A/B Testing + Eye Tracking

A/B Testing and Eye Tracking are two common methods of collecting data on user interaction with webpages; from this data, we can determine how best to organize our webpages to attract the user interaction we desire. Here, I've performed an A/B test and an eye tracking test with two versions of a test website, as well as analysis of the data from each test. Both versions can be viewed at stark-journey-12954.herokuapp.com; a Python script is used to randomize which version appears to the user.

A/B Testing Hypotheses

Click-Through Rate

Null Hypothesis: The click-through rate on Version A will be equal to Version B.
Alternative Hypothesis: The click-through rate on Version A will be less than the click-through rate on Version B because Version B has more images and objects to click.

Time to Click

Null Hypothesis: The time to click on Version A will be equal to Version B.
Alternative Hypothesis: The time to click on Version B will be greater than the time to click on Version A because the flashy images and colors will take longer for the user to digest, taking them longer to click. In addition, users may not read the text on Version A (more minimalist) if it isn't inviting and instead just click to see what happens.

Dwell Time

Null Hypothesis: The dwell time on Version A will be equal to Version B.
Alternative Hypothesis: The dwell time on Version A will be less than the dwell time on Version B because "Reserve" (on Version A) encourages users to just look for a reserve button on the external page, while "Get Started" (Version B) encourages users to look around for more information on the external page.

Return Rate

Null Hypothesis: The return rate on Version A will be equal to Version B.
Alternative Hypothesis: The return rate will be greater on Version A than on Version B because the time to click will likely be shorter for Version A, thus implying that the users did not read our page as carefully and need to return to read more about other taxi companies.

Metrics

Click-Through Rate is calculated by counting the number of unique session IDs, then counting the number of these sessions that had at least one click.

Version A: 15 unique clicks, 23 unique sessions = $\frac{15}{23} \approx 0.652$.

Version B: 9 unique clicks, 15 unique sessions = $\frac{9}{15} = 0.6$.

Time to Click is found by subtracting the page load time from the click time for each click (click time – page load time), then finding the average. Only a user's first click should be counted.

Version A: average time to click ≈ 9257.733 milliseconds.
Version B: average time to click = 27630.111 milliseconds.

Dwell Time is found by subtracting the click time from the second page load time (second page load time – click time) for every click that left and returned, then finding the average.

Version A: average dwell time ≈ 307394.546 milliseconds.
Version B: average dwell time = 9056.125 milliseconds.

Return Rate is found by counting the number of total clicks (non-unique session IDs), then counting the number of these clicks that left the page and returned (page was loaded again after a click, meaning user returned).

Version A: 11 returns, 26 total clicks = $\frac{11}{26} \approx 0.423$.

Version B: 16 returns, 25 total clicks = $\frac{16}{25} = 0.64$.

Continue to next page for Statistical Analysis

Statistical Tests

Click-Through Rate relies on categorical data, meaning we are grouping users into two separate categories: those who perform the action (clicking on a link or button), and those who don't. We should use a chi-squared test.

Figure 1: Calculations for Chi-Squared Test for Click-Through Rate

Click-Through Rate

Observed	Clicked	No Click	Total
Version A	15	8	23
Version B	9	6	15
Total	24	14	38

Expected	Clicked	No Click	Total
Version A	14.526	8.474	23
Version B	9.474	5.526	15
Total	24	14	38

$$\chi^2 = \sum_{\text{all cases}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

~~values found~~ *calculations through Excel

$$\chi^2 = \frac{(15-14.526)^2}{14.526} + \frac{(8-8.474)^2}{8.474} + \frac{(9-9.474)^2}{9.474} + \frac{(6-5.526)^2}{5.526} \approx 0.106$$

