1. Phase 1: Project Design

1.1. Problem Defining and Business Goal

1.1.1. Introduction

Our product is a subscription-based digital product that offers users a free (Free) basic tier, with the option to upgrade to Premium for access to advanced features. Historical data indicates that:

- A significant portion of Free users visit the Pricing Page,
- However, only a small percentage of them complete the upgrade to Premium.

This behavior suggests the presence of friction or lack of clarity on the pricing page, which may be preventing users from making an informed and confident purchase decision.

1.1.2. Business Problem

The low conversion rate from Free to Premium directly leads to:

- Reduced subscription revenue
- Lower user lifetime value (LTV)
- Inefficient utilization of inbound traffic to the Pricing Page

The product team hypothesizes that the current pricing page design does not sufficiently:

- Clearly communicate the value proposition of the Premium plan
- Guide users toward a quick and confident purchase decision

1.1.3. Experiment Objective

The objective of this experiment is to evaluate whether the new pricing page design (Treatment) can:

- Increase the conversion rate from Free to Premium
- Without negatively impacting user experience or other critical business metrics

More specifically, this experiment aims to answer the following question:

"Do the proposed changes to the Pricing Page lead to a statistically significant increase in the Free-to-Premium conversion rate compared to the current version?"

1.1.4. Why A/B Testing?

Since:

- The change is applied to a specific point in the funnel (the Pricing Page), and
- We aim to measure its causal impact on user behavior,

A/B testing is the most appropriate methodology to:

- Perform a fair comparison between the current version (Control) and the redesigned version (Treatment)
- Control for randomness and noise in user behavior
- Enable data-driven decision-making based on statistically valid results

1.1.5. Expected Outcome (Decision-Oriented)

At the conclusion of the experiment, one of the following decisions will be made:

- Launch: If a statistically significant and healthy uplift is observed
- Iterate: If positive signals are detected, but the results are not strong or conclusive enough
- Rollback: If the treatment shows a negative impact or no meaningful effect

1.2. Treatment Definition

- Control (A): The current version of the Pricing Page, serving as the baseline
- Treatment (B): A redesigned Pricing Page with explicit changes to plan presentation and call-to-action (CTA) elements

1.2.1. Objective of the Treatment

The Treatment is designed to reduce friction and increase clarity in order to:

- Communicate the value of the Premium plan more quickly and clearly
- Make the user decision-making process easier
- Increase CTA engagement and payment flow initiation

1.2.2. Elements Held Constant

To ensure a clean and unbiased experiment, the following elements are not modified in the Treatment:

- Price points
- Plan feature entitlements (actual included features)
- Checkout / payment flow (since the experiment focuses solely on the Pricing Page)
- User eligibility (who is allowed to purchase)

1.2.3. Exposure Definition

For accurate analysis, we must clearly define what it means for a user to have "seen" the pricing page:

- The user must visit the Pricing Page (page view is recorded)
- The assigned variant must be explicitly identified (variant = A or B)
- The timestamp of the first exposure per user must be captured

1.3. Population Definition

1.3.1. What is population in this project?

The experiment population includes all users who meet the following criteria:

- Their subscription status is Free
- They visit the Pricing Page
- The visit occurs within the experiment window
- They are exposed for the first time during the experiment to either the Control or Treatment variant

1.3.2. Preventing Effect Dilution

If all Free users are included in the experiment — even those who never visit the Pricing Page:

- The observed conversion rate becomes artificially low
- The treatment effect is diluted
- The required sample size increases unnecessarily

1.3.3. Preventing Bias

If we include only users who converted or highly active users in the experiment:

- Selection bias is introduced
- The results become non-generalizable to the broader Free user population

1.3.4. Exposure Window Definition

- For each user Exposure time is defined as the first timestamp at which the user views the Pricing Page
- Assignment to Control (A) or Treatment (B) occurs at this exact moment
- A conversion window of 7 days begins immediately following exposure

1.3.5. Edge Cases & Attribution Rules

- If a user views the Pricing Page multiple times, only the first exposure is counted
- If a user reverts back to Free after exposure, they remain in their original cohort
- If a user switches devices after exposure, the assigned variant remains fixed, using a persistent user_id

2. Phase 2: Metric Design

2.1. Primary Metric

2.1.1. Free → Premium Conversion Rate

Conversion Rate is defined as the percentage of eligible users (population) who upgrade to Premium within a specified time window after viewing the Pricing Page.

