Final Report: A/B Testing the Impact of an Automatic Tutorial on User Engagement

Team 14

Final Report: A/B Testing the Impact of an Automatic Tutorial on User Engagement

1. Introduction & Research Question

User interaction is an essential determinant of the success of interactive web applications, especially those that assist users in conducting complicated data operations. Our team created a web application that simplifies data cleaning and exploratory data analysis (EDA) of uploaded datasets. We introduced an automated click-through tutorial in a new app version to enhance user experience and assist new users in mastering the app. This study examines the effect of this tutorial on user interaction in a systematic A/B test. The central research question that led our project was:

"Does adding an automatic tutorial session to the app improve user engagement compared to the original version without the tutorial?"

2. Experimental Design & Methodology

2.1 Control and Treatment Groups

The treatment group employed a version of the original app with an automatic tutorial session, whereas the control group employed the original app without a tutorial. The tutorial was initiated at the beginning of the session. It led users through the most essential functions of the app: uploading data, seeing descriptive statistics, and employing several EDA tools.

2.2 Key Metrics

We selected the following key performance metrics to analyze:

- Task Completion: Measured by counts of clucks on “Clean Data”, “Apply Numeric”, and “Apply Categorical” buttons
- Page Leaves: Measured by “Page Leave” events with no associated label (interpreted as bounce behavior)
- Tab Engagement: Measured by the total userEngagementDuration for “Tab Duration” events, grouped by tab name
- Tutorial Usage: Measured by clicks on “Start Tutorial” and “Skip Tutorial” within the treatment group

2.3 Randomness

The 66 students in our class were randomly divided and assigned into two even groups, control and treatment. Both groups were sent a distinct application link through email.

3. Data Collection

3.1 Google Analytics Integration

First, we integrated Google Analytics event trackers into our code for both versions of our app. The 4 event actions we tracked were “Click”, “Tab Viewed”, “Tab Duration”, & “Page Leave.” Each “Click” action was labeled “Clean Data,” “Apply Numeric,” or “Apply Categorical” across both apps, specifying the button that was clicked. For the tutorial version, the click actions could additionally be labeled as “Start Tutorial” or “Skip Tutorial.” Each “Tab Viewed” action was labeled “User Guide”, “Raw Data Preview”, “Cleaned Data Preview”, “Feature Engineered Data”, or “Exploratory Data Analysis”, specifying the tab they viewed. Each “Tab Duration”

action was labeled with one of the aforementioned tab names along with the time in seconds the user spent on that tab.

3.1.1 Realization About Engagement Timing

We had initially considered analyzing tab-specific engagement durations by parsing custom labels such as “User Guide - 52s” from GA event parameters. However, we realized that this was unnecessary because:

- GA4 already tracks userEngagementDuration as part of each event group.
- The use of dimensions such as eventName and eventLabel provides sufficient context for interpreting the engagement data.
- Relying on userEngagementDuration ensures accuracy and consistency across all event types.

As a result, we stopped parsing time manually from event labels and focused solely on the userEngagementDuration metric that was already provided in the API response.

3.2 Google Analytics Strategy

Initially, our plan was to use Google Analytics 4 (GA4) raw data through BigQuery integration. However, we later discovered that BigQuery only begins collecting data after the integration is enabled. Since our web application had already been shared with users and data was actively being generated, we were unable to access historical event-level data through BigQuery.

To address this limitation, we turned to the Google Analytics Data API (v1beta), which allows for programmatic access to aggregated data GA4 metrics and dimensions.

3.3 API-Based Extraction Strategy

We used Python and the official google-analytics-data library to send a query to the GA4 API. The script is based on the BetaAnalyticsDataClient and collects a structured report of user interactions across a specific data range. The following were the key components of the extraction:

- PROPERTY_ID: The GA4 property ID of our app
- KEY_PATH: The path to our private service account credentials (for authentication)
- START_DATE / END_DATE: The full time span of the A/B testing period
- DIMENSIONS:
    - eventName: the type of interaction (e.g. Click, Page Leave, Tab Viewed)
    - eventLabel: where the interaction occurred (e.g. Clean Data, Raw Data Preview, Start Tutorial)
    - date: the date of the event
    - city, country, deviceCategory, browser: user-level metadata
    - sessionSourceMedium: how the user found the site
    - pagePath: which page the event occurred on
- METRICS:
    - eventCount: how many times that event occurred
    - userEngagementDuration: the total time the users were engaged (per dimension group)
    - engagedSessions: the number of sessions with engagement
    - screenPageViews: page views during the session

The API aggregates data across all provided dimensions. For example, if multiple users triggered the same event on the same day, city, and browser, the result would be in a single row with the summed metrics for that group

The results were written to two CSV files for our two applications, which contain aggregate analytics grouped by the above dimensions.