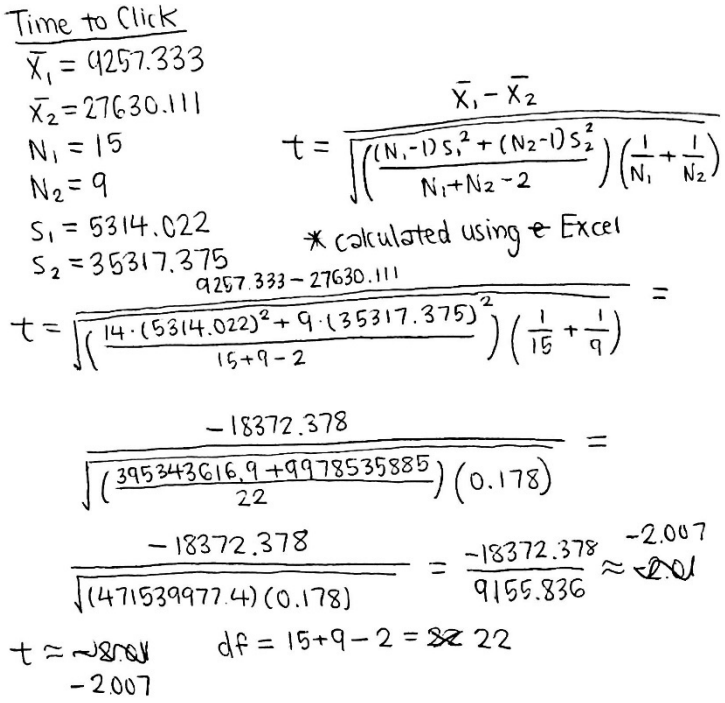
$$df = (\text{rows}-1)(\text{cols}-1) = (2-1)\cdot(2-1) = 1$$

$\chi^2 \approx 0.106$, $df = 1$

Looking at a χ^2 table and using $df = 1$, we can see that the critical value at $p = 0.05$ is approximately 3.84. My χ^2 of 0.106 is not greater than the critical value of 3.84. This means there is not a statistically significant difference in the number of users who click between versions. We fail to reject the null hypothesis that the click-through rate on Version A will be the same on Version B.

Time to Click uses a difference of means – we are determining whether there is a difference in the average amount of time a user takes to click on the different versions. We should use an independent samples t-test.

Figure 2: Calculations for T-Test for Time to Click



$t \approx$ -2.007, df = 22

Looking at a t-table and using df = 22, we can see that the critical value at p = 0.05 is approximately 1.717. My t of -2.007 is greater than the critical value of 1.717 (using absolute values). This means there is a statistically significant difference in the amount of time a user takes to click on the different versions. We thus can reject the null hypothesis that the time to click on Version A would be equal to Version B.

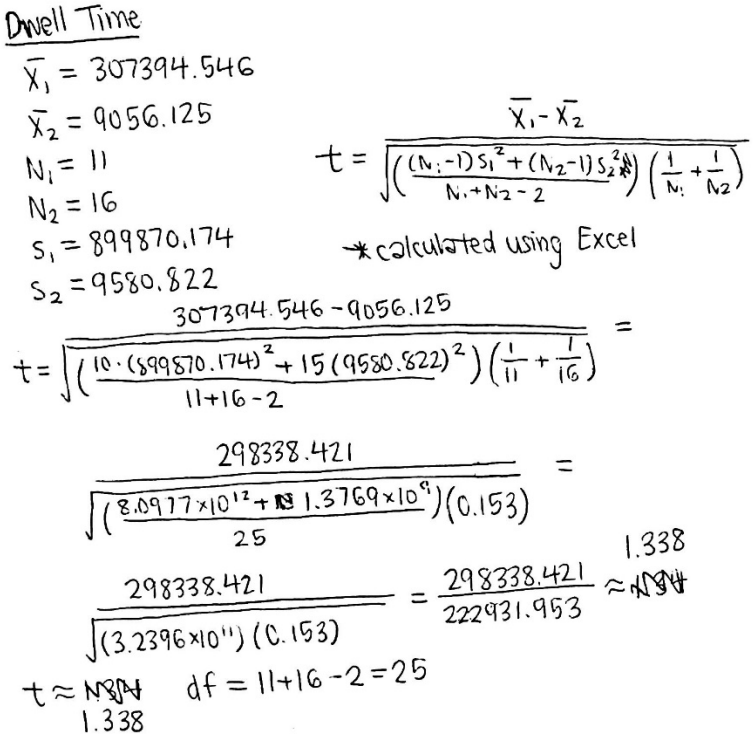
The value of t is negative, meaning the mean for Version A was less than the mean for Version B. This means the time to click on Version B is statistically significantly greater than that on Version A.

