EDA REPORT

TEAM 28

CONTRIBUTORS:

S/N	NAMES	DATASET WORKED ON
1.	Kayode Idris Adelakun (**ikadelakun@gmail.com**) **Team Leader**	**User Data (user_data.csv)** **Marketing Data (marketing_data.csv)**
2.	Most.Sonia Islam (saniaislamsava@gmail.com)	**Cognito Data (cognito_raw.csv)**
3.	Nahian Tasnim (nahian.tasnim@slu.edu)	**Learner Opportunity Data (learner_opportunity_raw.csv)**
4.	Mithun Kumar Dey (mithunkumardey789@gmail.com)	**Cohort Data (cohort_data.csv**
5.	Nirja Patel (nirja2501@gmail.com)	**Opportunity Data (opp_data.csv)**

EDA REPORT ON USER DATA (USER_DATA.CSV)

Overview of the Dataset

Features	Details
Rows	76,036
Columns	5
Column Names	Learner ID, Country, Degree, Institution, Major
Data Types	Texts, Numbers

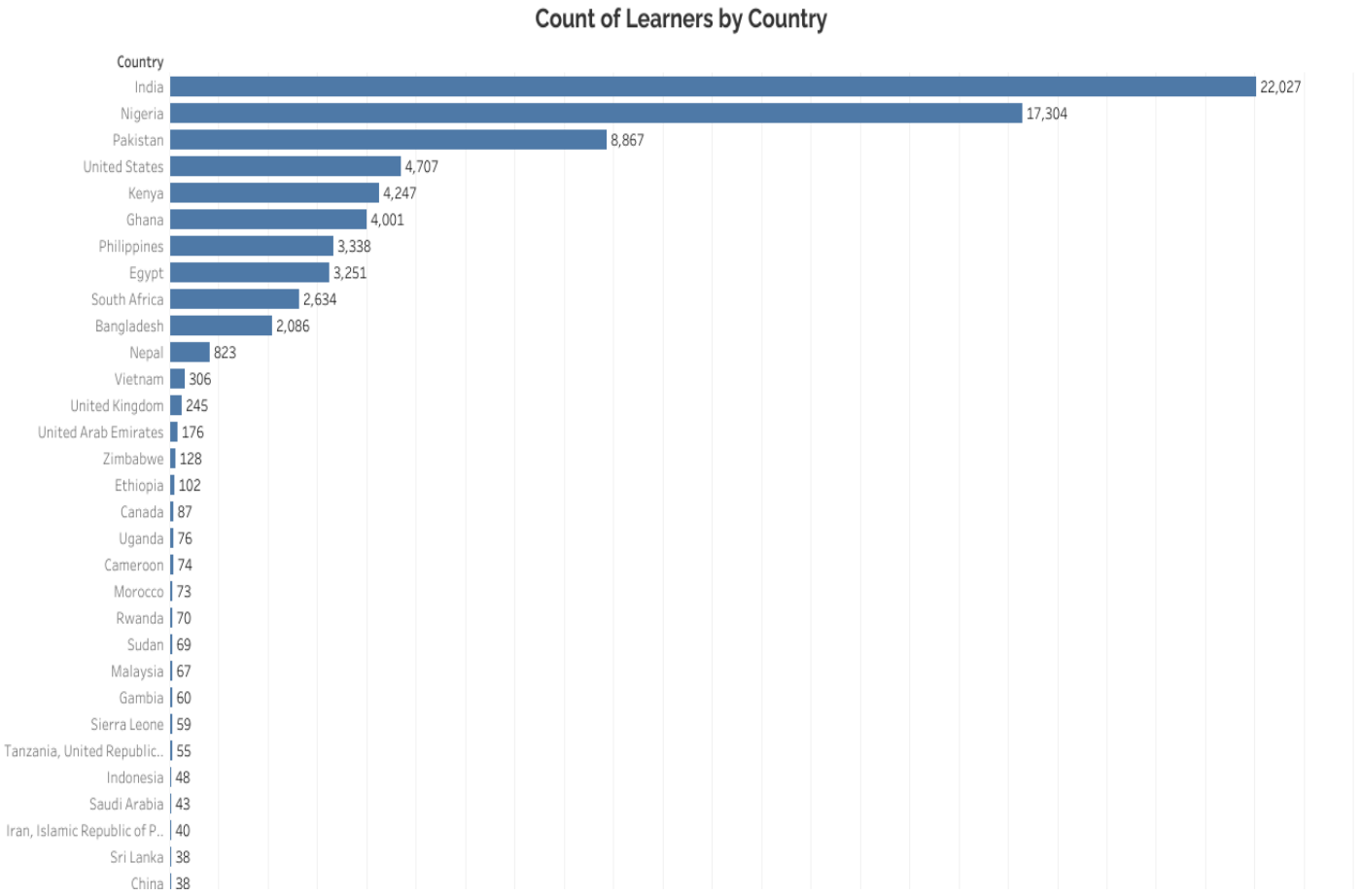
Summary Statistics

Column	Count	Unique Values	Most Frequent	Frequency
Learner ID	76,036	76,036	All Unique	1
Country	76,036	154	India	22,027
Degree	76,036	7	Graduate Student	31,736
Institution	76,036	34,488	Saint Louis University	2,163
Major	76,036	4,478	Computer Science	4,692

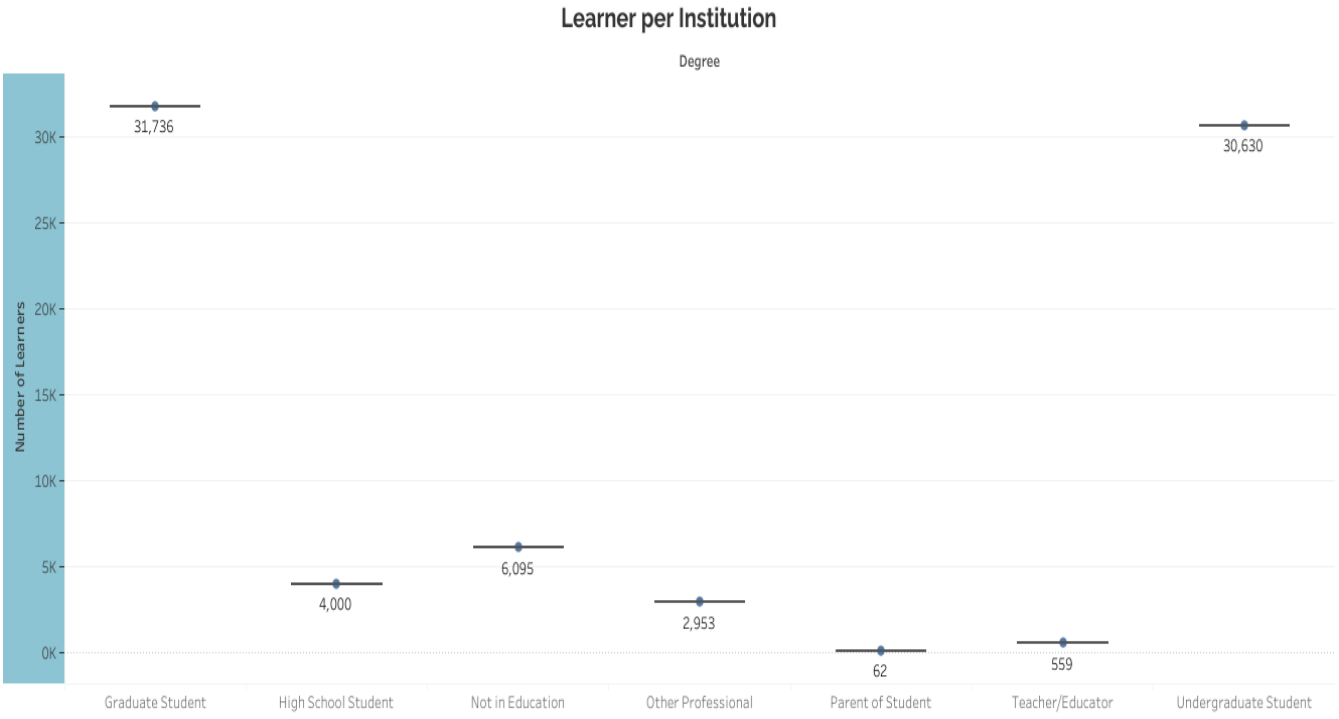
MISSING VALUES

COLUMN	MISSING COUNTS
Learner ID	0
Country	2,275
Degree	52,693
Institution	53,073
Major	52,871