4. Statistical Analysis Methodology

To evaluate the effectiveness of introducing a tutorial into our web application, we conducted a rigorous analysis grounded in both exploratory data analysis (EDA) and statistical hypothesis testing. The objective was to determine whether the tutorial influenced key user behaviours, such as task completion, user engagement, and bounce rates.

4.1 Exploratory Data Analysis

We began by preprocessing both datasets–one from the version with the tutorial and one without–by normalizing column values. This included:

- Cleaning whitespace from eventName and eventLabels
- Handling missing labels by filling with “Unlabeled”
- Truncating eventLabels at the first dash (-) to group similar interactions (e.g., “User Guide – 52s” was mapped to “User Guide”)

This allowed for more interpretable and meaningful aggregation of events, especially for analyzing engagement duration by content section (e.g., tab views). (See our section on data collection where we talk about why we had to remove the extra information at the end of the “Tab Duration” labels.)

We then aggregated the data by (eventName, eventLabel) to observe the most frequent interactions and their distribution across both versions of the app. This revealed key interaction patterns, such as:

- Frequency of clicks on buttons like “Clean Data” and “Apply Numeric”
- Engagement with various tabs or content panels through “Tab Duration” and “Tab Viewed” events
- Bounce behavior captured through the “Page Leave” event

This EDA phase helped to identify metrics of interest for statistical testing.

4.2 Statistical Analysis

Based on the exploratory findings and the experiment design, we selected the following key performance metrics:

- Task Completion: Measured by counts of clucks on “Clean Data”, “Apply Numeric”, and “Apply Categorical” buttons
- Page Leaves: Measured by “Page Leave” events with no associated label (interpreted as bounce behavior)
- Tab Engagement: Measured by the total userEngagementDuration for “Tab Duration” events, grouped by tab name
- Tutorial Usage: Measured by clicks on “Start Tutorial” and “Skip Tutorial” within the treatment group

We applied the following statistical methods to evaluate these metrics:

4.2.1 Chi-Squared Test of Independence

Used to assess whether there were statistically significant differences in categorical event outcomes (e.g., whether the user completed a task or left the page) between the control and treatment groups. For each test:

- A 2x2 contingency table was constructed with rows No Tutorial/With Tutorial and columns Event Occurred/Not Occurred.
- We compared raw event counts (e.g., task completions) relative to total events per group.
- The Chi-squared statistic and p-value were computed to determine significance.

4.2.2 Welch’s t-Test

Used to compare average engagement durations between the control and treatment groups for tab viewing behavior. This test was selected due to:

- Unequal sample sizes and potentially unequal variances
- Continuous nature of the metric (userEngagementDuration)

Engagement durations from “Tab Duration” events were extracted and compared using a two-tailed Welch’s t-test. This allowed us to test the null hypothesis that the mean engagement durations for the two groups were equal.

