WEEK 1 - Data Understanding, EDA, and Cleaning

By

HRITURAJ SAHA

(SLU 0704 DVA | Team 15)

DATA VISUALIZATION ASSOCIATE EARLY INTERNSHIP

DURATION - 1 MONTH

APRIL - 2025

USER DATA REPORT

EXPLORATORY DATA ANALYSIS (EDA) REPORT

Dataset Name - Learner_Raw.csv

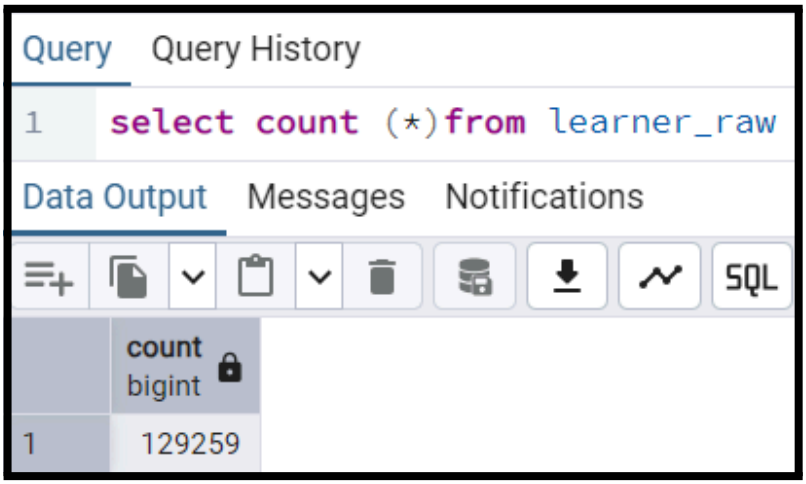
The dataset contains 129,259 learner records with 5 columns -

1. learner_id [Primary Key]
2. country
3. degree
4. institution
5. major

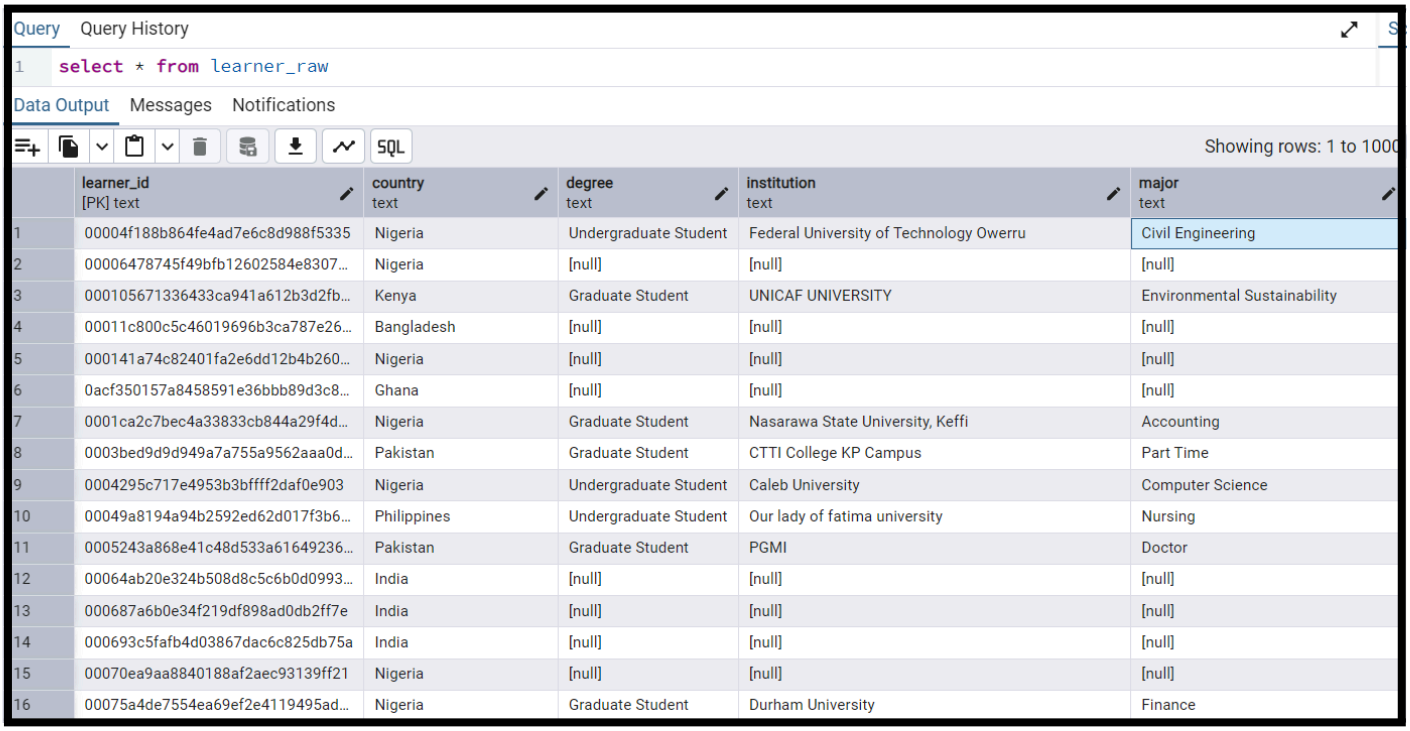
Created Table “learner_raw” in PostgreSQL and copied the original data into the table.

Initial Checks -

select count (*) from learner_raw



select * from learner_raw



	learner_id [PK] text	country text	degree text	institution text	major text
1	00004f188b864fe4ad7e6c8d988f5335	Nigeria	Undergraduate Student	Federal University of Technology Owerru	Civil Engineering
2	00006478745f49bfb12602584e8307...	Nigeria	[null]	[null]	[null]
3	000105671336433ca941a612b3d2fb...	Kenya	Graduate Student	UNICAF UNIVERSITY	Environmental Sustainability
4	00011c800c5c46019696b3ca787e26...	Bangladesh	[null]	[null]	[null]
5	000141a74c82401fa2e6dd12b4b260...	Nigeria	[null]	[null]	[null]
6	0acf350157a8458591e36bbb89d3c8...	Ghana	[null]	[null]	[null]
7	0001ca2c7bec4a33833cb844a29f4d...	Nigeria	Graduate Student	Nasarawa State University, Keffi	Accounting
8	0003bed9d9d949a7a755a9562aaa0d...	Pakistan	Graduate Student	CTTI College KP Campus	Part Time
9	0004295c717e4953b3bffff2daf0e903	Nigeria	Undergraduate Student	Caleb University	Computer Science
10	00049a8194a94b2592ed62d017f3b6...	Philippines	Undergraduate Student	Our lady of fatima university	Nursing
11	0005243a868e41c48d533a61649236...	Pakistan	Graduate Student	PGMI	Doctor
12	00064ab20e324b508d8c5c6b0d0993...	India	[null]	[null]	[null]
13	000687a6b0e34f219df898ad0db2ff7e	India	[null]	[null]	[null]
14	000693c5fafb4d03867dac6c825db75a	India	[null]	[null]	[null]
15	00070ea9aa8840188af2aec93139ff21	Nigeria	[null]	[null]	[null]
16	00075a4de7554ea69ef2e4119495ad...	Nigeria	Graduate Student	Durham University	Finance

DATA CLEANING AND VALIDATION

Missing Data -

- learner_id - 0, 0.000000%
- country - 2,275, 1.760032%
- degree - 52,693, 40.765440%
- institution - 53,073, 41.059423%
- major - 52,871, 40.903148%

Total Missing Values - 160912.

Top 10 Countries by Learner Count -

1. India - 33868
2. Nigeria - 30696
3. Pakistan - 14112
4. Kenya - 8246
5. United States - 7064
6. Ghana - 6874
7. Philippines - 6039
8. Egypt - 5301
9. South Africa - 4587
10. Bangladesh - 3525

Top 10 Degrees -

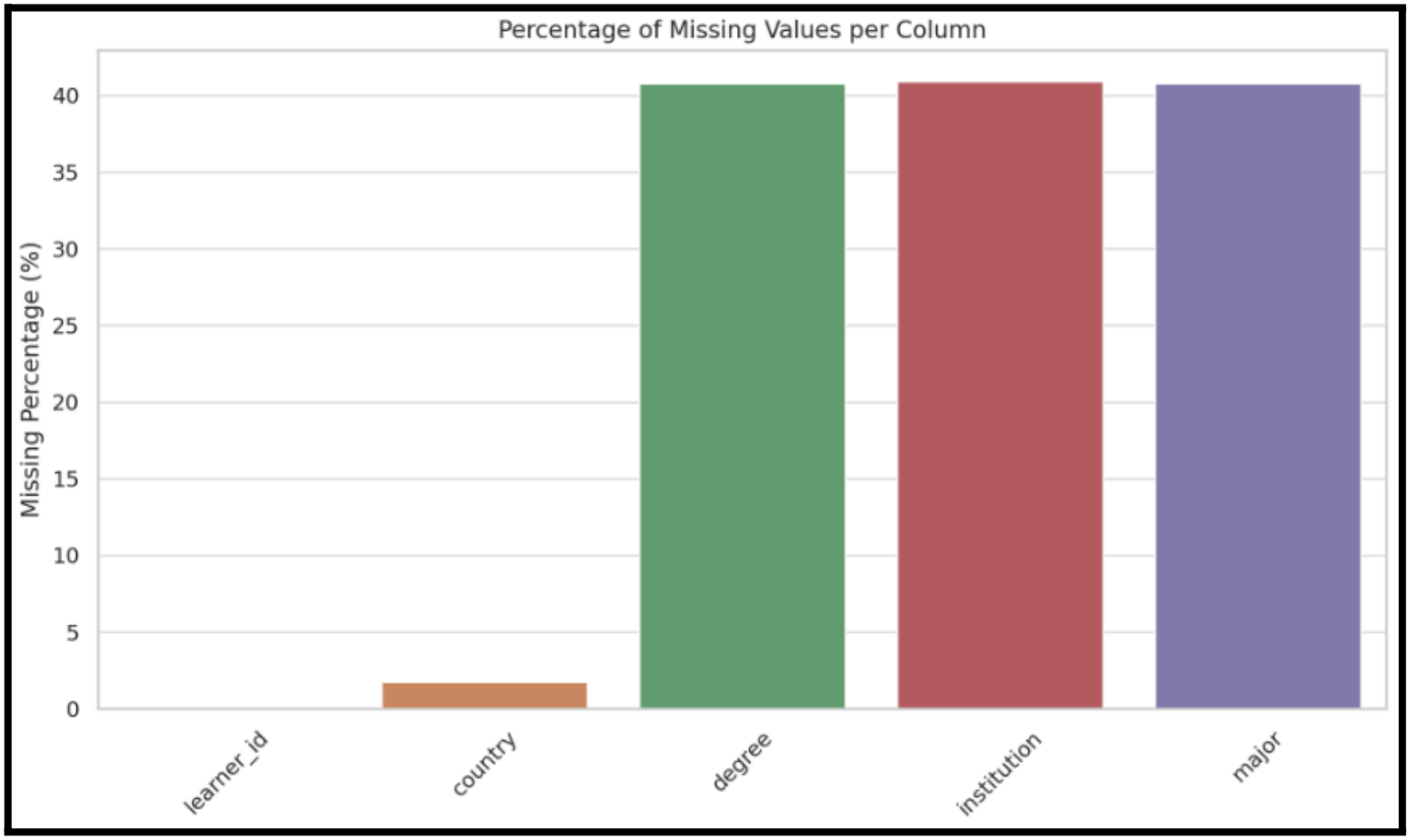
1. Graduate Student - 31806
2. Undergraduate Student - 30709
3. Not in Education - 6319
4. High School Student - 4109
5. Other Professional - 2997
6. Teacher/Educator - 562
7. Parent of Student - 64

Top 10 Institutions -

1. Saint Louis University - 2163
2. University of Lagos - 605
3. Illinois Institute of Technology - 553
4. University of Ghana - 524
5. saint louis university - 498
6. University of Ibadan - 461
7. University of Ilorin - 425
8. Kwame Nkrumah University of Science and Technology - 410
9. University of Benin - 400
10. Obafemi Awolowo University - 350

Top 10 Majors -

1. Computer Science - 4704
2. Business Administration - 1679
3. Computer Science and Engineering - 1653
4. Accounting and Finance - 1589
5. Data Science - 1483
6. Data Analytics - 1443
7. Bachelor of Science in Computer Science - 1434
8. Mechanical Engineering - 1361
9. Computer Engineering - 1356
10. Information Systems - 1339



This dashboard is created using Python libraries like Matplotlib & Seaborn in Google Colab.

