataset, consider diversifying efforts to strengthen underrepresented categories like Masterclasses and Engagement initiatives.

3. **Optimize for Demand:**
   Prioritize content creation and outreach in high-performing categories, but also evaluate the reasons behind lower participation in others—this could reveal gaps in visibility, relevance, or access.

4. **Assess Redundancy in Opportunity Names:**
   The repetition of certain opportunity names may suggest duplication or overlap. Consider merging similar offerings or clarifying distinctions to streamline the opportunity catalog.

5. **Track Long-Term Performance by Category:**
   Introduce metrics (e.g., conversion rates, completion rates) for each category to better evaluate which types of opportunities lead to meaningful outcomes.

6. **Build a Simple Dashboard**
   Create a visual dashboard (in Excel or Power BI) to explore the data more interactively.

Data Dictionary:

Column No.	Column Name	Description
1	**Opportunity ID**	**Unique Id assigned to each opportunity record**
2	**Opportunity Name**	**Title of the opportunity**
3	**Category**	**Type of the opportunity**
4	**Opportunity Code**	**A unique code used to identify the opportunity internally.**
5	**Tracking Questions**	**(REMOVED column) Contains Metadata and High Null values**

Conclusion:

In conclusion, the Week 1 EDA has helped us understand the structure, quality, and Key Performance Indicators (KPI's) in the dataset. These insights will guide the cleaning, transformation, and deeper analysis in Week 2.

Marketing Data

The dataset under analysis contains detailed information from a marketing campaign, providing insights into customer engagement, ad performance, and campaign efficiency. It includes key performance indicators (KPI's) such as:

- Ad Account Name
- Campaign Name
- Delivery Status and Level
- Reach
- Clicks and Views
- Result Type and Count
- Cost per Result
- Amount Spent (AED)
- CPC (Cost per Click)
- Campaign Start Date

This exploratory data analysis (EDA) aims to:

- Understand the structure and quality of the dataset
- Identify trends, correlations, and outliers
- Highlight key metrics driving campaign performance
- Detect any missing, inconsistent, or duplicate data
- Provide visual insights through charts, histograms, and heatmaps

The insights drawn will be instrumental in cleaning and transforming the data for modeling and in making data- driven decisions to optimize future marketing strategies.

Dataset Overview:

The table below tells the number of categorical and numerical fields in the dataset.

Column Category	Number of fields
Numerical Fields	8
Categorical Fields	5

The table below tells the datatype of each field/ column, no. of blanks and recommendation to handle.

Column No.	Column Name	Column Category	No. of Blanks	How to Handle? (Recommendations)
1.	Ad Account Name	Categorical	0	-
2.	Campaign name	Categorical	2	Remove
3.	Delivery status	Categorical	0	-
4.	Delivery level	Categorical	0	-
5.	Reach	Numerical	0	-
6.	Outbound clicks	Numerical	2	Replace with 0
7.	Landing page views	Numerical	2	Replace with 0
8.	Result type	Categorical	0	-
9.	Results	Numerical	0	-
10.	Cost per result	Numerical	0	-
11.	Amount spent (AED)	Numerical	0	-

