Assessing the Impact of Guided Tutorials on User Behavior: An A/B Test in Interactive Data Analysis App

Team 8: Dailin Song (ds4354), Yi Lu (yl5735), Ruoshi Zhang (rz2699)

Deployed Web App: https://ruoshi-zhang.shinyapps.io/Project3_AB_test/

GitHub Repository: https://github.com/JamesSSSong/5243Team8_Project3/tree/main

Introduction

This project implements an A/B test using a Shiny web application called Interactive Data Analysis App. Users were randomly assigned to each version via URL query parameters, and user engagement was tracked using Google Analytics. Key events, including button clicks, bounce rate, and average session time were recorded and compared to determine which version performed better.

Research Question: Does a guided tutorial interface (Group A) lead to greater user engagement compared to the simplified version (Group B)?

Experimental Design & Methodology

- **Independent Variable**: Random group assignment (A or B)
- **Dependent Variable**: Button click counts, bounce rate, average session duration

To fulfill random assignment, we used JavaScript to assign each user a group and redirect the URL using `?group=A` or `?group=B`. Furthermore, Google Analytics (GA) is integrated into R Shiny to track each user's activities. We customized events for tracking button clicks, such as `LoadData`, `ProcessData`, `SavePCA`, `ApplyFeatureSelection`, `SaveNewFeature`, `EnableHistogram`, and other EDA graphing buttons. In addition, two levels of grouping, `group` (event-scoped) and `ab_group` (user-scoped) were also sent to GA for further use.

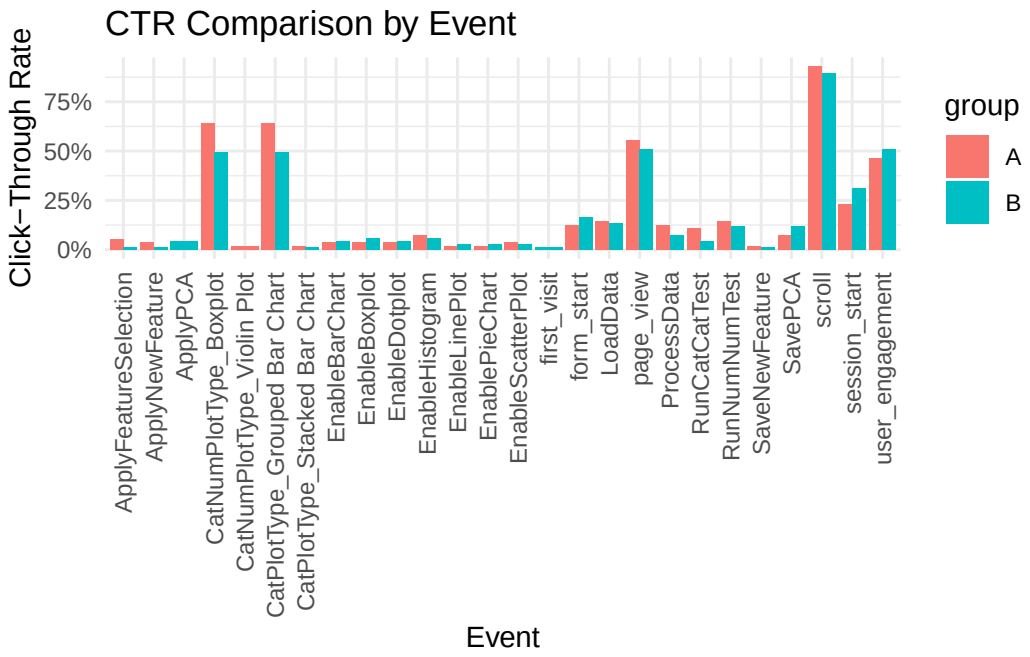
Data Collection

The deployed web link was emailed to students from STAT 4243/5243 and statistics majors. Data are collected through GA. It is important to note that GA only provides records in terms of groups and there are no individual user records. The metrics under the dimension of groups and events being considered are:

- Sessions: The number of sessions.
- Event count: The number of times a specific action was logged by users from different groups.
- Total users: The total count of distinct users triggered an event.
- Bounce rate: The percentage of sessions where users did not actively engage.
- Average session duration: The average time (in seconds) users spent per session.