Dwell Time uses a difference of means – we are determining whether there is a difference in the average amount of time a user spends on the external page before returning. We should use an independent samples t-test.

In organizing the data to calculate dwell times, I decided to calculate dwell time for every click that left the page and then returned, rather than just the first click for each user.

Additionally, there were several clicks which opened links in a new tab, meaning a click was registered, but they did not leave the page. This resulted in a "negative" dwell time, because a click time was registered, but the page load time on the next line was unchanged from the current line. These were all excluded from dwell time calculations.

Figure 3: Calculations for T-Test for Dwell Time



$t \approx 1.338$, $df = 25$

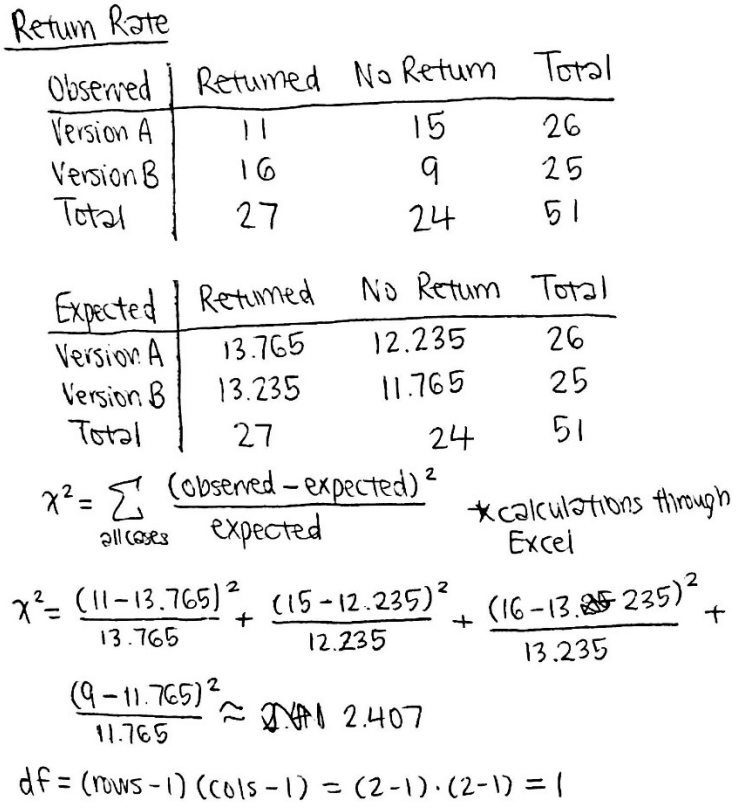
Looking at a t-table and using $df = 25$, we can see that the critical value at $p = 0.05$ is approximately 1.708. My t of 1.338 is not greater than the critical value of 1.708. This means there is not a statistically significant difference in the amount of time a user dwells on an external page between versions. We fail to reject the null hypothesis that the dwell time on Version A will be equal to Version B.

Return Rate relies on categorical data, meaning we are grouping clicks into two separate categories: those which return to the landing page after leaving to an external page, and those that don't. We should use a chi-squared test.

In organizing the data to calculate the number of clicks which returned to the page, I decided to count every click that left the page and then returned, rather than just counting every user who returned to the page once. Some users left and returned multiple times, and they were counted multiple times in these cases.