Key Findings-

1. Countries - India (26.2%), Nigeria (23.7%) are dominant.
2. Degrees - Graduate and undergraduate students make up the majority.
3. Institutions - Saint Louis University appears twice with different cases (data normalization needed).
4. Majors - Strong tech focus – Computer Science and related fields are most common.

OPPORTUNITY DATA REPORT

EXPLORATORY DATA ANALYSIS (EDA) REPORT

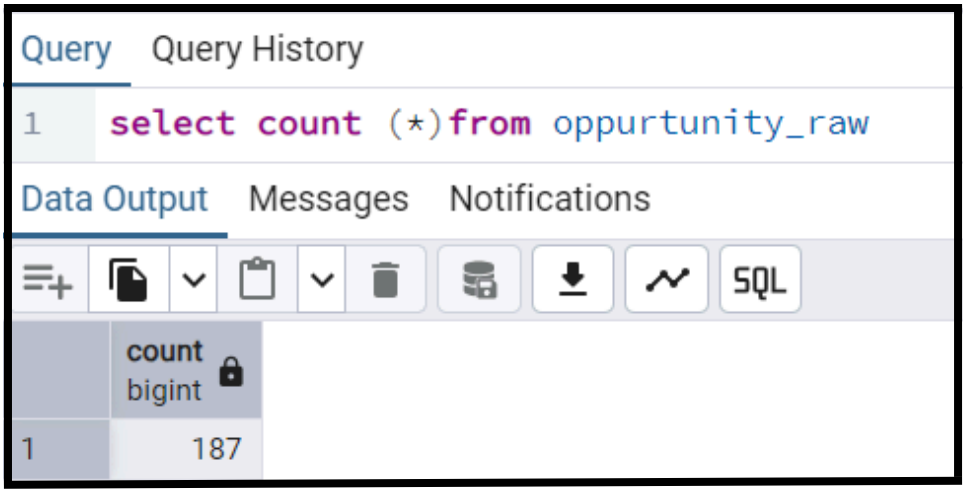
Dataset Name - Opportunity_Raw.csv

- Total Records - 187
- Columns - 5
  - opportunity_id [Primary Key]
  - opportunity_name
  - category
  - opportunity_code
  - tracking_questions

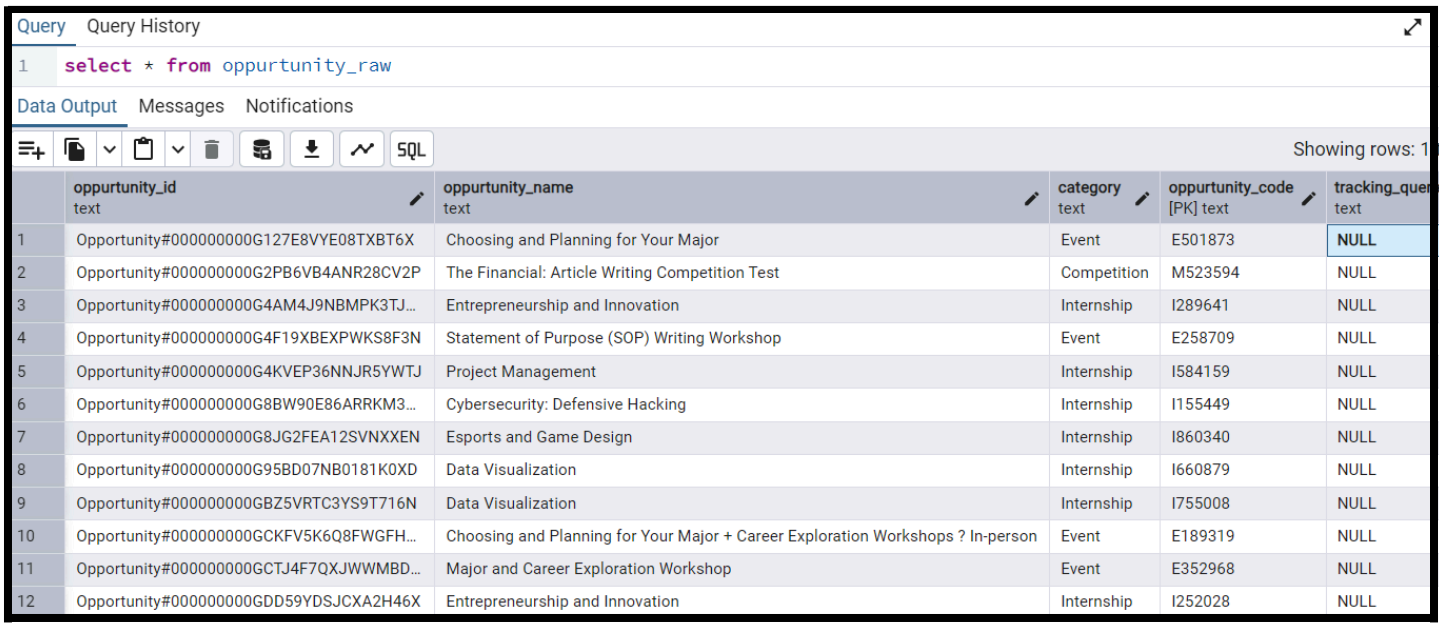
Created Table “oppurtunity_raw” in PostgreSQL and copied the original data into the table.

Initial Checks -

select count (*) from oppurtunity_raw



select * from oppurtunity_raw



	oppurtunity_id text	oppurtunity_name text	category text	oppurtunity_code [PK] text	tracking_quer text
1	Opportunity#000000000G127E8VYE08TXBT6X	Choosing and Planning for Your Major	Event	E501873	NULL
2	Opportunity#000000000G2PB6VB4ANR28CV2P	The Financial: Article Writing Competition Test	Competition	M523594	NULL
3	Opportunity#000000000G4AM4J9NBMPK3TJ...	Entrepreneurship and Innovation	Internship	I289641	NULL
4	Opportunity#000000000G4F19XBEXPWKS8F3N	Statement of Purpose (SOP) Writing Workshop	Event	E258709	NULL
5	Opportunity#000000000G4KVEP36NNJR5YWTJ	Project Management	Internship	I584159	NULL
6	Opportunity#000000000G8BW90E86ARRKM3...	Cybersecurity: Defensive Hacking	Internship	I155449	NULL
7	Opportunity#000000000G8JG2FEA12SVNXXEN	Esports and Game Design	Internship	I860340	NULL
8	Opportunity#000000000G95BD07NB0181K0XD	Data Visualization	Internship	I660879	NULL
9	Opportunity#000000000GBZ5VRTC3YS9T716N	Data Visualization	Internship	I755008	NULL
10	Opportunity#000000000GCKFV5K6Q8FWGFH...	Choosing and Planning for Your Major + Career Exploration Workshops ? In-person	Event	E189319	NULL
11	Opportunity#000000000GCTJ4F7QXJWWMBD...	Major and Career Exploration Workshop	Event	E352968	NULL
12	Opportunity#000000000GDD59YDSJCXA2H46X	Entrepreneurship and Innovation	Internship	I252028	NULL

DATA CLEANING AND VALIDATION

Missing Data Summary -

- opportunity_id - 0, 0.00%
- opportunity_name - 0, 0.00%
- category - 0, 0.00%
- opportunity_code - 0, 0.00%
- tracking_questions - 69, 36.898396%

Category Distribution -

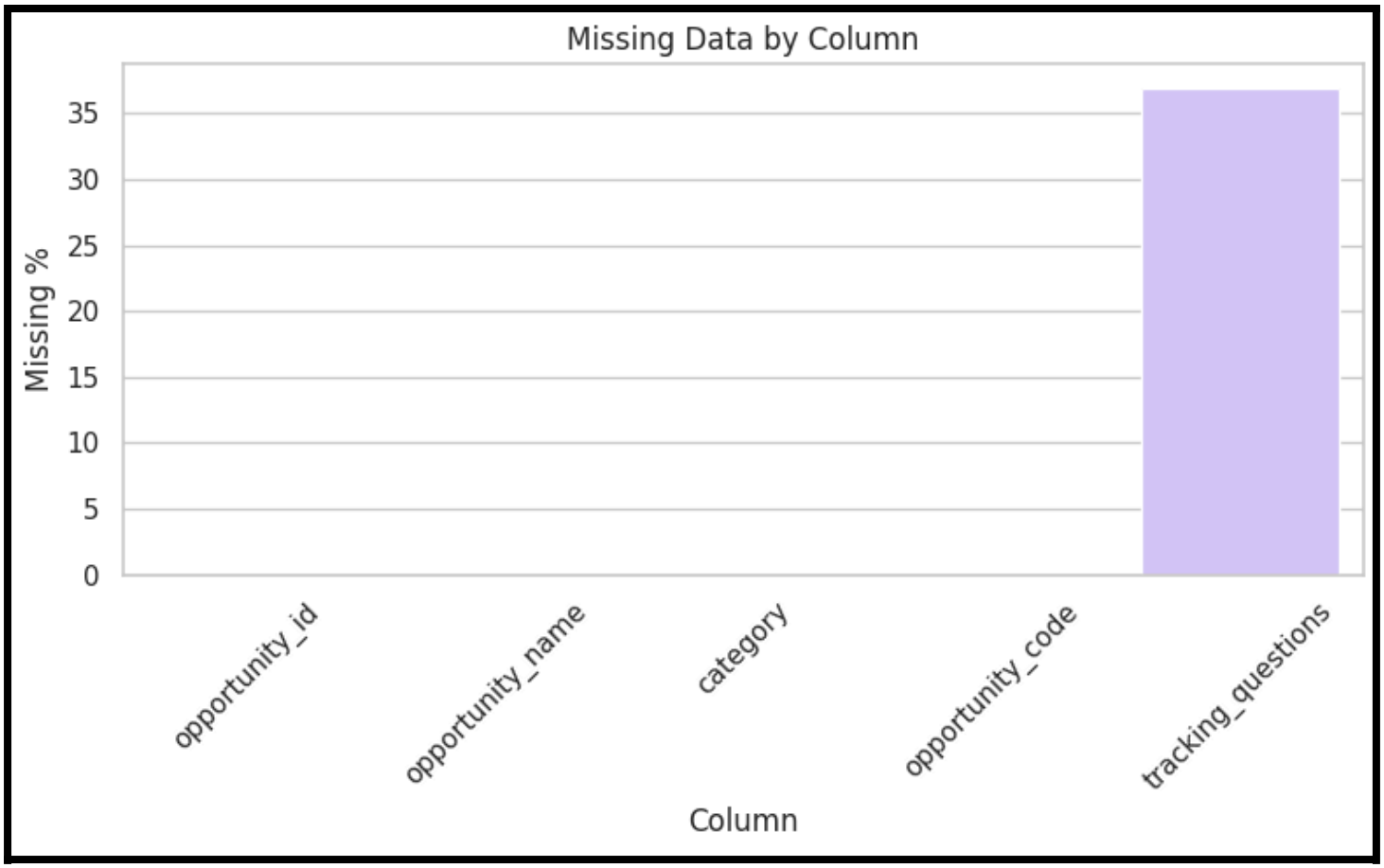
- Internship - 43
- Event - 41
- Competition - 41
- Career **-** 23
- Course - 18
- Masterclass - 11
- Engagement - 10

