Analytical Engineer Technical Assessment

General Instructions:

- We've designed this assessment to help you better understand your skill level with these tools. We encourage you to do your best on every section, and your experience level will be taken into account in our review.
- This assessment needs to be delivered in a GitHub, GitLab, or similar git repository provider. Regardless of the choice, we need to be able to access it in order to evaluate it.
- If you implement this project in GCP, you will need to give us access to your environment so that we can evaluate it.

Part 1: Data Understanding and Modelling

You are given the following sample data for three tables (csv files attached to email):

Transactions

id	timestamp	player_id	type	amount
1	2024-01-10 17:07:01+00:00	7	Withdraw	213.13
2	2024-02-09 19:19:27+00:00	4	Withdraw	182.64
3	2024-02-13 05:40:50+00:00	8	Withdraw	184.22
4	2024-02-24 04:35:15+00:00	5	Deposit	304.94
5	2024-03-27 16:28:29+00:00	7	Withdraw	525.23
6	2024-04-05 02:33:27+00:00	10	Deposit	432.51
7	2024-04-16 19:27:32+00:00	3	Withdraw	291.94
8	2024-09-05 07:53:23+00:00	7	Withdraw	612.24
9	2024-10-13 00:24:33+00:00	8	Withdraw	140.35
10	2024-10-14 23:45:05+00:00	5	Withdraw	292.85

Players

id	affiliate_id	country_code	is_kyc_approved	created_at	updated_at
1	1	DE	TRUE	2024-01-10 17:07:01+00:00	2024-02-08 09:05:26+00:00
2		BR	TRUE	2024-02-09 19:19:27+00:00	2024-03-05 08:06:20+00:00
3	2	BR	TRUE	2024-02-13 05:40:50+00:00	2024-02-17 10:58:17+00:00
4	3	GB	TRUE	2024-02-24 04:35:15+00:00	2024-03-05 19:19:09+00:00
5	4	AU	FALSE	2024-03-27 16:28:29+00:00	2024-04-01 05:38:15+00:00
6		FR	TRUE	2024-04-05 02:33:27+00:00	2024-04-06 13:16:25+00:00
7	5	CA	TRUE	2024-04-16 19:27:32+00:00	2024-04-18 08:29:18+00:00
8	6	BR	TRUE	2024-09-05 07:53:23+00:00	2024-10-01 19:17:16+00:00
9	7	AU	FALSE	2024-10-13 00:24:33+00:00	2024-10-24 07:28:43+00:00
10	8	CA	FALSE	2024-10-14 23:45:05+00:00	2024-11-04 11:19:33+00:00
10	8	CA	TRUE	2024-10-14 23:45:05+00:00	2024-11-06 12:31:09+00:00

Affiliate			
id	code	origin	redeemed_at
1	DWEHSP	YouTube	2024-01-10 17:07:01+00:00
2	QGKTBC	Discord	2024-02-13 05:40:50+00:00
3	YRLUJA	X	2024-02-24 04:35:15+00:00
4	ZXCVBN		2024-03-27 16:28:29+00:00
5	POIUYT	YouTube	2024-04-16 19:27:32+00:00
6	LKJHGF	Discord	2024-09-05 07:53:23+00:00
7	MNBVCX		2024-10-13 00:24:33+00:00
8	ASDFGH	X	2024-10-14 23:45:05+00:00
9	HGFDSA	YouTube	
10	TREWQS	Discord	

The data should comply with the following business rules (some inconsistencies may occur):

1. The players are not allowed to make any transactions before they are KYC verified.
2. Multiple players can use the same affiliate code, but each code redeemed should match 1:1 to the player with the affiliate ID.
3. An affiliate redeemed must be linked to a player.
4. Players may or may not use an affiliate code to join.
5. The IDs of all tables are unique.

Tasks:

1. Examine the sample data, determine which rows don't follow the business rules and clean the data.
2. Write a Python script that extends the sample data for these tables according to the provided schema to 1000 rows.

Part 2: DBT Models

Using the three tables above, write DBT models to create tables to fulfil these requirements:

Model 1: Create one row per player daily, with a column for deposits and another for withdrawals. Withdrawals should be negative.

Model 2: Sum and count of deposits per player country for players who are KYC approved and affiliate origin from Discord.

Model 3: One row per player with their three most considerable deposit amounts.

Optional Bonus Section: Automation with Airflow

For this section, you can use this boilerplate structure to help you get started with Airflow:
https://github.com/lavedonio-ancientgaming/technical-assessment

1. Orchestrate the DBT models created using Airflow, considering the pipeline will be refreshed daily.
2. Implement this pipeline in BigQuery and use Google Cloud Composer as the Airflow provider.