The denominator consists of all users who:

- Are on the Free plan
- Viewed the Pricing Page during the experiment window
- Were assigned to either variant A or B
- These users are referred to as Exposed Eligible Users

From the same exposed eligible users, the numerator includes users who:

- Within the conversion window (7 days after first exposure),
- Have at least one successful Premium upgrade event or completed purchase

$$Conversion\ Rate = \frac{Number\ of\ Users\ Who\ Converted\ to\ Premium}{Numer\ of\ Eligible\ Users\ Exposed}$$

2.2. Secondary Metric

Secondary metrics are metrics that:

- Are not the direct business end-goal
- Help us understand the mechanism of impact
- Explain why a change succeeded or failed

While the primary metric answers “Did we succeed?”, secondary metrics explain “Why did this outcome occur?”

2.2.1. Click-Through Rate (CTR) on CTA

The CTA Click-Through Rate (CTR) is defined as the proportion of users who:

- Viewed the Pricing Page, and
- Clicked the "Upgrade / Go Premium" CTA

This metric indicates whether the Treatment improves user attention and purchase motivation:

If conversion increases but CTR remains unchanged:

- The improvement is likely driven by a better checkout or payment experience

If CTR increases but conversion does not:

- There is likely friction after the CTA (e.g., in the checkout flow)

$$CTA\ /\ CTR = \frac{Number\ of\ Users\ Who\ Clicked\ on\ CTA\ Button}{Numer\ of\ User\ Who\ Viewed\ Pricing\ Page}$$

2.2.2. Checkout Start Rate

Checkout Start Rate is defined as the proportion of users who:

- Clicked the CTA, and
- Initiated the checkout process

This metric measures friction between the Pricing Page and the Checkout flow. It helps determine:

- The CTA is click-attractive but low-intent

$$Checkout\ Start\ Date = \frac{Number\ of\ Users\ Who\ Started\ Checkout}{Numer\ of\ User\ Who\ Cicked\ on\ CTA\ Button}$$

2.3. Guardrails Metric

Guardrail metrics are metrics that:

- They are not the direct objective of the experiment
- Must not degrade, even if the primary metric improves
- Help assess the risk associated with shipping a change

If the primary metric answers "Should we move forward?", guardrail metrics answer "At what cost?". Without guardrail metrics:

- Conversion may increase
- But user satisfaction may decline
- Or short-term revenue may improve while long-term LTV is damaged

2.3.1. Refund Rate

Refund Rate is defined as the percentage of users who:

- Upgraded from Free to Premium, and
- Requested a refund within a short time window (14 days)

A high refund rate indicates that:

- Users purchased with misaligned or incorrect expectations, or
- The pricing page oversold the value of the Premium plan
- Even if conversion increases, a high refund rate invalidates the true impact of the experiment

$$Refund\ Rate = \frac{Number\ of\ Refunded\ Subscription}{Numer\ of\ Premium\ Activation}$$

2.3.2. Early Churn Rate

Early Churn Rate is defined as the percentage of Premium users who:

- Cancel their subscription
- Within a short period after upgrading (30 days)

High conversion + high churn = artificial growth

Indicates that the promised value does not align with the actual user experience

$$Churn\ Rate = \frac{Number\ of\ Users\ Canceled\ Within\ 30\ days}{Numer\ of\ Premium\ Activation}$$

3. Setup Hypothesis and Testing

3.1. What is Hypothesis?

Hypotheses are the formal statements that define:

- What is being tested
- What constitutes statistical significance
- How statistical decisions will be made

An A/B test without hypotheses is merely a numerical comparison without interpretability or meaning.

3.1.1. Null Hypothesis (H_0)

H_0: The changes introduced in the Pricing Page have no effect on the Free-to-Premium conversion rate. In other words:

Any observed difference in conversion rates is purely due to random chance and noise and no true underlying effect exists

$$H_0 = CR.A_{control} = CR.B_{treatment}$$

3.1.2. Alternative Hypothesis (H_1)

H_1: The redesigned Pricing Page increases the Free-to-Premium conversion rate. In other words:

We expect the Treatment to drive an improvement not merely to be different from the Control

$$H_1 = CR.A_{control} < CR.B_{treatment}$$

3.2. Significance Level and Power of Test

3.2.1. Significance Level (α)

Significance level (α) represents the probability of making an error by concluding that the Treatment is better, when in reality no true improvement exists. In statistical terms, α corresponds to:

- A False Positive, or A Type I Error
- Rejecting H_0 when H_0 is actually true

A commonly used value is α = 0.05.