Any click which opened a link in a new tab did not count towards clicks which returned to the page. A click was registered, but the user never left the landing page, so it cannot be counted as a return.

Figure 4: Calculations for Chi-Squared Test for Return Rate

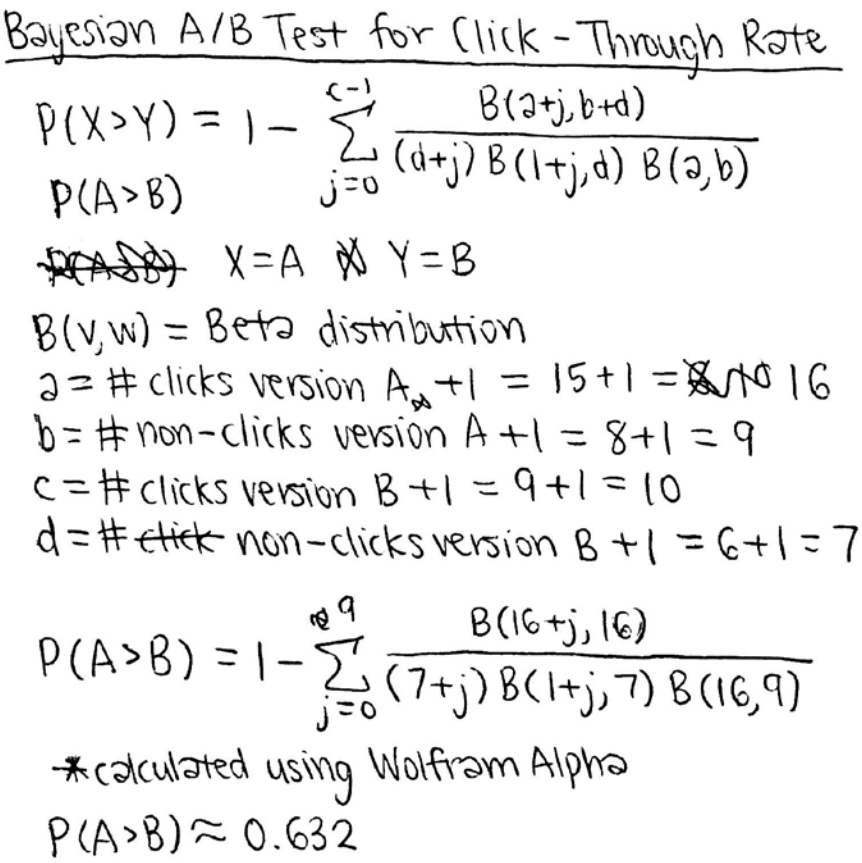


$\chi^2 \approx 2.407$, $df = 1$

Looking at a χ^2 table and using $df = 1$, we can see that the critical value at $p = 0.05$ is approximately 3.84. My χ^2 of 2.407 is not greater than the critical value of 3.84. This means there is not a statistically significant difference in the number of users who leave and then return between versions. We fail to reject the null hypothesis that the return rate on Version A will be the same on Version B.

Bayesian Probability Calculation

Figure 5: Bayesian Calculations for Click-Through Rate



The probability that Version A truly has a higher click-through rate than Version B, or $P(A > B)$, is approximately 0.632. I chose to calculate $P(A > B)$ rather than $P(B > A)$ because even though in our alternative hypothesis, we expect Version A to have a lower click-through rate than Version B, our data shows that Version A actually had a higher click-through rate (though the difference was not statistically significant).

Eye Tracking

Hypothesis: Version B will have a greater proportion of user eye gazes along the vertical center of the screen than Version A because the content on Version B is centered vertically on the screen, with vertical scrolling to additional content. Version A, on the other hand, separates the content into four quadrants on the screen.