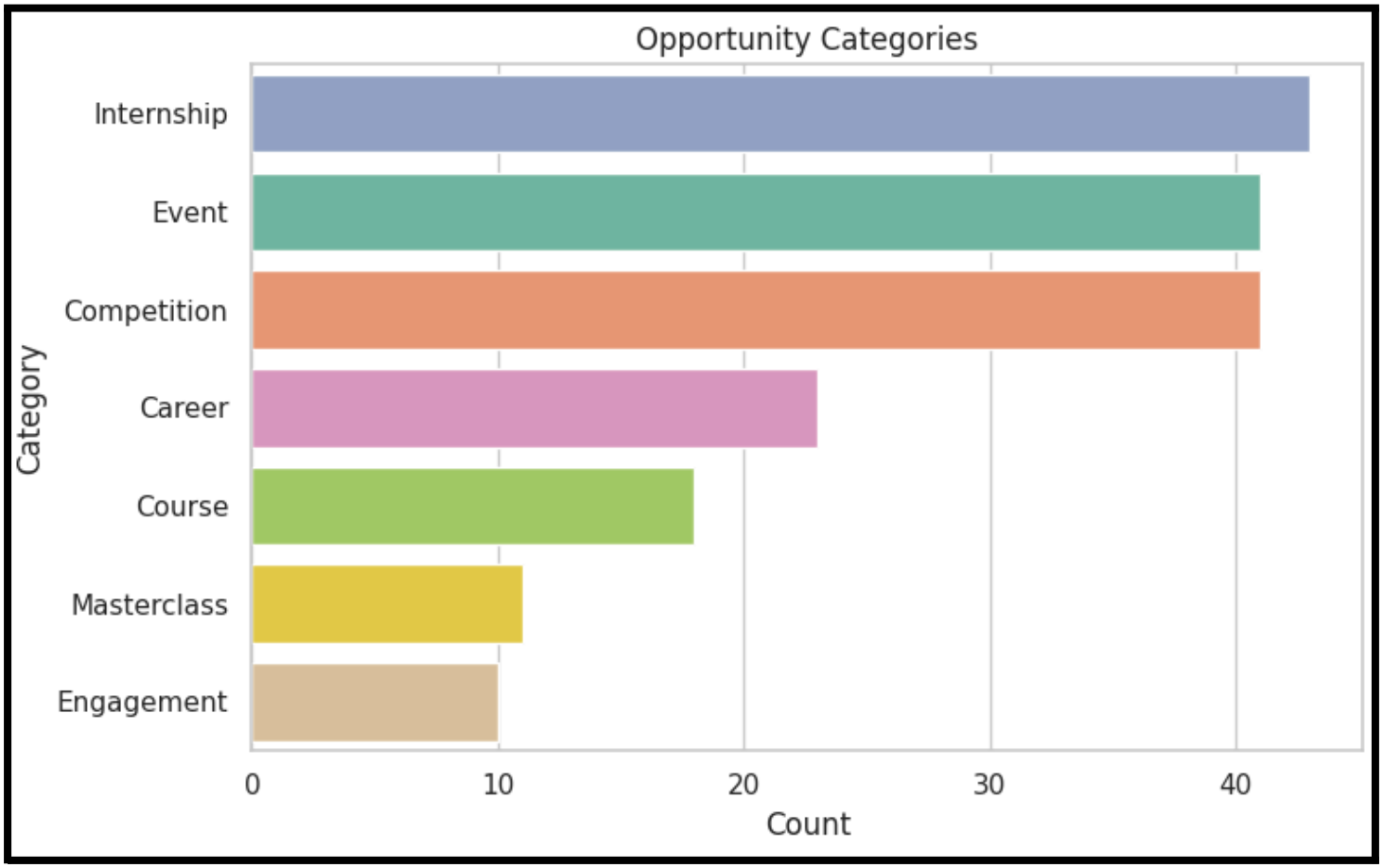
Uniqueness Check -

- opportunity_id is unique (187/187 are distinct)
- opportunity_code is also unique
- opportunity_name has some recurring patterns (e.g., names starting with "The Financial", "Entrepreneurship", etc.)

Key Findings -

- The high percentage of missing tracking_questions could indicate optional or inconsistently recorded data.
- Naming patterns in opportunity_name hint at repetition or templated naming conventions.
- Opportunities are skewed toward experiential offerings like internships and competitions, which may reflect platform or audience priorities.





These dashboards are created using Python libraries like Matplotlib & Seaborn in Google Colab.

COHORT DATA REPORT

EXPLORATORY DATA ANALYSIS (EDA) REPORT

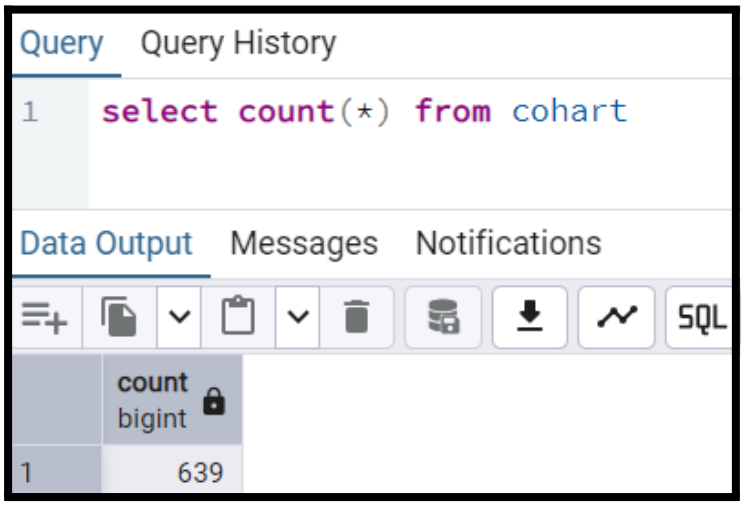
Dataset Name - CohortRaw.csv

- Rows - 639
- Columns - 5
  - cohort_id
  - cohort_code [Primary Key]
  - start_date
  - end_date
  - size

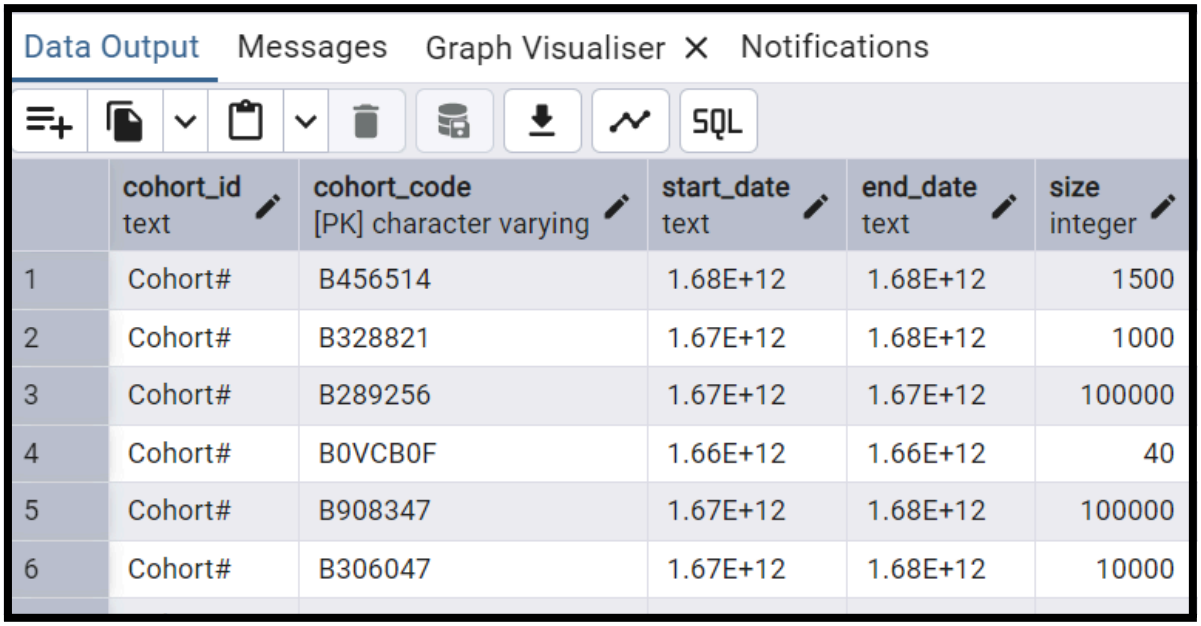
Created Table “cohart” in PostgreSQL and copied the original data into the table.

Initial Checks -

select count (*) from cohart



select * from cohart



	cohort_id text	cohort_code [PK] character varying	start_date text	end_date text	size integer
1	Cohort#	B456514	1.68E+12	1.68E+12	1500
2	Cohort#	B328821	1.67E+12	1.68E+12	1000
3	Cohort#	B289256	1.67E+12	1.67E+12	100000
4	Cohort#	B0VCB0F	1.66E+12	1.66E+12	40
5	Cohort#	B908347	1.67E+12	1.68E+12	100000
6	Cohort#	B306047	1.67E+12	1.68E+12	10000

DATA CLEANING AND VALIDATION

Key Statistics -

Feature	Mean	Std Dev	Min	25%	50%	75%	Max
Size	5,741	20,994	3	500	800	1500	100,000
Duration(d)	56	108	0	29	29	35	1,096

Explanation -

Size -

- Mean (Average) = 5,741 → On average, each cohort has about 5,741 individuals.
- Standard Deviation (Std Dev) = 20,994 → High variability, meaning cohort sizes vary a lot.
- Minimum (Min) = 3 → Smallest cohort has only 3 people.
- 25th Percentile (Q1) = 500 → 25% of the cohorts have fewer than 500 people.
- Median (Q2 or 50%) = 800 → Half of the cohorts have fewer than 800 people.
- 75th Percentile (Q3) = 1,500 → 75% of cohorts have fewer than 1,500 people.
- Maximum (Max) = 100,000 → Some cohorts are extremely large (possibly outliers or special cases).

Duration-

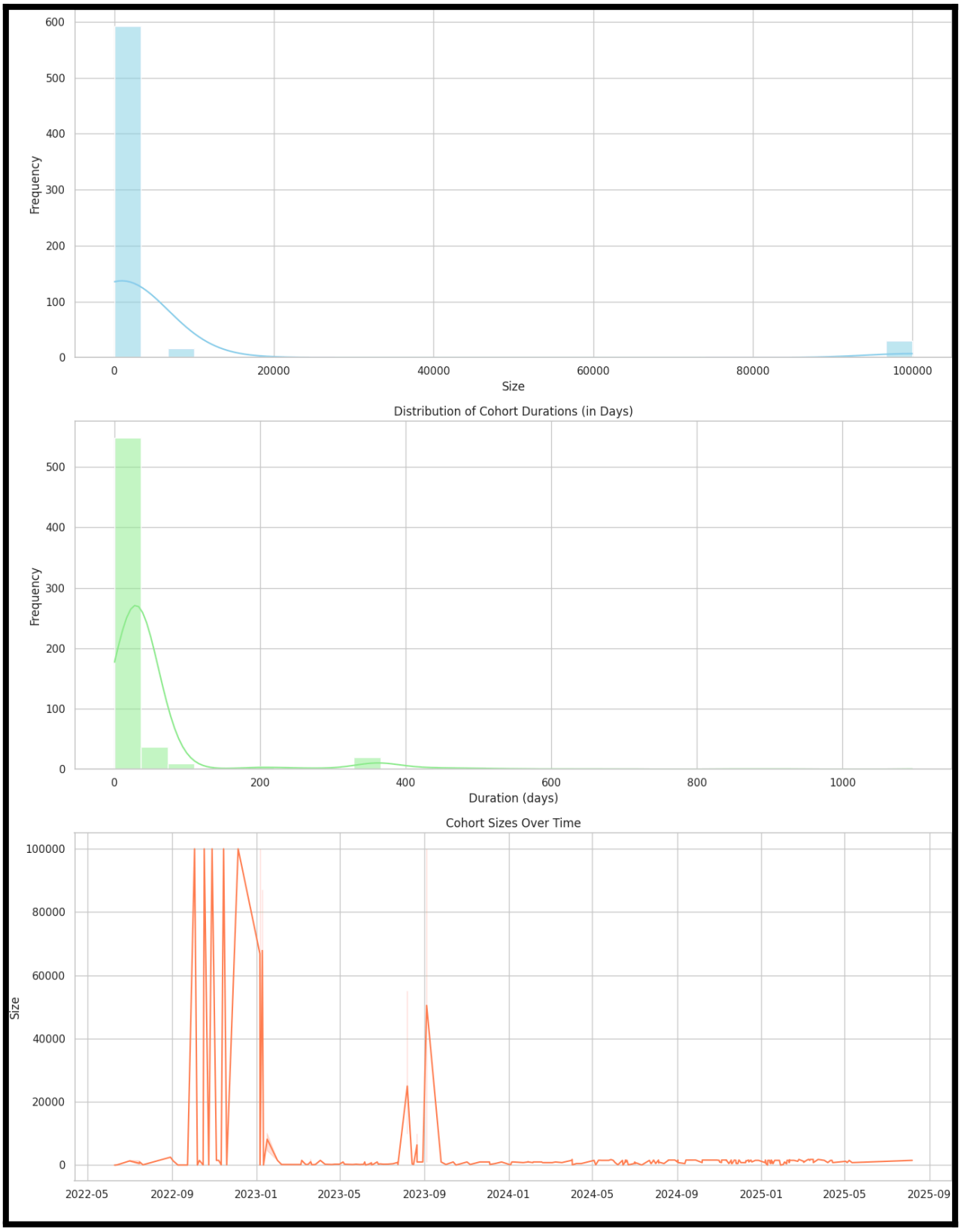
- Mean = 56 days → On average, cohorts last for about 2 months.
- Std Dev = 108 days → Large variation in how long cohorts last.
- Min = 0 days → Some cohorts have the same start and end date (maybe an error or 1-day event).
- 25th Percentile = 29 days
- Median = 29 days → 50% of cohorts last less than or equal to a month.
- 75th Percentile = 35 days → 75% are under 5 weeks.
- Max = 1,096 days → Longest cohort spans 3 years, clearly an outlier.

