A/B Testing Analysis

Harisankar Kartha

Introduction

- **Context:** An investment management company conducted an A/B experiment to enhance their user experience by implementing a new UI.

- **Objective**:
  - Analyze the effectiveness of the new UI design vs. the existing one by uncovering actionable insights.
  - Use rigorous statistical analysis and testing.

A/B Experiment

- An A/B test was set into motion from 3/15/2017 to 6/20/2017 by the company.
- **Control Group**: Clients interacted with the company's traditional online process.
- **Test Group**: Clients experienced the new, spruced-up digital interface.
- Both groups navigated through an identical process sequence: an initial page, three subsequent steps, and finally, a confirmation page signaling process completion.

Data

- **Client Profiles (df_final_demo):** Demographics like age, gender, and account details of the clients
  - Size and dimensionality: 70,609 rows × 9 columns
- **Digital Footprints (df_final_web_data):** A detailed trace of client interactions online, divided into two parts: pt_1 and pt_2
  - Size and dimensionality of the combined df: 755,405 rows × 5 columns
- **Experiment Roster (df_final_experiment_clients):** A list revealing which clients were part of the grand experiment
  - Size and dimensionality: 70,609 rows × 2 columns

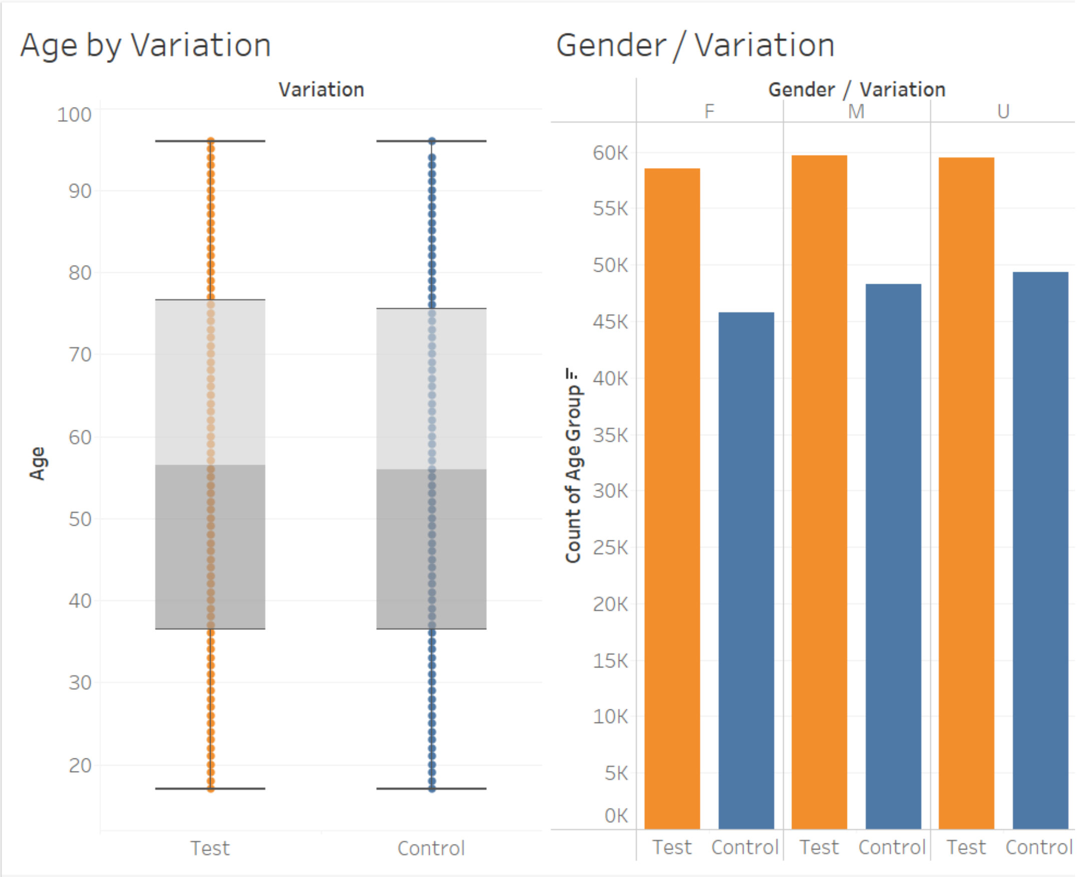
Data Preprocessing and Cleaning

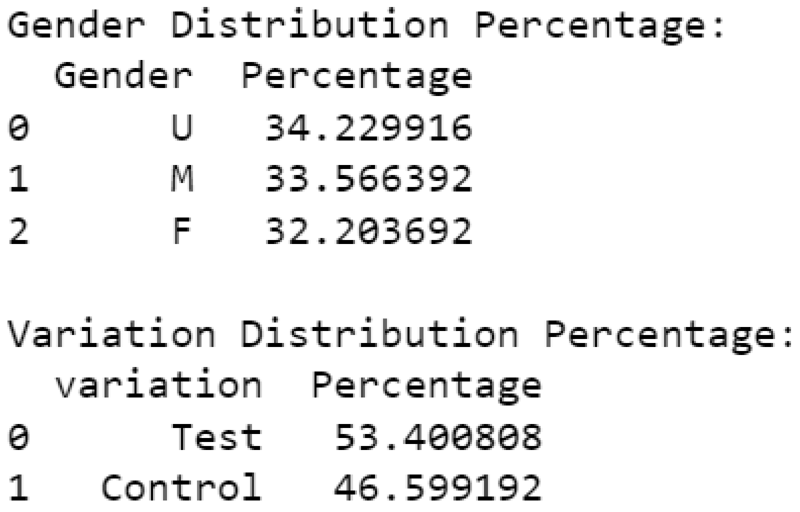
- Renamed columns for better readability
- Dropped clients with more than 5 null values (15 instances)
- Filled age null values with the mean value
- Assigned the 3 clients with gender ‘X’ to gender ‘U’
- Merged client profiles and experiment roster datasets

Exploratory Data Analysis

- On average, clients have been with the company for around **12 years** and 150 months. The standard deviation is 7 years (wide range).
- The average age of clients is approximately **47 years**. The standard deviation is 15 years (moderate spread).
- On average, clients have about **2 accounts**.
- The average balance across all clients is approximately **USD 149,515**. The standard deviation is USD 302,036.40 (wide range).
- On average, clients made around **3 calls** and logged in approximately **6 times** in the last 6 months.

Exploratory Data Analysis

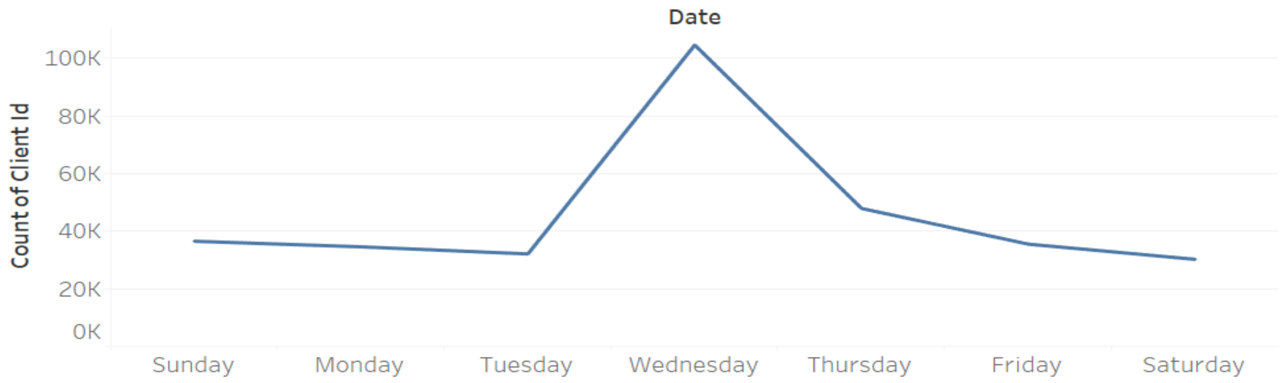




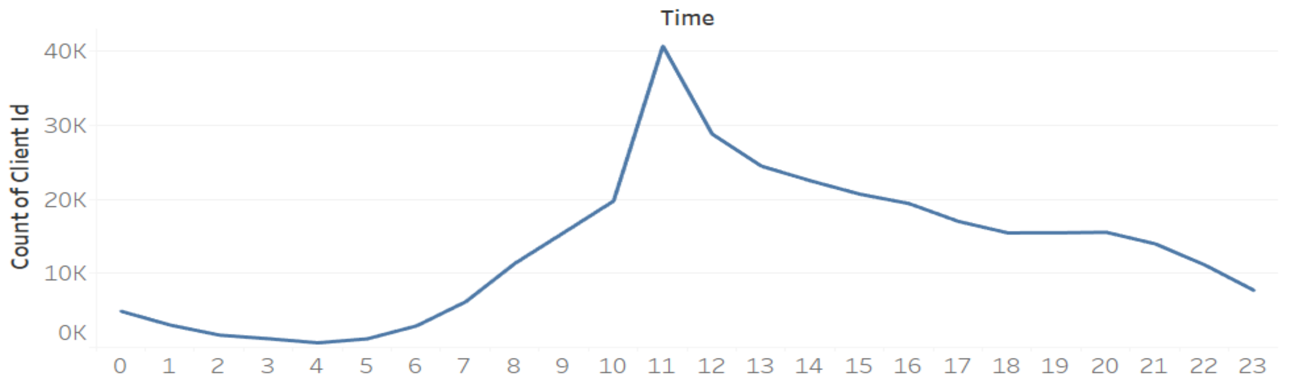
```
Gender Distribution Percentage:
  Gender  Percentage
0      U   34.229916
1      M   33.566392
2      F   32.203692

Variation Distribution Percentage:
  variation  Percentage
0      Test   53.400808
1   Control   46.599192
```