This means that if there is truly no effect, we accept a maximum 5% chance of incorrectly concluding that the Treatment has an effect.

- Being too strict (e.g., α = 0.01) → Requires a much larger sample size
- Being too lenient (e.g., α = 0.10) → Increases the risk of false positives

3.2.2. Statistical Power

Statistical power is the probability that the experiment will correctly detect an improvement when the Treatment truly has a positive effect. In other words, power represents:

- The likelihood of true positive detection
- The ability to avoid a Type II Error (False Negative)

A commonly used target is Power = 80%.

This means that if a real and practically meaningful improvement exists, the experiment has an 80% chance of detecting it.

- Higher power (e.g., 90%) → Requires a larger sample size
- Lower power → Increases the risk of missing real improvements

3.3. Define MDE

3.3.1. Minimum Detectable Effect

The smallest improvement in the primary metric that, if it truly exists, we both want and are able to detect using a statistical test. Without a clearly defined MDE:

- Sample size calculations become meaningless
- The test may run too short and fail to detect real effects
- Or run too long without delivering meaningful business value

3.3.2. Business Framing MDE

"What is the smallest increase in conversion rate that is actually meaningful for the business?". In this project:

- 0.1% uplift → Likely noise
- 5% uplift → Often unrealistically large
- 1–2% uplift → Typically reasonable for a Pricing Page experiment

3.4. Calculate Sample Size

Here we calculate the sample size:

$Current\ Conversion\ Rate = p_1 = \mathbf{0.05}$
$MDE\ (\Delta) = \mathbf{0.01}$
$Desired\ Conversion\ Rate = p_1 + MDE = 0.05 + 0.01 = \mathbf{0.06}$
$\propto = \mathbf{0.05}$
$\beta = \mathbf{0.8}$

We need to do one-tail test. Since, we have variance population known and number of data is more than 30, we use Z-score:
$Z(\propto \%5) = \mathbf{1.645}$
$Z(\beta\ \%80) = \mathbf{0.842}$

Now, we use two-proportion sample size formula:

$$n = \frac{(\sqrt{p_1(1-p_1) + p_2(1-p_2)} \times \beta \times Z_{\propto} \times \sqrt{(\bar{p}-1)2\bar{p}} + z)^2}{(1p_2 - p)^2}$$

After calculation, for each group we will have 6,426 samples. So finally, the file that we will apply test on it should include 12,852 records.

4. Data & Instrumentation

4.1. Define Data Model

In our database, we have 3 different tables.

4.1.1. ab.ssignments

The source of truth for the experiment defines:

- **user_id**: Who entered the experiment
- **variant**: Which variant was exposed
- **assignment_ts**: From what point in time the user has been in the experiment

4.1.2. ab.events

To measure the funnel and secondary metrics:

- **event_id**: Unique event identifier
- **user_id**: Who did this event
- **event_ts**: Event timestamp
- **event_name:** Event type (*pricing_page_view, click_upgrade_cta, start_checkout*)

4.1.3. ab.subscriptions

To measure the primary metric and guardrails, the following user-level fields are required

- **user_id:** User identifier
- **activated_ts**: Timestamp when the user upgraded to Premium (NULL if not converted)
- **canceled_ts**: Subscription cancellation timestamp (used for early churn)
- **refunded_flag:** Indicates whether a refund occurred (true / false)

4.2. Sanity Check Data

Sanity checks are a set of basic but essential validations to ensure that:

- The experiment was implemented correctly
- The data is not corrupted or biased
- Subsequent results will not be misleading

4.2.1. Sample Ratio Mismatch (SRM)

SRM checks whether the number of users assigned to Control (A) and Treatment (B) is approximately balanced (close to 50/50).

```sql
WITH counts AS (
    SELECT
        variant,
        COUNT(*) AS users
    FROM ab.assignments
    GROUP BY variant
),
total AS (
    SELECT SUM(users) AS total_users FROM counts
)
SELECT
    c.variant,
    c.users,
    CAST(1.0 * c.users / t.total_users AS DECIMAL(5,4)) AS pct_users
FROM counts c
CROSS JOIN total t;
```

4.2.2. Temporal Balance Between Variants

This check verifies whether users in Control (A) and Treatment (B) entered the experiment uniformly over time.

```sql
WITH daily AS (
    SELECT
        CAST(assignment_ts AS DATE) AS assignment_date,
        variant,
        COUNT(*) AS users
    FROM ab.assignments
    GROUP BY CAST(assignment_ts AS DATE), variant
),
daily_total AS (
    SELECT
        assignment_date,
        SUM(users) AS total_users
    FROM daily
    GROUP BY assignment_date
)
SELECT
    d.assignment_date,
    d.variant,
    d.users,
    CAST(1.0 * d.users / t.total_users AS DECIMAL(5,4)) AS pct_users
FROM daily d
JOIN daily_total t
    ON d.assignment_date = t.assignment_date
ORDER BY d.assignment_date, d.variant;
```