12.	CPC (cost per link click)	Numerical	0	-
13.	Reporting starts	Numerical	0	-

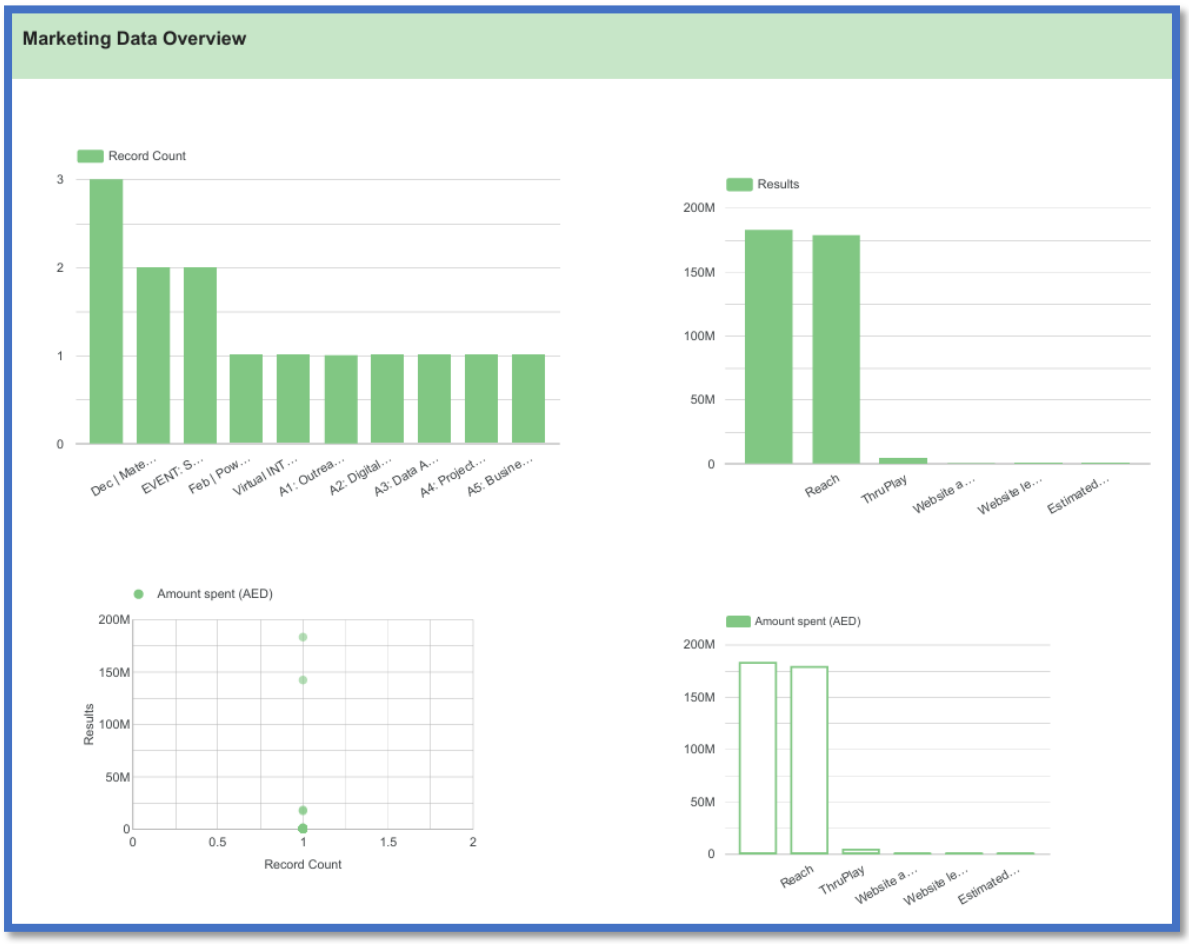
No. of Duplicate Rows in dataset: 0

The starting date of campaign for whole dataset: 1 January 2023.

The Statistical Summary of dataset is as follows:

Column No.	Column Name	Mean	Median	Mode
1	Reach	1720052.5	148357	#N/A
2	Outbound clicks	3557.345324	1579	42
3	Landing page views	2026.100719	818	200
4	Results	1313045.9	371	200
5	Cost per result	4.167181043	2.920683	#N/A
6	Amount spent (AED)	2312.947266	1222.5	1840.17
7	CPC (cost per link click)	1.061585572	0.7637075	#N/A
8	Reporting starts	#N/A	#N/A	#N/A

Visualizations:



These charts show that a few campaigns and result types dominate the dataset in both record volume and performance. High investment in "Reach" and "ThruPlay" aligns with high results, suggesting current ad strategies are geared towards awareness and video engagement. Lesser-performing result types may require targeted improvements or reallocation of budget.

