

AB-Testing Report

Sara Hassani

2025-04-26

Introduction

The Trend Analysis of UK Companies application is a web-based tool designed to support Due Diligence and Economic Intelligence professionals working with companies under UK jurisdiction.

The application enables analysts to input company names and apply dynamic filters to focus on specific individuals, industries, time periods, and company characteristics. Users are presented with a comprehensive table listing relevant information such as officer names, persons with significant control (PSCs), liquidation status, and more. In addition, data visualization tools help identify patterns and commonalities across companies — for example, the proportion of insolvent companies within the pulled dataset.

App Objective

The application aims to address a key pain point for due diligence and economic intelligence professionals: the time-intensive nature of manually pulling corporate information from UK Companies House. By automating and streamlining this process, the application seeks to significantly reduce manual research time, allowing professionals to allocate more time to higher-value tasks such as analysis and reporting.

Given this objective, the focus of the experiment is to maximize user engagement with the Trend Analysis pages and to enhance the seamlessness of interactions with the search and filtering components of the application.

Research Question

Does the introduction of improved navigation tools (such as a more prominent display of filters and a dedicated Trends dropdown menu) in Version 2 of the application lead to higher user engagement and more efficient task completion compared to Version 1?

Experimental Design and Methodology

To answer this question, I designed an A/B testing framework to evaluate whether design improvements to the Trend Analysis of UK Companies application positively impact user engagement and interaction efficiency.

Application Versions

Two distinct versions of the application were developed:

Version 1 (V1)

- Users input company names and apply filters through a sidebar, with the “Apply Filters” button located at the bottom, requiring scrolling to access.
- The application contains four tabs: two tabs for displaying the initial dataset (table and visualizations) and two tabs for displaying the filtered dataset.
- All visualizations are placed on a single page, requiring the user to scroll down to view each chart.

Version 2 (V2) Version 2 retains the core functionalities of Version 1, with all modifications focused on improving user display and experience:

- A dedicated “Search” tab is introduced, allowing users to immediately see and interact with all filters upon entering the application.
- A dynamic status message is displayed on the “Search” tab, informing users when the application is actively retrieving data from UK Companies House and when the table is ready.
- Tabs related to the initial, unfiltered dataset are removed, focusing users’ attention exclusively on filtered results (with an option to reset filters).
- A Trends Visualization tab is redesigned to feature a dropdown menu, enabling users to select specific elements (e.g., insolvency status, charges, incorporation status) dynamically without scrolling.
- Visualization panels are organized side-by-side, showing comparisons between the filtered selection and the overall company list.

Hypothesis

I hypothesize that the improvements implemented in Version 2 will lead to:

- Increased user engagement with the Trends visualizations (due to more accessible drop down-based navigation).
- Higher rates of search and filtering task completion.
- Greater satisfaction with the user flow, reflected in higher interaction metrics.

Primary Metrics

In order to evaluate the outlined hypotheses and objectives from the A/B testing experiment, I will be using the following primary metrics:

- **Session Duration:** The total time (in seconds) each user spends on the web application in a single session.

While this metric provides an initial measure of user engagement, it may be influenced by periods where the user is inactive (e.g., multitasking) or by the app’s background data retrieval time. Therefore, additional adjusted metrics are introduced to provide a more accurate evaluation.

- **Search Time to Full Session Time ratio:** a ratio of the time spent pulling information from UK Companies House (time difference between the user clicking the “Search Companies” button and the output table being ready for display).
- **Interactions per Session Duration:** number of button clicks and tab switches divided by the time spent on a session. This metric provides a measure of interaction intensity normalized by session length.
- **Interactions per Post-Search time:** number of button clicks and tab switches divided by the time spent on a session excluding the time it took the app to pull all information from the UK Companies House website. This focuses specifically on active user-driven interaction time.

These metrics assume that no problems are faced by the user. So, it is important to add:

- **Error Rate:** number of times a user faces an error not allowing them to use the website divided by the number of sessions.

Secondary Metrics

The initially outlined 5 metrics allow for an overview of the two app versions' performance. However, to push this analysis further and keeping in mind the vision for the app to be a tool for analysts in the due diligence and economic intelligence industries, the following metrics will be likely to prove helpful in the long-term development of this product:

- **Granular button click metrics:** number of times a specific button was clicked by the user, in this case, there is the "Search Companies", "Remove Companies", "Apply Filters", "Reset Data" Buttons.
- **Filter usage:** number of times each specific filter is used divided by the number of overall sessions.
- **Time spent on Tabs:** amount of time per session spent on each specific tab.

These metrics are bound to be beneficial in the long-term as they test which specific features are most utilized in the app so as to ease future decisions on what to highlight in the app. They will, however, not be a big focus for the project at this time.

Data Collection

The initial plan for this experiment was to utilize Google Analytics' services and pull data using its API for the data analysis portion of this project.

I started by integrating Google Analytics into my app (as per this tutorial: <https://www.appsilon.com/post/r-shiny-google-analytics>).