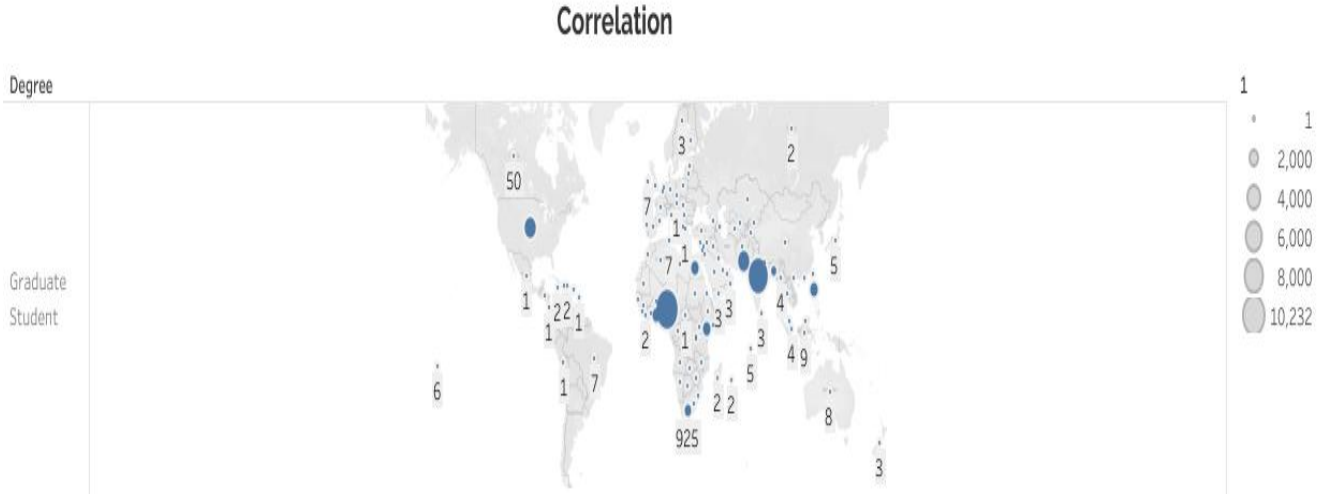
COUNT OF LEARNERS BY COUNTRY



LEARNERS PER INSTITUTION

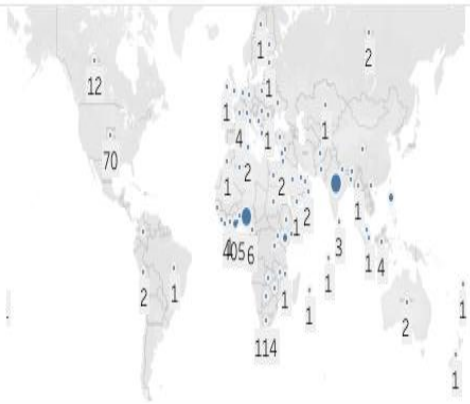


CORRELATIONS



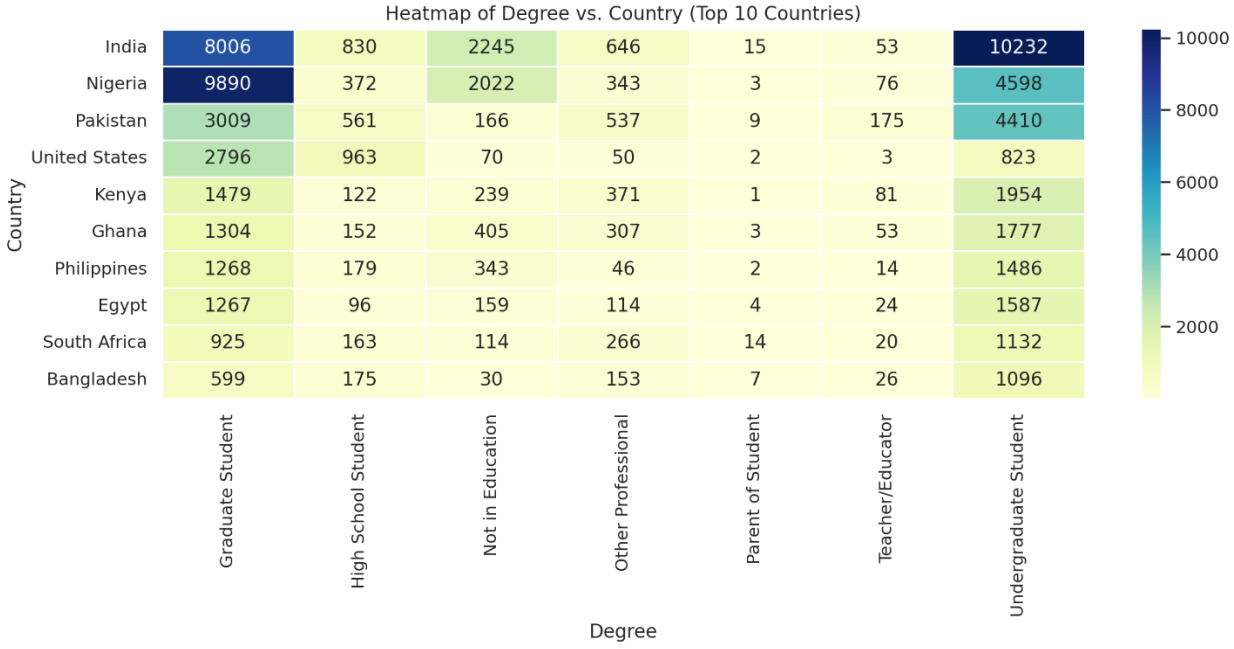
High School Student

Not in Education



Other Professional


Teacher/Ed..
Undergradua
te Student
© 2025 Mapbox © OpenStreetMap


Heatmap of Degree vs. Country (Top 10 Countries)
Country
India 8006 830 2245 646 15 53 10232
Nigeria 9890 372 2022 343 3 76 4598
Pakistan 3009 561 166 537 9 175 4410
United States 2796 963 70 50 2 3 823
Kenya 1479 122 239 371 1 81 1954
Ghana 1304 152 405 307 3 53 1777
Philippines 1268 179 343 46 2 14 1486
Egypt 1267 96 159 114 4 24 1587
South Africa 925 163 114 266 14 20 1132
Bangladesh 599 175 30 153 7 26 1096
Graduate Student
High School Student
Not in Education
Other Professional
Parent of Student
Teacher/Educator
Undergraduate Student
Degree
10000
8000
6000
4000
2000

KEY FINDINGS

- The dataset contains 76,036 records, each representing a unique learner.
- It spans 154 countries, 34,488 institutions, and 4,478 unique majors, indicating global and academic diversity**.**
- India accounts for the largest share of learners

EDA REPORT ON OPPORTUNITY DATA (OPP_DATA.CSV)

1. Overview of the Dataset

Features	Details
Rows	187
Columns	5
Column Names	opportunity_id, opportunity_name, category, opportunity_code, tracking_questions
Data Types	All columns are of type character varying except tracking_questions (data type- text)

2. Summary Statistics

Column	Count	Unique	Most Frequent	Frequency
opportunity_id	187	187	Internship	43
opportunity_name	187	170	Event	41
category	187	7	Competition	41
opportunity_code	187	187	Career	23
tracking_questions	187	119	Course	18

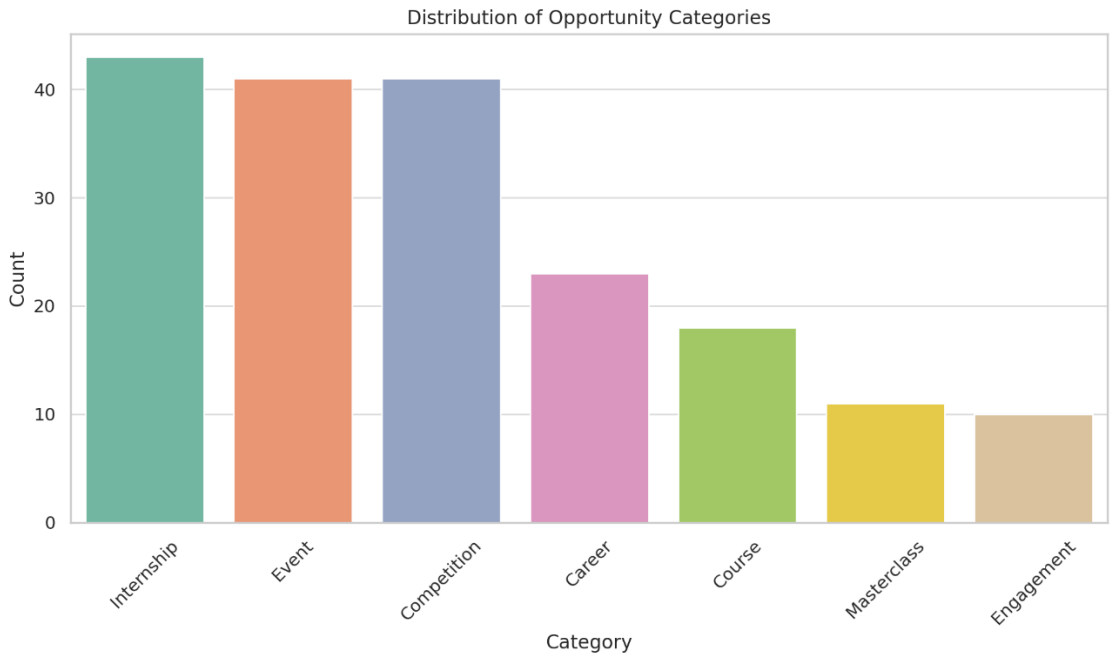
3. Missing Values

Column	Missing Count	Missing percentage
opportunity_id	0	0.00

opportunity_name	5	3.45
Category	10	6.90
Opportunity_code	7	4.83
tracking_questions	68	47.24

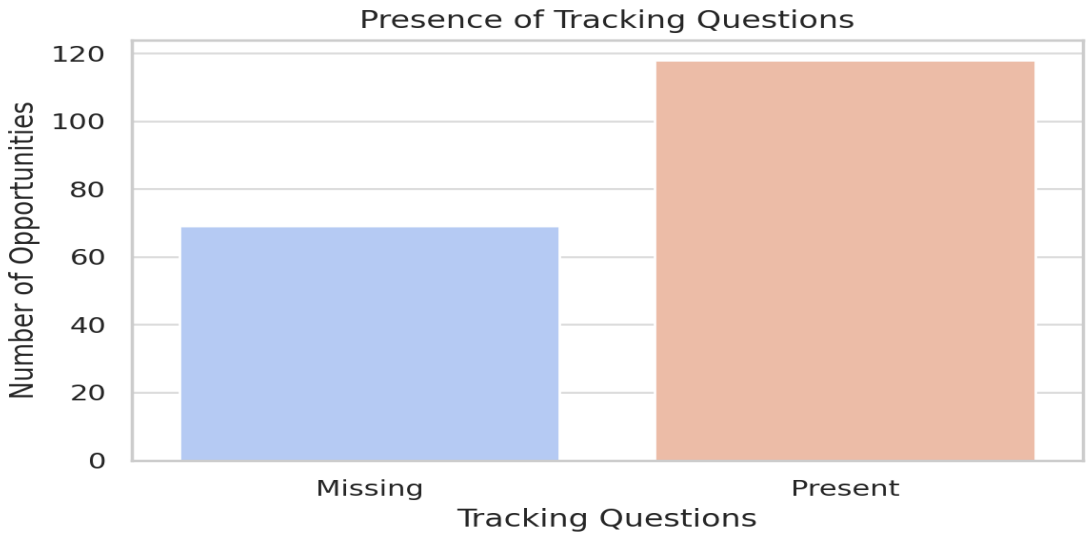
4. Duplicate Records
    - Duplicate Rows:0
    - All Records are unique by opportunity_id

5. Data Visulization
    1) Category-wise distribution
        - Internship dominates with highest number of opportunities.