Recommendations

Here are some key takeaways from our analysis to improve future marketing performance:

1. **Focus on What Works**
   Invest more in result types like *Reach* and *ThruPlay*—they're delivering the best outcomes.
2. **Look Into Low Performers**
   Check campaigns that had high spend but low results or zero clicks/views. Something might be off in targeting or content.
3. **Clean Up Campaign Names**
   Keep campaign names consistent to make tracking and analysis easier.
4. **Keep an Eye on Missing Data**
   Watch for blanks in landing page views or outbound clicks—they could lead to missed insights.
5. **Track Trends Over Time**
   Use the start dates to spot performance trends by month or quarter.
6. **Build a Simple Dashboard**
   Create a visual dashboard (in Excel or Power BI) to explore the data more interactively by campaign or date.

Data Dictionary:

Column No.	Column Name	Description
1	Ad Account Name	Name of the advertiser's account
2	Campaign Name	Title of the marketing campaign
3	Delivery Status	Indicates if the ad was delivered or not
4	Delivery Level	Specifies the level at which delivery was tracked (e.g., ad, set, campaign)
5	Reach	Number of unique users who saw the ad
6	Outbound Clicks	Clicks that led users away from the platform
7	Landing Page Views	Number of times the landing page was actually viewed
8	Result Type	Type of result tracked (e.g., Reach, ThruPlay, Leads)
9	Results	Total number of outcomes from the campaign based on the result type
10	Cost per Result	Cost incurred to generate one result

11	**Amount Spent (AED)**	**Total spending on the campaign in AED**
12	**CPC (Cost per Link Click)**	**Average cost per link click**
13	**Reporting Starts**	**Start date of data collection/reporting**

Conclusion:

In conclusion, the Week 1 EDA has helped us understand the structure, quality, and key performance indicators in the dataset. These insights will guide the cleaning, transformation, and deeper analysis in Week 2.

Learning Opportunity Data

The dataset under analysis contains detailed information about learner enrolments in a program or opportunity database. It captures learner identifiers, application dates, cohort assignments, and a numerical status indicator. Key performance indicators (KPI's) that it includes are such things as:

- Enrolment Id
- Learner Id
- Assigned Cohort
- Apply Date
- Status

The aim in this exploratory data analysis (EDA) is:

- Understand the structure and quality of the dataset
- Identify trends, missing data, or inconsistencies
- Highlight key attributes influencing learner status or behavior
- Provide visual insights and statistical summaries

The insights drawn will be instrumental in cleaning and transforming the data for modelling, visualization, and driving data-informed learner engagement strategies.

Dataset Overview:

The table below tells the number of categorical and numerical fields in the dataset.

Column Category	Number of fields
Numerical Fields	1
Categorical Fields	4

The table below tells the datatype of each field/ column, no. of blanks and recommendation to handle.

Column No.	Column Name	Column Category	No. of Blanks	How to Handle? (Recommendations)
1.	Enrolment_id	Categorical	0	-
2.	Learner_id	Categorical	0	-
3.	Assigned_cohort	Categorical	13318	Replace with "Unknown" or "NaN"
4.	Apply_date	Categorical	188	Replace with "Unknown" or "NaN"
5.	Status	Numerical	186	Replace with mode (1070.0)

No. of Duplicate Rows in dataset:

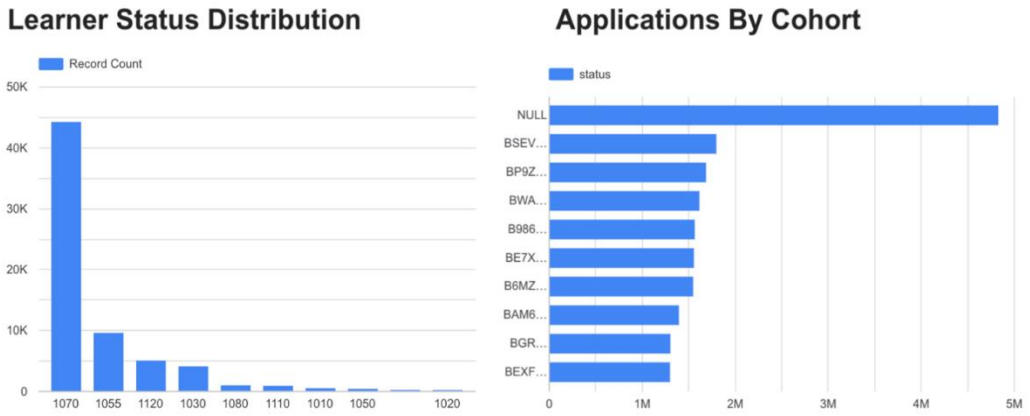
None occurred.