Statistical Analysis & Results

1. Compare click-through rate (CTR) between two groups



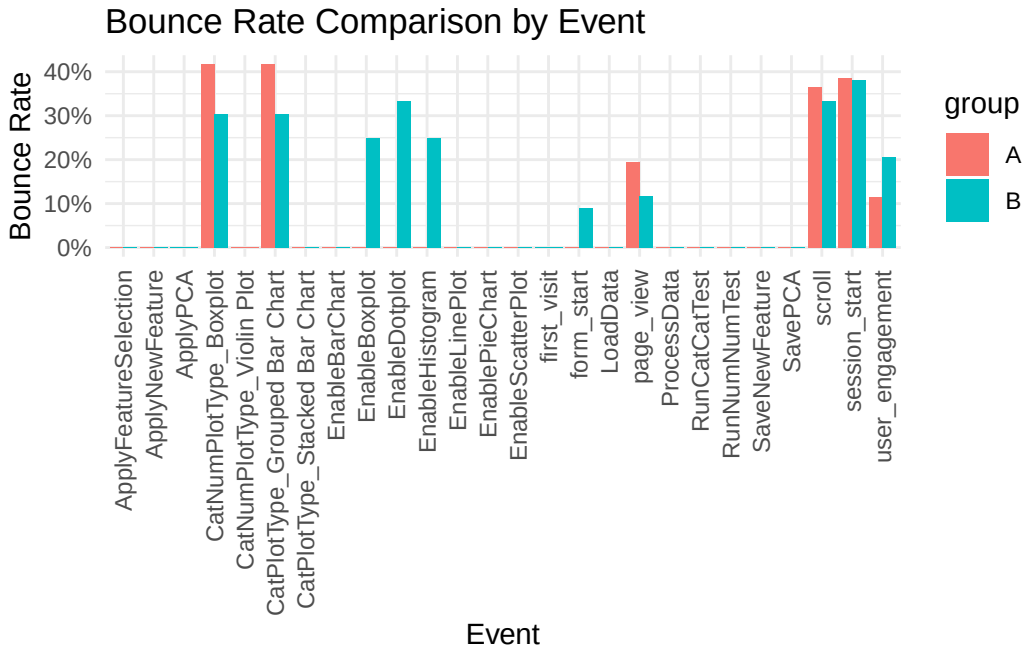
Given $CTR = \frac{session}{totalnumberofsession}$, the graph above displays Group A and B's CTR of every event. `user_engagement` and `scroll` have high rates in both groups, indicating that users are generally engaged with the content. Users from Group A who click through the data

preprocessing section with a higher rate. Some buttons from the EDA section, such as boxplot for categorical vs. numerical variables and bar chart for bivariate categorical variables, also show higher rates in Group A. The other events seem to have similar CTRs for both groups.

A chi-squared test was applied to whether whether CTRs were different for Group A and B. The results show that only two buttons from the EDA section, `CatNumPlotType_Boxplot` and `CatPlotType_Grouped Bar Chart`, have some significant differences while all the others did not.

Since the sample size is quite small, we further tried Fisher's exact test and Bootstrap to see the difference but the results were the same as the chi-squared test. We can conclude that there is no evidence supporting that the general CTR is different between the two groups whereas only some events were clicked at significantly different rates.