4.2.3 Tutorial Uptake Rate

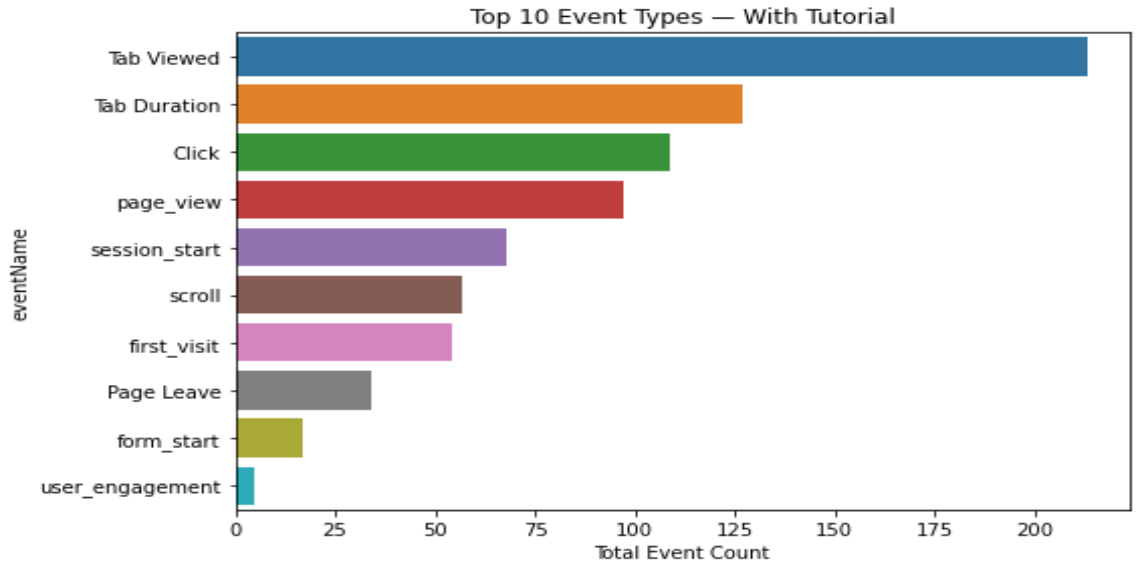
Although not a formal hypothesis test, we also calculated the uptake rate of the tutorial:

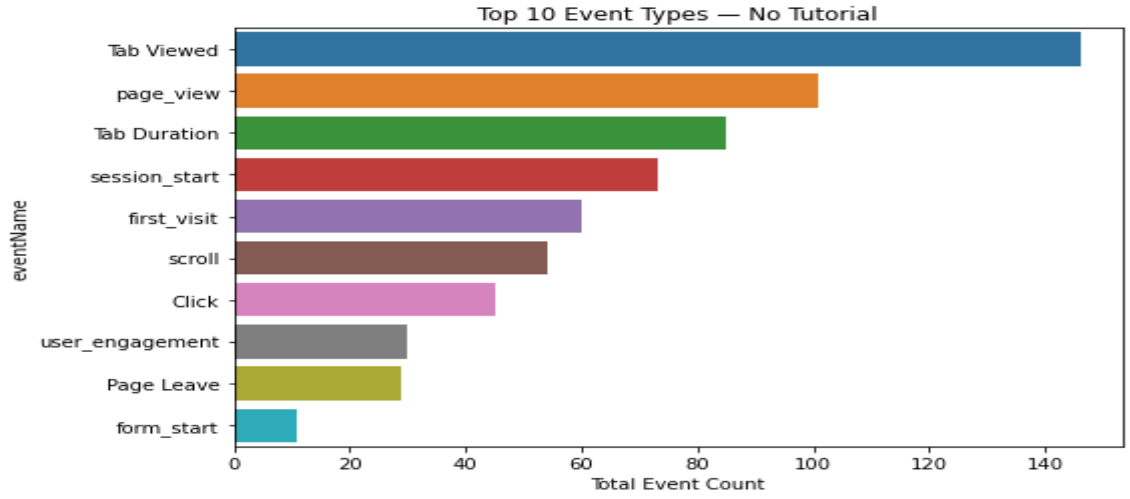
- The denominator was the total number of users who triggered a first_visit event (used as a proxy for unique users).
- The numerator was the number of users who clicked “Start Tutorial”.
- This was expressed as a proportion to describe how widely the tutorial was adopted.

This comprehensive approach ensured that both surface-level behaviour (via EDA) and deeper causal relationships (via statistical inference) were accounted for in evaluating the impact of the tutorial.