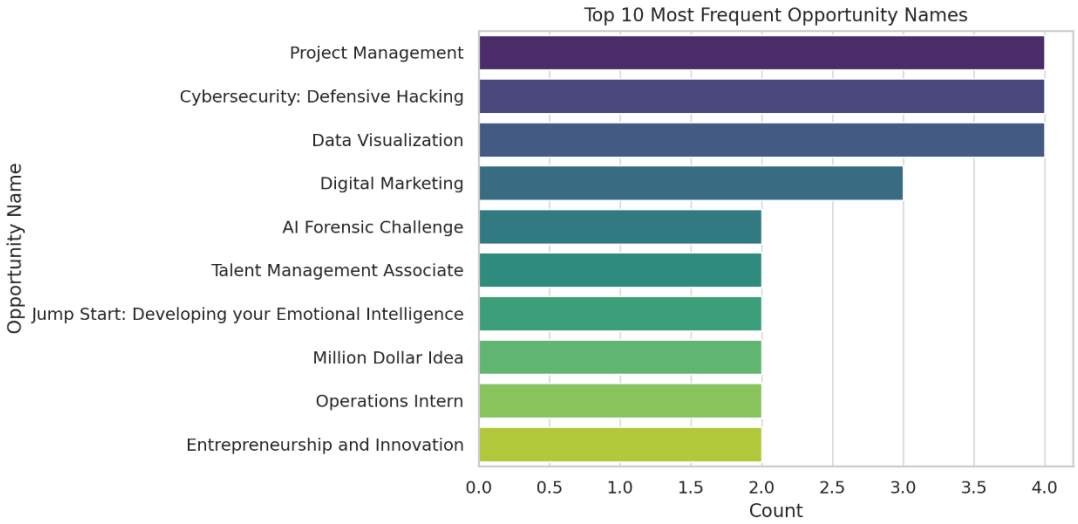


2) Presence of tracking questions

- Present- This bar shows that 111 opportunities have valid entries in tracking_questions
- Missing- This bar shows that 69 **opportunities** lack information in the tracking_questions column — they are either empty or null.



3) Most Frequent Opportunity distribution



6. Key Findings

- Around **36% of the opportunities** have missing tracking questions, indicating incomplete entries that may require attention before analysis or reporting.
- The dataset includes a variety of opportunity types, with categories like **Workshop, Internship, Scholarship**, and **Event** being the most common — indicating a diverse set of opportunities offered.

- Critical fields such as opportunity_name and opportunity_code are **mostly complete**, but **some entries have missing category or tracking information**, which may affect filtering or classification tasks

EDA REPORT ON COHORT DATA (COHORT_DATA.CSV)

I. Overview of the Dataset

Features	Details
Rows	639
Columns	5
Column Names	Cohort Id, cohort code,Size,Start date,End date
Data Types	Number,Text,Date

II. Summary Statistics

Column	Count	Unique
Cohort Id	639	639
cohort code	639	639
Size	639	53
Start Date	639	263
End Date	639	275

III. Missing Values

Column	Missing Count	Missing percentage

Cohort Id	0	0.00
Cohort code	0	0.00
Size	0	0.00
Start Date	0	0.00
End Date	0	0.00

IV. Duplicate Records

- Duplicate Rows:0
- All Records are unique by opportunity_id

V. Average Duration 56.41

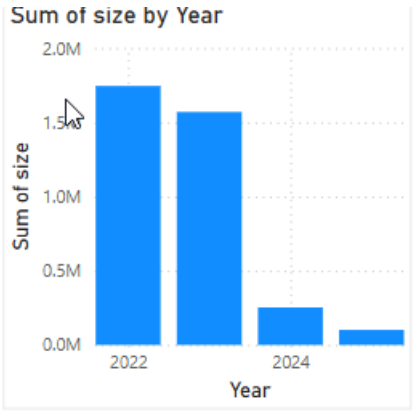
VI. Average Size 5741.42

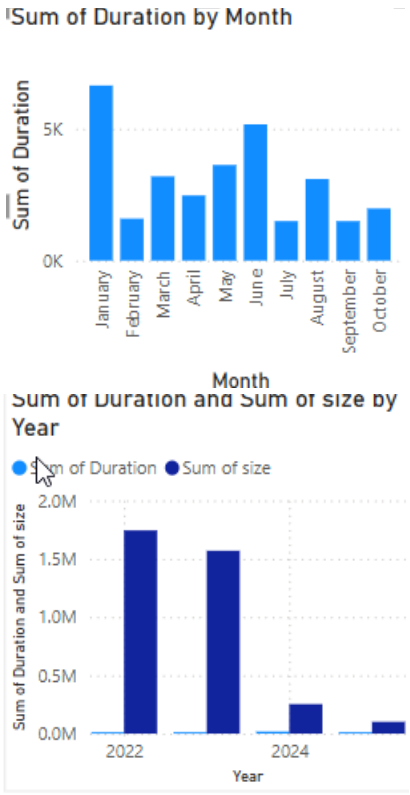
VII. Maximum Size 100000

VIII. Minimum Size 3

IX. Maximum Duration 1096.06

Visualization





X. Key Findings

- The dataset is **well-structured and clean**, with no missing or duplicate entries.
- There is **significant diversity** in both cohort size and duration:

➢ Some are very small, others extremely large.

➢ Durations range from a few days to **over 3 years**.

- The high number of **unique start and end dates** suggests a **steady flow of cohort activity** over time.

EDA REPORT ON MARKETING DATA (MARKETING_DATA.CSV)

XI. Overview of the Dataset

Features	Details
Rows	141
Columns	➢ Ad account name ➢ Campaign Name, ➢ Delivery Status ➢ Delivery level ➢ Reach ➢ Outbound Clicks ➢ Landing Page Views ➢ Result Type ➢ Results ➢ Cost per Result ➢ Amount Spent (AED) ➢ CPC (Cost per Link Click) ➢ Reporting Starts
Data Types	Number, Text, Date

Summary Statistics

- Reach ranges from 1.3K to 141.8M
- Results have a mean of 1.29M, but also a very high standard deviation, suggesting large outliers or high variation across campaigns.
- Cost per Result varies significantly (min: AED 0.0036, max: AED 47.09)

Missing Data

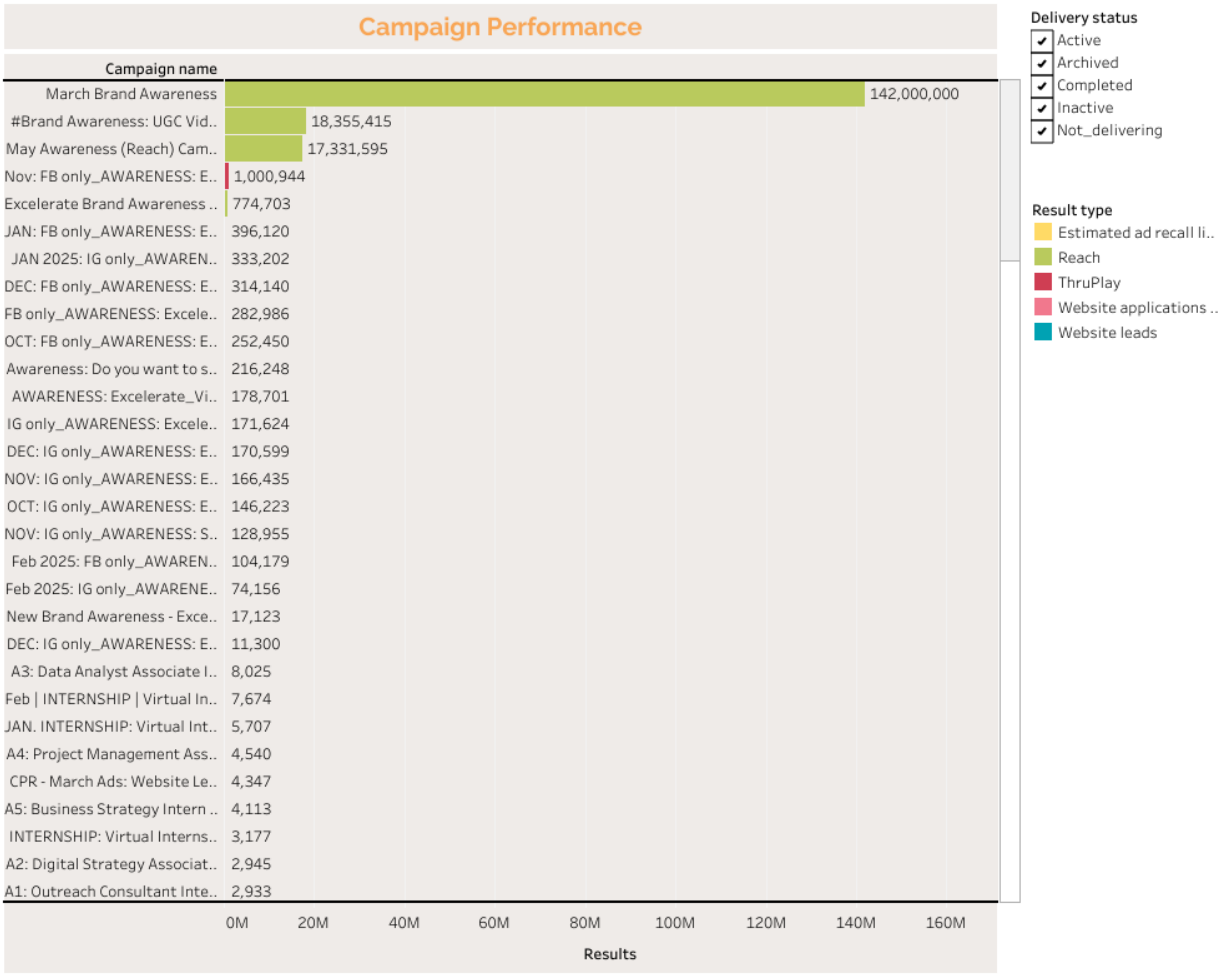
Column	Missing Values
Outbound Clicks	1604
Landing page view	1604
CPC (Cost per link Click)	0.7637075