Exploratory Data Analysis

Visits per day of the week

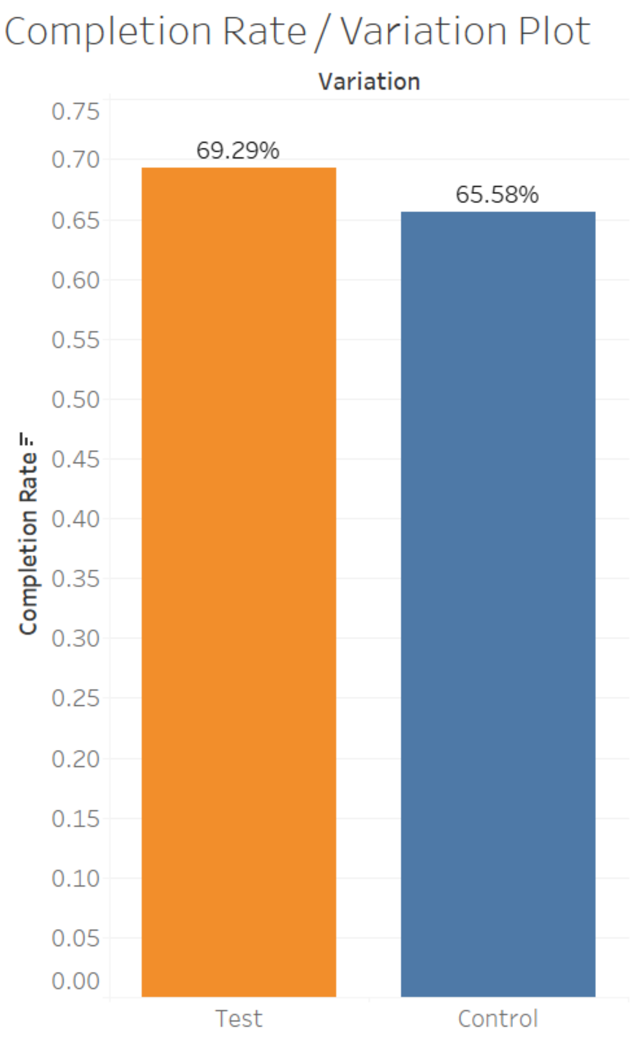


Visits per hour of the day



Performance Metrics

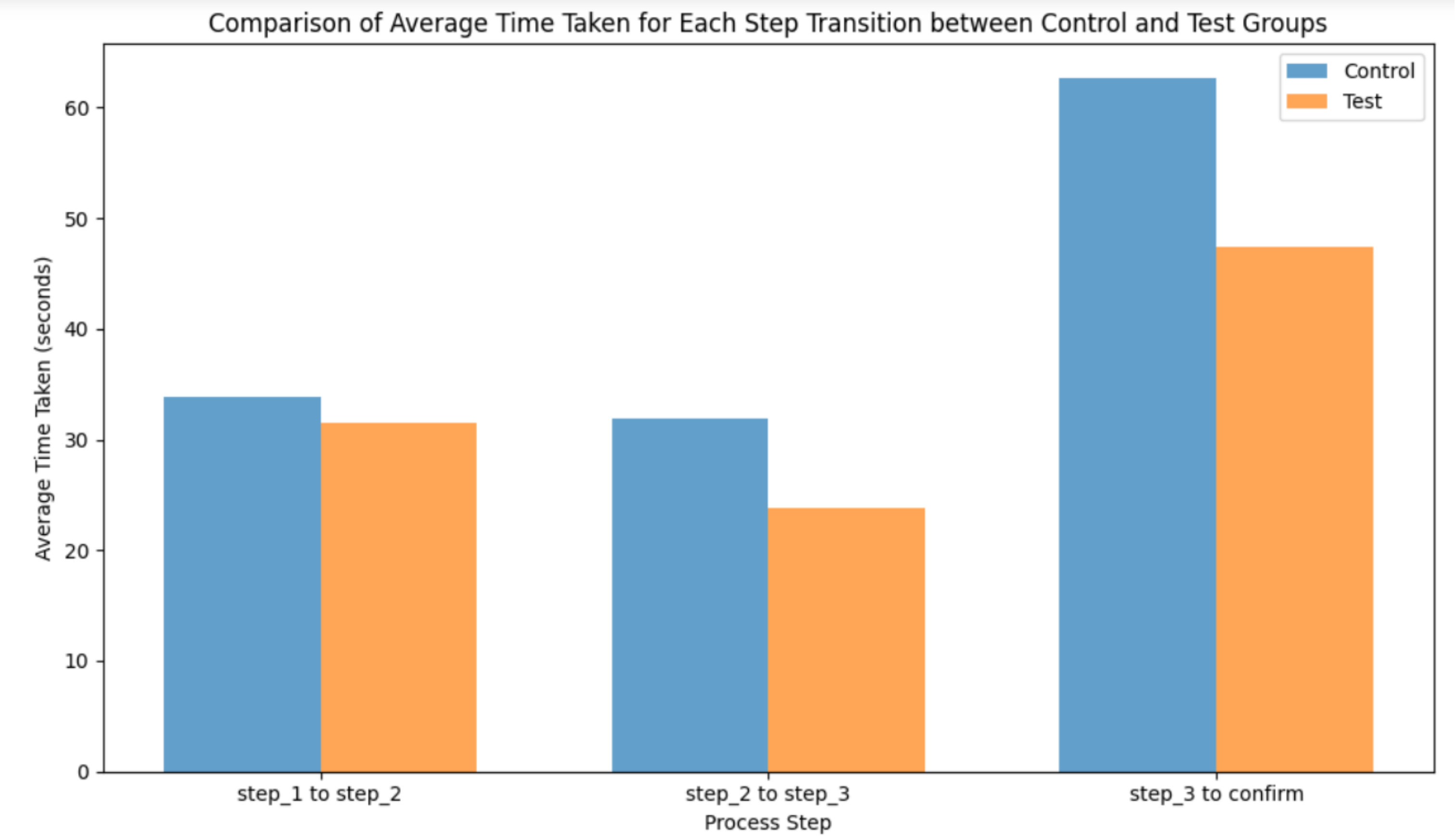
- **Completion Rate:** The proportion of users who reach the final ‘confirm’ step.
  - Completion Rate by Age Group
  - Completion Rate by Gender
- **Time Spent on Each Step:** The average duration users spend on each step.
- **Error Rates:** If there’s a step where users go back to a previous step, it may indicate confusion or an error (clients moving from a later step to an earlier one).
- **Step Abandonment Rate:** The proportion of users who abandon the process at each step. This will help identify specific steps where users are most likely to drop off.