2. Compare bounce rate between two groups

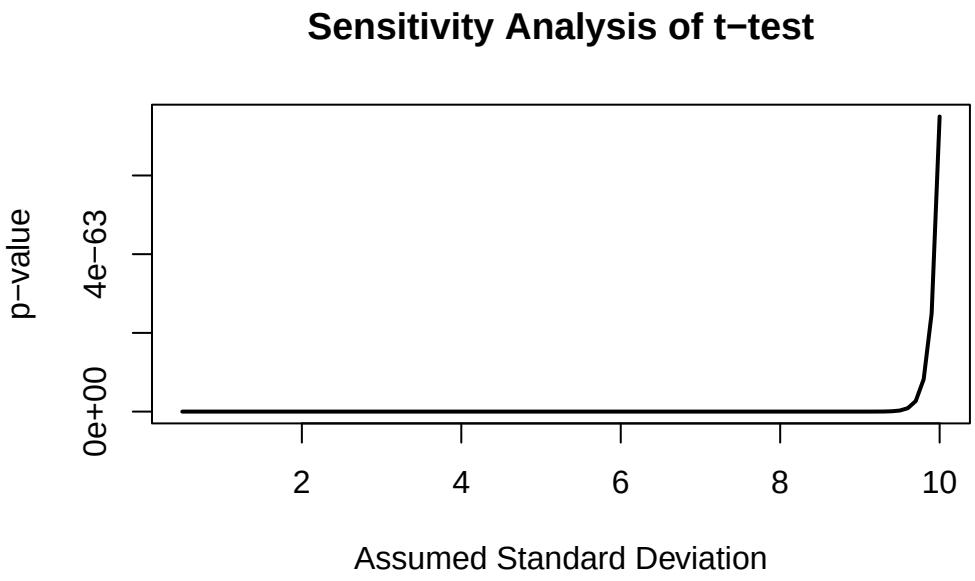


The plots reveal that Group A has noticeably higher bounce rates on events like `CatNumPlotType_Boxplot` and `CatNumPlotType_Violin Plot` and Group B shows higher bounce rates on events `EnableBoxplot`, `EnableDotplot`, and `EnableHistogram`. Other customized events all remain at a rate of 0 meaning there is no bouncing happened.

By implementing the chi-squared test and Fisher's exact test, a significant difference in bounce rates can be observed on events: `CatNumPlotType_Boxplot`, `CatPlotType_Grouped Bar Chart`, `EnableBoxplot`, `EnableDotplot`, `EnableHistogram`, and `form_start`.

Group B shows a clearly higher bar for Group B on `form_start` than Group A, which conveys that users in Group B are more likely to abandon the page shortly after their first interaction with a form. In the EDA section, both groups have users who leave the page after looking at certain graphical displays.

3. Compare total average session duration between two groups



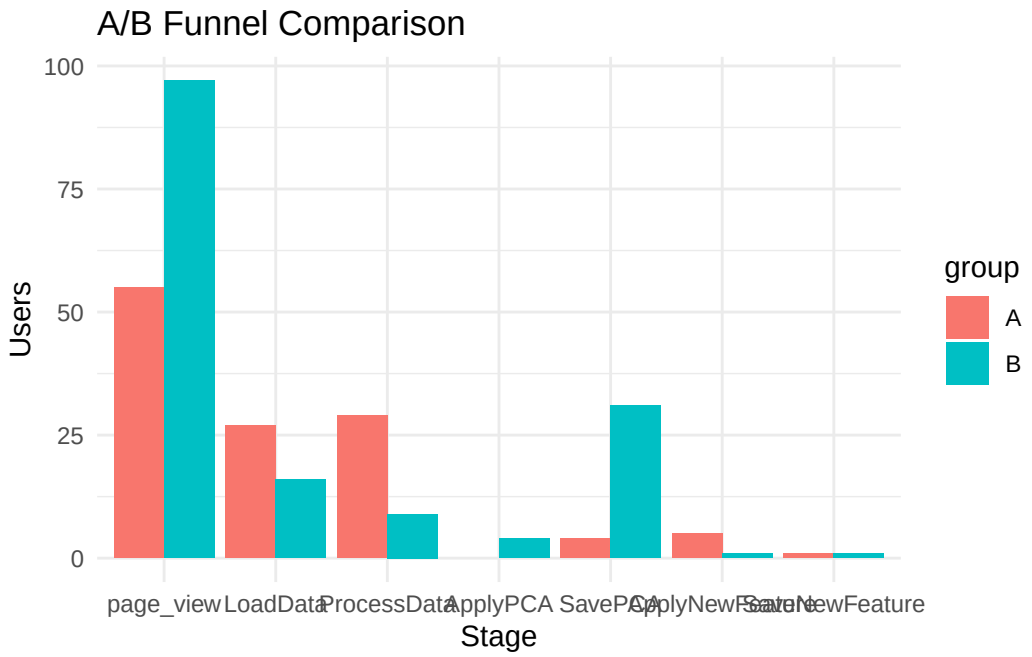

```
    Welch Two Sample t-test

data:  x and y
t = -173.77, df = 116.15, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -61.00606 -59.63108
sample estimates:
mean of x mean of y
 269.2713  329.5899
```