Duplicates: None

DATA VISUALIZATION

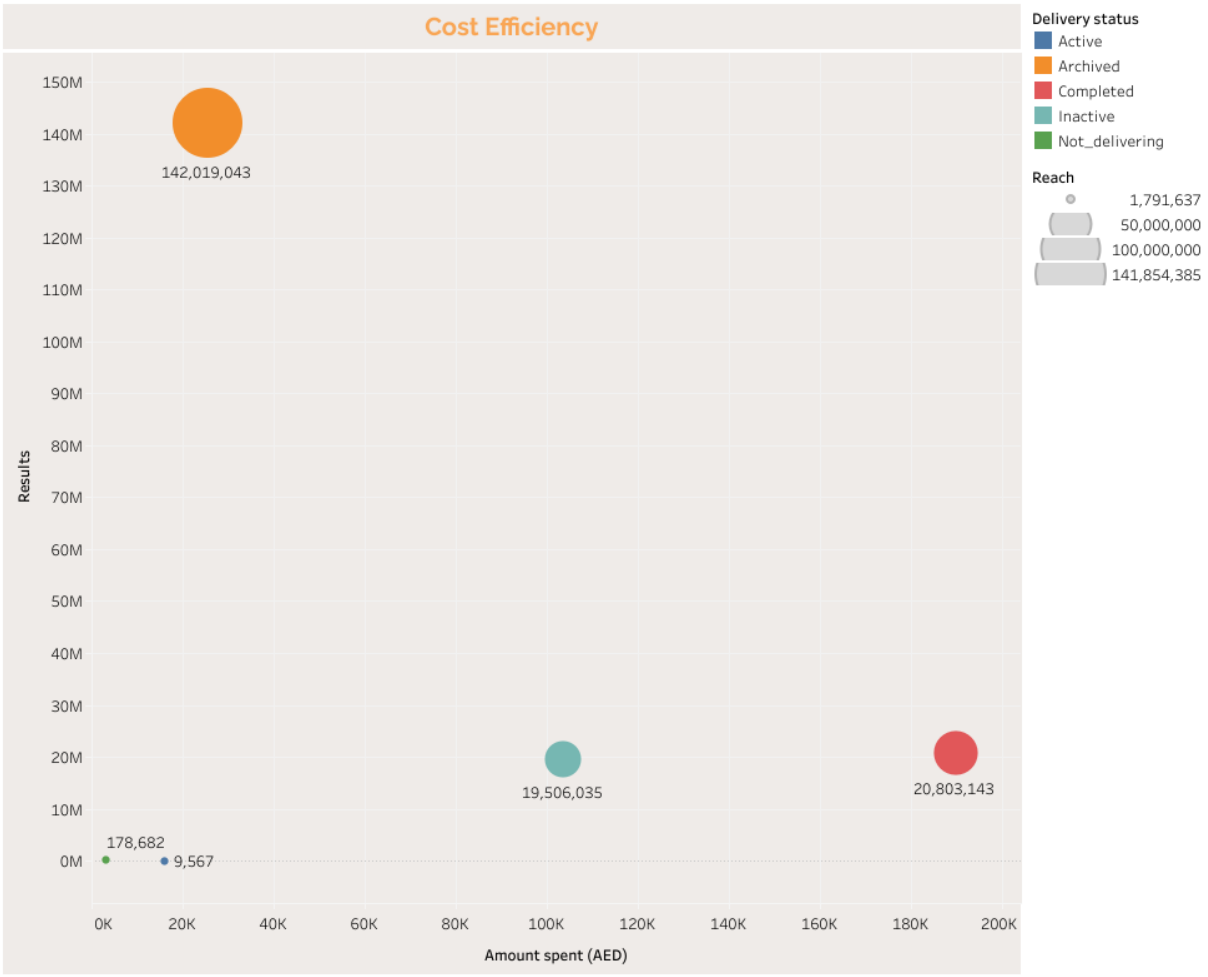
i. CAMPAIGN PERFORMANCE

- Identifies the top-performing campaigns
- The number of results generated by each campaign.
- Campaigns are listed on the Y-axis, and results are on the X-axis.



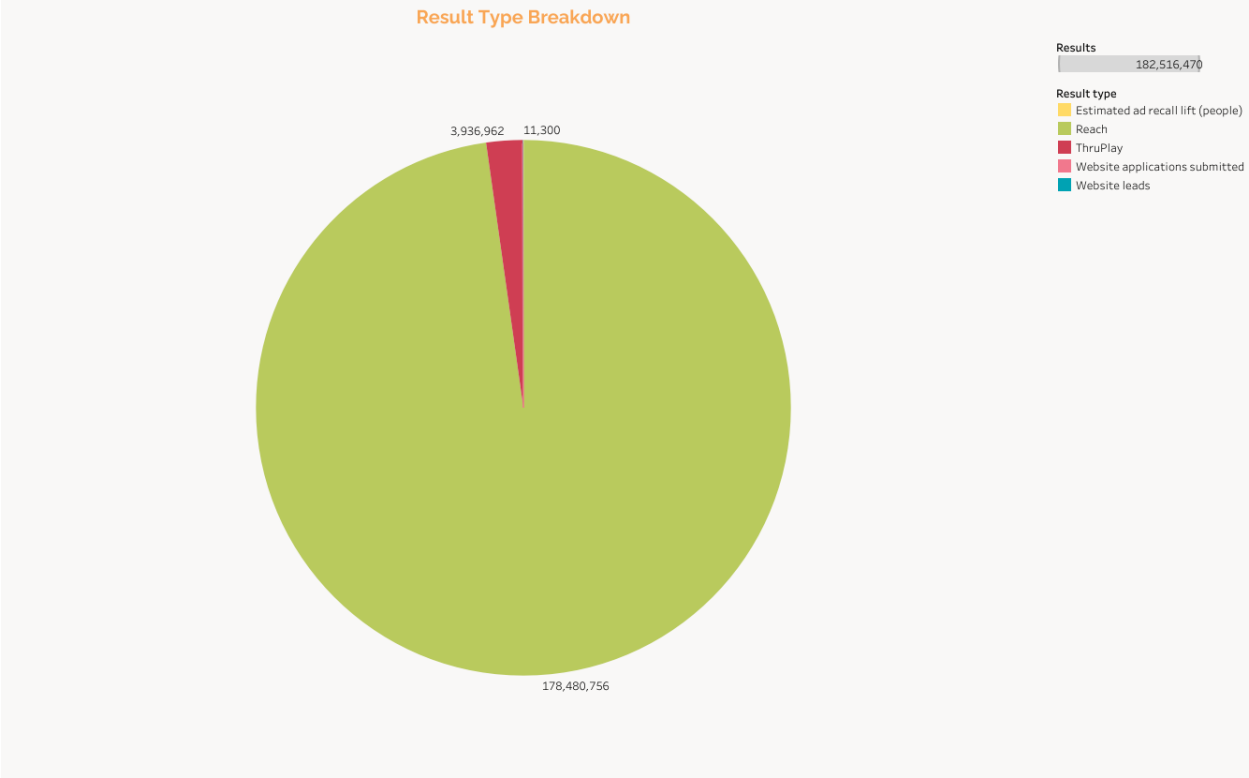
XII. COST EFFICIENCY

- Relationship between Amount Spent (AED) and Results, with bubble size representing results or another cost metric.
- A big bubble with low spend = great ROI.
- A small bubble with high spend = poor ROI.