Missing Data Summary -

- cohort_id - 0, 0.00%
- cohort_code - 0, 0.00%
- start_date - 0, 0.00%
- end_date - 0, 0.00%
- size - 0, 0.00%

Key Findings -

- Cohort Size Distribution
  - Positively skewed - Most cohorts have sizes under 2000, but a few go up to 100,000.
- Cohort Duration Distribution
  - Majority have a standard duration around 29–35 days.
  - A few cohorts last over a year (max ~ 3 years).
- Cohort Sizes Over Time
  - No strong trend, but some bursts of large cohorts observed periodically.



These dashboards are created using Python libraries like Matplotlib & Seaborn in Google Colab.

LEARNER OPPORTUNITY DATA REPORT

EXPLORATORY DATA ANALYSIS (EDA) REPORT

Dataset Name - LearnerOpportunity_Raw.csv

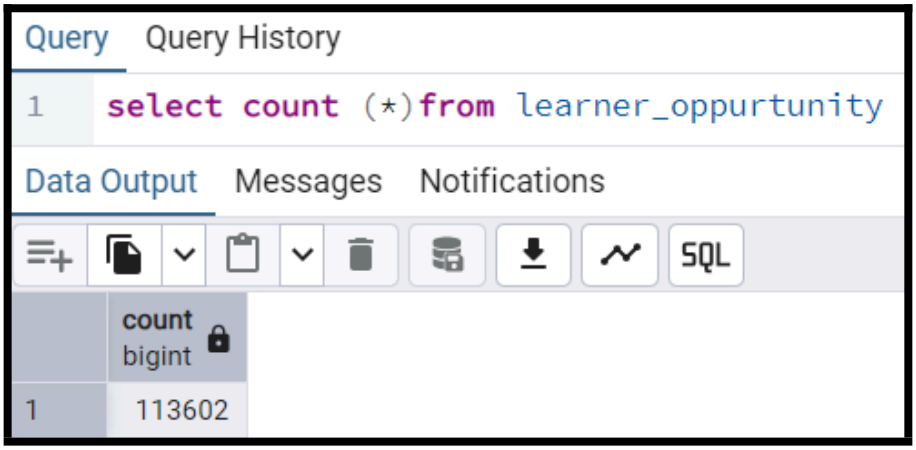
- Total Records - 113,602
- Total Features (Columns) - 5

1. enrollment_id [Primary Key]
2. learner_id
3. assigned_cohort
4. apply_date
5. status

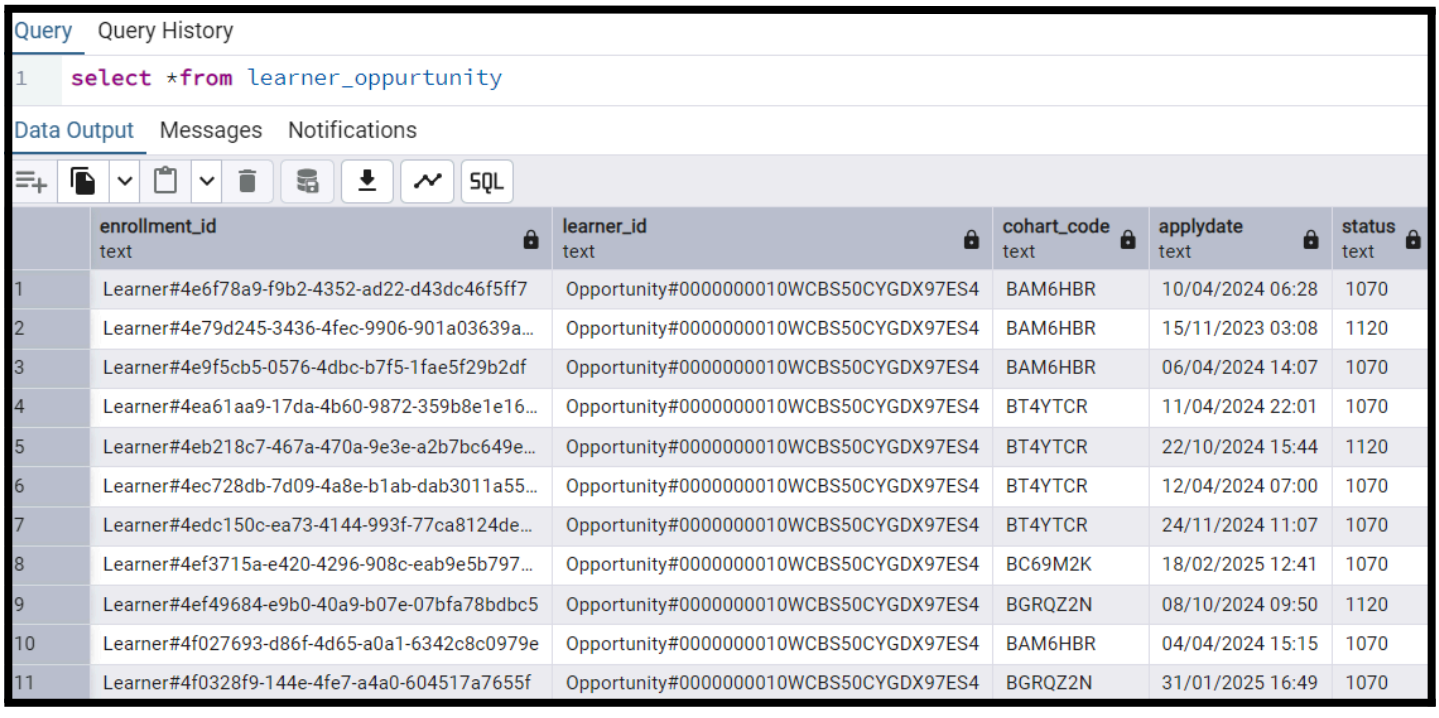
Created Table “learner_oppurtunity” in PostgreSQL and copied the original data into the table.

Initial Checks -

select count (*) from learner_oppurtunity



select * from learner_oppurtunity



	enrollment_id text	learner_id text	cohart_code text	applydate text	status text
1	Learner#4e6f78a9-f9b2-4352-ad22-d43dc46f5ff7	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	10/04/2024 06:28	1070
2	Learner#4e79d245-3436-4fec-9906-901a03639a...	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	15/11/2023 03:08	1120
3	Learner#4e9f5cb5-0576-4dbc-b7f5-1fae5f29b2df	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	06/04/2024 14:07	1070
4	Learner#4ea61aa9-17da-4b60-9872-359b8e1e16...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	11/04/2024 22:01	1070
5	Learner#4eb218c7-467a-470a-9e3e-a2b7bc649e...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	22/10/2024 15:44	1120
6	Learner#4ec728db-7d09-4a8e-b1ab-dab3011a55...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	12/04/2024 07:00	1070
7	Learner#4edc150c-ea73-4144-993f-77ca8124de...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	24/11/2024 11:07	1070
8	Learner#4ef3715a-e420-4296-908c-eab9e5b797...	Opportunity#0000000010WCBS50CYGDX97ES4	BC69M2K	18/02/2025 12:41	1070
9	Learner#4ef49684-e9b0-40a9-b07e-07bfa78bdbc5	Opportunity#0000000010WCBS50CYGDX97ES4	BGRQZ2N	08/10/2024 09:50	1120
10	Learner#4f027693-d86f-4d65-a0a1-6342c8c0979e	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	04/04/2024 15:15	1070
11	Learner#4f0328f9-144e-4fe7-a4a0-604517a7655f	Opportunity#0000000010WCBS50CYGDX97ES4	BGRQZ2N	31/01/2025 16:49	1070

DATA CLEANING AND VALIDATION

Missing Values -

- assigned_cohort - 13,318, 11.72%
- apply_date - 188, 0.17%
- status - 186, 0.16%

Possible Observations -

1. Uniqueness:

- Unique Enrollments - 57,966
- Unique Learners - 187 (On average, each learner has ~ 310 enrollments! This indicates multiple applications per learner).

2. Duplicate Check:

- Duplicate Records - 0

Status Distribution (Top 10 Codes) -

- 1010 - 659
- 1020 - 161
- 1030 - 12,236
- 1040 - 24
- 1050 - 1,003
- 1055 - 11,471
- 1070 - 76,109
- 1080 - 1,191
- 1110 - 1,514
- 1120 - 9,048

Status 1070 dominates with over 76k entries, indicating it's a key milestone.
Other notable statuses - 1030, 1055, and 1120.
Some status codes like 1040 and 1020 are rare, possibly edge or transitional cases.

Assigned Cohorts (Top 10) -

- BAM6HBR - 1,805
- BSEV9QO - 1,733
- BGRQZ2N - 1,719
- BP9ZV19 - 1,611
- BWAG78I - 1,564
- BT4YTCR - 1,532
- BEXFE8O - 1,525
- B986905 - 1,522
- B6MZ4HK - 1,502
- BE7X8PZ - 1,497

Cohorts like BAM6HBR, BSEV9QO, and BGRQZ2N have the highest enrollments (~1700+ each).
Most top cohorts have similar sizes (~1500–1800), hinting at a structured batch distribution.