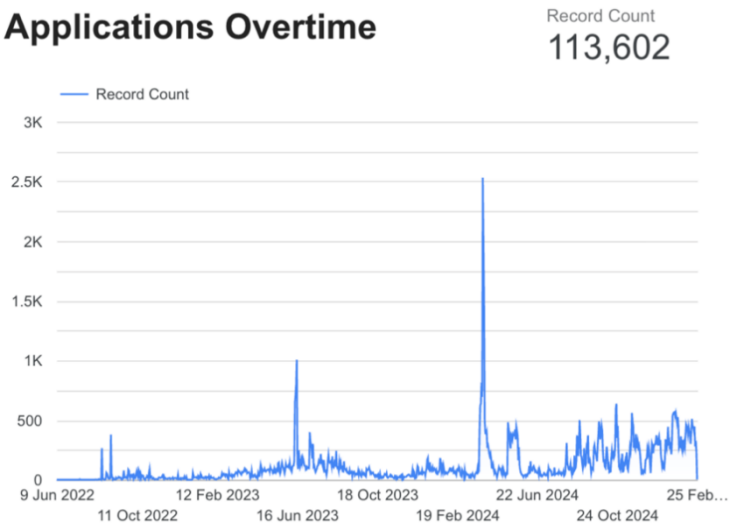
The Statistical Summary of dataset is as follows:

Column No.	Column Name	Mean	Median	Mode	Min	Max	Std Dev
1	Status	1068.19	1070.0	1070.0	1010.0	1120.0	21.03

The status field is a numerical representation of learner progress or classification. Its average value is 1068.19, with the most frequent value (mode) being 1070.0. The distribution is fairly concentrated as seen in the low standard deviation (21.03), suggesting that most learners fall into a narrow status band. Further clarification on what status codes represent would enhance interpretation.

Visualization





The first chart shows how learners are distributed based on their current outcomes, making it easy to identify if most are in one stage or spread across different results. This helps assess whether learners are progressing uniformly or if there are potential bottlenecks. The second chart tracks how applications change over time,

revealing trends such as spikes during certain periods or steady growth. These patterns are useful for understanding when learners are most active and planning outreach accordingly. The third chart compares how many learners are assigned to each group, helping to highlight which programs are more popular and where more support or attention may be needed.

Recommendations and Key Takeaways:

Few recommendations from our analysis to enhance performance,

- **Focus on Cohort Engagement**
  Invest in analyzing cohorts with a high number of applications or success rates. These cohorts may reflect effective outreach, better onboarding, or more motivated learners.
- **Investigate Learner Status Distribution**
  Review why the majority of learners have the same status code (most are 1070). If this status reflects a bottleneck or default state, the learning pipeline might need re-evaluation.
- **Standardize Data Entry for Cohorts and Dates**
  A significant number of entries have missing assigned_cohort and apply_date values. This suggests inconsistent data logging, which can hurt long-term analysis and tracking.
- **Fill Gaps in Application Data**
  Missing apply_date entries may indicate offline or delayed applications. These should be reconciled to maintain accurate timelines and detect trends in application behavior.
- **Track Learner Trends Over Time**
  Once apply_date is cleaned and formatted as a date, use it to analyze application volume and learner engagement by month, quarter, or academic cycle.
- **Create a Monitoring Dashboard**
  Build a dashboard (in Excel, Power BI, or Streamlit) to visualize learner distribution by cohort, daily or weekly application counts, and status transitions. This can support outreach, scheduling, and data quality checks.