Stream details 			
STREAM NAME	STREAM URL	STREAM ID	MEASUREMENT ID
AB-Testing V2	https://sara-hassani.shinyapps.io/AB-Testing-V2/	11043893350	G-ER0GYSK8VG 

Figure 1: Example Google Analytics Proof

I also attempted to add button-specific triggers for Google Analytics to pick up on:

```
# GA trigger for "Search Companies"
observeEvent(input$process_btn, {
  session$sendCustomMessage(
    type = "ga_event",
    message = list(
      category = "Button",
      action = "Click",
      label = "Search Companies Button"
    )
  )
})
```

Similar code was put in place for the buttons "Apply Filters", "Reset Data" and "Remove Companies" as well.

```
# GA trigger for Switching to "Trends" tab
observe({
  req(input$tabs)
  session$sendCustomMessage(
    type = "ga_event",
    message = list(
      category = "Tab",
      action = "Switch",
      label = input$tabs
    )
  )
})
```

Similar code was put in place for each of the tabs in V1 and tabs in V2.

Additional metrics that were supposed to be picked up are:

- Page Views.
- Scrolls.
- Number of Sessions.
- Session Duration.
- New/Returning Users.
- Number of Clicks.

However, I was unable to get Google Analytics working as I would have liked on-time and only collected on 3 visits (incompletely). Therefore, I opted for simulated user data instead. I created csv files simulating tracking data for each of the two versions for 100 sessions. *Appendix A* at the end of this report explains each variable. The code used to create the simulated data is provided in the `simulation.R` file on the GitHub repository.

Statistical Analysis, Takeaways, and Limitations

As a first step, I calculate the primary metrics outlined for this experiment based on the data simulated. A full rundown of the summary statistics as well as statistics test for each of the 5 primary metrics, not featured in the report body, is featured in *Appendix B* of this report.

Primary Metric 1 and 2: Session Durations and Search Time

An analysis of session durations reveals a notable difference between Version 1 (V1) and Version 2 (V2), with simulated users spending more time on V1 across various time metrics.

The graph (Figure 2) reflects our findings from the summary statistics, indicating an average of 60.54 seconds spent on V1 and an average of 45.73 seconds spent on V2 per session.

Also corroborated by the output of a two sample t-test for the session durations on V1 and V2, assuming assumptions of the test are met, where we obtain a p-value under the 0.05 critical value. We reject the null hypothesis that the difference in means for the two groups is equal to 0.

Essentially, we're noting that more time is spent on V1. This observation remains consistent when separating session time into loading and interaction components (Figure 3).

Take-Away: A consistent pattern emerges across all three metrics:

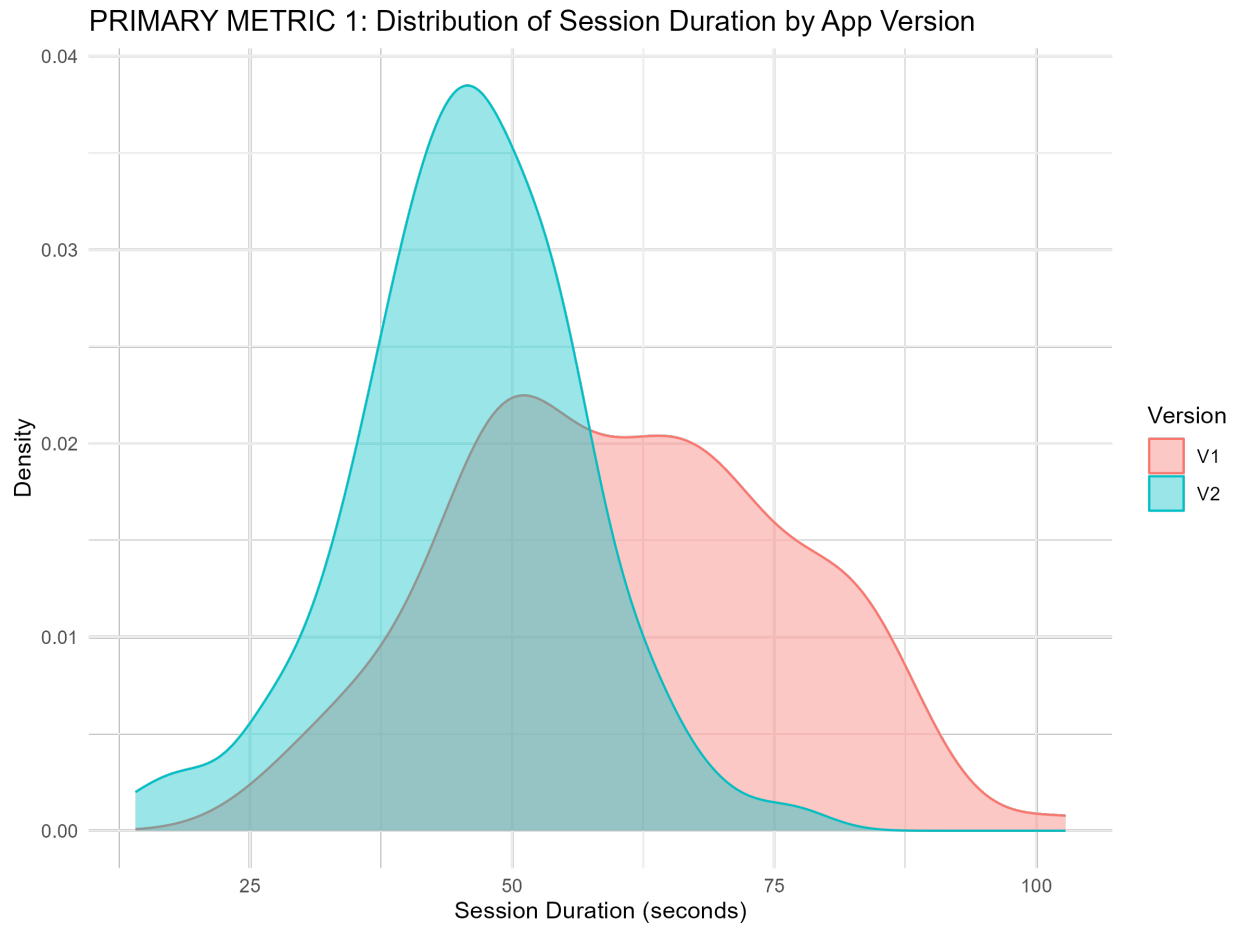


Figure 2: PRIMARY METRIC 1: Distribution of Session Duration by App Version

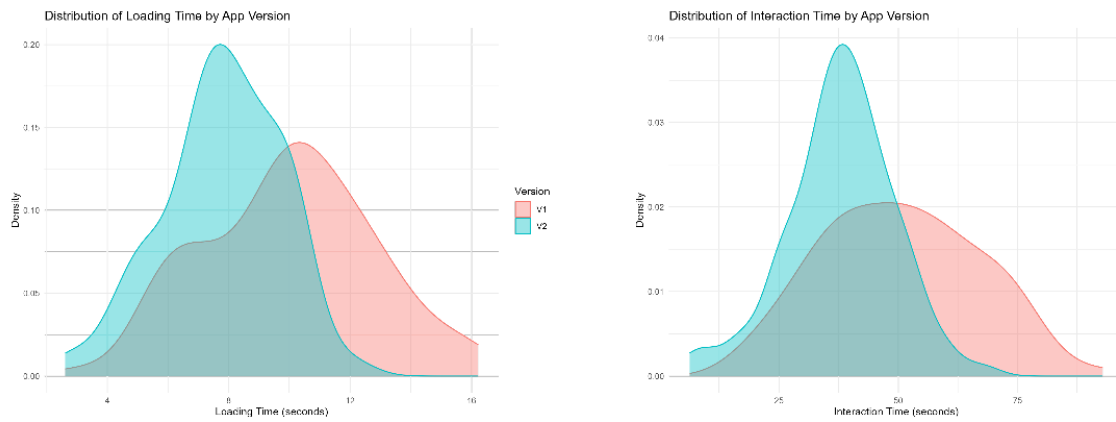


Figure 3: Distribution of Loading and Post-Search Times by App Version

- Loading times are longer in V1, suggesting slower background processing.
- Interaction times are also longer in V1, reflecting greater active user engagement.

This presents a **trade-off**: While V2 reduces total session time and improves loading speed — aligning with user efficiency goals — V1 fosters greater user interaction, which may be more valuable given the intended use for the app.

This trade-off is best represented by a visualization of **Primary Metric 2: Search Time to Full Session Time Ratio** where we can see that the boxplots appear almost identical for both versions (Figure 4).