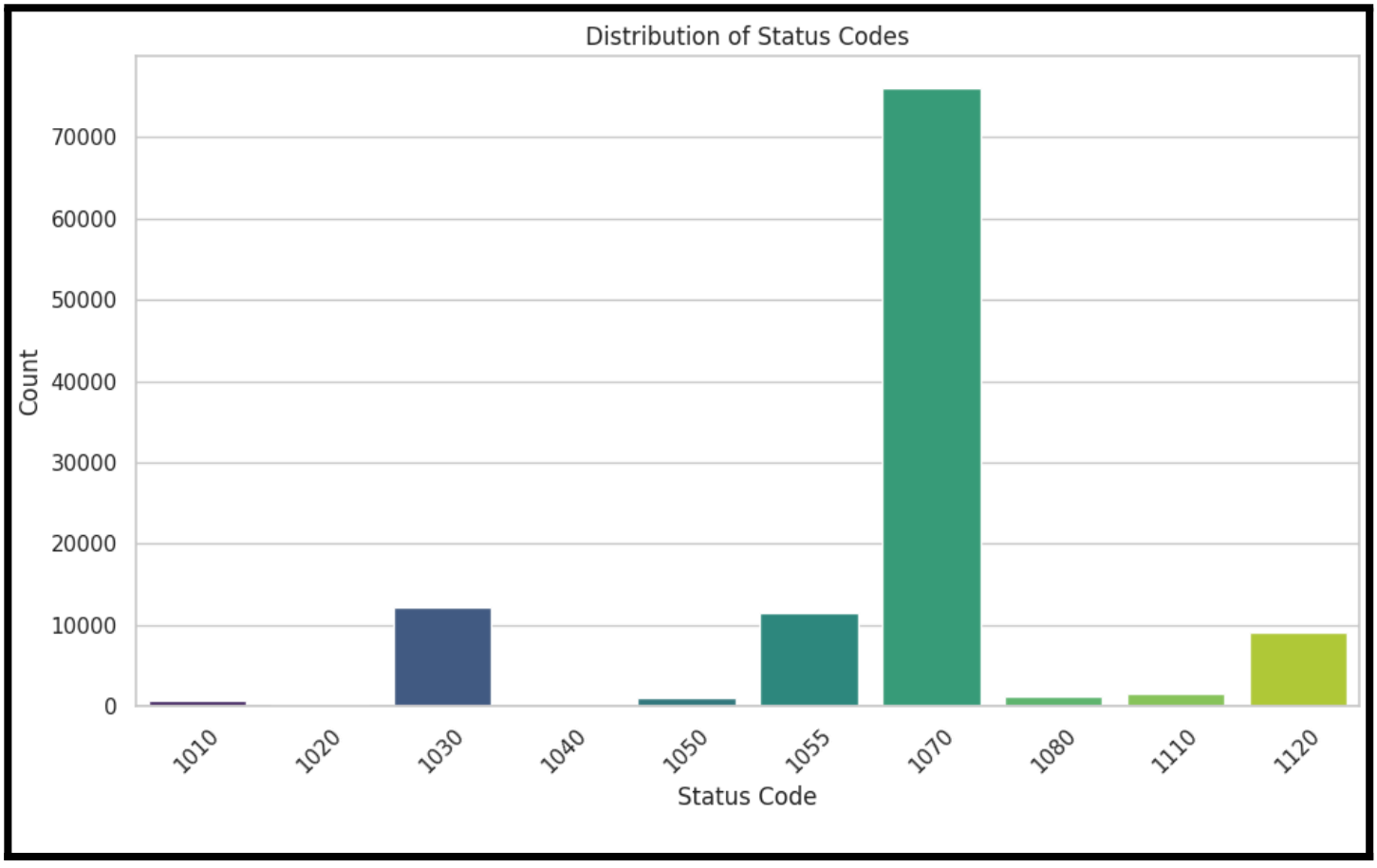
Application Timeline -

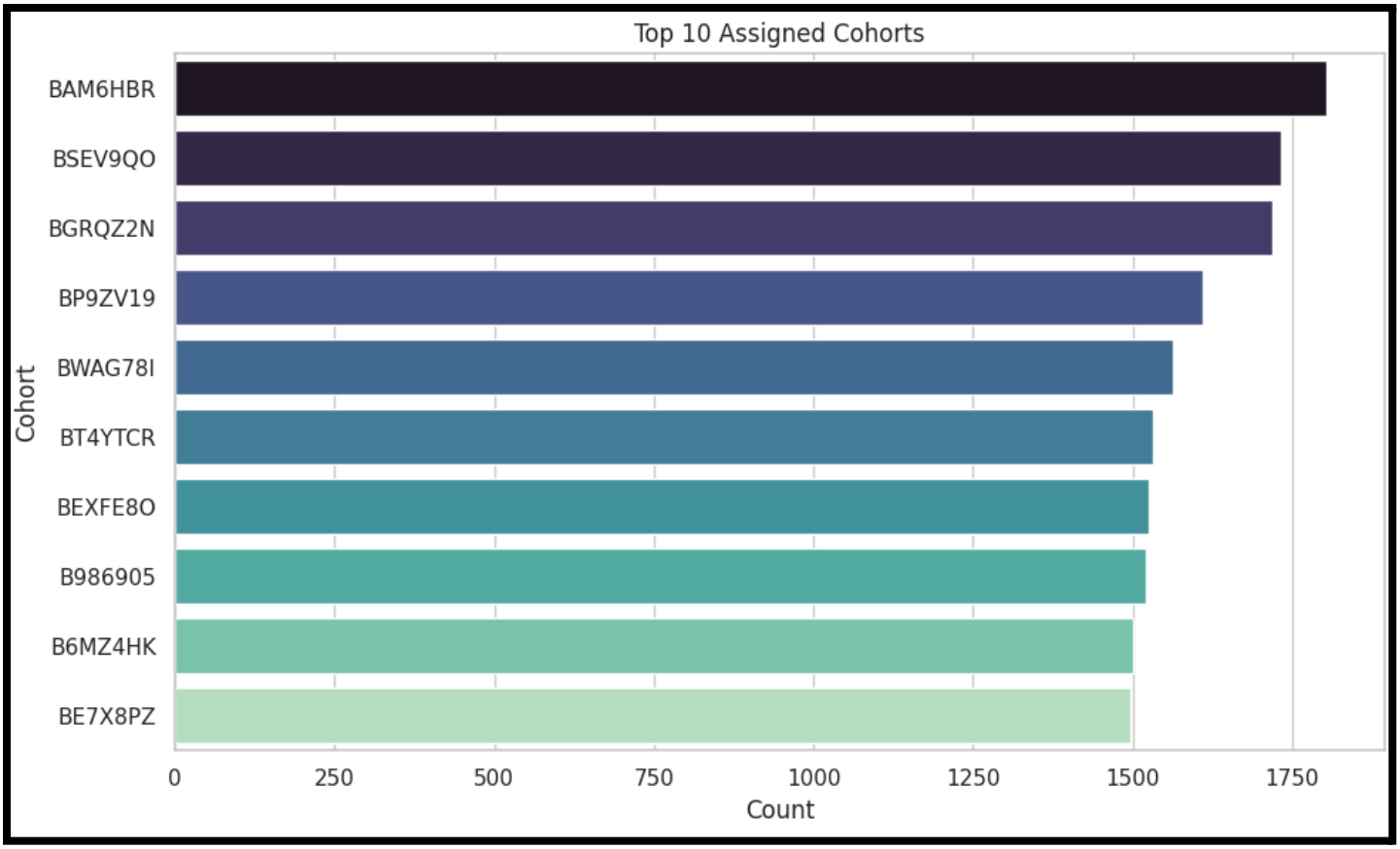
- Earliest Apply Date - June 9, 2022
- First 10 Application Days -

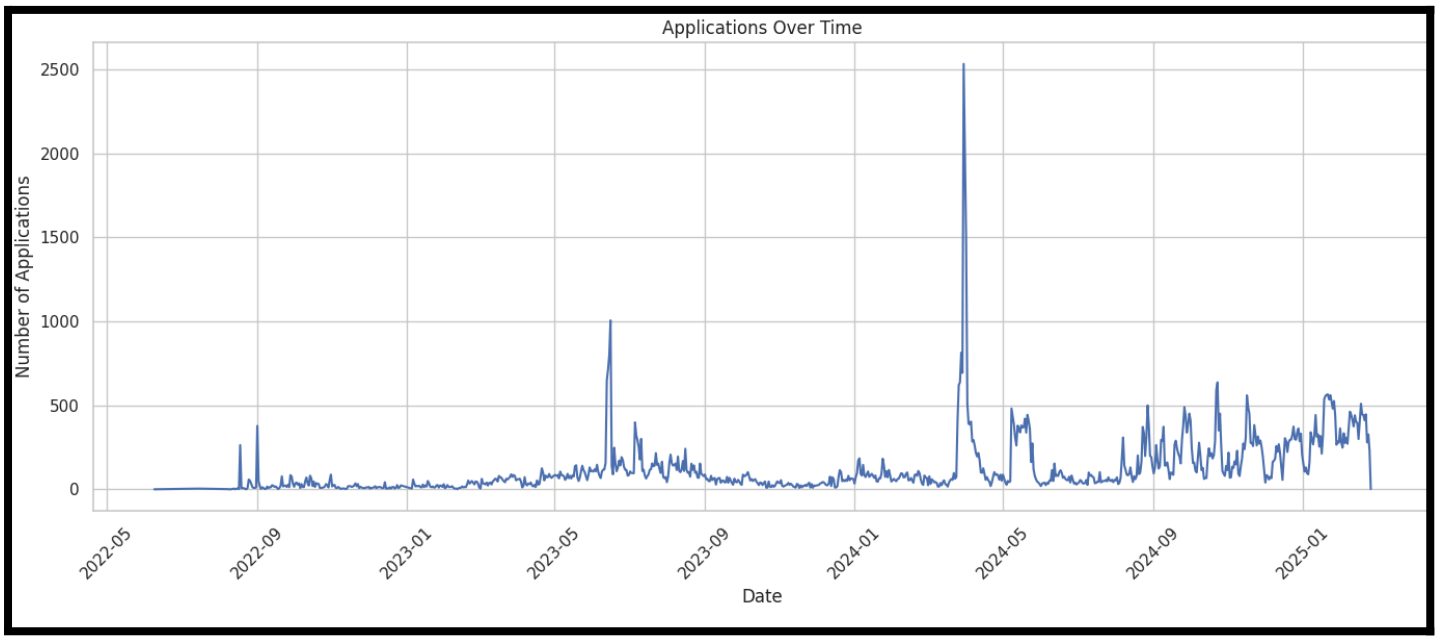
1. 2022-06-09 : 1
2. 2022-07-15 : 5
3. 2022-08-08 : 2
4. 2022-08-09 : 1
5. 2022-08-10 : 1
6. 2022-08-12 : 3
7. 2022-08-13 : 5
8. 2022-08-14 : 1
9. 2022-08-15 : 4
10. 2022-08-16 : 7

Key Findings -

- 187 learners generated 113,602 enrollments → ~310 enrollments per learner, which means learners are applying multiple times across opportunities/programs or system could be tracking re-attempts, reapplications, or phases of a process.
- Status 1070 makes up 67% of all records which likely represents a major milestone or default system-generated status (e.g., "Application Complete" or "Accepted").
- 11.7% of records are missing assigned_cohort.
- Top cohorts range from 1,497 to 1,805 enrollments which suggests a deliberate strategy for batch sizing and distribution.







These dashboards are created using Python libraries like Matplotlib & Seaborn in Google Colab.

COGNITO DATA REPORT

EXPLORATORY DATA ANALYSIS (EDA) REPORT

Dataset Name - Cognito_Raw2.csv

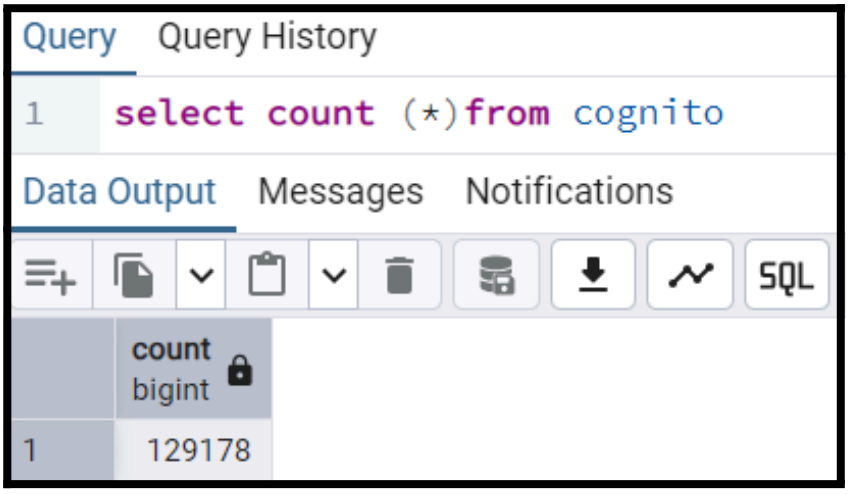
The dataset contains 129,178 user records with 9 columns.