Data Dictionary:

Column No.	Column Name	Description
1	Enrolment id	Unique identifier for each enrolment
2	Learner Id	Unique identifier for each learner
3	Assigned cohort	Cohort to which learner was assigned
4	Apply date	Date the learner applied
5	status	Status code representing learner's application outcome

Conclusion:

In conclusion, the Week 1 EDA has helped us understand the structure, quality, and key performance indicators in the dataset. These insights will guide the cleaning, transformation, and deeper analysis in Week 2.

Cohort Data

Overview

The dataset cohort_data tracks cohort-based learning programs. Each record represents a unique cohort with associated metadata, such as cohort code, timeline, and size. The primary objective of this EDA is to understand the structure, timeline distribution, scale, and anomalies in cohort enrollments.

Key Columns in the Dataset:

- Cohort ID
- Cohort Code
- Start Date
- End Date
- Size (Number of learners)

The aim in this exploratory data analysis (EDA) is:

- Understand dataset structure and types
- Detect missing or inconsistent records
- Explore distribution of cohort sizes and timelines
- Spot anomalies (e.g., start = end date, extremely large cohorts)
- Recommend data handling and transformation steps

Dataset Overview:

The table below tells the number of categorical ,alpha-numerical,text fields in the dataset.

Column Category	Number of fields
Numerical Fields	1(Size)
Categorical Fields	2(Cohort Id and Cohort Code)
Date Fields	2(Start and End)

The table below tells the datatype of each field/ column, no. of blanks and recommendation to handle.

Column No.	Column Name	Column Category	No. of 'NULLs'	How to Handle? (Recommendations)
1.	Cohort_id	Categorical	0	-
2.	Cohort_code	Categorical	0	-

3.	Start date	Date (UNIX)	0	Convert from UNIX to readable format
4.	End date	Date (UNIX)	0	Convert from UNIX to readable format
5.	Size	Numerical	0	Review for extreme values

No. of Duplicate Rows in dataset: 0.

Statistical Summary

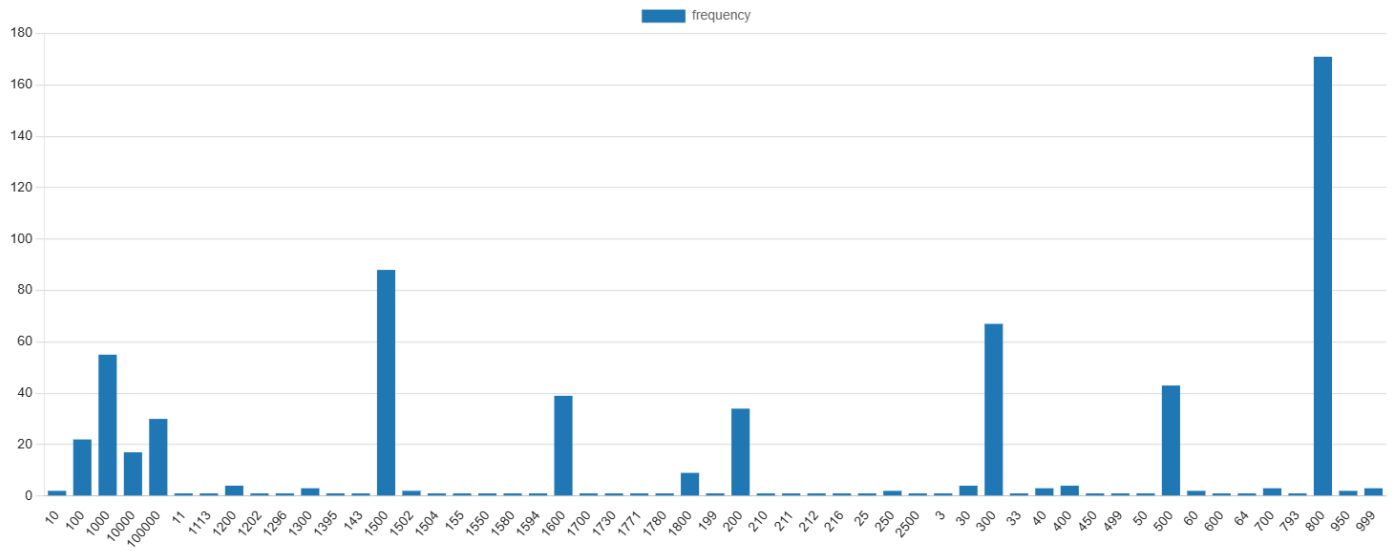
Metric	Size
Mean	5741.4241001564945227
Median	800
Max	100000
Min	3
Standard Dev.	20994.26661203