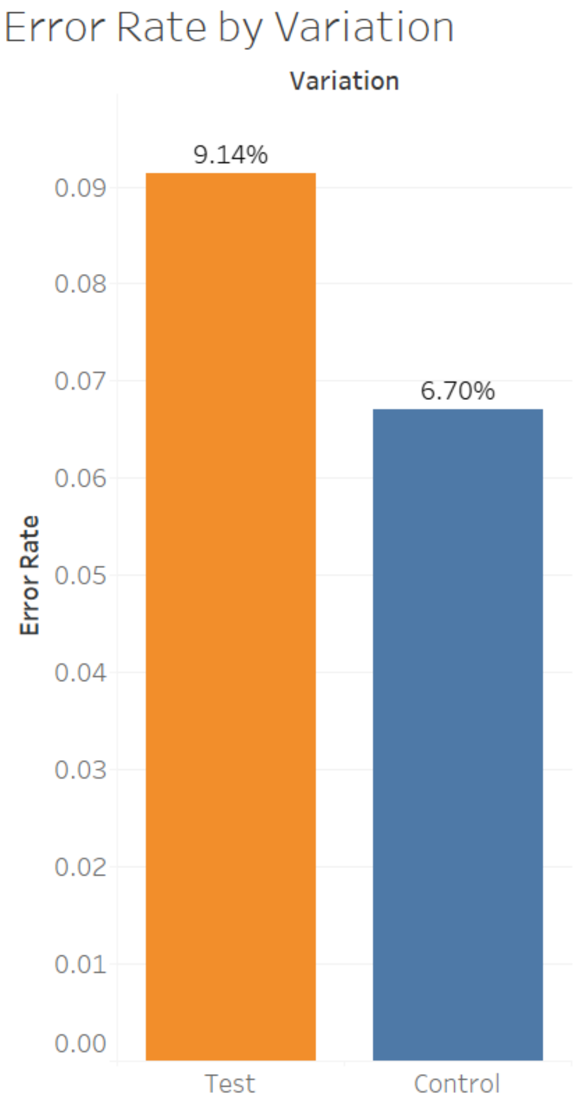
Completion Rate

- The test group has a higher completion rate (69.29%) compared to the control group (65.58%). This indicates that the variation being tested is more effective in guiding users to complete the process and reach the final 'confirm' step.
- Percentage increase in the completion rate from the control group to the test group: 3.71%

Time Spent on Each Step



The Test variation shows a decrease in the average total time taken compared to the Control variation (102.79 vs 128.36 seconds)