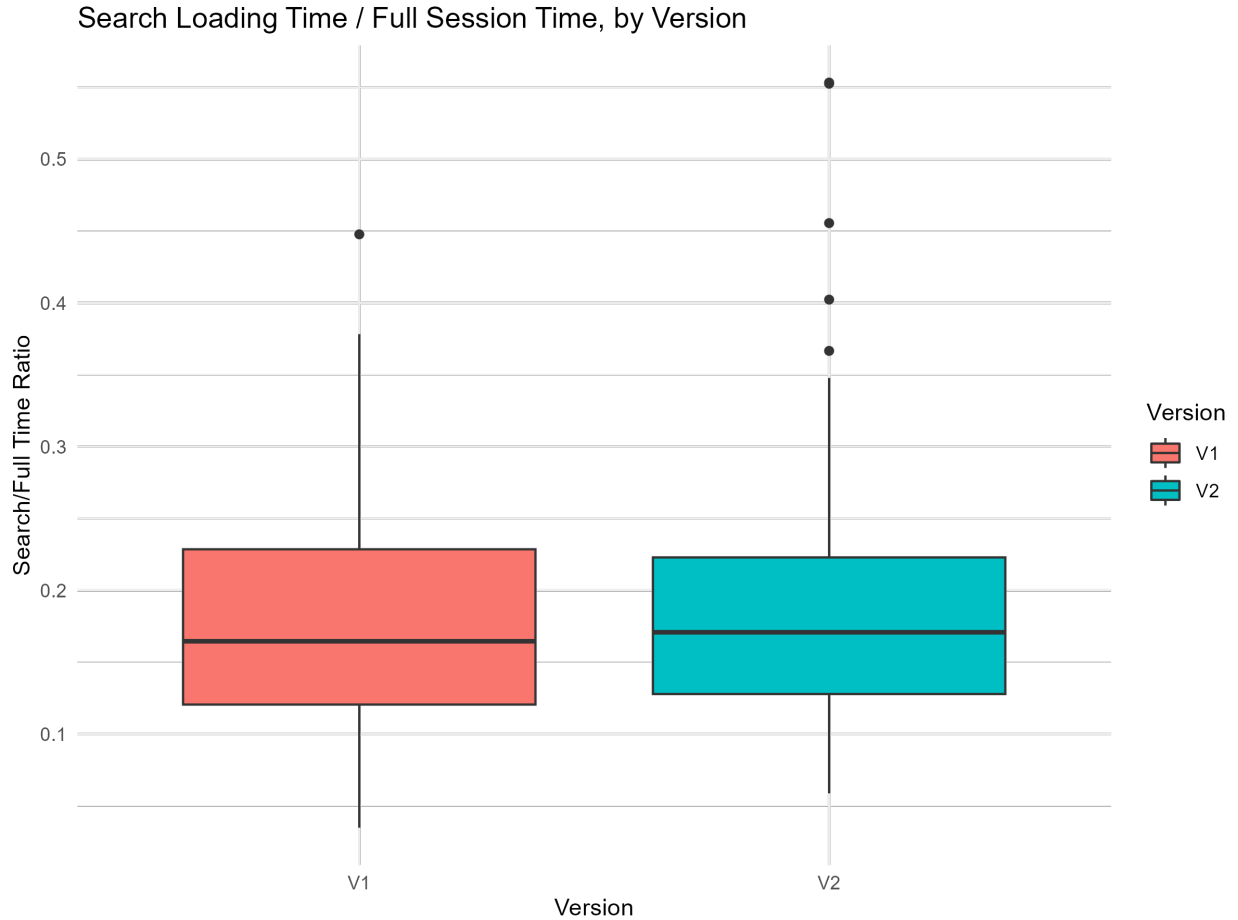


Figure 4: PRIMARY METRIC 2: Search Loading Time / Full Session Time, by Version

In order to inform this decision further, I look deeper into the utilization of the app through the analysis of interactions.

Primary Metrics 3 and 4: Interaction Rates

An evaluation of the Number of Interactions per Second, whether it be over the full session duration (metric 3) or only the post-search duration (metric 4) elucidates our finding from the analysis of the first and second primary metrics. We note, that while we spend more time for interaction on V1 then on V2, the overall interactions rate (per second) is lower for V1 than V2 (Figure 5).

This is corroborated by the t-tests conducted for both metrics, where p-values obtained were below the 0.05 critical value in both cases. So, we reject the null hypothesis that the difference in means of the two groups is equal to 0.

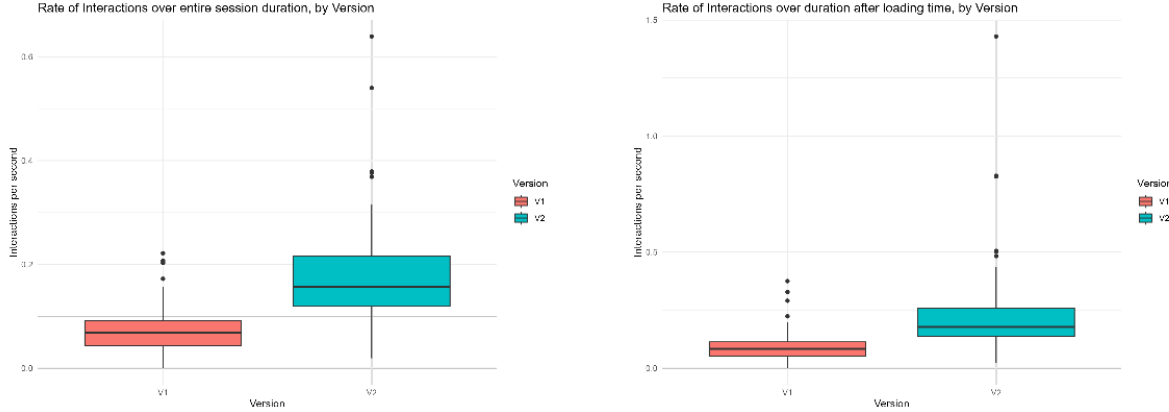


Figure 5: PRIMARY METRIC 3 and 4: Interaction Rates, by Version

Limitations: However, given the difference in session times established when analyzing the first two metrics, this difference could be due to the time spent on the app for users in V1 being significantly higher than in V2. Meaning, while the interaction rate might be higher, the raw number of interactions could be the same or even lower for V2. Figure 6 indicates that is not the case. The number of interactions is clearly significantly higher in V2 than it is in V1. But this is not a metric we can trust, which also makes primary metrics 2 and 3 less reliable. The reason is that there are more “interactions” to be had on V2 given the inclusion of the drop-down menu (5 additional possible interactions) and only the reduction of 1 tab compared to V1.