Figure 6: Version A Heatmap

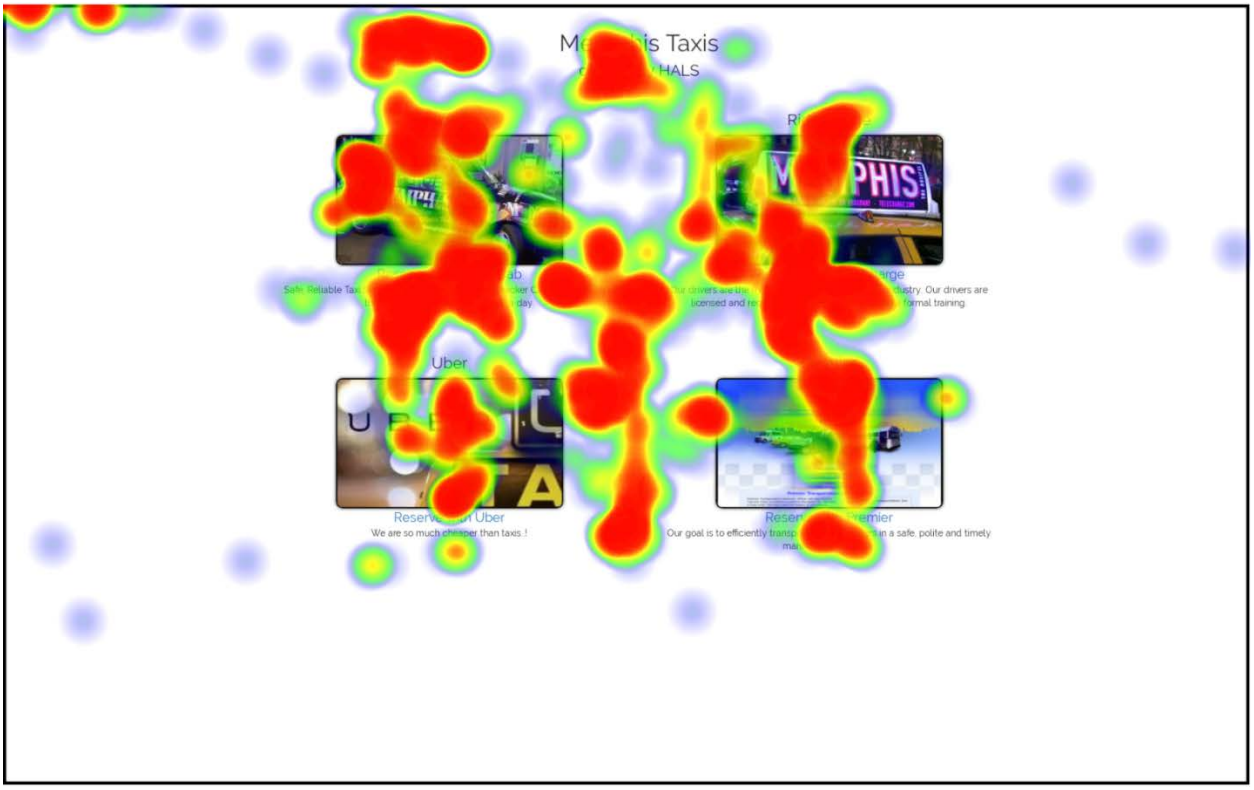


Figure 7: Version A Action Shot of Replay

Figure 8: Version A Final Shot of Replay

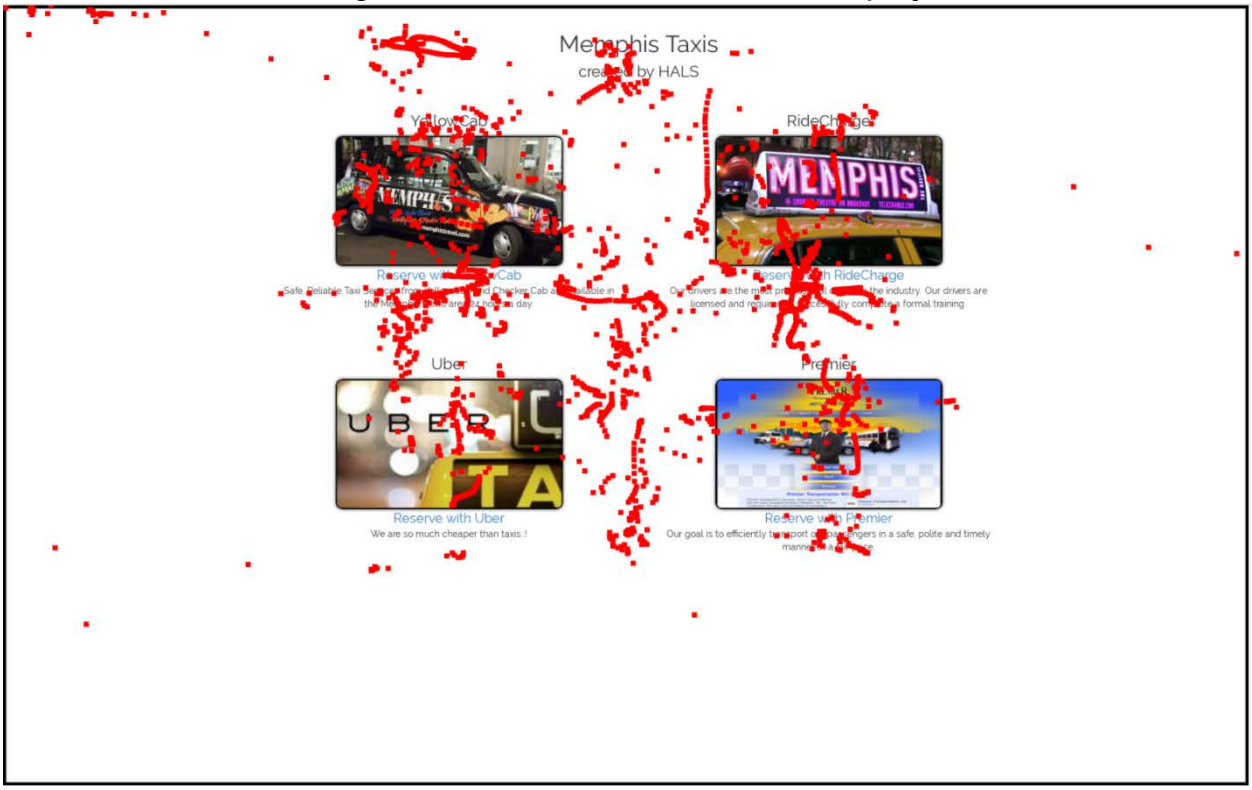


Figure 9: Version B Heatmap

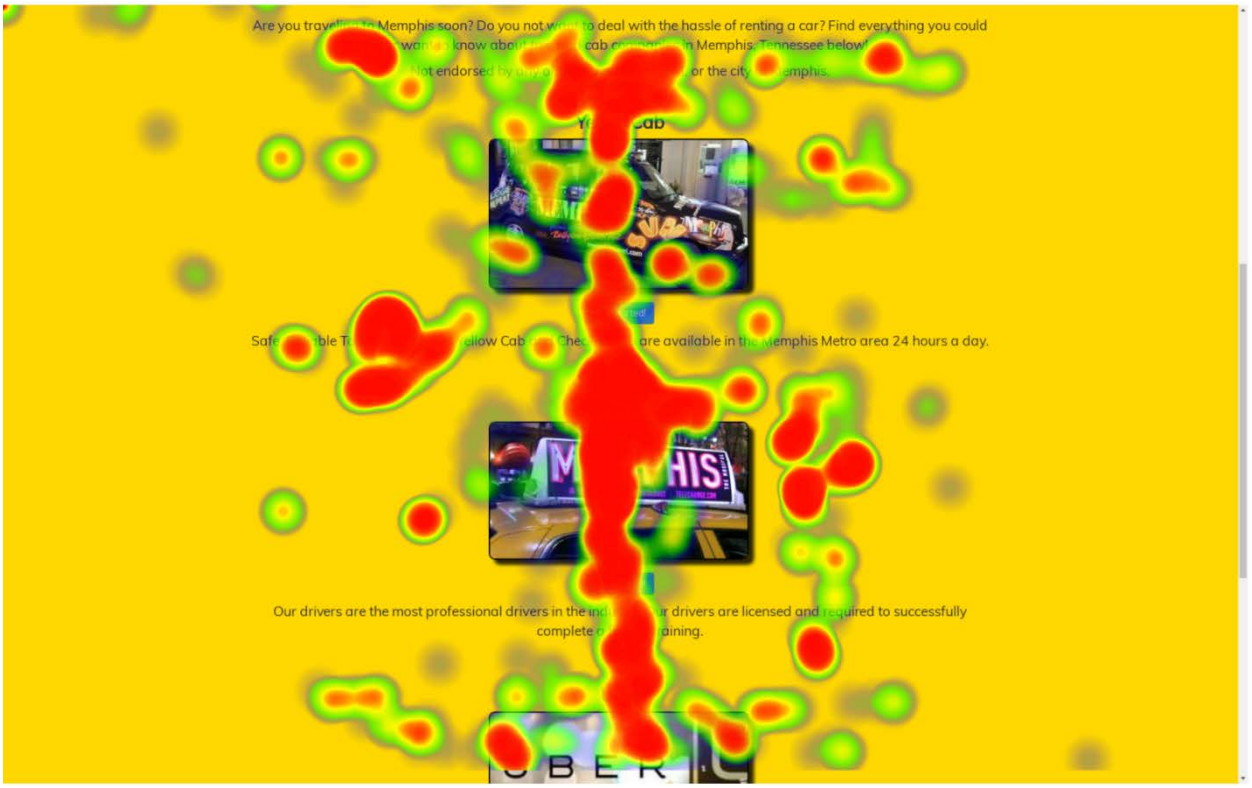


Figure 10: Version B Action Shot of Replay

Figure 11: Version B Final Shot of Replay

We can clearly see through Figures 9, 10, and 11 that on Version B, the user's gaze is mostly focused on the vertical center of the screen, following the vertically centered layout of the content. This is different from Figures 6, 7, and 8, which show that in Version A, the user's gaze is slightly more scattered, following the four quadrants of content laid out here. This closely matches the expectations we had in our qualitative eye tracking hypothesis.

Comparison