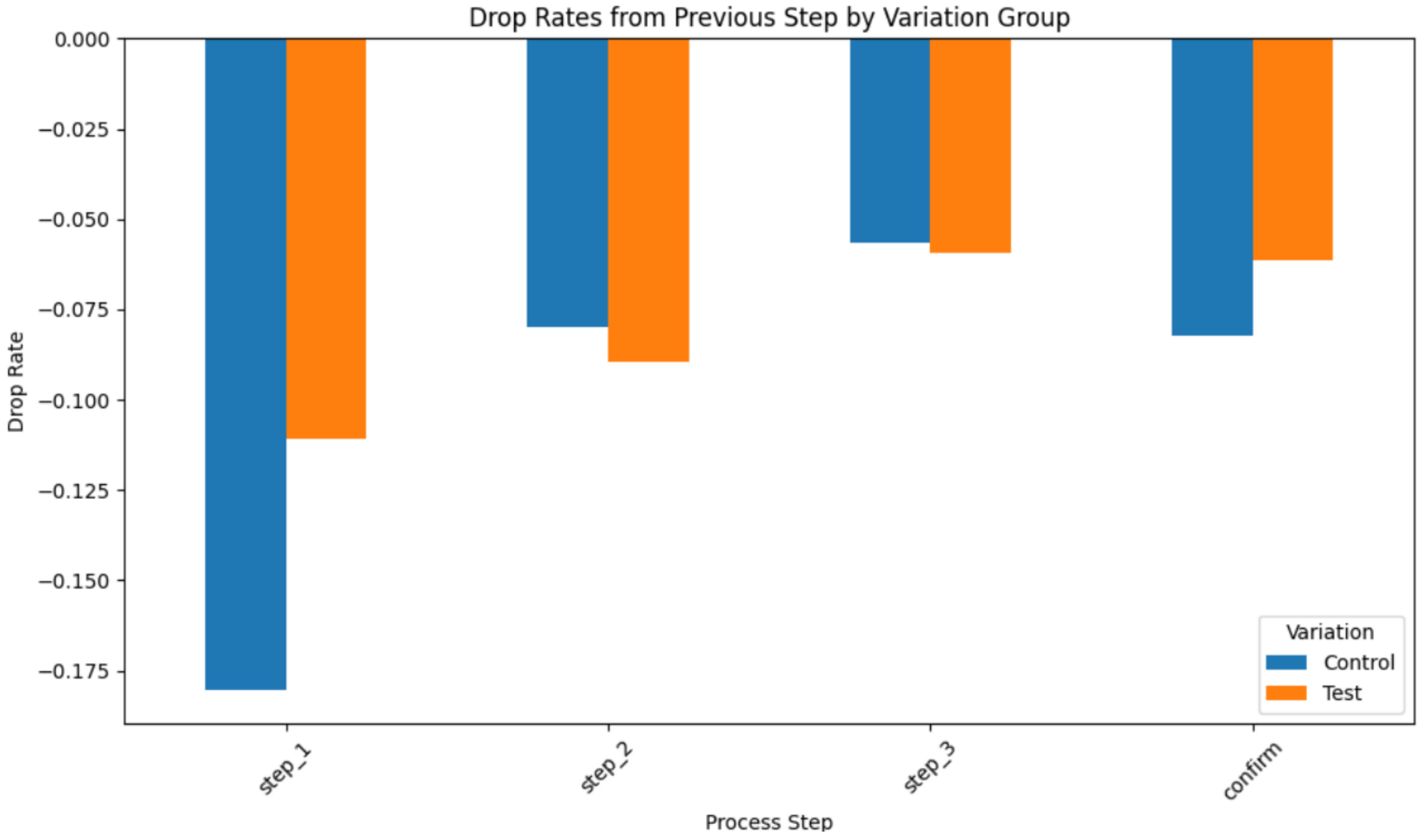


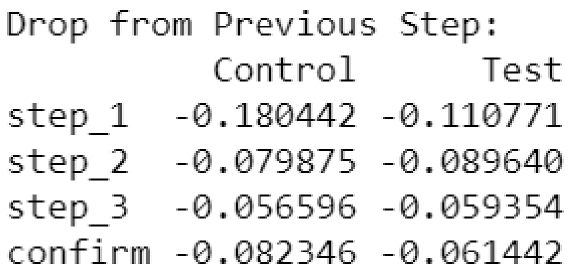
Error Rate

- **Higher Error Rate in Test Group:**
The test group has a higher error rate (9.14%) compared to the control group (6.70%).

- **User Experience Insight:**
The higher error rate in the test group is a signal that the new design, process, or feature being tested could need adjustments to improve user experience and reduce confusion.

Step Abandonment Rate





```
Drop from Previous Step:
             Control      Test
step_1   -0.180442 -0.110771
step_2   -0.079875 -0.089640
step_3   -0.056596 -0.059354
confirm  -0.082346 -0.061442
```