Observations & Insights

- Some cohorts have **very large sizes** (e.g., 100,000), which may be valid for MOOC-style programs but should be verified.
- A few cohorts have **start and end dates the same**, possibly indicating 1-day events or data entry errors.
- Distribution of cohort start dates seems to center around specific periods — this can be visualized with a histogram or line chart.

Visualizations:

Histogram of cohort sizes



Data Dictionary

Column Name	Description
cohort_id	Unique identifier of the cohort
cohort_code	Alphanumeric label for the cohort
start_date	Starting date of the cohort (UNIX timestamp)
end_date	Ending date of the cohort (UNIX timestamp)
size	Number of learners enrolled in the cohort

Conclusion

This EDA provides a foundational understanding of the cohort-based learning dataset. It highlights important trends, identifies anomalies, and offers steps for data cleaning and enhancement. The cleaned dataset can be further used for performance tracking, forecasting, and dashboard creation.

Cognito Data

The Cognito dataset is comprised of raw user profile data collected to help identify users, classify users, and validate their data. The dataset contains various demographic and registration-based properties of each user. Key Performance Indicators (KPI's) that it includes are such things as :

- User ID
- Email
- Gender
- UserCreateDate
- UserLastModifiedDate
- Birthday
- City
- Zip
- State

The aim in this exploratory data analysis (EDA) is:

- Evaluating quality, structure, and completeness of data overall, and specifically for critical fields.
- Identifying and addressing missing, null, or non-standard values (e.g, 'null', 'NA', invalid dates).
- Standardizing fields (eg, Gender and Zip) for consistency.
- Disaggregating data for elements such as time from Created Date.
- Validating uniqueness using ID and identifying any duplicates.
- Creating statistical and visual summaries to assist in the preparation, enhancement, and analysis of the data downstream.

The analysis delivered will prepare a cleaned and reliable version of the data set that can be used for user segmentation, profile validation and decision making based on user characteristics and registration behaviour.

Dataset Overview:

The table below tells the number of categorical ,alpha-numerical,date fields in the dataset.

Column Category	Number of fields
Alpha-Numerical Fields	3
Categorical Fields	3
Date	3

The table below tells the datatype of each field/ column, no. of blanks/null values and recommendation to handle.

Column No.	Column Name	Column Category	No. of Blanks/Null	How to Handle? (Recommendations)
1.	User ID	Alpha-Numerical	0	-
2.	Email	Alpha-Numerical	0	-
3.	Gender	Categorical	42862	NA
4.	UserCreateDate	Date	0	-
5.	UserLastModifiedDate	Date	0	-
6.	Birthday	Date	42862	Remove column (Because it has No use)
7.	City	Categorical	42863	NA
8.	Zip	Alpha-Numerical	42867	0
9.	State	Categorical	42864	NA

No. of Duplicate Rows in dataset:

There are so many duplicate values in Gender and Date column but there user_id is different that's why I didn't delete those columns.