5. Results & Interpretations

5.1 Top 10 Event Types (EDA)





Key Findings:

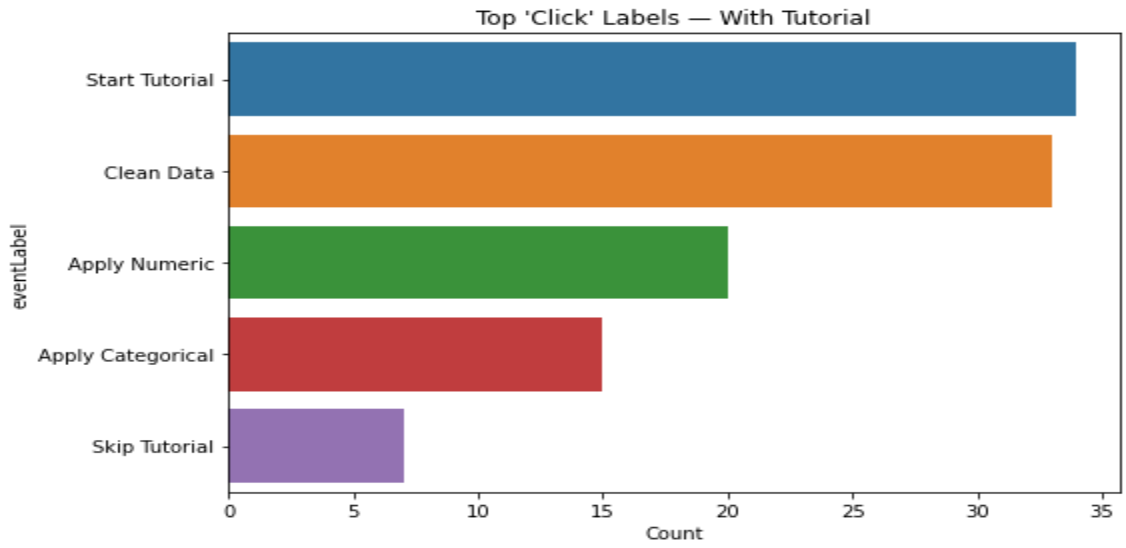
1. More engagement with the tutorial
    a. Users in the tutorial version generated more overall Tab Viewed, Tab Duration, and Click events compared to users in the non-tutorial version.
    b. This suggests users were navigating more, spending longer in tabs, and interacting more with features when the tutorial was present.
2. Higher “Click” and “Tab Duration” counts with tutorial

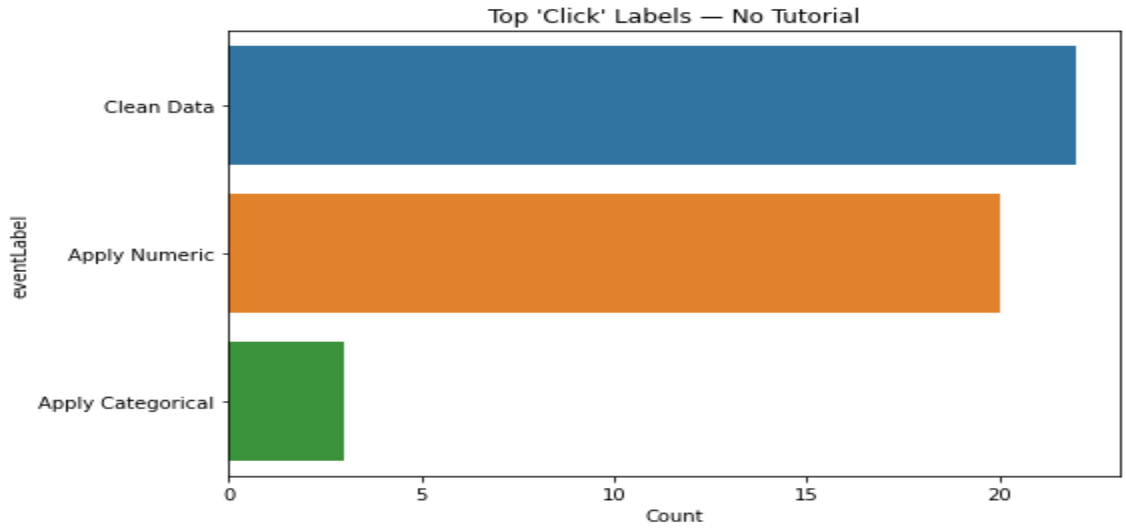
   a. The Click event was noticeably higher with the tutorial, indicating that users were more likely to complete tasks when guided.
   b. Tab Duration was also higher, suggesting users were spending more time exploring or understanding each section with the help of the tutorial
3. More “Page Views” without tutorial
   a. Interestingly, users in the non-tutorial version had a relatively higher count of page_view events, which may reflect more aimless navigation or refreshing without structured guidance.
4. Lower drop-off in the tutorial group
   a. Page Leave and form_start events were slightly lower in the tutorial version, which could suggest better task retention and reduced early exits due to improved onboarding.