Hypothesis Testing

Hypothesis I: Test and Control Group Completion Rate Equality Assessment

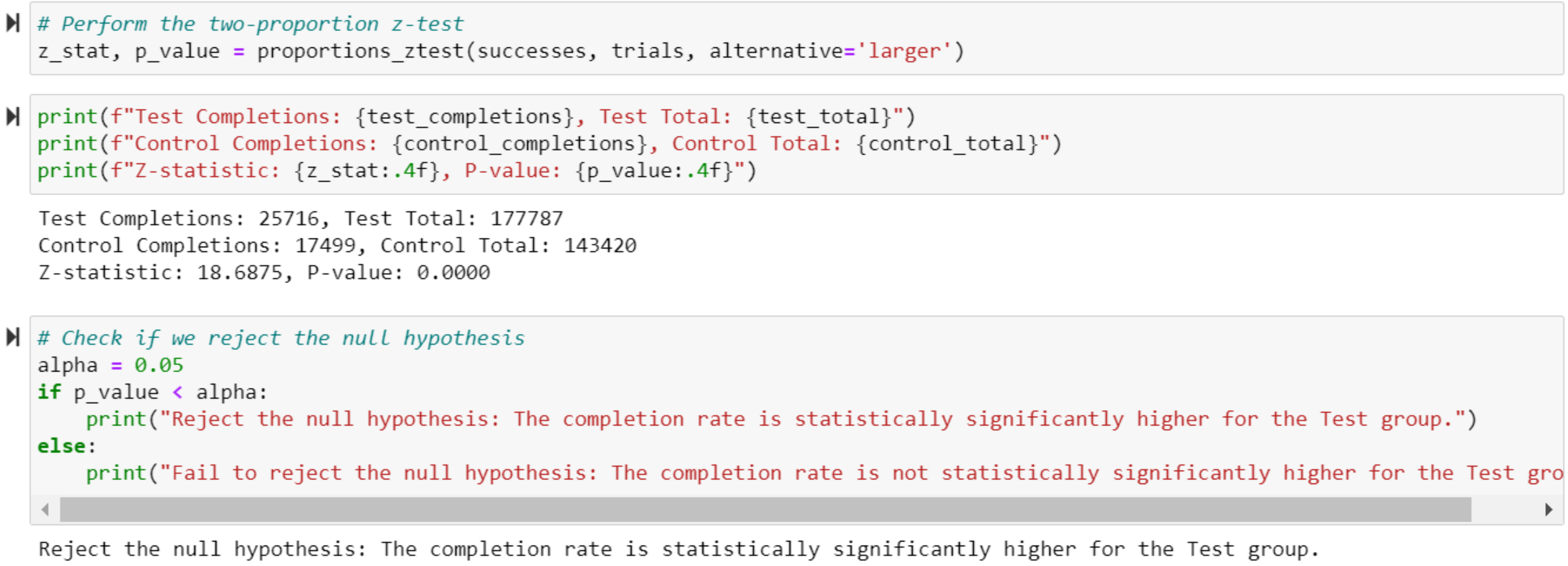
- **Null Hypothesis (H0):**

The completion rates for the Test and Control groups are equal.

- **Alternative Hypothesis (H1):**

The completion rates for the Test group are higher than those for the Control group.

Hypothesis I



```
# Perform the two-proportion z-test
z_stat, p_value = proportions_ztest(successes, trials, alternative='larger')
```

```
print(f"Test Completions: {test_completions}, Test Total: {test_total}")
print(f"Control Completions: {control_completions}, Control Total: {control_total}")
print(f"Z-statistic: {z_stat:.4f}, P-value: {p_value:.4f}")
```

```
Test Completions: 25716, Test Total: 177787
Control Completions: 17499, Control Total: 143420
Z-statistic: 18.6875, P-value: 0.0000
```

```
# Check if we reject the null hypothesis
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The completion rate is statistically significantly higher for the Test group.")
else:
    print("Fail to reject the null hypothesis: The completion rate is not statistically significantly higher for the Test gro
```

```
Reject the null hypothesis: The completion rate is statistically significantly higher for the Test group.
```

Cost of New UI

- The introduction of a new UI design comes with its associated costs, To justify these costs, the company has determined that any new design should lead to a minimum increase in the completion rate at 5% to be deemed cost-effective.

- If the new design doesn’t lead to at least this level of improvement, it may not be justifiable from a cost perspective, regardless of its statistical significance.

Hypothesis Testing

Hypothesis II: Evaluating Completion Rate Differential Between Test and Control Groups (5% threshold)