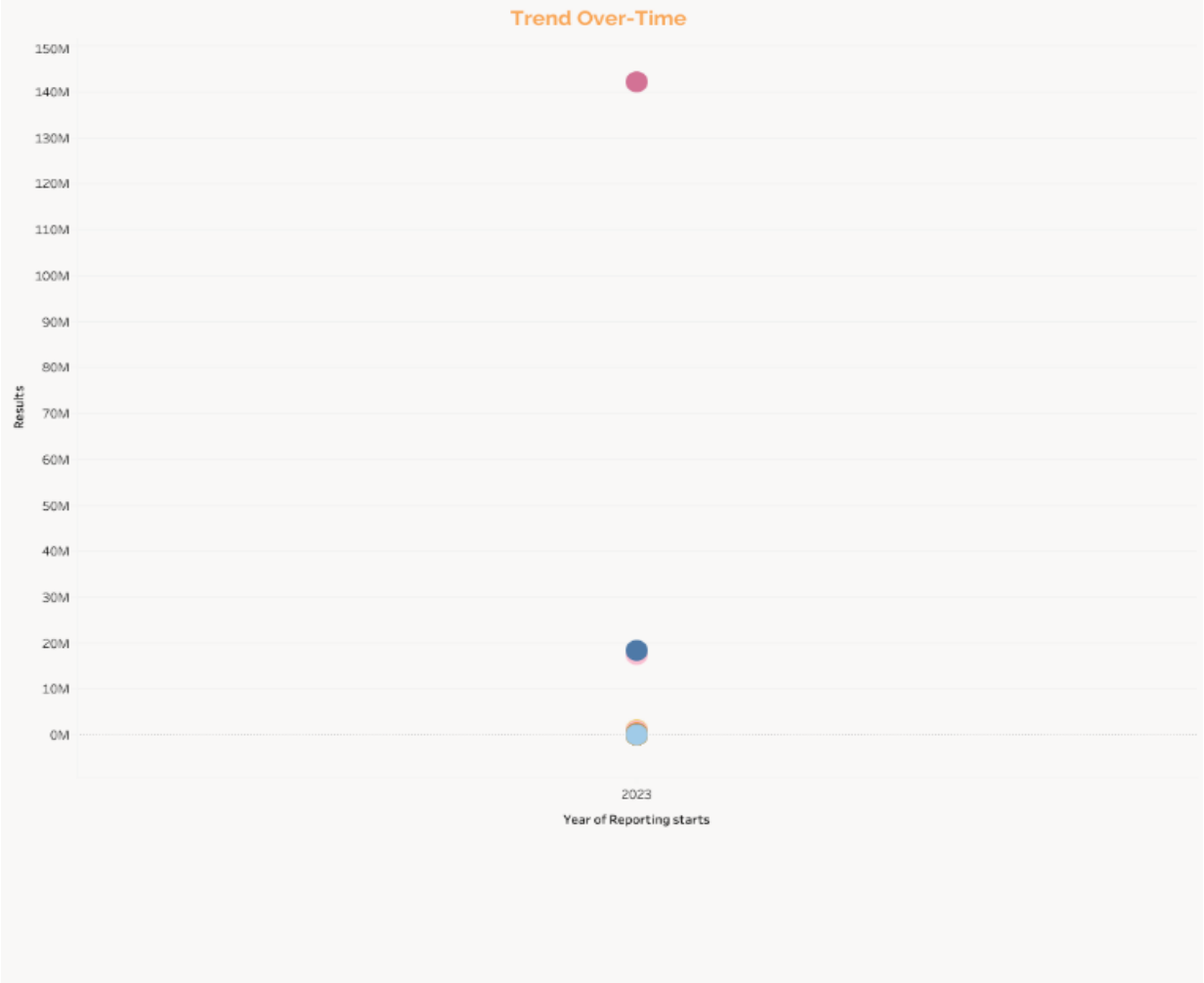
XIII. RESULT TYPE BREAKDOWN

- Proportions of different result types
-



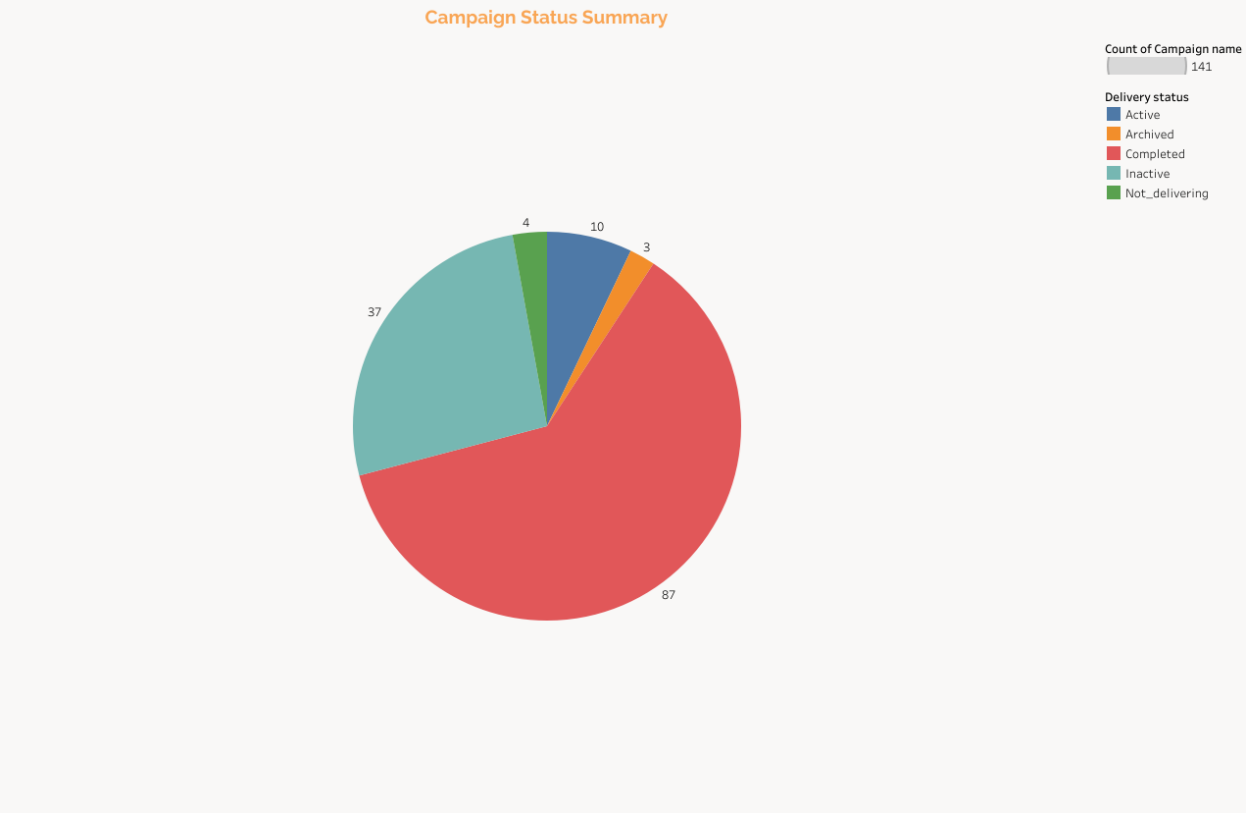
XIV. TREND OVERTIME

- Results by year



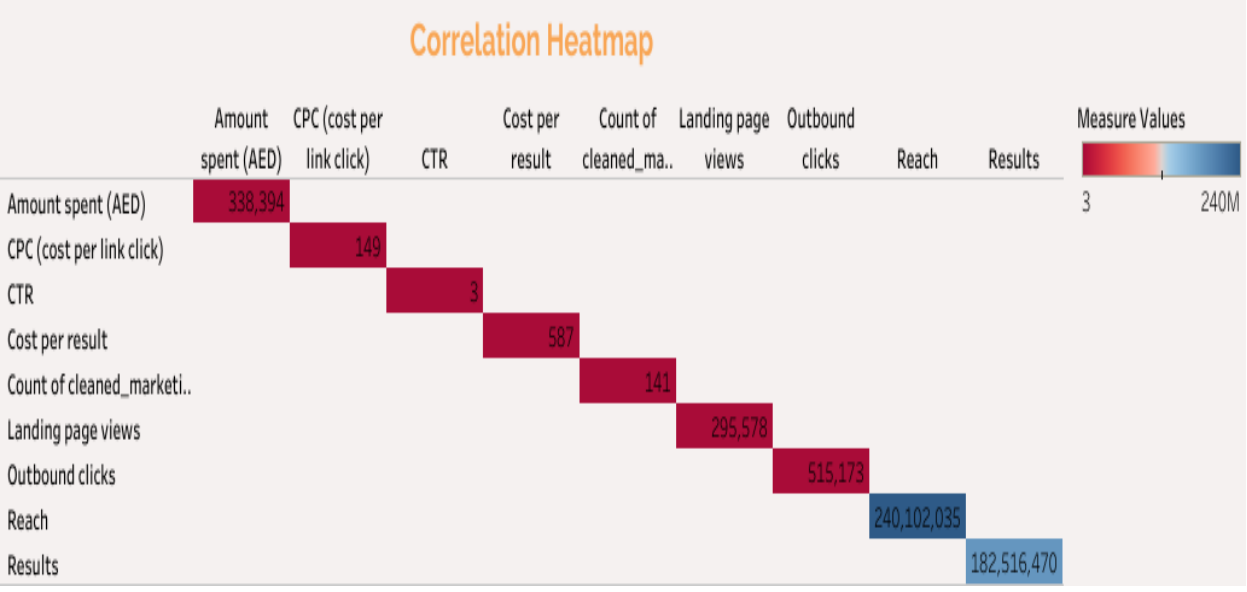
V. CAMPAIGN STATUS SUMMARY

- Distribution of campaign statuses: Completed, Inactive, Deleted, Learning, etc.
- Helps assess whether campaigns are actively running, paused, or retired.



V. **CORRELATION HEATMAP**

- This shows relationships between numeric variables, it is useful for spotting what influence my results the most.



KEY FINDINGS

- Efficiency varied significantly by campaign.
- Cost per Result spans AED 0.003 to AED 47.09, showing varying cost-efficiency across campaigns
- Results also range from 1 to 142 million, indicating a few extremely high-performing campaigns.