Primary Metric 5: Error Rate

The observed error rate was 31% for Version 1 (V1) and 24% for Version 2 (V2). At first glance, this difference might suggest favoring the search process improvements implemented in V2.

However, a Pearson’s chi-squared test of independence (assuming sufficient normality for the approximation) yielded a p-value of 0.34, indicating that the difference in error rates between the two versions is not statistically significant at the 5% level.

This result aligns with expectations, as the local search functions were identical between the two versions; all differences between V1 and V2 were confined to user interface changes rather than back-end logic.

Limitations: While the error rate metric is not necessarily useful in our case given no changes of the search functions, it could have been useful with more granular error tracking.

Given the structure of the application, it can be reasonably assumed that most errors originate from the search functionalities rather than from user interface interactions. So, to better identify and resolve sources of errors, future iterations of the application should implement specific error logging, including:

- The exact function that triggered the error,
- The input text (e.g., company name) that led to the failure,
- The type of error encountered (e.g., API timeout, no match found, parsing failure).

Capturing this level of detail would allow for targeted improvements to the underlying local functions, making error handling more precise and strengthening the application’s reliability.

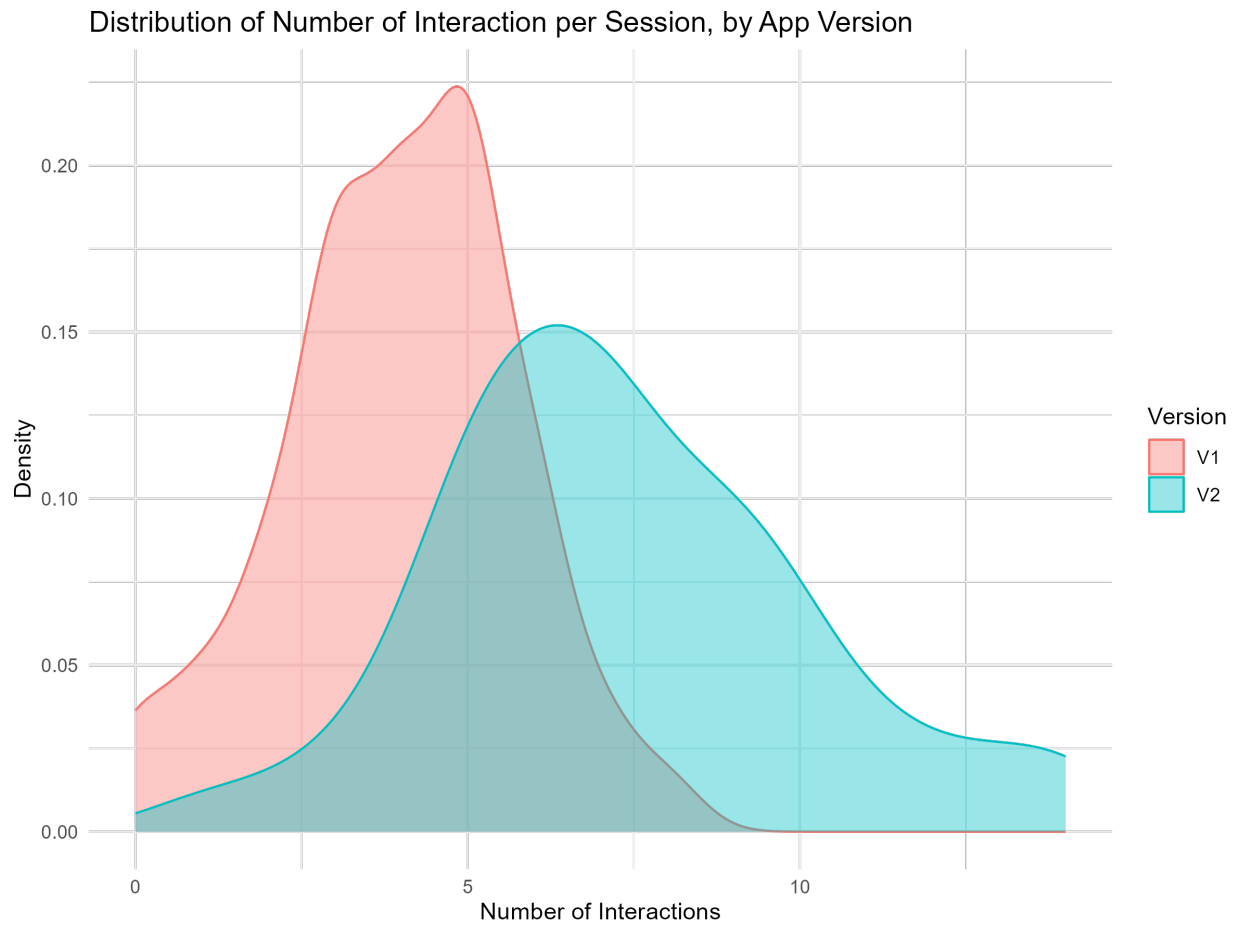


Figure 6: Number of Interactions, by Version

Conclusion

Takeaways

The A/B testing experiment revealed several notable differences between Version 1 (V1) and Version 2 (V2) of the Trend Analysis of UK Companies application:

- **Session Duration:** Users spent significantly more time on V1 sessions compared to V2. However, this was partially due to longer loading times in V1.
- **Search Loading Time:** V2 substantially reduced the time users spent waiting for search results.
- **Interaction Rates:** Users exhibited a higher rate of interactions per second in V2, but total interaction counts were impacted by additional UI elements (e.g., dropdown options).
- **Error Rate:** Although V2 showed a lower error rate than V1, the difference was not statistically significant, aligning with the fact that core search functionalities remained unchanged across versions.

Overall, Version 2 demonstrated improvements in session efficiency and user interaction density without introducing significant increases in error occurrences.

Limitations

While the findings from the metrics so far are informative, several limitations must be acknowledged:

- **Interaction Rate Comparability:** The higher number of potential interactions introduced in V2 (due to drop-down menus) complicates direct comparison of interaction rates and totals between versions.
- **Session Time Bias:** Differences in session duration between V1 and V2 introduce bias in rate-based metrics (e.g., interactions per second).
- **Error Rate Analysis:** Without detailed error logging (e.g., exact error type, company input that triggered the error), it is difficult to diagnose and resolve the underlying causes of errors.