Interpretation:

- The tutorial likely improved user onboarding, task clarity, and overall engagement by helping users navigate and act with more confidence.
- Increased Click and Tab Duration events in the tutorial version point toward greater feature usage and stickiness, which is a positive outcome for UX design.
- These results suggest that even a simple walkthrough or guide can meaningfully increase engagement and task completion in data-heavy apps.

5.2 Top “Click” Labels (EDA)





Key Findings:

1. Tutorial was frequently started
    a. In the tutorial version, “Start Tutorial” was the most clicked action, indicating that most users opted into the guided experience.
    b. By contrast, “Skip Tutorial” was clicked much less times, suggesting that most users were interested in using the tutorial rather than bypassing it.
2. Increased engagement with app functionality
    a. Tutorial users performed more total clicks across all main tasks:

i. Clean Data
ii. Apply Numeric
iii. Apply Categorical

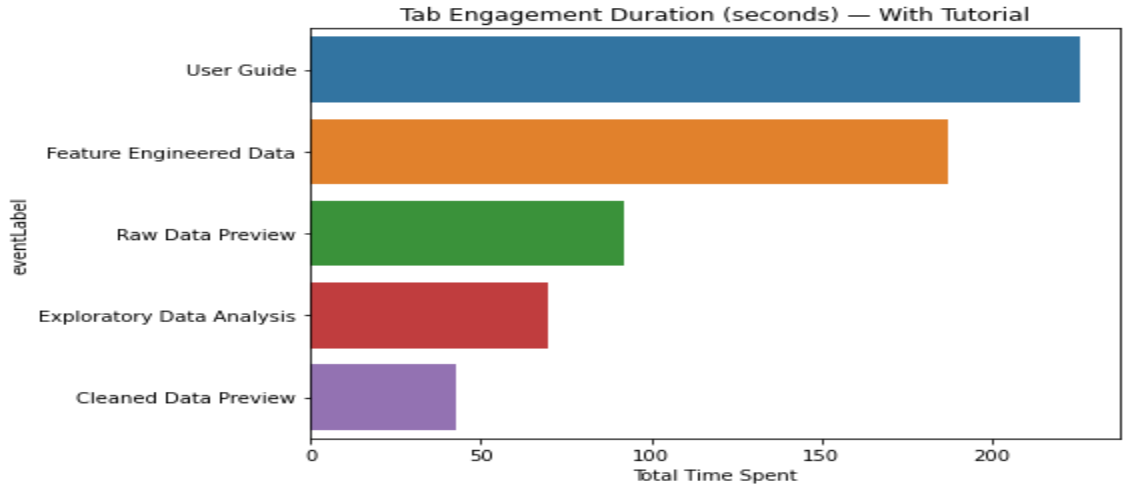
b. Compared to the control group:
    i. The “Clean Data” action was clicked more times with the tutorial vs without.
    ii. Similarly, “Apply Numeric” and “Apply Categorical” actions were also performed more often with the tutorial.