EDA REPORT ON LEARNER OPPORTUNITY DATA (LEARNER_OPPORTUNITY_RAW.CSV)

Overview of the Dataset:

No. of rows		113602				
No. of columns		5				
Columns	Name	enrollment_id	learner_id	assigned_cohort	apply_date	status
	Data type	object	object	object	object	int64
No. of Duplicate rows		0				

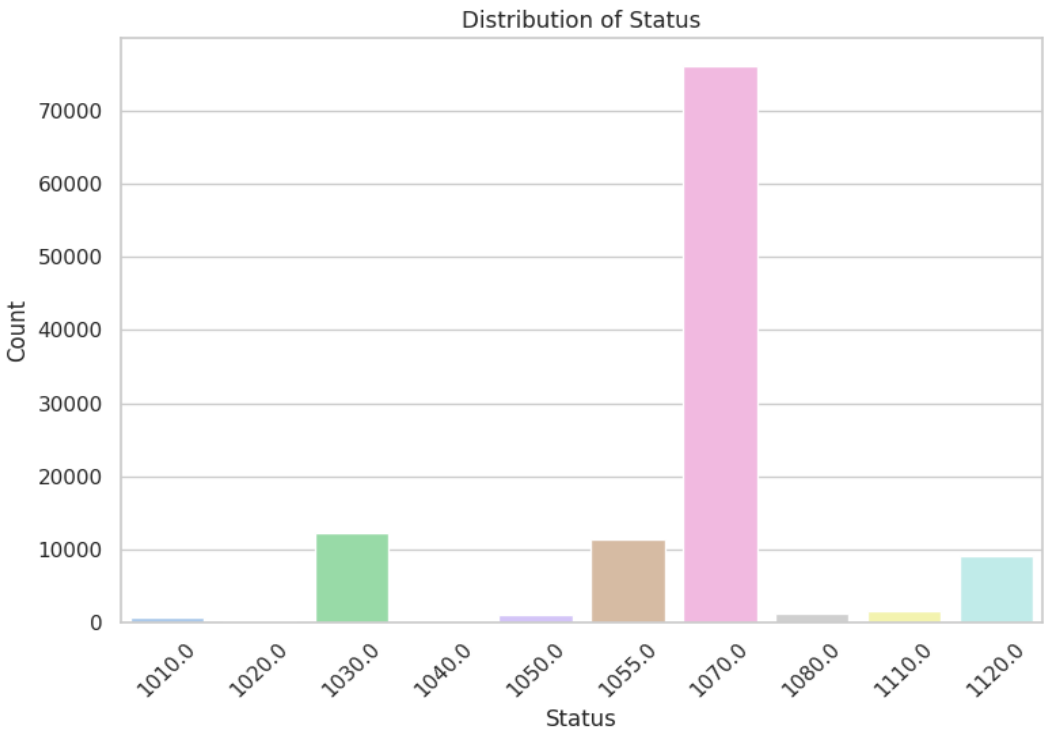
Statistical Summary:

Column	Missing values	Percentage of missing values (%)	Unique values	Duplicate values	Most frequent value	Frequency
enrollment_id	0	0	57966	55636	Opportunity#	186
learner_id	0	0	187	113415	Opportunity# 000000000G WQAXC5X 45C2MHJ28	10772
assigned_cohort	13318	11.72	575	113026	BAM6HBR	1805
apply_date	188	0.17	112623	978	2022-09-01 09:56:25.417 000+00:00	348
status	186	0.16	10	113591	1070	76109

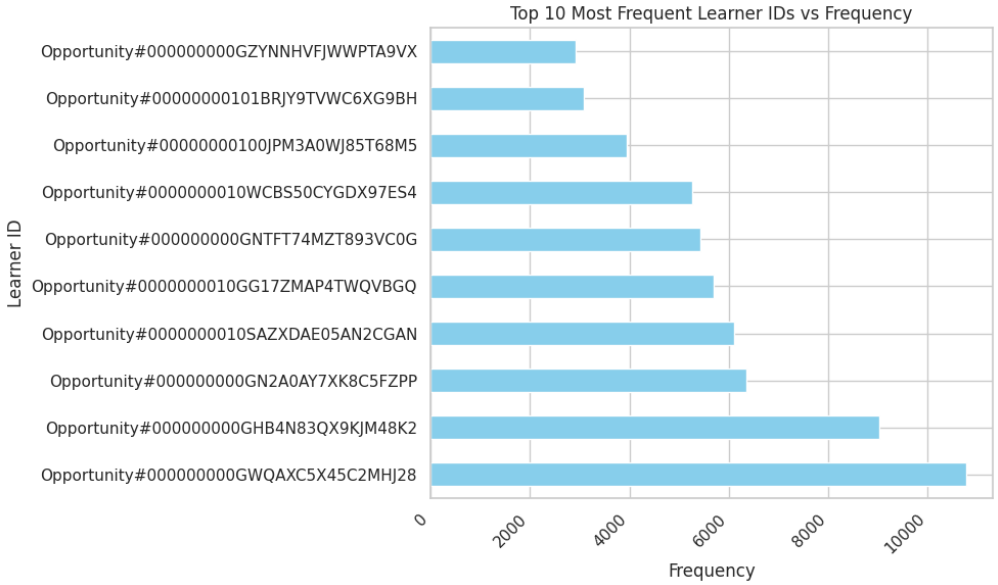
Visualization:

- **Distribution of status column:**

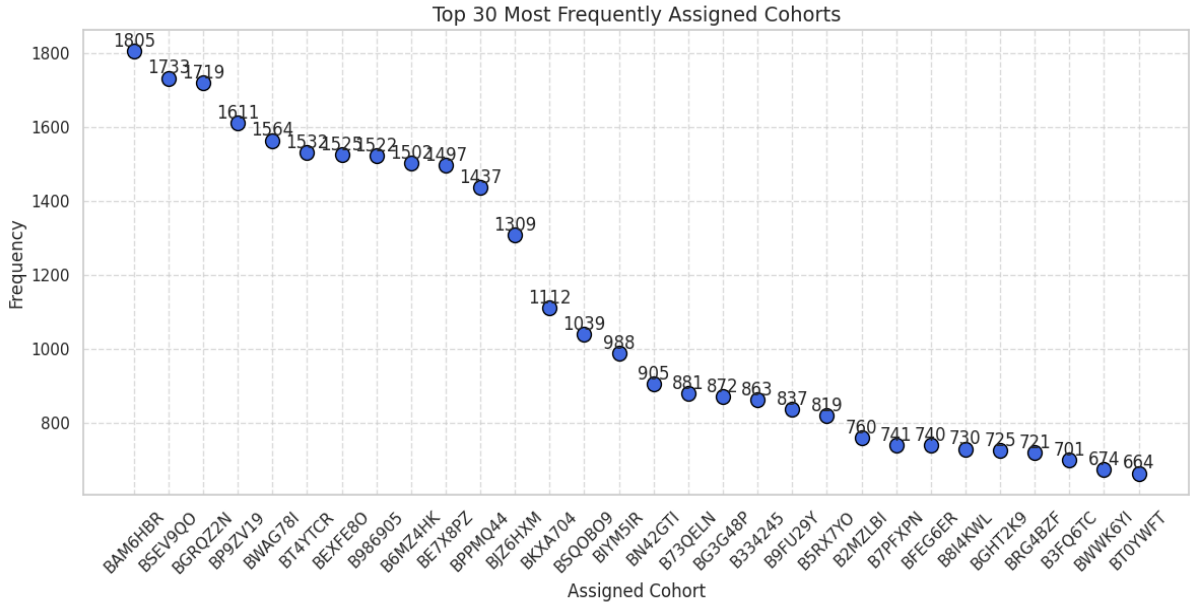
Most of the entries have a status of 1070 (Frequency – 76109). Other status values have frequency lower than 13000.



- **Top 10 most frequent learner_id:**

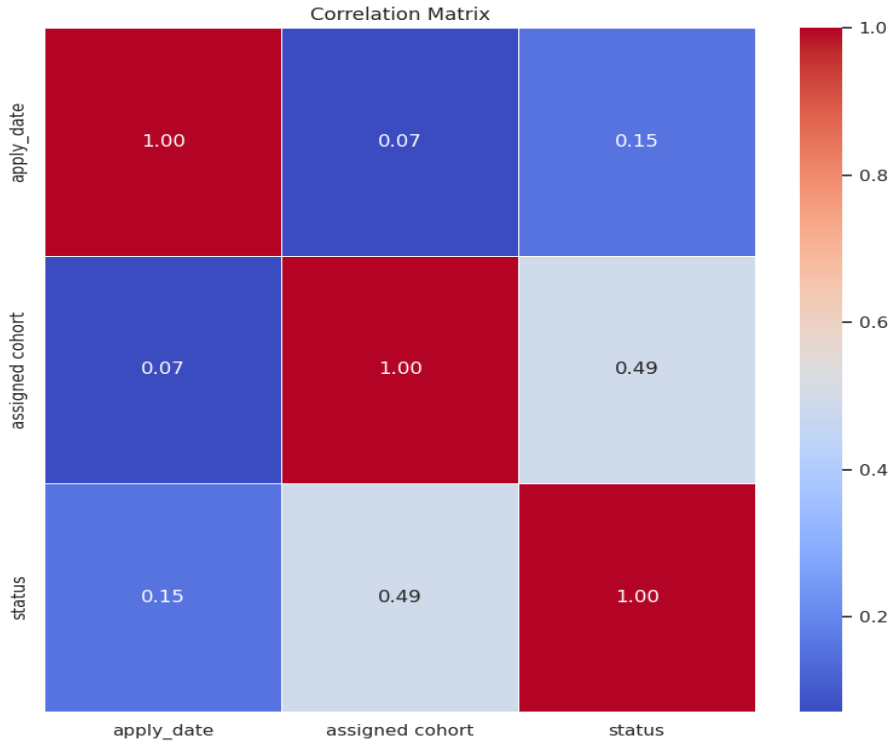


- **Top 30 most frequently assigned cohort:**



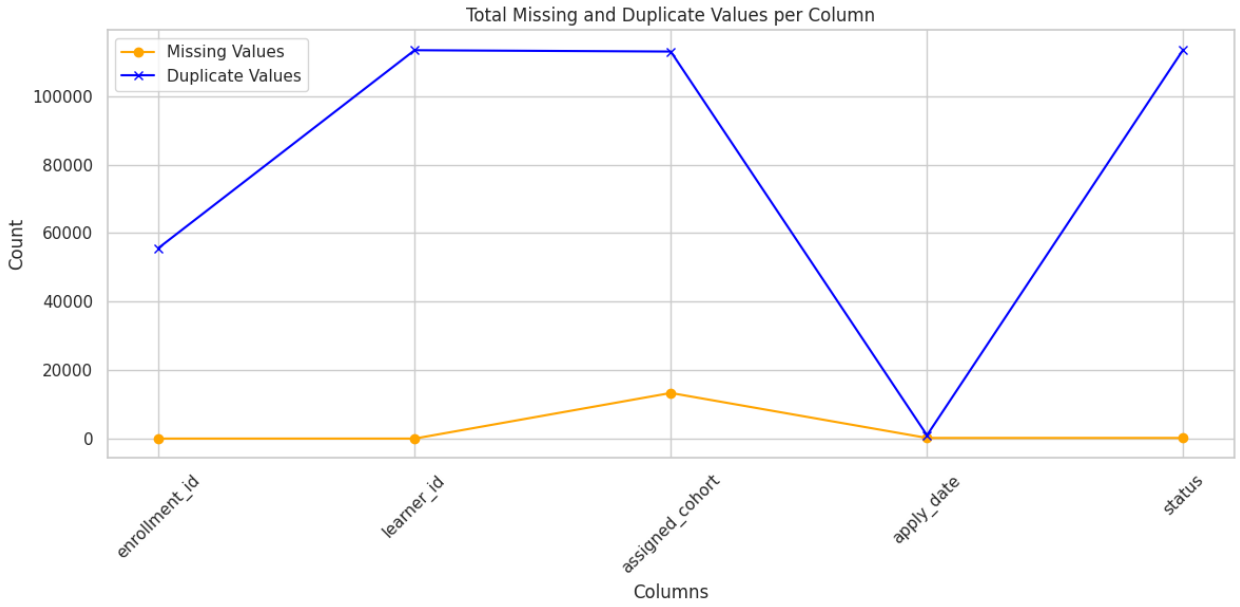
- **Correlation:**

It shows the correlation of apply_date, assigned_cohort & status column using a matrix. Assigned_cohort and status columns have correlation coefficients of 0.49. It means they both have a little influence on each other. However, the other columns have almost no influence on each other.



