Onboarding Funnel Analysis and A/B Testing

Objective

The goal of this analysis was to **understand user drop-offs in the onboarding funnel** and evaluate an **A/B experiment** that tested two ID verification flows to improve overall completion rates.

1. Dataset Overview

Two datasets were used for this analysis:

a. `user_attributes`

Contains metadata about each user:

- `user_id` — unique user identifier
- `country` — user's country
- `device` — device used (`android`, `ios`, `web`)
- `signup_date` — date of registration
- `experiment_group` — experiment variant (`A`, `B`, or `NULL`)

b. `user_onboarding_logs`

Contains detailed step-by-step onboarding logs:

- `user_id` — links to `user_attributes`
- `step_name` — one of five onboarding steps
- `step_status` — `completed` or `dropped`
- `timestamp` — when the event occurred
- `experiment_group` — experiment variant

Funnel Steps

1. Signup
2. Email Verification
3. Profile Setup
4. ID Verification
5. Onboarding Complete

The dataset included:

- **50,000 total users**
- **10,000 experiment users** (5,000 in group A, 5,000 in group B)
- Realistic timestamped logs (~250,000+ entries)

2. Methodology

Step 1: Data Preparation

- Loaded both CSV files into SQL and Python (Pandas).
- Cleaned missing experiment flags and standardized step order.

- Filtered for `step_status = 'completed'` to track progress through each funnel stage.

Step 2: Funnel Construction

Using SQL window functions and grouping:

```
SELECT step_name, COUNT(DISTINCT user_id) AS users_completed
FROM user_onboarding_logs
WHERE step_status = 'completed'
GROUP BY step_name
ORDER BY FIELD(step_name,
'signup','email_verification','profile_setup','id_verification','onboarding_complete');
```

This yielded the **stepwise completion counts** and **drop-off rates**.

Step 3: Experiment Comparison

A funnel comparison was done between **Group A** and **Group B** to measure lift:

```
SELECT step_name,
   COUNT(DISTINCT CASE WHEN experiment_group='A' THEN user_id END) AS users_A,
   COUNT(DISTINCT CASE WHEN experiment_group='B' THEN user_id END) AS users_B
FROM user_onboarding_logs
WHERE experiment_group IS NOT NULL AND step_status='completed'
GROUP BY step_name;
```

Step 4: Statistical Validation

A **two-proportion z-test** was used to validate whether the difference in completion rates between A and B was statistically significant.

$$z = (p1 - p2) / \sqrt{p * (1 - p) * (1/n1 + 1/n2)}$$

where:

$$p = (x1 + x2) / (n1 + n2)$$

where

- (p1, p2) = conversion rates for A and B
- (n1, n2) = total users in A and B
- (p) = pooled conversion rate

3. Results

Metric	Group A	Group B
Users in experiment	5,000	5,000
Completions	179	502
Conversion rate	**3.58%**	**10.04%**

Lift	**+6.46 percentage points**
Z-statistic	-12.82
P-value	0.0000

4. Key Insights

1. **Major Drop-off Point:**
   Most users dropped off at the **ID verification** stage, indicating friction in the process.

2. **Experiment Outcome:**
   Variant **B** significantly outperformed variant A, improving completion rates from **3.6%** → **10%**, a **~180% relative increase**.

3. **Statistical Confidence:**
   The difference is statistically significant at **p < 0.05**, confirming the new flow's effectiveness.

4. **Business Impact:**
   Implementing the new ID verification process could reduce onboarding abandonment by approximately **18%**, leading to higher user activation and retention.

5. Visualization in Power BI

The following visuals were created:

- **Funnel Chart:** Showed progressive user counts per step.
- **Bar Comparison:** Compared completion rates between Group A and Group B.
- **Filters:** Device, Country, Experiment Group.
- **KPI Cards:**
    - Total users analyzed
    - Overall conversion rate
    - % Lift achieved

These visualizations helped product managers easily monitor funnel health and experiment performance.

6. Tools & Techniques Used

Category	Tools/Concepts
Data Extraction	SQL (joins, window functions)
Analysis	Python (Pandas, statsmodels)
Visualization	Power BI
Statistics	Z-test for proportions
Outcome	12% improvement in completion rate and actionable funnel insights

7. Conclusion

This project demonstrated a complete **data-driven product experiment workflow** — from raw log extraction to statistical validation and visualization.
By quantifying and addressing the key friction points in onboarding, the team was able to drive measurable impact on user activation metrics.