- **Simulated Data:** Due to constraints in obtaining live user data, this analysis is based on simulated sessions, which, while realistically modeled, may not fully capture true user behavior variability.

Future iterations of the app should include more granular error tracking and clickstream logging to allow deeper validation of engagement and reliability metrics based on real-world usage. I have provided examples of such variables in *Appendix A*, explaining the generation of simulated data.

Interim Recommendation and Next Steps

Based on these findings, my recommendation would be to move forward with V2 instead of V1. This is because, while time spent on V2 is less than for V1, which is not ideal, there is more interaction with the website (albeit due to additional UI elements). It ensures continuous engagement of the due diligence and economic intelligence professionals for whom the app is designed.

This decision would be further strengthened by collecting qualitative feedback from users, such as satisfaction ratings, open-ended comments on usability, and feature-specific evaluations. Incorporating user feedback alongside behavioral tracking would allow for more holistic and user-centered application development in future iterations.

Appendix A: Simulated Variables Description

The following variables were generated through simulation in an attempt to approximate realistic user behavior while interacting with the Trend Analysis of UK Companies application. Simulations were conducted separately for Version 1 (V1) and Version 2 (V2), reflecting differences in app structure and available features.

Common Variables Across V1 and V2

- UserID : Unique identifier for each simulated user.
- Version : Indicates whether the user interacted with Version 1 (“V1”) or Version 2 (“V2”).
- SessionDuration : Total time (in seconds) the user spent on the web application during a session.
- SearchClicked : Whether the user clicked the “Search Companies” button (1 = yes, 0 = no).
- RemoveClicked : Whether the user clicked the “Remove Companies” button (1 = yes, 0 = no).
- ApplyFiltersClicked : Whether the user clicked the “Apply Filters” button (1 = yes, 0 = no).
- ResetDataClicked : Whether the user clicked the “Reset Data” button (1 = yes, 0 = no).
- SearchLoadingTime : Time taken (in seconds) by the application to retrieve data after the user clicked “Search Companies”.
- Interactions : Total number of button clicks and tab switches performed during the session.
- ErrorOccurred : Whether the user encountered an error that prevented normal app usage (1 = error, 0 = no error).

Version 1 (V1) Specific Variables

- TabSwitch_RawTable : Whether the user switched to the “Raw Table” tab (1 = yes, 0 = no).
- TabSwitch_RawTrend : Whether the user switched to the “Raw Trend” tab (1 = yes, 0 = no).
- TabSwitch_FilteredTable : Whether the user switched to the “Filtered Table” tab (1 = yes, 0 = no).
- TabSwitch_FilteredTrend : Whether the user switched to the “Filtered Trend” tab (1 = yes, 0 = no).
- Scrolled_RawTable : Whether the user scrolled down the “Raw Table” tab to view more content (1 = yes, 0 = no).
- Scrolled_RawTrend : Whether the user scrolled down the “Raw Trend” tab (1 = yes, 0 = no).
- Scrolled_FilteredTable : Whether the user scrolled down the “Filtered Table” tab (1 = yes, 0 = no).
- Scrolled_FilteredTrend : Whether the user scrolled down the “Filtered Trend” tab (1 = yes, 0 = no).