- **Missing and Duplicate values:**

Enrollment_id and learner_id don't have any missing values while apply_date and status have negligible number of missing values. Other than apply_date, all of the columns have a high number of duplicate values.



Key Findings:

- There are no duplicate rows.
- There is no outlier value in any column.
- There is no column with unique values and this dataset has a high number of duplicate values per column.
- The correlation among columns is very weak.
- Most of the entries have a status of 1070.

EDA REPORT ON COGNITO DATA (COGNITO_RAW.CSV)

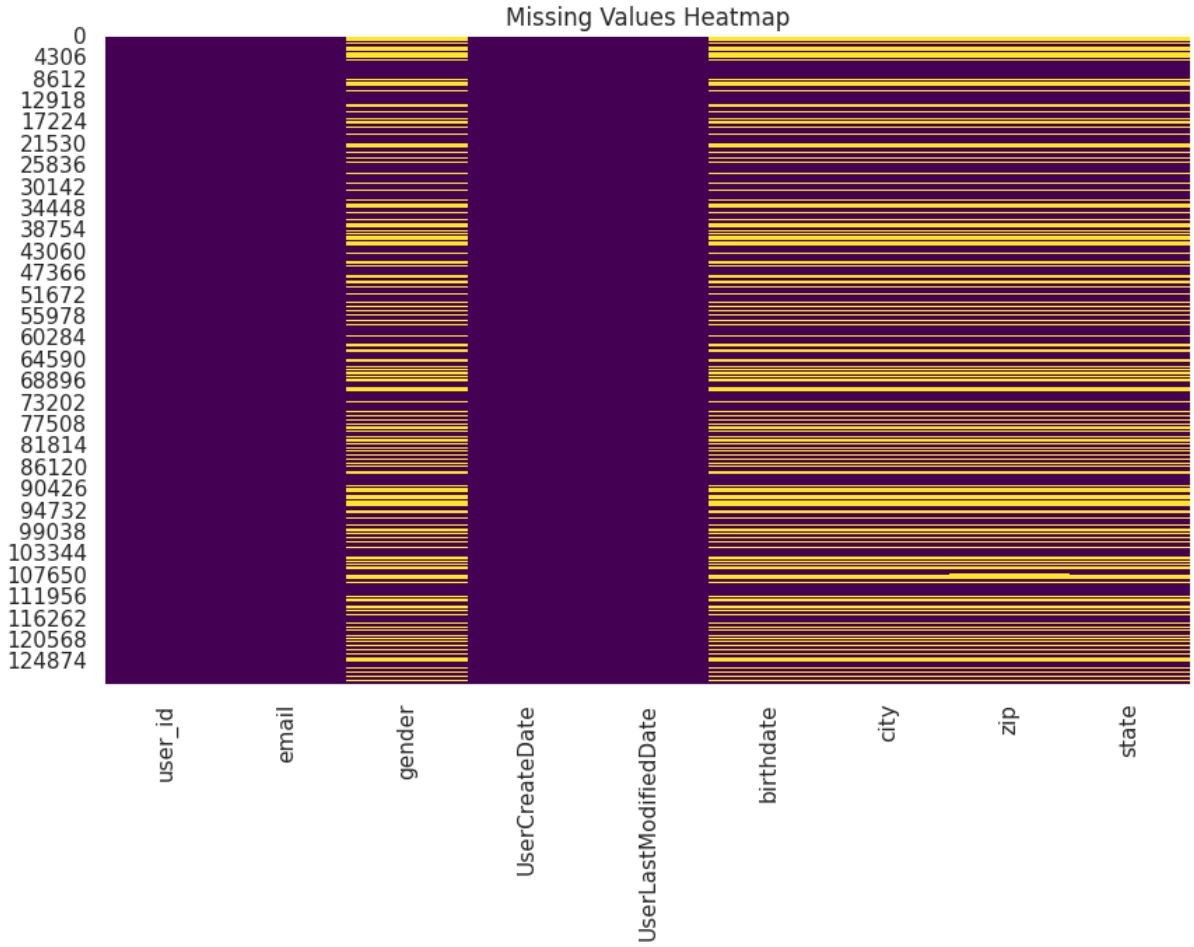
Features	Details
Rows	129178
Columns	9
Column Names	User_id, email, gender, UserCreateDate, UserLastModifiedDate, birthdate, city, zip, state.
Data Types	text

Summary Statistics

index	**count**	**unique**	**freq**
user_id	129178	129178	1
email	129178	129169	2
gender	86316	4	49344
UserCreateDate	129178	127424	14
UserLastModifiedDate	129178	129177	2
birthdate	86316	9729	115
city	86305	13430	3031
zip	86251	20375	2646
state	86154	6174	6154

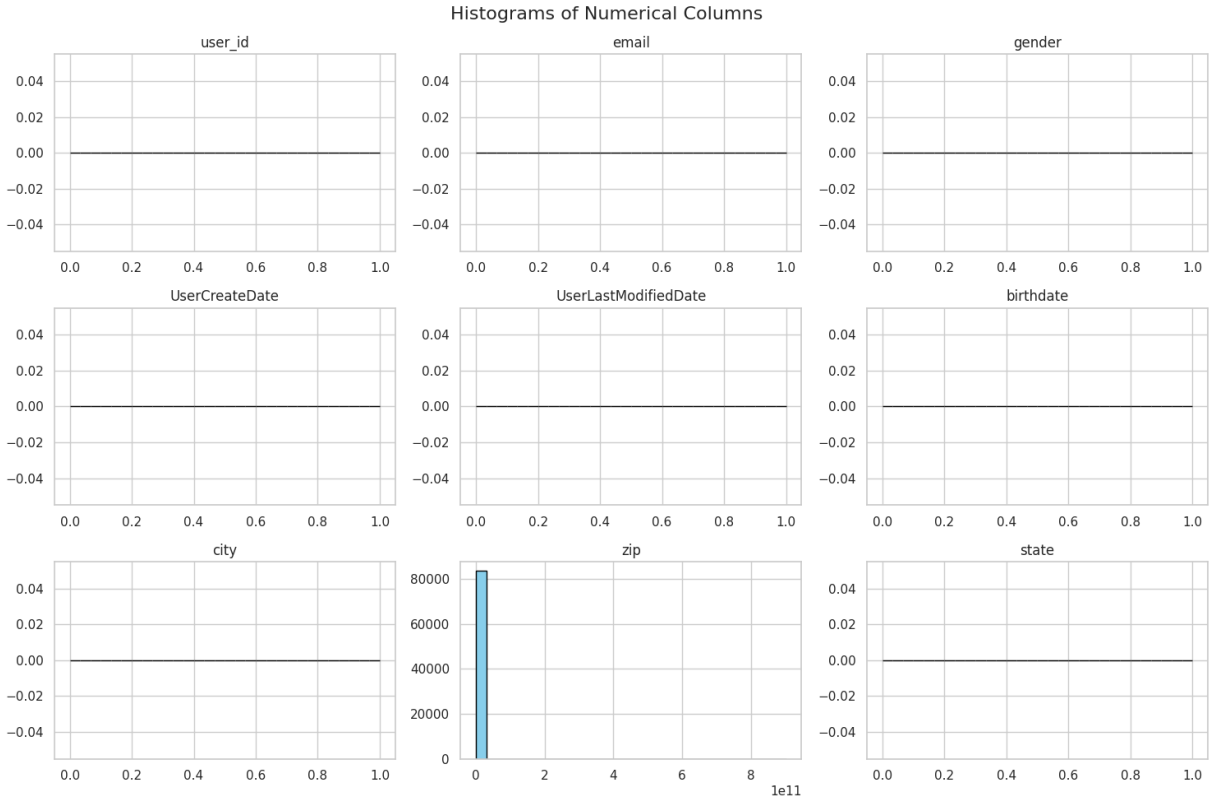
Missing Values, Duplicates & Inconsistencies