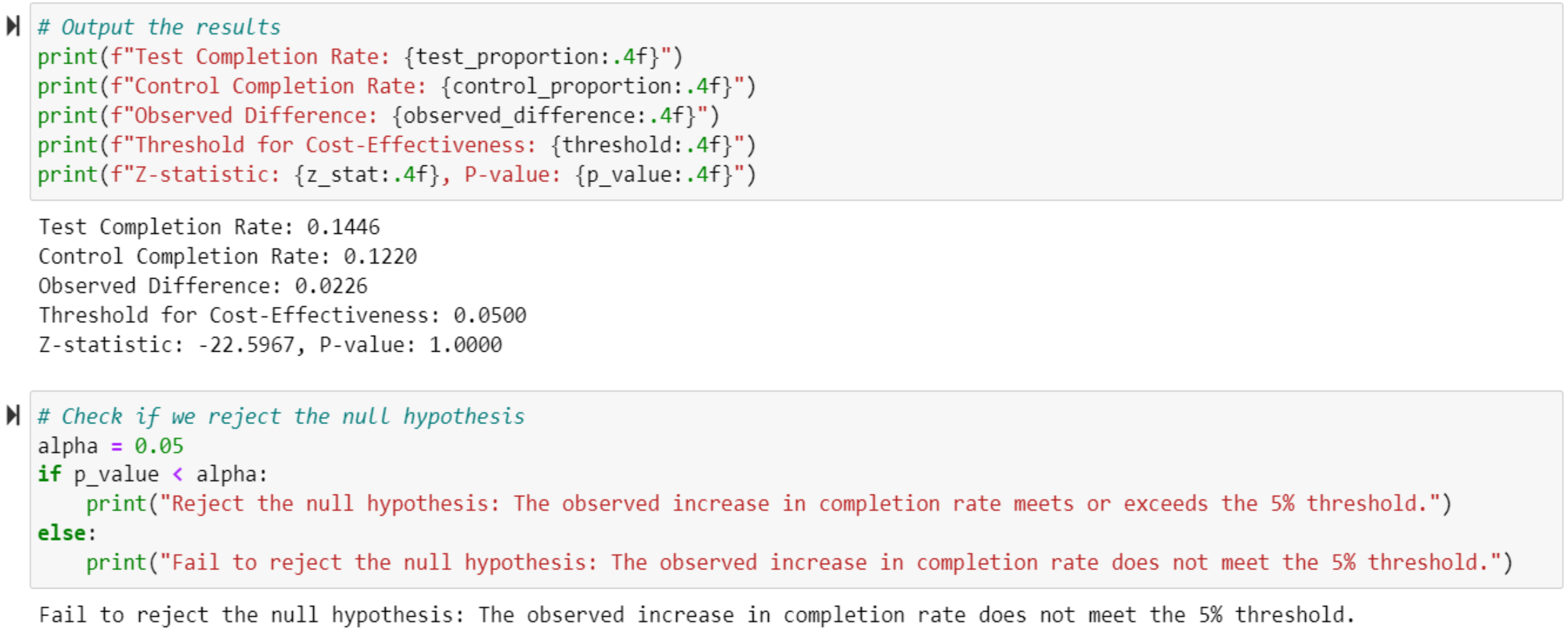
- **Null Hypothesis (H0):**

The increase in completion rate for the Test group compared to the Control group is less than 5%.

- **Alternative Hypothesis (H1):**

The increase in completion rate for the Test group compared to the Control group is at least 5%.

Hypothesis II



```
# Output the results
print(f"Test Completion Rate: {test_proportion:.4f}")
print(f"Control Completion Rate: {control_proportion:.4f}")
print(f"Observed Difference: {observed_difference:.4f}")
print(f"Threshold for Cost-Effectiveness: {threshold:.4f}")
print(f"Z-statistic: {z_stat:.4f}, P-value: {p_value:.4f}")
```

```
Test Completion Rate: 0.1446
Control Completion Rate: 0.1220
Observed Difference: 0.0226
Threshold for Cost-Effectiveness: 0.0500
Z-statistic: -22.5967, P-value: 1.0000
```

```
# Check if we reject the null hypothesis
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The observed increase in completion rate meets or exceeds the 5% threshold.")
else:
    print("Fail to reject the null hypothesis: The observed increase in completion rate does not meet the 5% threshold.")
```

```
Fail to reject the null hypothesis: The observed increase in completion rate does not meet the 5% threshold.
```