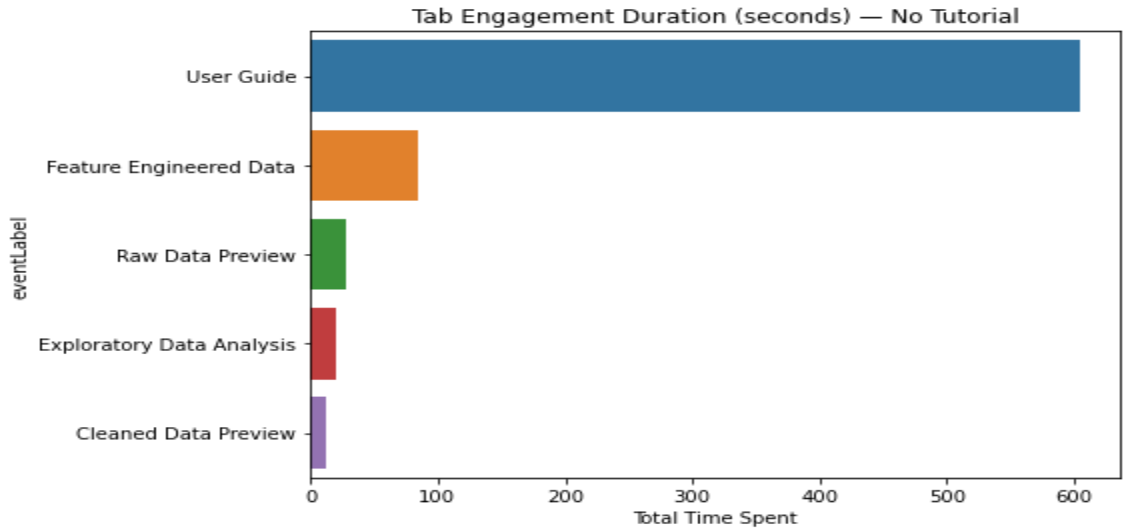
3. Apply Categorical was neglected in control version
    a. The “Apply Categorical” action was almost unused in the no-tutorial version.
    b. In contrast, tutorial users clicked it many more times, suggesting that guided instructions encouraged broader feature usage, especially for features users may be less familiar with.

Interpretation:

- The tutorial appears to have positively influenced engagement, not just by helping users get started, but by encouraging them to explore and use more features.
- A significant number of users willingly opted into the tutorial, validating the need for embedded onboarding in data-centric applications.
- Guided exploration leads to better feature coverage. Without the tutorial, users tended to stick to basics (Clean Data, Numeric transformations). With the tutorial, they were more confident using Categorical transformations as well.

5.3 Tab Engagement Duration (EDA)





Key Findings:

1. User Guide dominated time in both versions
    a. In both versions, the User Guide tab had the highest total time spent.
    b. However, non-tutorial users spent far more time on the User Guide tab compared to tutorial users, suggesting they may have spent more time figuring things out on their own.
2. Tutorial encouraged more even engagement across tabs
    a. In the tutorial version, users spent meaningful time across all core workflow tabs
3. Without tutorial, tab exploration was shallow

a. In the no-tutorial version, time spent in:
    i. Feature Engineered Data was minimal
    ii. EDA, Cleaned Data Preview, and Raw Data Preview had very little engagement

Interpretation:

- The tutorial appeared to guide users through each step, leading to more balanced engagement across the full app experience.
- Users without the tutorial over-relied on the User Guide and spent significantly less time interacting with the data transformation or analysis features.
- These findings suggest that the tutorial not only helped users understand what to do, but also motivated them to take action in the right places.

5.4 Chi-squared Tests (Statistical Analysis)

Results & Interpretations for Task Completions:

- No Tutorial: 42 completions out of 635 total events (~6.6%)
- With Tutorial: 53 completions out of 781 total events (~6.8%)
- Chi-squared test p-value = 0.9826

There is no statistically significant difference in task completion rates between the two versions. The p-value is far above the conventional threshold (0.05), meaning any difference observed is very likely due to chance. While the tutorial group had slightly more completions numerically, this result is not strong enough to conclude that the tutorial caused the increase

Results & Interpretations for Page Leaves:

- No Tutorial: 29 page leaves out of 635 total events (~4.6%)
- With Tutorial: 34 page leaves out of 781 total events (~4.4%)
- Chi-squared test p-value = 0.9488

There is no statistically significant difference in the rate of users exiting the page between the two groups. The near-identical percentages and high p-value suggest that the presence of a tutorial did not noticeably reduce early exits.