4.2.3. Variant Exclusivity Check

This check ensures that each user is exposed to only one variant. Verifies that no user has seen both Control (A) and Treatment (B)

```sql
SELECT
    user_id,
    COUNT(DISTINCT variant) AS variant_count
FROM ab.assignments
GROUP BY user_id
HAVING COUNT(DISTINCT variant) > 1;
```

4.2.4. Event Timing & Ordering Validation

This check verifies that:

- All events occur after variant assignment
- No conversion events happen before exposure

Are there any events that occurred before variant assignment

```sql
SELECT TOP 100
    e.user_id,
    a.variant,
    a.assignment_ts,
    e.event_ts,
    e.event_name
FROM ab.events e
JOIN ab.assignments a
    ON a.user_id = e.user_id
WHERE e.event_ts < a.assignment_ts
ORDER BY e.event_ts;
```

Did any user upgrade to Premium before exposure to the experiment

```sql
SELECT TOP 100
    s.user_id,
    a.variant,
    a.assignment_ts,
    s.activated_ts
FROM ab.subscriptions s
JOIN ab.assignments a
    ON a.user_id = s.user_id
WHERE s.activated_ts IS NOT NULL
  AND s.activated_ts < a.assignment_ts
ORDER BY s.activated_ts;
```

4.2.5. Funnel Logging Validation

This check ensures that the funnel is properly logged:

- Do all users have a pricing_page_view event?
- Are CTA clicks and checkout start events logically consistent?

Do all assigned users have at least one *pricing_page_view* event

```sql
SELECT COUNT(*) AS users_missing_pricing_view
FROM ab.assignments a
LEFT JOIN ab.events e
  ON e.user_id = a.user_id
 AND e.event_name = 'pricing_page_view'
WHERE e.user_id IS NULL;
```

Are there any users who clicked the CTA without first viewing the Pricing Page

```sql
SELECT COUNT(*) AS clicks_without_view
FROM ab.events c
WHERE c.event_name = 'click_upgrade_cta'
  AND NOT EXISTS (
      SELECT 1
      FROM ab.events v
      WHERE v.user_id = c.user_id
        AND v.event_name = 'pricing_page_view'
        AND v.event_ts <= c.event_ts
  );
```

Are there any users who started the checkout process without clicking the CTA

```sql
SELECT COUNT(*) AS checkouts_without_click
FROM ab.events s
WHERE s.event_name = 'start_checkout'
  AND NOT EXISTS (
      SELECT 1
      FROM ab.events c
      WHERE c.user_id = s.user_id
        AND c.event_name = 'click_upgrade_cta'
        AND c.event_ts <= s.event_ts
  );
```

For each user, does the event sequence follow the expected order: *pricing_page_view* → *click_upgrade_cta* → *start_checkout*?

```sql
WITH steps AS (
  SELECT
    user_id,
    MIN(CASE WHEN event_name='pricing_page_view' THEN event_ts END) AS view_ts,
    MIN(CASE WHEN event_name='click_upgrade_cta' THEN event_ts END) AS click_ts,
    MIN(CASE WHEN event_name='start_checkout' THEN event_ts END) AS checkout_ts
  FROM ab.events
  GROUP BY user_id
)
SELECT TOP 100 *
FROM steps
WHERE
  (click_ts IS NOT NULL AND view_ts IS NOT NULL AND click_ts < view_ts)
  OR
  (checkout_ts IS NOT NULL AND click_ts IS NOT NULL AND checkout_ts < click_ts)
ORDER BY user_id;
```

4.2.6. Duplicate & Missing Data Checks

These checks validate basic data quality to ensure experiment results are not biased by corrupted or incomplete data.