- user_id [Primary Key]
- email
- gender
- UserCreateDate
- UserLastModifiedDate
- birthdate
- city
- zip
- state

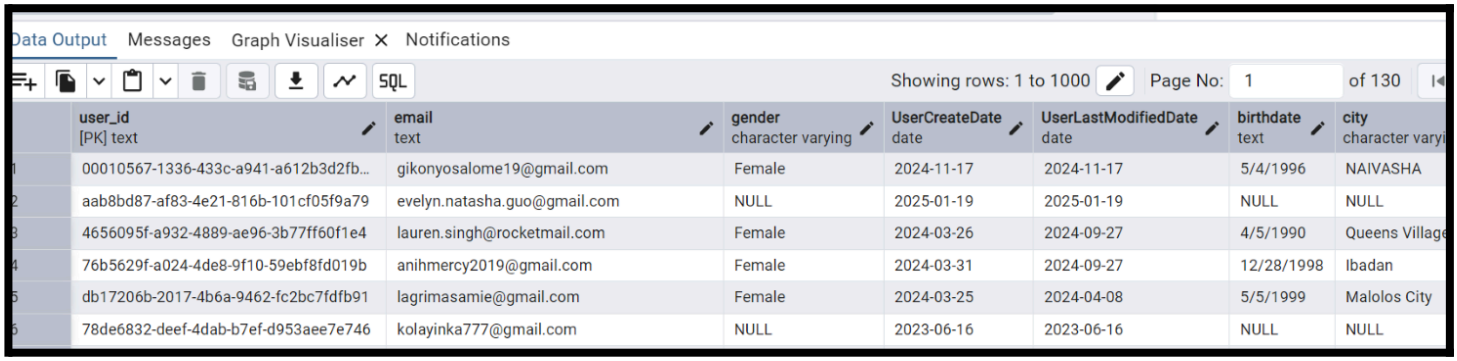
Created Table “cognito” in PostgreSQL and copied the original data into the table.

Initial Checks -

select count (*) from cognito



select * from cognito



	user_id [PK] text	email text	gender character varying	UserCreateDate date	UserLastModifiedDate date	birthdate text	city character varyi
1	00010567-1336-433c-a941-a612b3d2fb...	gikonyosalome19@gmail.com	Female	2024-11-17	2024-11-17	5/4/1996	NAIVASHA
2	aab8bd87-af83-4e21-816b-101cf05f9a79	evelyn.natasha.guo@gmail.com	NULL	2025-01-19	2025-01-19	NULL	NULL
3	4656095f-a932-4889-ae96-3b77ff60f1e4	lauren.singh@rocketmail.com	Female	2024-03-26	2024-09-27	4/5/1990	Queens Villag
4	76b5629f-a024-4de8-9f10-59ebf8fd019b	anihmercy2019@gmail.com	Female	2024-03-31	2024-09-27	12/28/1998	Ibadan
5	db17206b-2017-4b6a-9462-fc2bc7fdfb91	lagrimasamie@gmail.com	Female	2024-03-25	2024-04-08	5/5/1999	Malolos City
6	78de6832-deef-4dab-b7ef-d953aee7e746	kolayinka777@gmail.com	NULL	2023-06-16	2023-06-16	NULL	NULL

DATA CLEANING AND VALIDATION

Missing Data -

- state - 42,937 missing
- zip - 42,869 missing
- city - 42,866 missing
- gender - 42,862 missing
- birthdate & derived age - 42,862 missing

~33% of the data has missing demographic/location details.

Gender Distribution -

- Female - Dominant category
- Male - Present in smaller number
- Missing/Unknown - A large chunk (~33%) is missing

Age Distribution -

- Mean Age - 26 years
- Median Age - 25 years
- Age Range - 3 to 101 years
  - Likely some outliers (e.g., age 3 or 101)

Top 10 Cities -

- Lagos - 3031
- Nairobi - 2675
- Accra - 2136
- Karachi - 1841
- Hyderabad - 1836
- Abuja - 1532
- Lahore - 1381
- Ibadan - 1221
- Dhaka - 1101
- Saint Louis - 1056

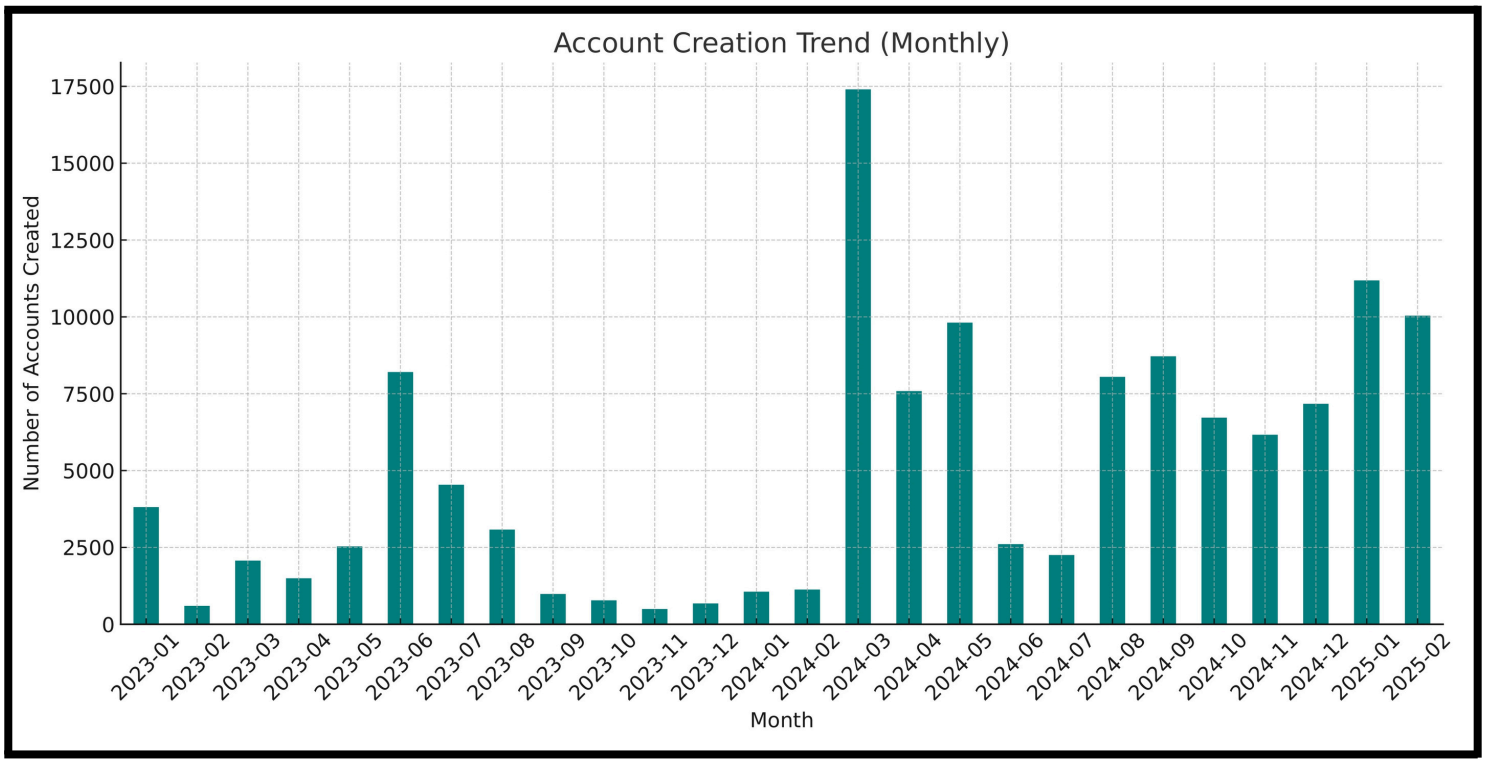
Top 10 States -

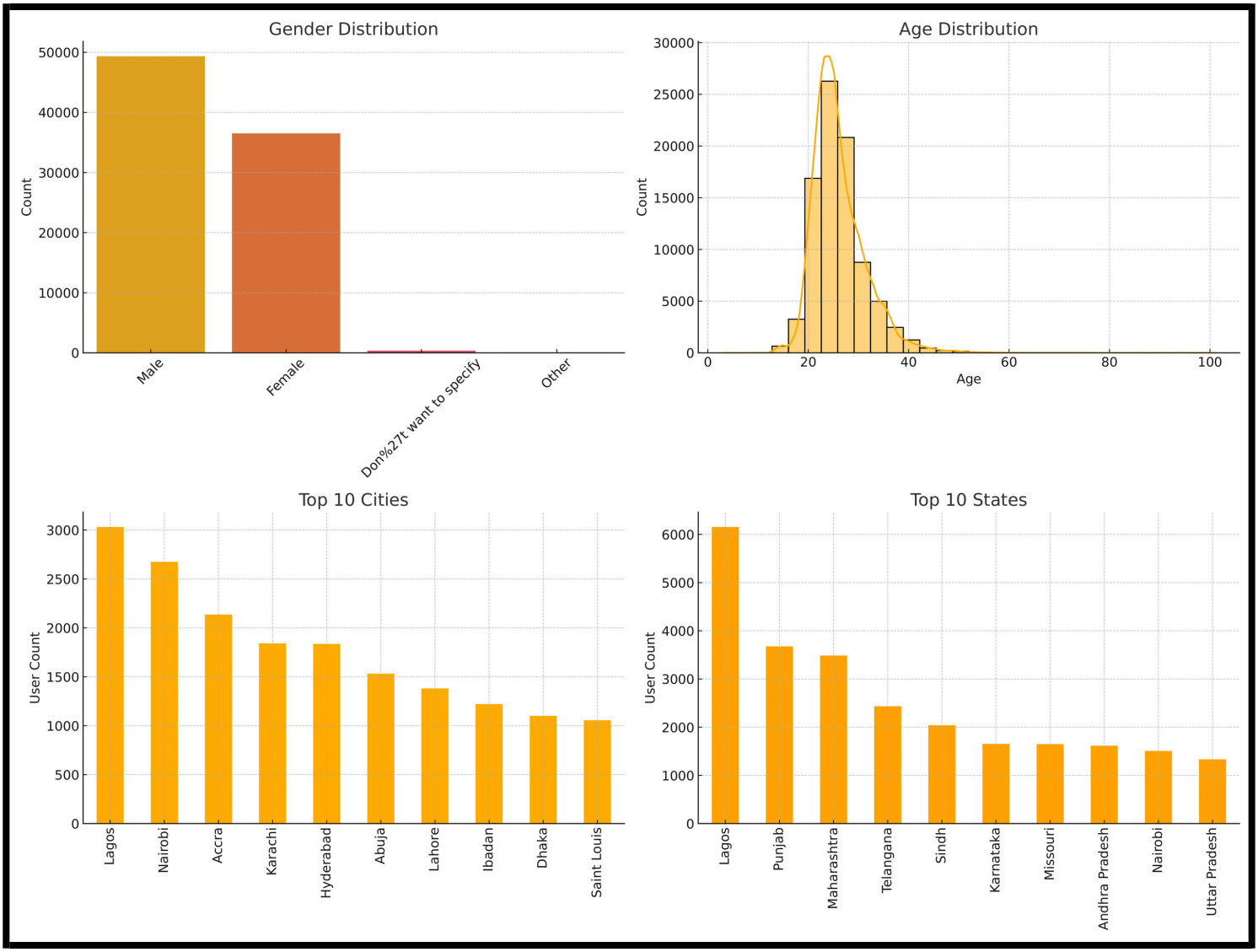
- Lagos - 6154
- Punjab - 3677
- Maharashtra - 3487
- Telangana - 2434
- Sindh - 2040
- Karnataka - 1653
- Missouri - 1648
- Andhra Pradesh - 1616
- Nairobi - 1507
- Uttar Pradesh - 1333

User Account Creation Trend (First 10 Months of 2023) -

- Peaks - June 2023 (8,206 accounts), March 2023 (2,071)
- Lowest activity - February and October 2023

This suggests a surge in platform engagement mid-year.





These dashboards are created using Python libraries like Matplotlib & Seaborn in Google Colab.