To Check Dataset Importation:

To check the data is imported successfully or not.There are some steps that we have to follow :

1) Click on the new database which you have created on POSTGRE SQL and select Query Tool.
2) Write a query

   CREATE TABLE cognito_raw (user_id TEXT,email TEXT,gender TEXT, usercreatedate TEXT, userlastmodifieddate TEXT, birthdate TEXT, city TEXT,zip TEXT, state TEXT);

   To import the raw data we use this query

3) To check table is created or not ,write a query

SELECT * FROM cognito_raw;

OR

SELECT user_id ,email ,gender , usercreatedate , userlastmodifieddate , birthdate , city ,zip , state

FROM cognito_raw;

If you see the table,which means you created the table successfully but it is empty because we don't insert any data yet.

4) After importing the data, you have to check that our data is imported successfully or not .For that we use query

SELECT user_id ,email ,gender , usercreatedate , userlastmodifieddate , birthdate , city ,zip , state

FROM cognito_raw;

If you see the data in the form of rows and columns ,it means the data is imported successfully.

5) To import data with proper datatypes we use this query

Create table cognito(user_id varchar(100),email varchar(100),gender varchar(50),usercreatedate Timestamp,userlastmodifieddate Timestamp,birthdate DATE,city varchar(50),zip varchar(50),state varchar(50))

6) Repeat Steps 3 & 4 , if these steps successfully worked then delete the cognito_raw table

Using a query

DROP TABLE cognito_raw

To Handle Duplicates and Null Values:

To check there is no null and duplicate values we use a query

SELECT DISTINCT user_id FROM cognito;

If the No. of Columns = No. of Columns in Output ,it means there is no duplicate and null values.

To check all the columns we use a query SELECT DISTINCT column_name FROM opportunity;

If the No. of Columns ≠ No. of Columns in Output ,it means there is duplicate and null values.

In this dataset, there are so many duplicate or null values in gender and date columns but there user_id and email are different so that's why I didn't delete those values.

I delete birthdate column because it have so many null values and it is useless column with using the query

ALTER TABLE cognito

DROP COLUMN birthdate;

I replaced null values with either "NA" or "0" according to their preferences,with using the query

UPDATE cognito

SET gender = 'NA'