I would recommend Memphis Taxis Co. to use Version B as a starting point to make modifications and conduct more tests. My statistical tests show that there is a significant difference between Version A and Version B in time to click, and Version B had a longer time to click than Version A. Assuming Memphis Taxis Co. makes money through ads, and assuming the company makes more money when users stay on the page longer, using Version B as a starting point for future improvements would allow the company to make more money. Additionally, based on the eye tracking data, Version B is easier for users to focus on – the content is centered vertically, meaning users can focus their gaze on the vertical center of the page and scroll for more content rather than shifting their focus around the page. If Memphis Taxi Co. wants to improve other metrics such as click-through rate, however, additional modifications and tests will need to be conducted, as there were no other statistically significant differences between Version A and Version B.

Our A/B testing data tells us that Version B has a statistically significantly longer time to click than Version A, which is likely because Version B requires users to scroll down before they can view the content; this isn't visible in the eye tracking data, as it looks like the user's gaze in Version B was much more focused, implying the user would click faster. While eye tracking provides a more concrete view into how users might interact with the webpage, A/B testing allows us to gather more data about more users at one time, which is more useful for designing interfaces that will best suit the wider population. It is probably most beneficial to do eye tracking in the preliminary stages of testing to understand more closely how an individual user might interact with the webpage, then move on to A/B testing when we need to better understand how this interface will affect our consumer base as a whole.