Hypothesis Testing

Hypothesis III: Assessing Systematic Differences in Group Distribution

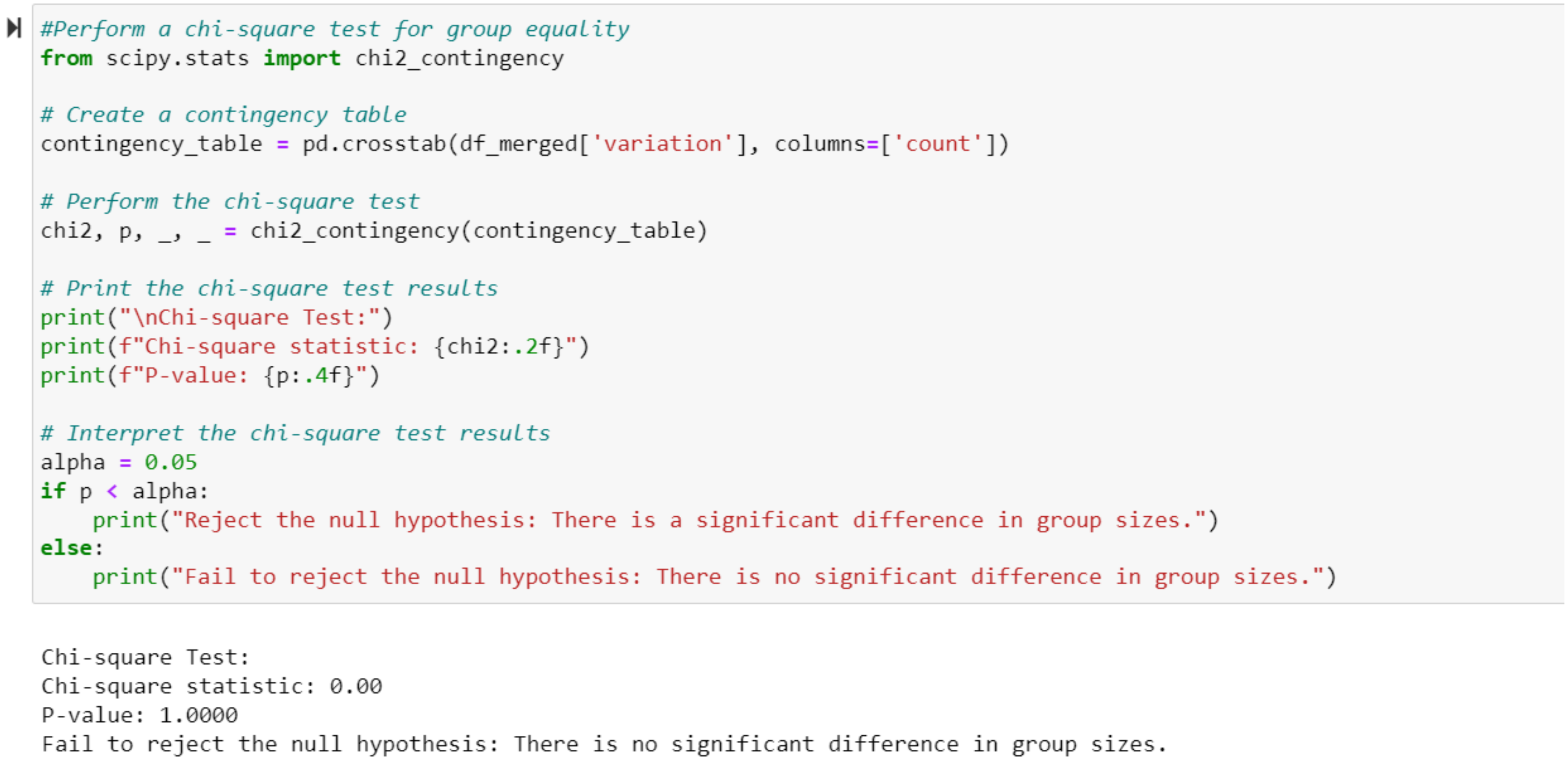
- **Null Hypothesis (H0):**

The distribution of clients between the Test and Control groups is equal, indicating that there is no systematic difference in the allocation of clients to the two groups.

- **Alternative Hypothesis (H1):**

The distribution of clients between the Test and Control groups is not equal, suggesting that there is a systematic difference in the allocation of clients, possibly indicating non-random assignment or unequal group sizes.

Hypothesis III



```
#Perform a chi-square test for group equality
from scipy.stats import chi2_contingency

# Create a contingency table
contingency_table = pd.crosstab(df_merged['variation'], columns=['count'])

# Perform the chi-square test
chi2, p, _, _ = chi2_contingency(contingency_table)

# Print the chi-square test results
print("\nChi-square Test:")
print(f"Chi-square statistic: {chi2:.2f}")
print(f"P-value: {p:.4f}")

# Interpret the chi-square test results
alpha = 0.05
if p < alpha:
    print("Reject the null hypothesis: There is a significant difference in group sizes.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference in group sizes.")
```

```
Chi-square Test:
Chi-square statistic: 0.00
P-value: 1.0000
Fail to reject the null hypothesis: There is no significant difference in group sizes.
```

Summary

- The test variation demonstrates higher completion rates, shorter average time spent, and higher step completion rates compared to the control group, suggesting the test design is more effective in guiding users through the process and reducing drop-offs. However, the higher error rate in the test group indicates a need for adjustments to improve user experience and reduce confusion.

Conclusion

- The new design demonstrates a statistically significant higher completion rate, but it does not meet the required 5% increase threshold set by the organization. Despite the improvements in completion rates and user engagement, the new design may not be justifiable from a cost perspective, given that it does not meet the minimum increase threshold for cost-effectiveness. Further optimization and evaluation are necessary to ensure that the new design can achieve at least a 5% increase in completion rate to justify the associated costs.

Experiment Evaluation

- **Well-Structured Experiment**: Clear objectives and defined metrics for success, with key performance indicators (KPIs) like completion rates selected for measurement in an A/B test setup.

- **Randomization and Equality**: Random assignment ensured equal division between test and control groups, mitigating selection bias, and hypothesis testing showed no significant difference in group sizes.

- **Duration Assessment**: Conducted over three months, from March 15 to June 20, 2017, the experiment's timeframe could potentially reveal temporal effects and seasonality, offering deeper insights into the long-term impact of design changes on user behavior.