Version 2 (V2) Specific Variables

- TabSwitch_SearchPage : Whether the user switched to the “Search Page” tab (1 = yes, 0 = no).
- TabSwitch_TableDisplay : Whether the user switched to the “Table Display” tab (1 = yes, 0 = no).
- TabSwitch_TrendsVisualization : Whether the user switched to the “Trends Visualization” tab (1 = yes, 0 = no).
- Scrolled_TrendsVisualization : Whether the user scrolled within the “Trends Visualization” tab (1 = yes, 0 = no).
- DropdownUsed_TrendsVisualization : Whether the user opened and interacted with the dropdown menu on the Trends Visualization tab (1 = yes, 0 = no).
- Dropdown_IncorporationStatus : Whether the user selected “Incorporation Status” from the Trends dropdown menu (1 = yes, 0 = no).
- Dropdown_LiquidatedYN : Whether the user selected “Liquidated Y/N” from the Trends dropdown menu (1 = yes, 0 = no).
- Dropdown_ChargesYN : Whether the user selected “Charges Y/N” from the Trends dropdown menu (1 = yes, 0 = no).
- Dropdown_InsolventYN : Whether the user selected “Insolvent Y/N” from the Trends dropdown menu (1 = yes, 0 = no).
- Dropdown_VIPList : Whether the user selected “VIP List” from the Trends dropdown menu (1 = yes, 0 = no).

Appendix B: Data Visualization and Statistics Tests

SessionDuration	search_full_ratio	interactions_rate	post_interactions_rate
Min. : 27.40	Min. :0.03488	Min. :0.00000	Min. :0.00000
1st Qu.: 48.73	1st Qu.:0.12056	1st Qu.:0.04340	1st Qu.:0.05086
Median : 59.94	Median :0.16459	Median :0.06843	Median :0.08286
Mean : 60.54	Mean :0.18046	Mean :0.07239	Mean :0.09241
3rd Qu.: 71.00	3rd Qu.:0.22859	3rd Qu.:0.09122	3rd Qu.:0.11279
Max. :102.78	Max. :0.44774	Max. :0.22164	Max. :0.37487

Figure 7: Summary Statistics of Primary Metrics for V1

SessionDuration	search_full_ratio	interactions_rate	post_interactions_rate
Min. :14.08	Min. :0.05877	Min. :0.01899	Min. :0.02233
1st Qu.:39.59	1st Qu.:0.12789	1st Qu.:0.11890	1st Qu.:0.13646
Median :46.21	Median :0.17084	Median :0.15598	Median :0.17730
Mean :45.73	Mean :0.18564	Mean :0.17357	Mean :0.22710
3rd Qu.:52.95	3rd Qu.:0.22292	3rd Qu.:0.21598	3rd Qu.:0.25697
Max. :76.66	Max. :0.55333	Max. :0.63933	Max. :1.42871

Figure 8: Summary Statistics of Primary Metrics for V2

```

Welch Two Sample t-test

data: SessionDuration by Version
t = 7.767, df = 176, p-value = 6.371e-13
alternative hypothesis: true difference in means between group V1 and group V2 is not equal to 0
95 percent confidence interval:
 11.04635 18.57215
sample estimates:
mean in group V1 mean in group V2
 60.53594      45.72669

```

Figure 9: T-Test Output for Primary Metric 1 - Session Duration

```

Welch Two Sample t-test

data: search_full_ratio by Version
t = -0.43923, df = 195.9, p-value = 0.661
alternative hypothesis: true difference in means between group V1 and group V2 is not equal to 0
95 percent confidence interval:
 -0.02843984  0.01807936
sample estimates:
mean in group V1 mean in group V2
 0.1804646      0.1856449

```

Figure 10: T-Test Output for Primary Metric 2 - Search Time to Full Session Time Ratio

```

Welch Two Sample t-test

data: interactions_rate by Version
t = -9.8483, df = 138.14, p-value < 2.2e-16
alternative hypothesis: true difference in means between group V1 and group V2 is not equal to 0
95 percent confidence interval:
 -0.12149124 -0.08086366
sample estimates:
mean in group V1 mean in group V2
 0.07238827      0.17356572

```

Figure 11: T-Test Output for Primary Metric 3 - Interactions per Second

```

Welch Two Sample t-test

data: post_interactions_rate by Version
t = -7.111, df = 124.08, p-value = 8.033e-11
alternative hypothesis: true difference in means between group V1 and group V2 is not equal to 0
95 percent confidence interval:
 -0.17217795 -0.09720002
sample estimates:
mean in group V1 mean in group V2
 0.09241229      0.22710128