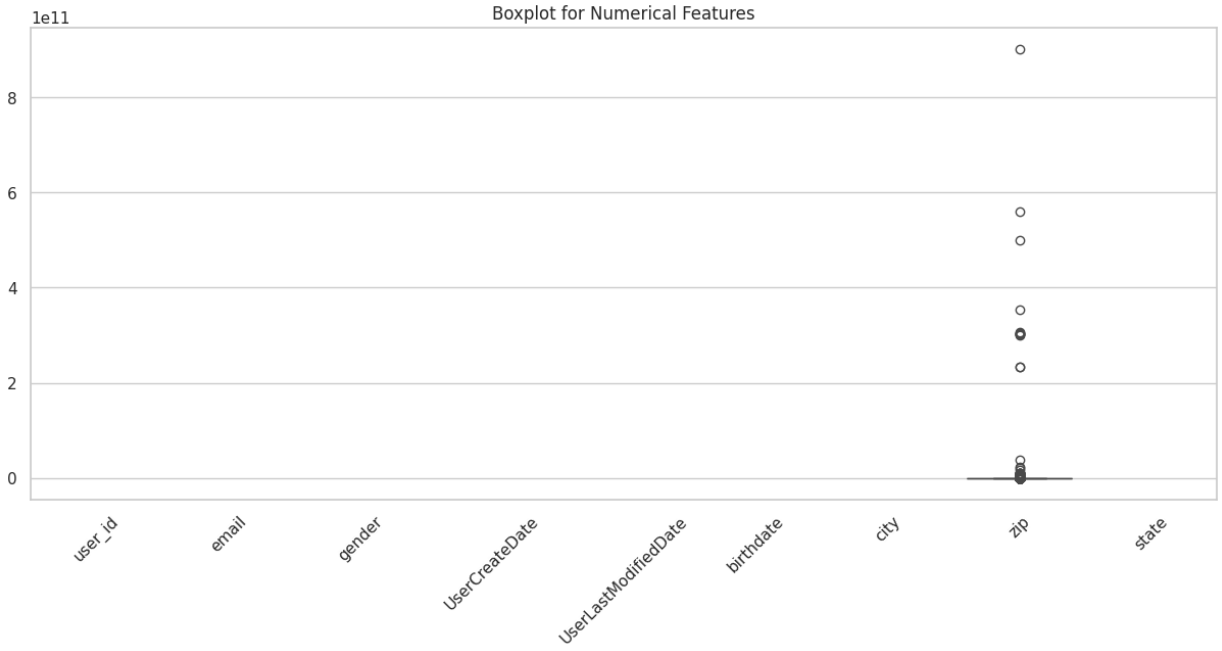
index	Missing count	Missing %	Duplicate count
user_id	0	0.0	0
email	0	0.0	9
gender	42862	33.18	5
UserCreateDate	0	0.0	0
UserLastModifiedDate	0	0.0	0
birthdate	42862	33.18	7666
city	42873	33.19	4726
zip	42927	33.23	7916
state	43024	33.31	2189

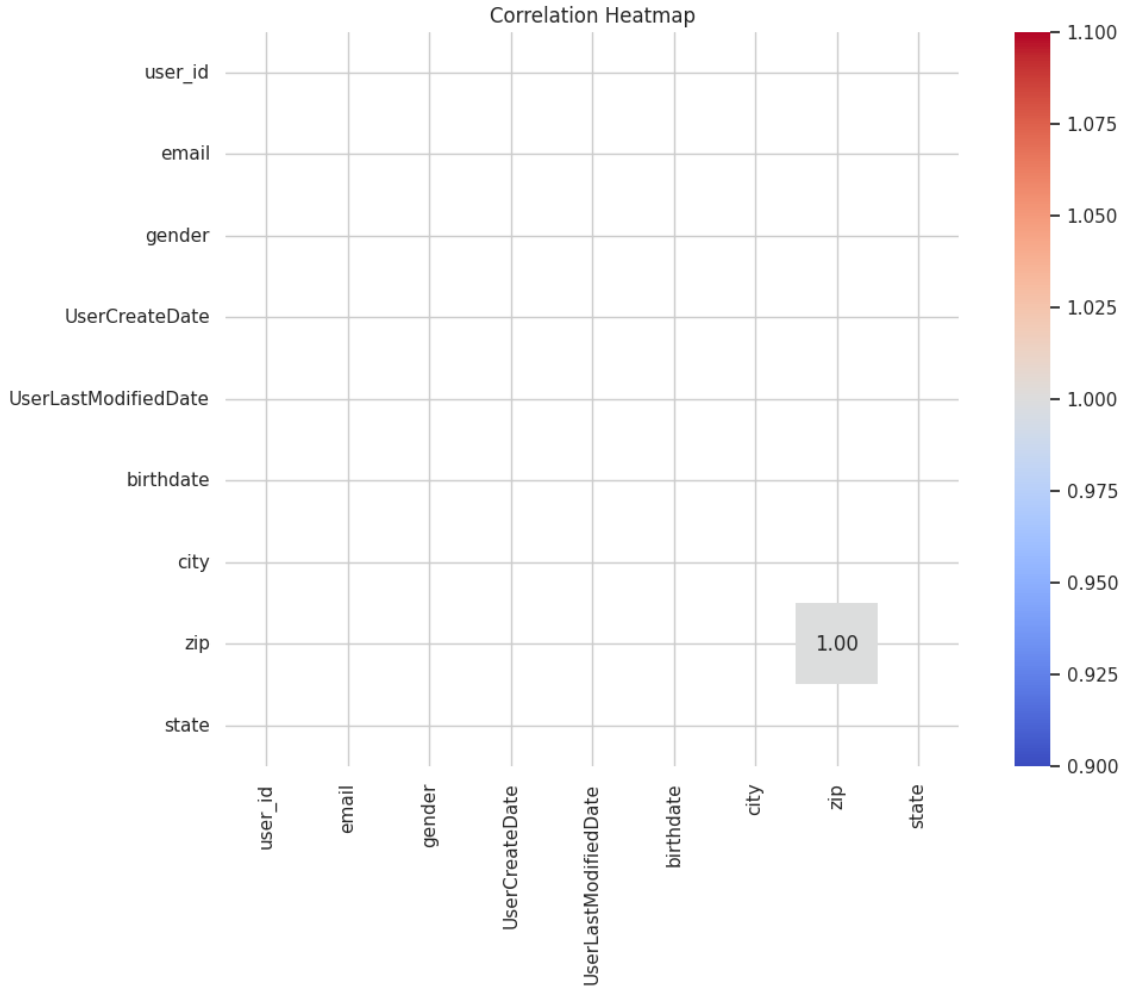


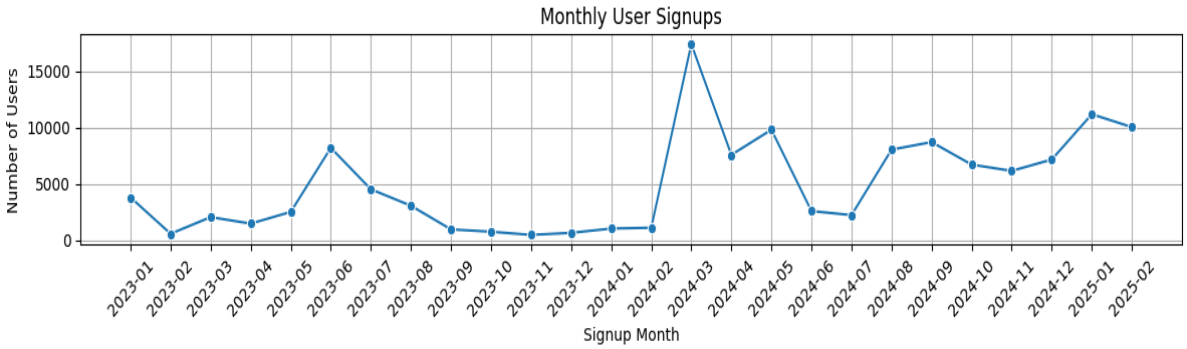
Number of duplicate rows: 0

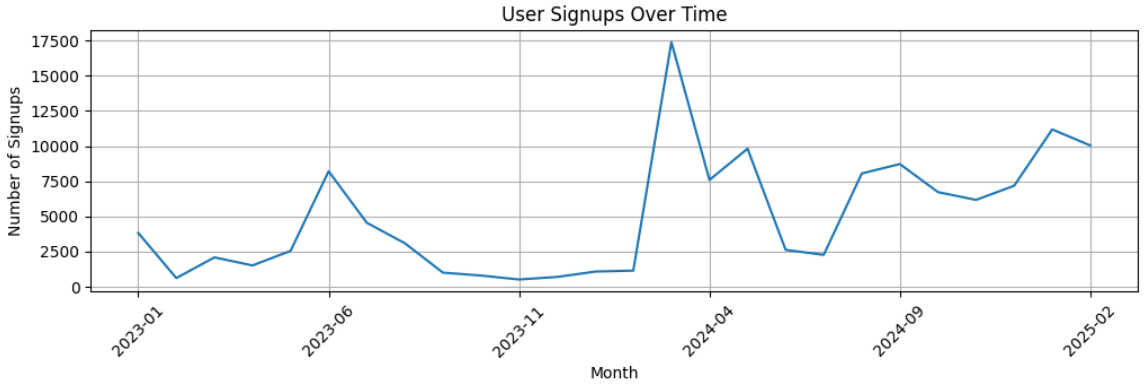
Data Visualization


Histograms of Numerical Columns
user_id
email
gender
UserCreateDate
UserLastModifiedDate
birthdate
city
zip
state
1e11


Boxplot for Numerical Features
1e11
user_id
email
gender
UserCreateDate
UserLastModifiedDate
birthdate
city
zip
state


Correlation Heatmap
user_id
email
gender
UserCreateDate
UserLastModifiedDate
birthdate
city
zip
state
1.00
1.100
1.075
1.050
1.025
1.000
0.975
0.950
0.925
0.900


Monthly User Signups
Number of Users
15000
10000
5000
0
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
2025-01
2025-02
Signup Month



Key Findings Summary

❖ Key EDA Findings:

➢ Total Rows and Columns: (129178, 9)

➢ Columns with missing values: ['user_id', 'email', 'gender', 'UserCreateDate', 'UserLastModifiedDate', 'birthdate', 'city', 'zip', 'state']

➢ Duplicate values present: True

➢ Strongly correlated pairs (abs(correlation) > 0.75)

❖ Next Steps for Data Cleaning:

➢ Handle missing values (impute/drop based on domain knowledge).

➢ Remove duplicate rows.

➢ Address outliers (based on boxplots).

➢ Encode categorical variables if needed for ML.

➢ Normalize/standardize numerical features if necessary.