Before implementing statistical tests for means, a sensitivity analysis was first conducted by varying the assumed standard deviation between 0.5 and 10. Across the full range, the p-value remained effectively zero, indicating that the t-test result is highly robust to uncertainty in the variance estimate.

A two-sample t-test is applied to assess whether the average session time differs between Group A and B with a standard deviation setting to 2. The extremely small p-value indicates a highly significant difference between the two group means. Moreover, the negative t-statistics and interval mean confirm that Group B has a significantly higher mean than Group A.

4. Other observations



We further built a funnel analysis to compare how users in Groups A and B progress through key stages of feature engineering related interaction. The graph shows that within users in Group A, no one really touched the `Apply PCA` button but only touched the `Save PCA` button.

Interpretation & Conclusion

While Group A displayed marginally higher CTRs in some stages, especially in EDA, the overall interaction patterns between groups were similar. Only two EDA tools showed a statistically significant difference in usage rates. This suggests that the tutorial may guide users to specific features, but it does not drastically change users' overall engagement or behavior patterns.

Bounce rate analysis further supports this finding. Overall, the presence of a tutorial does not lead to a substantial reduction in bounce rates. While Group A may exit more frequently after completing structured steps, and Group B may disengage due to exploration fatigue, these tendencies are event-specific rather than widespread behavioral differences.

While the tutorial may guide user actions, it does not appear to increase overall session time. In fact, users without the tutorial (Group B) remained active for longer, potentially due to more open-ended exploration or less structured task completion. This suggests that tutorials may lead users to accomplish tasks more efficiently, resulting in shorter but more purposeful sessions.

One possible explanation is that some users in Group A may look through the tutorial superficially, and further "engage" with the app, driven by the understanding their interaction was part of the data collection process. This is further illustrated in our funnel analysis, where many users skipped clicking Apply PCA (which generates PCA results) and proceeded directly to Save PCA, which stores results. This confirms that some users may have followed instructions passively without truly interacting with the analysis steps.

In summary, the tutorial successfully directed users to certain features and made their navigation more efficient, but it had little impact on overall click-through rates or bounce rates. While Group A completed tasks more efficiently, Group B users spent more time exploring. This suggests that tutorials enhance task orientation but may limit deeper, self-driven exploration.

Challenges & Limitations

The main challenge of this project was the delay in data updates from Google Analytics, particularly when using custom dimensions. This slows down debugging, A/B testing validation, and general development feedback procedures. Since most participants were classmates familiar with the project's purpose, their behavior may have been influenced by prior expectations, potentially limiting the representativeness of the results and introducing bias. Last but not least, the sample size in this study was relatively small, which limits the statistical power of the analysis. With fewer participants, it becomes harder to detect subtle differences between groups, increasing the likelihood of Type II errors.