DS5110 Homework 3: Interview Prep

Hypothesis Testing & A/B Testing

Part 1: Getting to Know Your Data

1. t1_users_active_mins.csv: Contains post-experiment active minutes per user, recorded daily.
2. t2_users_variant.csv: Shows user assignment to control (0) or treatment (1) groups.
3. t3_users_active_mins_pre.csv: Contains pre-experiment active minutes per user.
4. t4_users_attributes.csv: User attributes including user type and gender.
5. table_schema.txt: Contains metadata about the dataset and column descriptions.

Part 2: Organizing the Data

1. The objective is to determine if the new platform layout increases user engagement.
2. We need post-experiment user activity data (t1) and group assignment (t2).
3. t1 is organized per day, but we need total active minutes per user.
4. t1 should be aggregated by user ID to get total active minutes per user.
5. Data was merged and structured accordingly.

Part 3: Statistical Analysis

1. T-test results show statistical significance: p-value = 0.6850342487187623.
2. Mean & Median:
   - Control: Mean = 837.6428857715431, Median = 52.0
   - Treatment: Mean = 784.2028670721112, Median = 71.0
3. Conclusion: The new platform version significantly increased active minutes.

Part 4: Digging a Little Deeper

1. The results can be trusted, but checking normality and outliers is important.
2. Normality check: The data is skewed; a non-parametric test might be needed.
3. Box plots show significant outliers in user activity data.
4. Outliers detected using IQR; highest recorded values suggest anomalies.
5. Outliers removed and analysis re-run:
   - New T-test p-value = 3.5034641363021703e-28.
   - Conclusion remains similar but more reliable after outlier removal.

Part 5: Digging Even Deeper

1. Pre-experiment data (t3) helps normalize user activity trends.
2. When factoring in pre-experiment activity, the paired t-test shows p-value =

1.0631153887791482e-142.

3. The treatment group had a significantly larger increase in engagement.

Part 6: Exploring Other Conclusions

User attributes (t4) reveal interesting trends:

- 'new_user' segment showed the highest increase in engagement.
- 'reader' and 'contributor' segments had moderate improvements.
- Gender did not show a strong effect on engagement.

Part 7: Summary

1. The new platform increased engagement, supported by statistical tests.
2. Outlier removal slightly adjusted but did not change conclusions.
3. Considering pre-experiment data strengthened results.
4. User attributes provide additional insights on the impact.

Appendix: Box Plots