Key Findings -

1) Demographic Data Gaps

~33% of users have missing values in -

- gender
- birthdate (and derived age)
- city, zip, and state

2) Gender Imbalance

- Female users are the dominant group.
- Male users are fewer.
- A significant portion (33%) is missing or unknown.

3) Age Distribution Observations

- Mean Age - 26
- Median Age - 25
- Range - 3 to 101 years

4) Geographic Concentration

Top Cities -

- Lagos, Nairobi, Accra, Karachi, Hyderabad dominate the city list.
- Saint Louis is the only major city from the US in the top 10.

Top States -

- Dominated by regions in Nigeria, Pakistan, and India.
- Strong representation from West Africa and South Asia.

5) User Onboarding Trends (2023)

- Peak sign-ups in -
  - June 2023 (8,206 users)
  - March 2023 (2,071 users)
- Lowest activity - February and October 2023

MARKETING DATA REPORT

EXPLORATORY DATA ANALYSIS (EDA) REPORT

Dataset Name - Marketing Campaign Data All Accounts (2023-2024)(Detail1).csv

Total Records - 155
Total Features (Columns) - 13

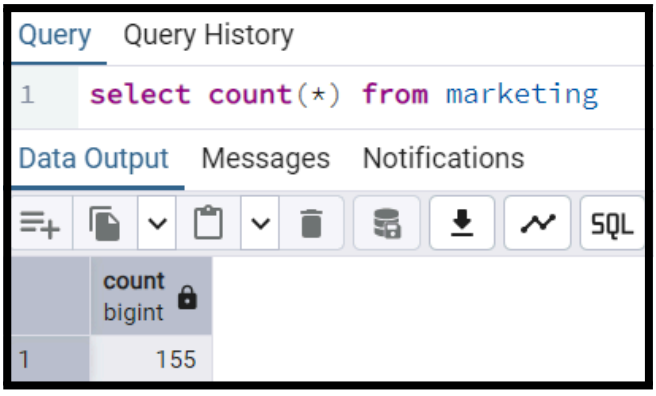
- Ad Account Name
- Campaign name
- Delivery status
- Delivery level
- Reach
- Outbound clicks
- Landing page views
- Result type
- Results
- Cost per result
- Amount spent (AED)
- CPC (cost per link click)
- Reporting starts

In this dataset, there does not appear to be a single column that uniquely identifies each row, i.e., no obvious primary key.

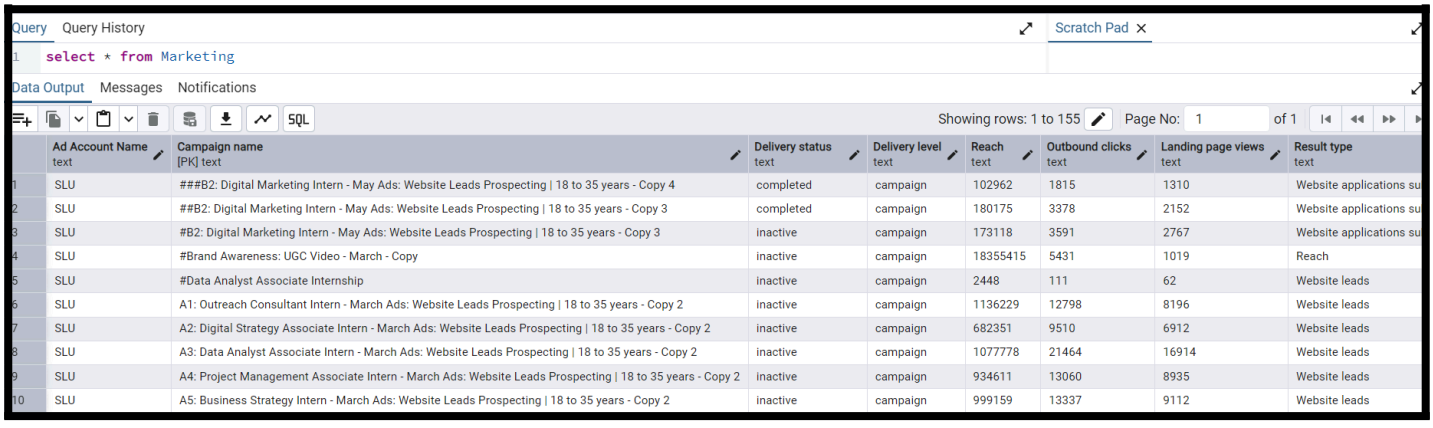
Created Table “Marketing” in PostgreSQL and copied the original data into the table.

Initial Checks -

select count (*) from Marketing



select * from Marketing



	Ad Account Name text	Campaign name [PK] text	Delivery status text	Delivery level text	Reach text	Outbound clicks text	Landing page views text	Result type text
1	SLU	###B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 4	completed	campaign	102962	1815	1310	Website applications su
2	SLU	##B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	completed	campaign	180175	3378	2152	Website applications su
3	SLU	#B2: Digital Marketing Intern - May Ads: Website Leads Prospecting | 18 to 35 years - Copy 3	inactive	campaign	173118	3591	2767	Website applications su
4	SLU	#Brand Awareness: UGC Video - March - Copy	inactive	campaign	18355415	5431	1019	Reach
5	SLU	#Data Analyst Associate Internship	inactive	campaign	2448	111	62	Website leads
6	SLU	A1: Outreach Consultant Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	1136229	12798	8196	Website leads
7	SLU	A2: Digital Strategy Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	682351	9510	6912	Website leads
8	SLU	A3: Data Analyst Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	1077778	21464	16914	Website leads
9	SLU	A4: Project Management Associate Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	934611	13060	8935	Website leads
10	SLU	A5: Business Strategy Intern - March Ads: Website Leads Prospecting | 18 to 35 years - Copy 2	inactive	campaign	999159	13337	9112	Website leads

DATA CLEANING AND VALIDATION

Missing Values -

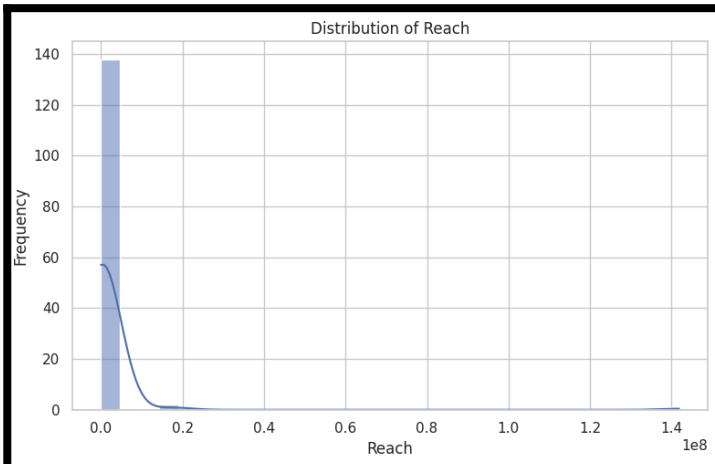
- Ad Account Name - 7
- Campaign name - 8
- Delivery status - 7
- Delivery level - 7
- Reach - 7
- Outbound clicks - 9
- Landing page views - 9
- Result type - 7
- Results - 6
- Cost per result - 7
- Amount spent (AED) - 6
- CPC (cost per link click) - 8
- Reporting starts - 7

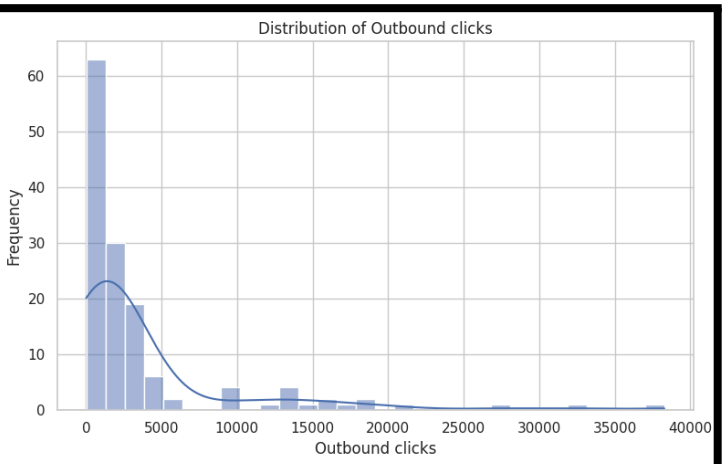
Data Cleaning -

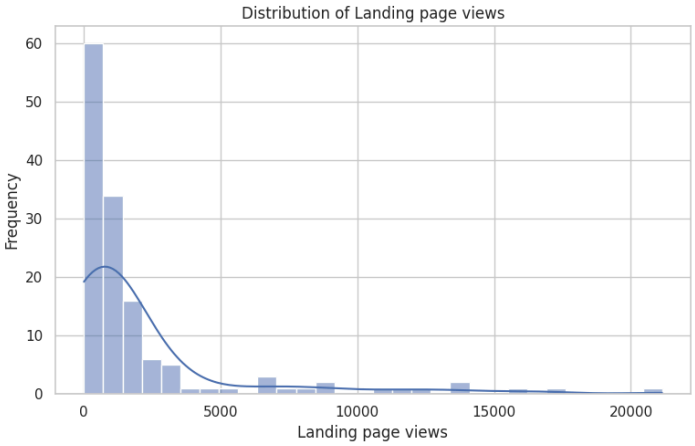
index	Reach	Outbound clicks	Landing page views	Results	Cost per result	Amount spent (AED)	CPC (cost per link click)
count	141.0	139.0	139.0	142.0	141.0	142.0	140.0
mean	1702851.3120567375	3683.201438848921	2113.971223021583	2570608.3591549294	4.163969588652482	2400.2125102605632	1.0566550642857144
std	12085460.523213452	6161.006572640047	3634.3954981326733	19419805.617420424	4.982311779566414	3937.369148029759	0.930421004493036
min	1315.0	14.0	1.0	1.0	0.003665	0.99	0.048919
25%	44420.0	531.0	260.5	161.25	1.2831	480.71500000000003	0.42874350000000006
50%	148357.0	1604.0	868.0	384.0	2.920683	1253.225	0.7637075
75%	422292.0	3202.5	1788.5	2022.75	5.5095	2206.3424999999997	1.3596395000000001
max	141835342.0	38284.0	21140.0	182509917.0	47.08949	25531.47	5.597563

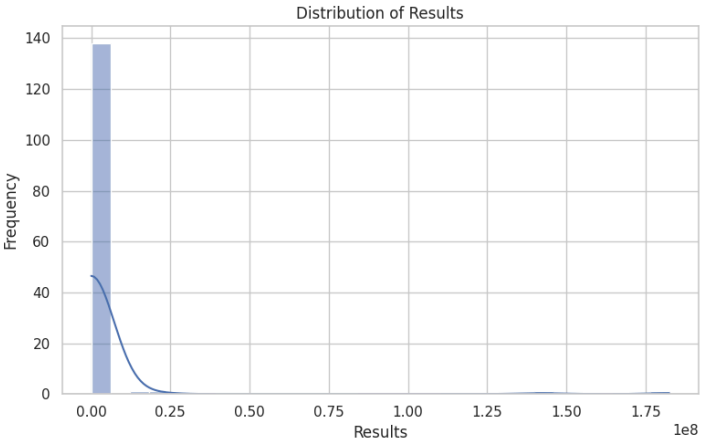
Custom Summary Statistics

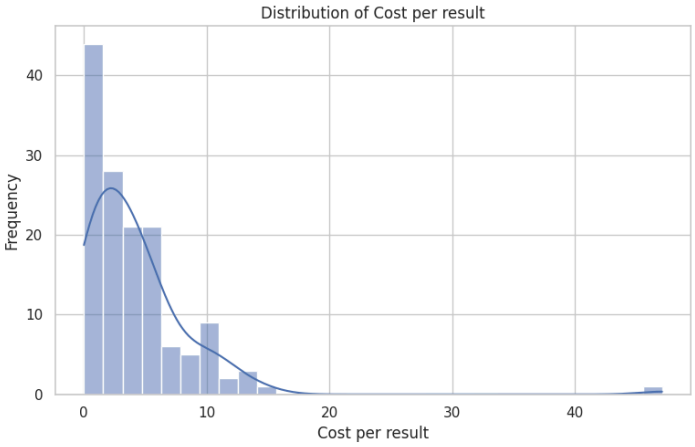
Metric	Reach	Results	Amount Spent (AED)	CPC	Cost/Result
Mean	1,702,851.31	2,570,608.36	2,400.21	1.06	4.16
Median (50%)	148,357.00	384.00	1,253.22	0.76	2.92
Max	141,835,342.00	182,509,917.00	25,531.47	5.60	47.09
Min	1,315.00	1.00	0.99	0.05	0.00

