Week 1 — Data Familiarization & KPI Validation

Objective: Build foundational understanding of the data and validate all essential KPIs using SQL and Python.
Outcome: A complete Week 1 analysis package containing data overview, SQL validations, event EDA, and a structured report.

Day 1 — Environment Setup and Initial Data Exploration

1. What You Do

Set up your local analysis environment, explore the raw data, confirm schema understanding, and document basic observations.

2. Why This Matters

Every DA project begins with a strong understanding of the data.

Interviewers expect that you can clearly describe:

- What tables exist
- What each field means
- How large the dataset is
- What potential issues exist

This is the foundation for all future SQL and EDA work.

3. Step-by-Step Tasks

(A) Install Tools

Install or confirm availability of:

SQL Tools

- MySQL Server
- MySQL Workbench or DBeaver

Python Tools

- Python 3.10+
- Jupyter Notebook
- pandas, numpy, matplotlib, seaborn

Run:

Code block

```
pip install pandas numpy matplotlib seaborn
```

(B) Inspect Each CSV File

Open these files (either in VSCode, Excel, or pandas):

- `users.csv`
- `subscriptions.csv`
- `user_events.csv`

Check:

1. Total rows
2. Total columns
3. Null counts
4. Duplicate user_ids
5. Field types (string, int, date)
6. High-level distributions
    - gender
    - status (active vs cancelled)
    - event_type

(C) Validate Foreign Keys

Manually check:

- Every user_id in subscriptions should appear in users.
- Every user_id in events should also appear in users.

Document any mismatches.

(D) Check Data Time Range

Confirm earliest and latest:

- signup_date
- cancel_date
- event_time

This ensures your retention windows later will make sense.

---

4. Deliverable

Create a markdown document:

`01_data_overview.md`

Sections:

1. Table summary (row counts, column counts)
2. Field descriptions (based on your Data Dictionary)
3. Data types and format confirmation
4. Key observations (e.g., "events table contains 31M rows")
5. Potential issues (e.g., out-of-order timestamps)

Day 2 — Basic KPI Validation Using SQL

1. What You Do

Use SQL to compute baseline KPIs:

- total users
- subscribed users
- actively subscribed users
- cancelled users
- monthly churn count

2. Why This Matters

These are the core KPIs used by any streaming business.

Interviewers expect you to define and compute these correctly.

3. Step-by-Step Tasks

(A) Connect to MySQL

Import `users` and `subscriptions` if not already imported.

In MySQL:

Code block

```
USE churn_analytics;
```

(B) Calculate Total Users

Code block

```
SELECT COUNT(*) AS total_users FROM users;
```

(C) Calculate Subscribed Users

Users with subscriptions:

Code block

```
SELECT COUNT(*) AS subscribed_users FROM subscriptions;
```

(D) Active vs Cancelled Users

Code block

```
SELECT status, COUNT(*)
FROM subscriptions
GROUP BY status;
```

(E) Monthly Churn Count

Code block

```sql
SELECT    DATE_FORMAT(cancel_date, '%Y-%m') AS cancel_month,COUNT(*) AS churned_users
FROM subscriptions
WHERE cancel_date IS NOT NULL
GROUP BY 1
ORDER BY 1;
```

---

4. Deliverable

`02_basic_kpi_check.sql`

Include all your SQL queries and a short note after each query explaining what it calculates.

Day 3 — Validate Churn & Retention Logic in SQL

1. What You Do

Build SQL logic for:

- monthly churn rate
- monthly retention rate
- monthly active users

2. Why This Matters

Retention and churn are the core of this project.

You must be able to compute:

- how many users churned
- how many users stayed
- what percentage retained

Interviewers often ask you to derive retention from raw subscription tables.

3. Step-by-Step Tasks

(A) Active Users at Start of Month (SQL)

(B) Monthly Churn Rate (SQL)

(C) Retention Rate (SQL)

---

4. Deliverable

`03_churn_retention_validation.sql`

Day 4 — Event Table EDA (Python)

1. What You Do

Perform initial behavior analysis.

2. Why This Matters

Event data drives:

- engagement analysis
- early churn signals
- cohort behavior patterns
- model features

You cannot do ML later without this EDA.

3. Step-by-Step Tasks

(A) Load Events Data

Code block

```
events = pd.read_csv('user_events.csv')
```

(B) Event Type Distribution

Code block

```
events['event_type'].value_counts()
```

(C) Daily Event Trend

Code block

```
events['date'] = pd.to_datetime(events['event_time']).dt.date
events.groupby('date').size().plot(figsize=(12,5))
```

(D) Device Type Distribution

Code block

```
events['device_type'].value_counts().plot(kind='bar')
```

(E) Events per User

`events.groupby('user_id').size().describe()`

Evaluate whether:

- some users have zero events
- some users are extremely active
- event volume decreases before cancel_date

4. Deliverable

`04_events_eda.ipynb`

Include:

- event summary
- plots
- top observations (1–2 lines each)

Day 5 — Week 1 Summary & Insights

1. What You Do

Consolidate all findings into a concise but professional Week 1 report.

2. Why This Matters

Interviewers expect you to be able to summarize:

- what your data looks like
- what initial metrics show
- what potential issues exist
- what next steps you will take

Week 1 report becomes part of your portfolio.

3. Required Sections

(1) Data Overview

- number of users
- number of subscriptions
- number of events
- key columns
- data ranges

(2) KPI Findings

- churn distribution
- retention initial insight
- active vs cancelled count

(3) Events EDA Results

- event type proportions

- device mix
- average events per user
- early observations

(4) Potential Data Issues

Examples:

- timestamp disorder
- heavy users
- silent users

(5) Next Step After Week 1

- build retention SQL
- build cohort table
- start funnel analysis

---

4. Deliverable

`05_week1_report.md`