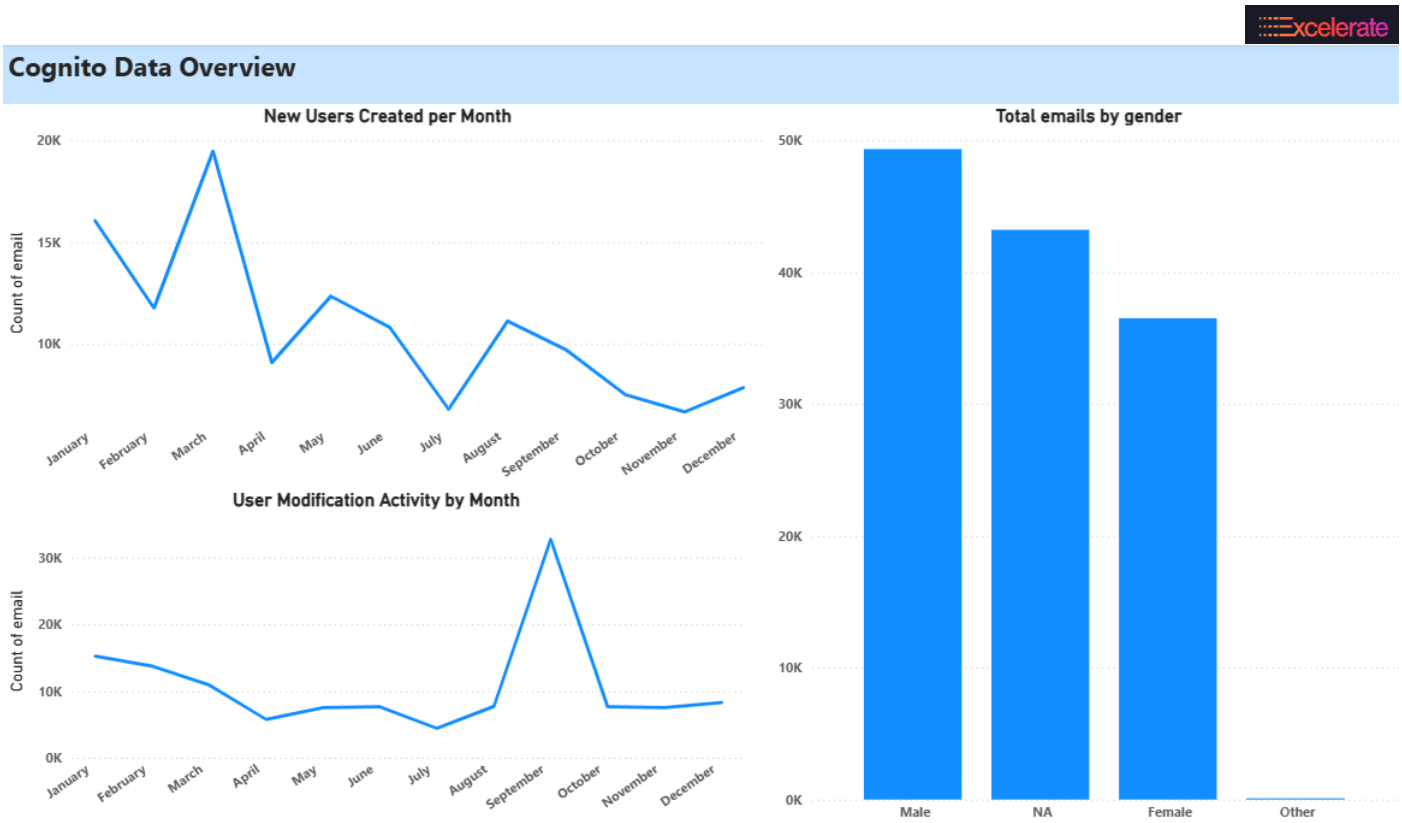
WHERE gender IS NULL

I use the same query for city,state and in the case of zip I replace "NA" with "0".

To check how many null values or blank values in my dataset I use this query

SELECT COUNT (*) FROM cognito WHERE city IS NULL;

Visualizations:



These charts show that user creation peaked in March, while profile updates spiked in October, indicating periods of high engagement or system activity. The gender distribution is dominated by Male and NA entries, with fewer Female users and minimal representation of "Other," pointing to gaps in demographic completeness. Overall, the dataset reflects strong user growth and activity patterns, but also highlights areas for improvement in data quality and inclusivity.

Recommendations

Here are some key takeaways from our analysis to improve data quality, user engagement, and future user management strategies:

1) **Improve Gender Data Collection**
   Encourage users to provide gender information during registration to reduce “NA” values and improve demographic insights.
2) **Address Data Gaps and Standardization**
   Clean and standardize inconsistent entries (e.g.,NA, missing zip codes) to ensure better accuracy in analysis and reporting.
3) **Investigate Activity Spikes**
   Analyze the causes of March (signups) and October (modifications) spikes to identify patterns—e.g., marketing campaigns or policy changes.
4) **Enhance Inclusivity**
   Offer more inclusive gender options and make users aware of them to ensure broader representation.
5) **Segment and Target Based on Trends**
   Use the time-based trends in creation/modification activity to plan targeted outreach or engagement efforts during high-traffic months.
6) **Build a Simple Dashboard**
   Create a visual dashboard (in Excel or Power BI) to explore the data more interactively.

Data Dictionary:

Column No.	Column Name	Description
1	User ID	Unique Id assigned to each user
2	Email	User's email address used for login or communication
3	Gender	Gender of the user
4	UserCreateDate	Date and time the user account was created.
5	UserLastModifiedDate	Date and Time of the last profile update
6	Birthdate	(Removed Column)it has high null values and no use in visualisation. Shows Birthdate of User
7	City	City of the user
8	Zip	Postal code of the user
9	State	State of the user

Conclusion:

In conclusion, the Week 1 EDA has helped us understand the structure, quality, and Key Performance Indicators (KPI's) in the dataset. These insights will guide the cleaning, transformation, and deeper analysis in Week 2.

END