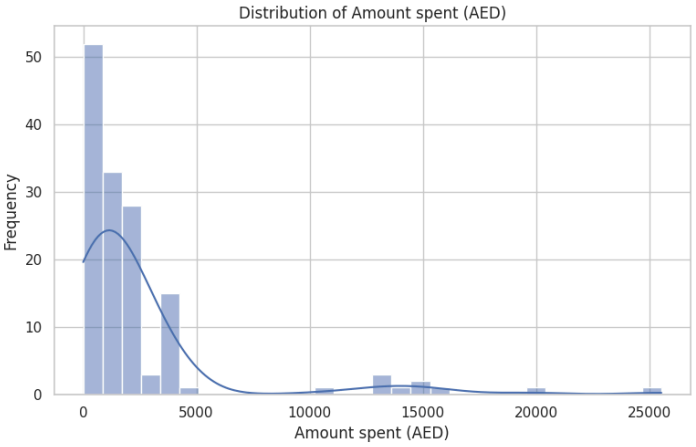

Distribution of Reach
Frequency
Reach
1e8

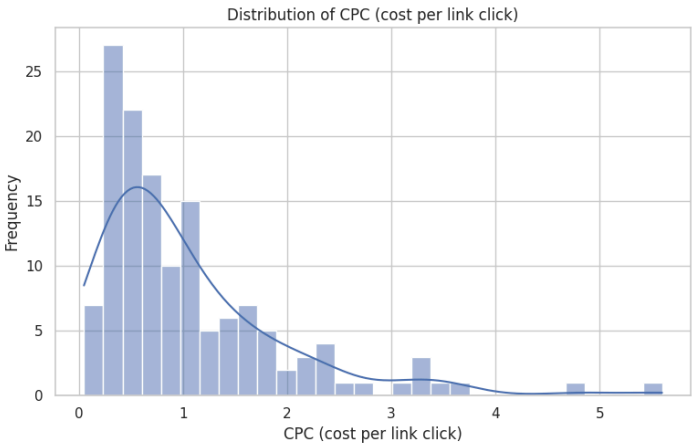

Distribution of Outbound clicks
Frequency
Outbound clicks

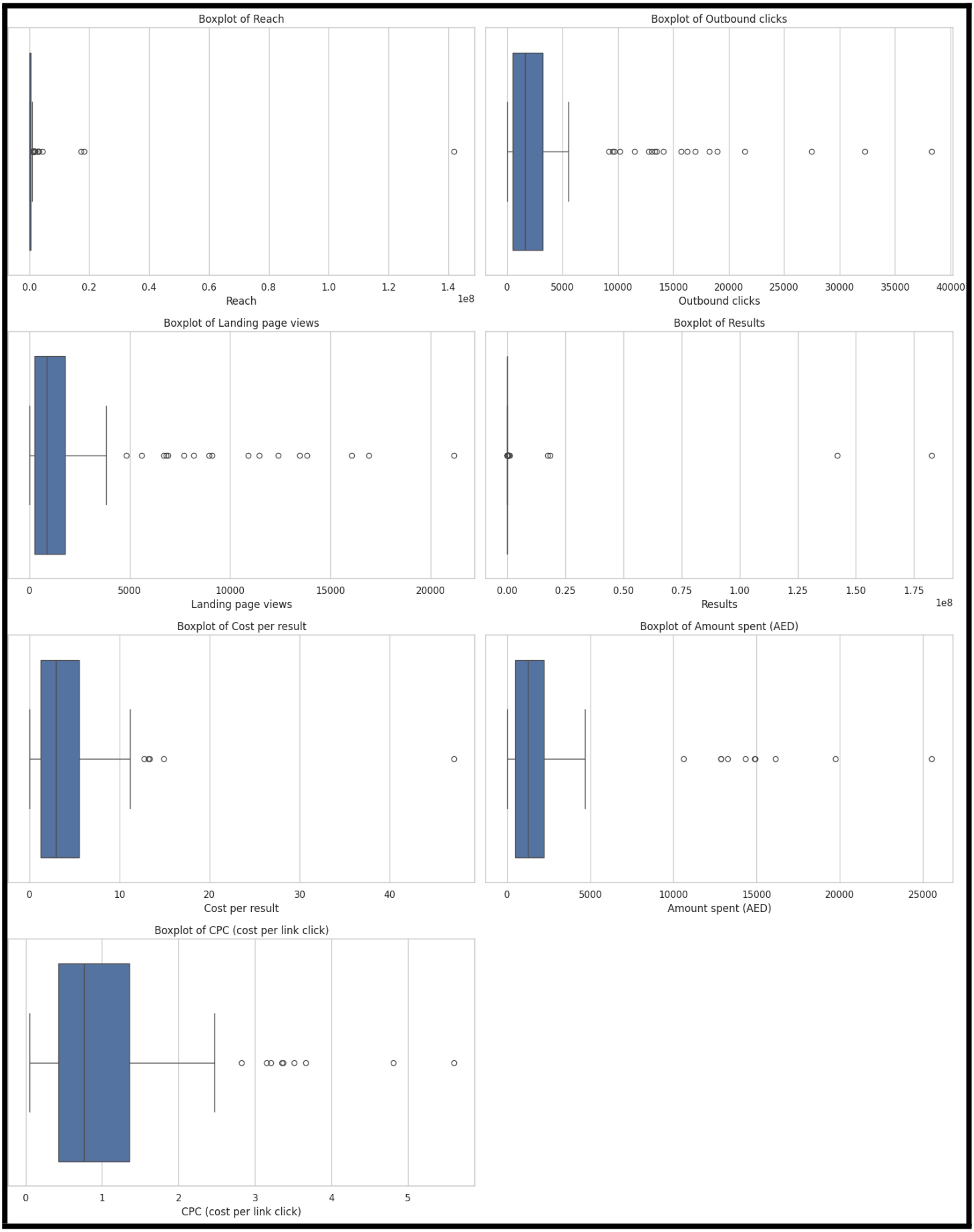

Distribution of Landing page views
Frequency
Landing page views


Distribution of Results
Frequency
Results
1e8

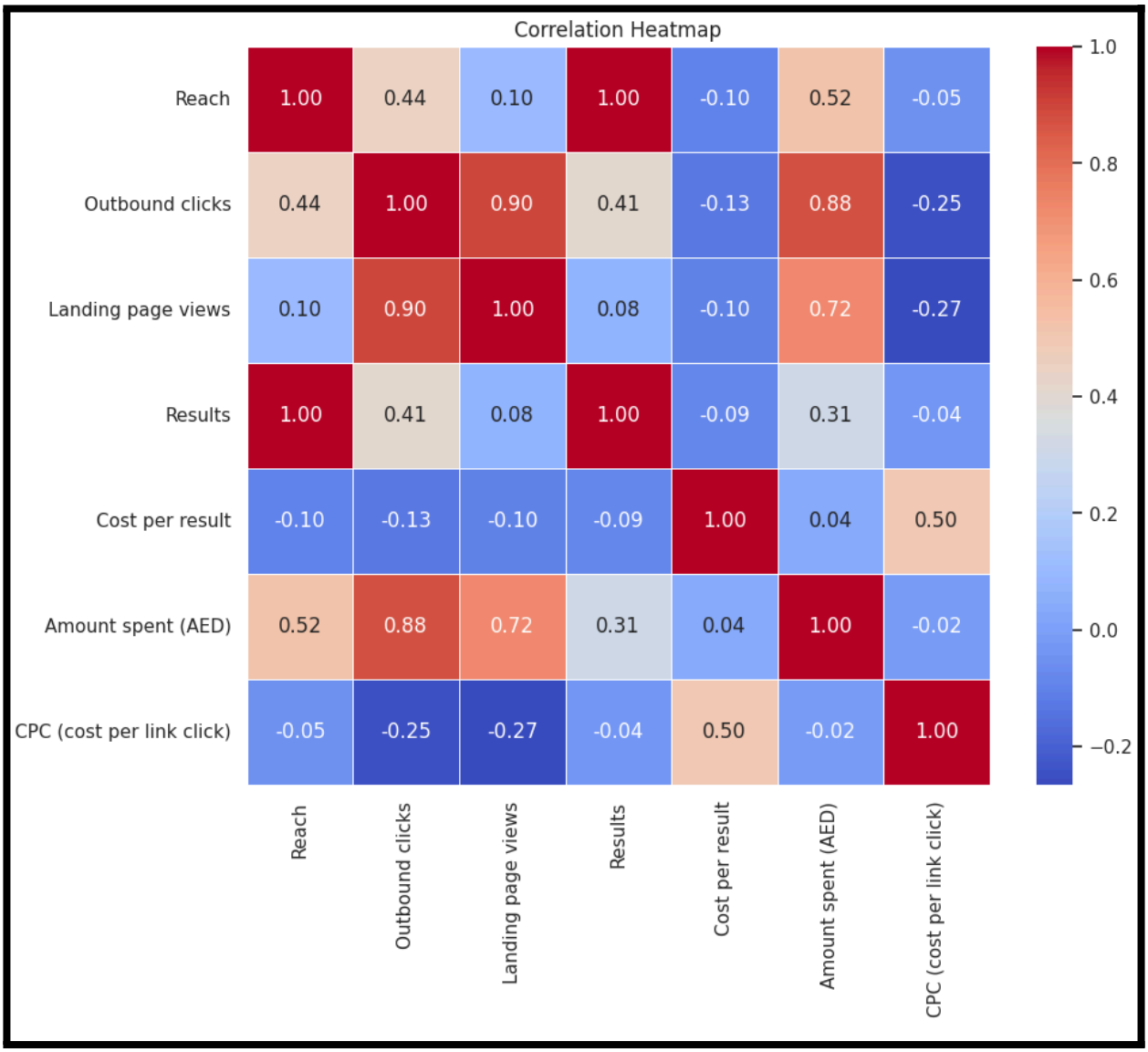

Distribution of Cost per result
Frequency
Cost per result


Distribution of Amount spent (AED)
Frequency
Amount spent (AED)


Distribution of CPC (cost per link click)
Frequency
CPC (cost per link click)



These dashboards are created using Python libraries like Matplotlib & Seaborn in Google Colab.



Key Findings -

1. Data Size & Structure -
   - The dataset includes detailed campaign metrics like Reach, Results, Amount Spent, CPC, and Cost per Result across multiple ad accounts.
   - Presence of both high-level aggregate campaigns and granular daily entries.
2. Missing Data -
   - Several columns contain missing values (e.g., Cost per result, CPC, Landing page views), which can affect modeling or trend analysis if not handled properly.
3. Outliers & Skewed Data -
   - Extreme values were detected in Reach, Results, and Amount Spent (AED) (e.g., max reach exceeds 140M).
   - Boxplots and histograms reveal heavy right skew in spending and results, indicating a few campaigns dominate the data.
4. Currency & Formatting Issues -
   - Currency formatting is inconsistent (some values show as text like “AED 1,200” in earlier stages).
   - Some fields should be strictly numerical for analysis (Amount spent, CPC, etc.).
5. Duplicates & Irrelevant Records -
   - Duplicate records were present and have been removed.
   - Guest contributors were filtered out as they are not relevant for business analysis.
6. Correlations -
   - Amount Spent (AED) shows moderate to strong positive correlation with Reach, Results, and Cost per Result.
   - CPC and Cost per Result have weaker correlation with other metrics — good indicators to analyze efficiency independently.