Is any *user_id* assigned more than once in the experiment assignment table

```sql
SELECT
    user_id,
    COUNT(*) AS cnt
FROM ab.assignments
GROUP BY user_id
HAVING COUNT(*) > 1;
```

Are there any events with missing critical fields such as *event_ts* or *event_name*

```sql
SELECT
    COUNT(*) AS invalid_events
FROM ab.events
WHERE event_ts IS NULL
   OR event_name IS NULL
   OR user_id IS NULL;
```

Are there any events that do not belong to a valid user

```sql
SELECT COUNT(*) AS orphan_events
FROM ab.events e
LEFT JOIN ab.assignments a
  ON a.user_id = e.user_id
WHERE a.user_id IS NULL;
```

Are there any Premium activation records (*activated_ts*) that do not belong to a valid user?

```sql
SELECT COUNT(*) AS orphan_subscriptions
FROM ab.subscriptions s
LEFT JOIN ab.assignments a
  ON a.user_id = s.user_id
WHERE a.user_id IS NULL;
```

Are there any users who have a cancellation or refund recorded without ever having a Premium activation?

```sql
SELECT COUNT(*) AS inconsistent_subscriptions
FROM ab.subscriptions
WHERE activated_ts IS NULL
  AND (canceled_ts IS NOT NULL OR refunded_flag = 1);
```

5. Apply A/B Testing & Analysis

5.1. Analysis Primary Metric

First, we extract the needed data from different tables in database:

Metric Name	Technical Definition (Binary)
converted_7d	1 if the user: Upgraded to Premium• After assignment• Within 7 days 0 otherwise
clicked_cta_7d	1 if the user: Clicked "Upgrade / Go Premium" CTA After assignment Within 7 days 0 otherwise
started_checkout_7d	1 if the user: Started the checkout process After assignment Within 7 days 0 otherwise
refunded_14d	1 if the user: Upgraded to Premium Requested a refund Within 14 days
early_churn_30d	1 if the user: Canceled their subscription Within 30 days after upgrading 0 otherwise

First of all, we calculate the conversion rate for each group. All the analysis are being done, in Python.

variant	users	converted	conversion_rate
A	6426	302	0.046997
B	6426	329	0.051198

Now, we calculate the Absolute uplift and confidence interval:

$$CR_B - CR_A$$

```
np.float64(0.0042016806722689065)
```

So, result shows, conversion rate increased, but we should test it with A/B testing to make sure this increase is real and is not by chance.

Alright—here we want to use a two-proportion z-test to evaluate whether the difference in conversion rates between variants A and B is truly statistically significant or could simply be due to random chance. Since the alternative hypothesis (H_1) is defined as directional (B > A), we conduct a one-tailed test.

```
the p-value is: 0.1351759686677776
```

Since p-value is greater than alpha (0.05), we there is not enough evidence to reject the null hypothesis and it means, that difference is by chance and changing in pricing page, does not improved the conversion rate.

5.2. Analysis Guardrail Metric

5.2.1. Refund Rate

The null hypothesis and alternate hypothesis for this metric is:

$$H_0 = Refund\ Rate.A_{control} = Refund\ Rate.B_{treatment}$$
$$H_1 = Refund\ Rate.A_{control} \neq Refund\ Rate.B_{treatment}$$

```
Two-tailed p-value for refunded_14d: 0.02531883647068489
```

	variant	users	refunded	refund_rate
0	A	6426	0	0.000000
1	B	6426	5	0.000778

As we see, p-value is less then alpha (0.05), we reject null hypotheses.

- *The premium users do not experience as they expected and in 14 days, they would like to refund their money. It seems the new pricing page, is telling many things which are beyond what product is in action.*

5.2.2. Churn Rate

The null hypothesis and alternate hypothesis for this metric is:

$$H_0 = Churn\ Rate.A_{control} = Churn\ Rate.B_{treatment}$$
$$H_1 = Churn\ Rate.A_{control} \neq Churn\ Rate.B_{treatment}$$

Two-tailed p-value for early_churn_30d: 0.25336314490851963

	variant	users	churned	churn_rate
0	A	6426	302	0.046997
1	B	6426	330	0.051354

As we see, p-value is more then alpha (0.05), so we cannot reject null hypothesis.

- *The new design page did not affect on churn rate.*

5.3. Analysis Secondary Metric

5.3.1. CTA Click-through Rate (CTR)

The null hypothesis and alternate hypothesis for this metric is:

$$H_0 = Click\ Rate.A_{control} = Click\ Rate.B_{treatment}$$
$$H_1 = Click\ Rate.A_{control} \neq Click\ Rate.B_{treatment}$$

Two-tailed p-value for clicked_cta_7d: 0.0

	variant	users	clicked	click_rate
0	A	6426	2583	0.401961
1	B	6426	3133	0.487551

As we see, p-value is less then alpha (0.05), we reject null hypotheses.

- *The click rate after changing the pricing page, increased.*