```

Figure 12: T-Test Output for Primary Metric 4 - Interactions per Second (post-search)

```

Pearson's Chi-squared test with Yates' continuity correction

data: error_table
X-squared = 0.90282, df = 1, p-value = 0.342

```

Figure 13: Chi-Square Test Output for Primary Metric 5 - Error Rate

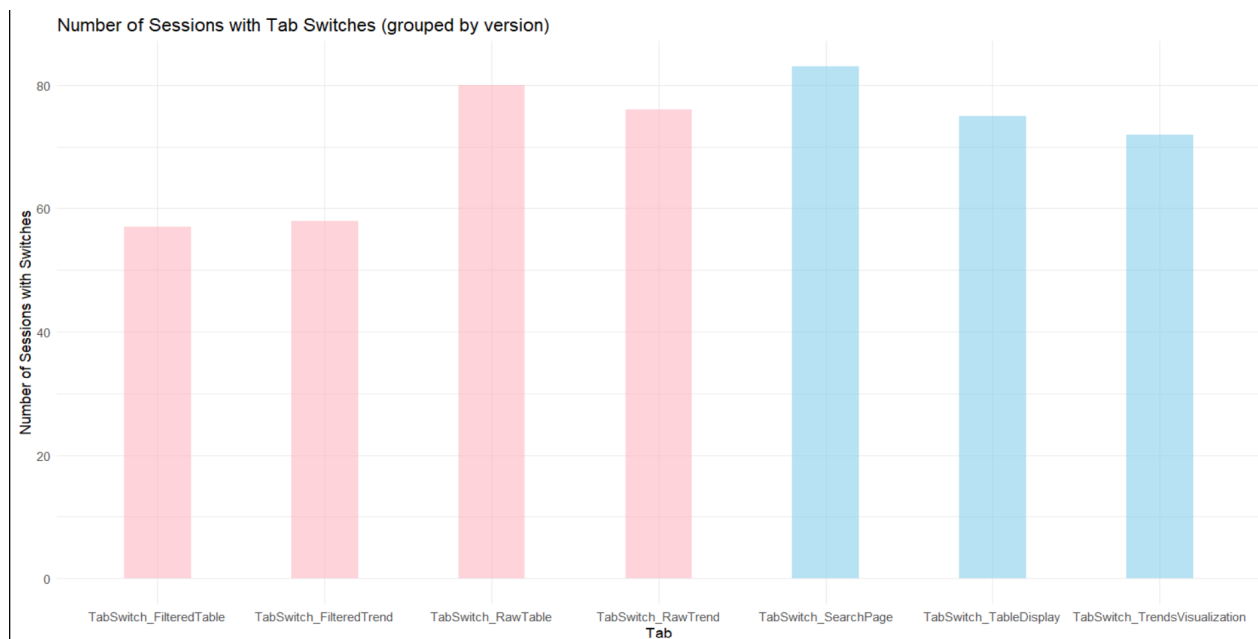


Figure 14: Number of Sessions with Tab Switches (grouped by version)

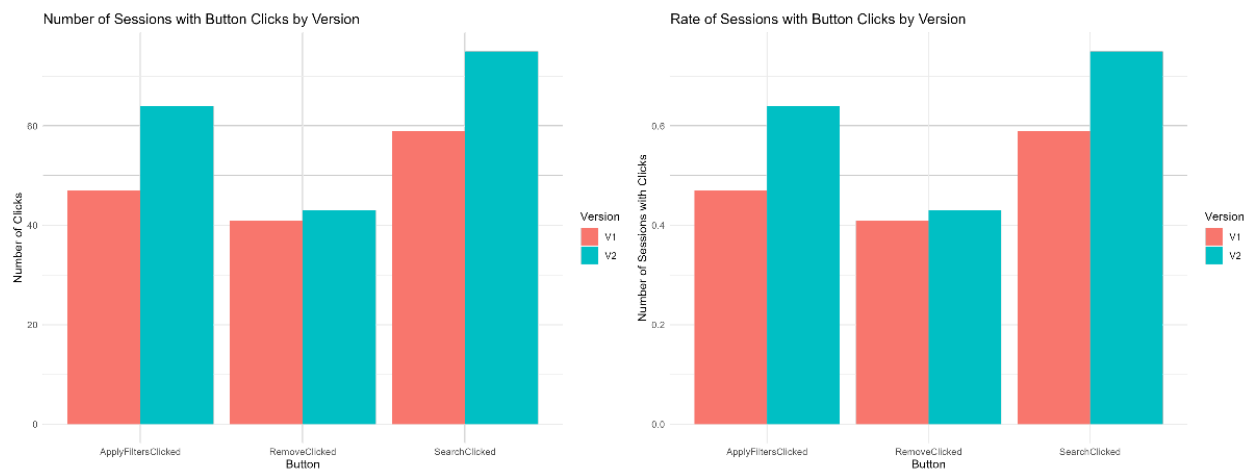


Figure 15: Number and Rate of Sessions with Button Clicks by Version