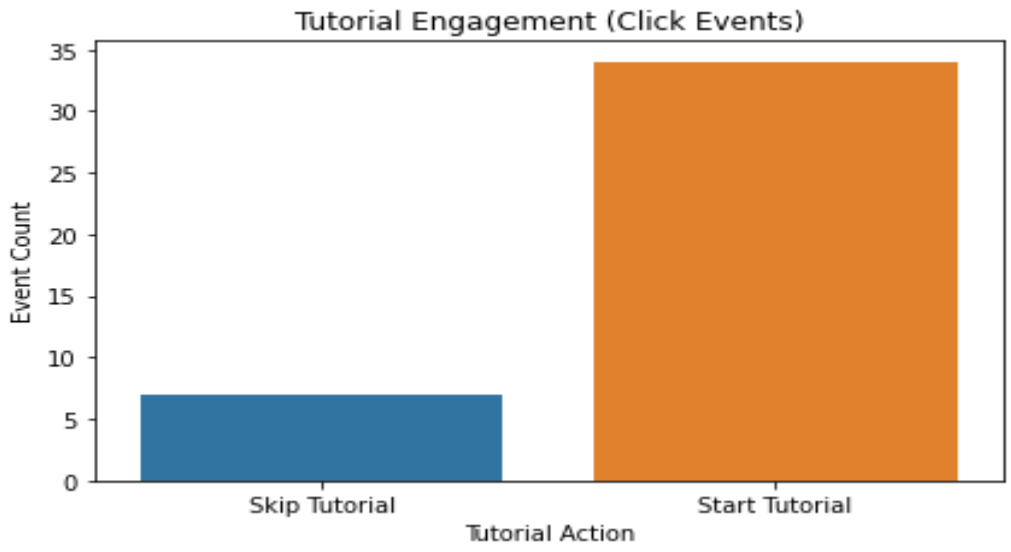
5.5 *Welch's t-Test (Statistical Analysis)*

Results & Interpretations for Tab Engagement:

- Mean Tab Duration with Tutorial: 41.20 seconds
- Mean Tab Duration without Tutorial: 62.58 seconds
- p-value: 0.6290

The average time users spent in each tab was higher in the no-tutorial version, but this difference was not statistically significant. The p-value suggests that the observed difference could easily be due to random variation in user behavior, and there's no evidence of a real difference caused by the tutorial.

5.6 Tutorial Uptake Rate (Statistical Analysis)



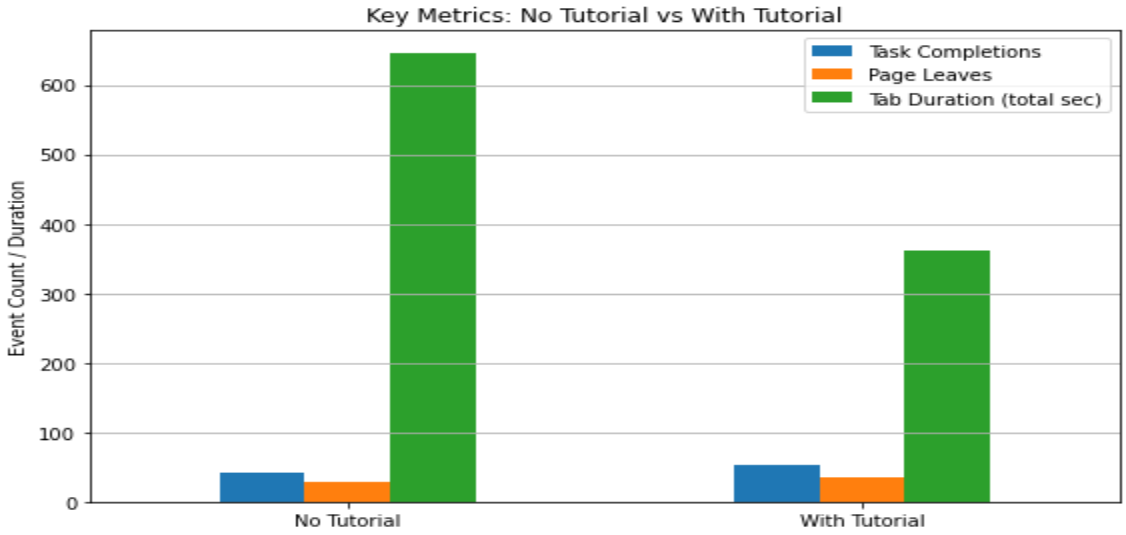
Results & Interpretation of Tutorial Usage

- Out of 54 users in the tutorial version:
    - 34 users (62.96%) actively chose to start the tutorial

- Only 7 users chose to skip the tutorial

This is a strong tutorial uptake rate, indicating that a majority of users valued or were interested in guided onboarding. The fact that "Start Tutorial" was the most clicked event label in the tutorial group further reinforces that the tutorial captured attention and action.

5.7 Aggregated Key A/B Metrics Comparison



Task completions were higher in the tutorial group, but statistical testing showed the difference was not significant ($p = 0.98$). Page leaves were also slightly higher with the tutorial, but again not statistically significant ($p = 0.95$). Interestingly, total tab duration was higher without the tutorial, but this might reflect users spending more time figuring things out on their own (especially in the User Guide tab), not necessarily better engagement.

6. Conclusion

This study aimed to evaluate whether integrating an automatic tutorial session into a data-processing application would improve user engagement compared to the original version without the tutorial. Using a combination of event tracking, visualization, and statistical testing, we assessed various engagement metrics including task completions, time spent on different app sections, click behavior, and tutorial uptake.

While statistical tests revealed no significant differences in task completion rates, exits, or average time spent per tab between the two groups, the behavioral patterns uncovered offer valuable insights. Users in the tutorial version were more likely to interact with core features of the app, including “Clean Data,” “Apply Numeric,” and “Apply Categorical” transformations. They also exhibited more balanced engagement across the app’s tabs, suggesting that the tutorial guided them through a fuller exploration of the app’s functionality.

Importantly, over 60% of users in the tutorial version chose to start the tutorial, while only a small fraction opted to skip it. This strong uptake rate indicates that users were receptive to guided onboarding. The tutorial’s ability to direct user attention and increase feature engagement–even without significant increases in task completions–suggests it successfully enhanced the quality of user interaction, if not the quantity of outcomes.

In conclusion, while the tutorial did not lead to statistically significant improvements in measured performance outcomes, it positively influenced user behavior in meaningful ways. The findings support the use of embedded, opt-in tutorials as a low-friction way to improve onboarding and encourage broader engagement with app functionality.

7. Challenges & Limitations

Despite yielding valuable insights, this study faced several limitations that constrain the generalizability and granularity of the findings.

First, the sample size was small. The app was distributed to members of our class over approximately one week, which limited both the number and diversity of users. As a result, our statistical power was low, making it difficult to detect significant differences between the tutorial and non-tutorial groups even if practical differences existed.

Second, we were unable to access event-level user data due to the timing of our analytics setup. Although Google Analytics Data API (v1beta) allowed us to track GA4 events and dimensions, we did not integrate BigQuery before collection had begun. This meant we could only analyze aggregated metrics and summary reports, rather than perform more detailed, user-level or

session-based analyses. As a result, we were unable to track complete user journeys, attribute behaviors to individual users, or analyze event sequences over time.

Additionally, some behavioural metrics–such as task complexity, user confusion, or satisfaction–are qualitative in nature, and were not captured in our tracking. Future studied would benefit from incorporating user surveys or interviews to complement event data with subjective feedback.

Together, these limitations suggest that while our results provide useful exploratory insights, further research with a larger and more diverse sample and access to granular behavioural data would be necessary to extend our findings.

GitHub Link

Link

Group Contributions

Zhisheng Yang worked on implementing the user tutorial.

Shreya Prabu worked on integrating Google Analytics & writing the report.

Ziming Zhang worked on deploying both apps, implementing pop-up windows for the tutorial, writing the readme file, & writing experiment design methods.

Ruijia Ge worked on sending out the application to users & writing the report.

Anika Kathuria worked on the EDA, statistical analysis, & writing the report.

References

Google Analytics Data API v1beta. (n.d.). Retrieved from

https://developers.google.com/analytics

Jeblick, K. et al. (2019). "The value of onboarding in UX: Increasing user engagement through guided tutorials." *International Journal of Human-Computer Interaction*.

Lahiri, D. & Rai, R. (2021). "User onboarding and retention: The psychology behind product tours." *Journal of Digital Experience Research*.

Stat-analysis.ipynb (2025). [Unpublished Python notebook containing experiment results and analysis.]

Wickham, H., & Grolemund, G. (2016). *R for Data Science*. O'Reilly Media.