

AB TESTING

CONTENTS

INTRODUCTION:

In this UI/UX study we are applying the powerful tools of A/B testing and eye-tracking to analyze and develop an understanding of whether different interface changes applied to a website create a statistically significant improvement. The metrics this study take into consideration for statistical significance in A/B testing include click-through rate, time to click, dwell time, and return rate.

LINK TO A/B WEBSITES: <https://vast-river-15187.herokuapp.com>

1. DEVELOPMENT OF HYPOTHESES:

Before collecting data for each of the metrics being evaluated, we developed null and alternative hypotheses based on what we expected to occur.



3. STATISTICAL TESTS AND EVALUATION

We conducted four statistical tests depending on the type of data we collected. We applied either a chi-squared test for categorical data or a t-test for continuous data.



2. EXPLANATION OF METRICS

We evaluate our statistical test results taking into consideration the p-values for each metric and also the confidence interval for time to click.



4. EYETRACKING AND ANALYSIS

Used eye-tracking data to develop heat-maps and visual representations of how a user interacted with each version of the sites being evaluated.



HYPOTHESES

CLICK THROUGH RATE:

Null: The click through rate I get on my Version A will be equal to the click through rate on my Version B. The differences between Version A and Version B are minor and contains the same number of elements so I expect users to interact with the elements similarly. While the text are aligned differently, users will more likely click the buttons in Version A than the pictures because of the grey background.

Alternative: The click through rate on my Version A is going to be less than that of Version B, because of the contrast with the grey background that increases the visibility of the buttons. You're more likely to click the buttons with the darker background as opposed to the white background. With Version B, there will be a focus on the images rather than the buttons.

TIME TO CLICK:

Null: The time to click on my Version B is going to be equal than that of Version A.

Alternative: The time to click I get on my Version B is going to be less than that of Version A because I expect users to experience reading the text in site A, but the layout of the text are emphasized. The text and “reserve” buttons are aligned to the left of the pictures of the companies because users read from left to right.

DWELL TIME:

Null: The dwelltime on my Version B is going to equal to the dwelltime of Version A.

Alternative: The dwell time on my Version A is going to be greater than that of Version B, because the user will want to know more information about the different car providers since they most likely took their time to read the description.

RETURN RATE:

Null: The return rate I get on my Version A will be equal to the return rate on my Version B.

Alternative: The return rate I get on my Version B is going to be greater than my return rate of Version A because the user is more likely to scan the entire page because of the distribution of the images and text and I expect them to come back and explore different companies.

METRICS

CLICK THROUGH RATE:

$$\text{Click Through Rate} = \frac{\text{Number of Unique Clicks}}{\text{Number of Unique Sessions}}$$

<timestamp>	<ver. a/b>	<page load time>	<click time>	<id of clicked element>	<sessionID>	<Click Through Rate>
2019-10-23T00:37:00.681492+00:00	app[web.1]: AB_TESTING:	A	1571801821964	0	0	mkgwvpuvv
2019-10-23T00:37:04.610748+00:00	app[web.1]: AB_TESTING:	A	1571801821964	1571801825888	mp3	mkgwvpuvv
2019-10-23T00:37:10.141291+00:00	app[web.1]: AB_TESTING:	A	1571791030125	0	0	djngwkpn
2019-10-23T00:37:35.016472+00:00	app[web.1]: AB_TESTING:	A	1571791030125	1571791055000	mp2	djngwkpn
2019-10-23T00:37:42.953965+00:00	app[web.1]: AB_TESTING:	A	1571791062934	0	0	djngwkpn
2019-10-23T00:37:44.30985+00:00	app[web.1]: AB_TESTING:	A	1571791062934	1571791064293	mp3	djngwkpn
2019-10-23T00:42:15.586811+00:00	app[web.1]: AB_TESTING:	A	1571791335564	0	0	djngwkpn
2019-10-23T00:42:27.805718+00:00	app[web.1]: AB_TESTING:	A	1571791335564	1571791347789	mp4	djngwkpn
2019-10-23T00:37:10.51026+00:00	app[web.1]: AB_TESTING:	A	1571791030252	0	0	noadnrqd
2019-10-23T00:37:16.509513+00:00	app[web.1]: AB_TESTING:	A	1571791030252	1571791036452	mp2	noadnrqd
2019-10-23T00:37:17.192134+00:00	app[web.1]: AB_TESTING:	A	1571791037082	0	0	wagijhjp

$$\begin{aligned} \text{Click Through Rate} &= \frac{\text{Number of Unique Clicks}}{\text{Number of Unique Sessions}} \\ &= \frac{3}{4} \\ &= 0.75 \end{aligned}$$

In the example above the yellow cells correspond to sessions that have clicks and the green cell to a session that didn't have a click.

TIME TO CLICK:

$$\text{Time to click} = \text{click (leave time)} - \text{initial page load time}$$

<timestamp>	<ver. a/b>	<page load time>	<click time>	<id of clicked element>	<sessionID>	<Time to Click>
2019-10-23T00:36:04.116267+00:00	app[web.1]: AB_TESTING:	A	1571790963705.00	0	0	xyovdf
2019-10-23T00:36:11.699169+00:00	app[web.1]: AB_TESTING:	A	1571790963705.00	1571790971275.00	mp2	xyovdf
2019-10-23T00:37:33.681566+00:00	app[web.1]: AB_TESTING:	A	1571791053267.00	0	0	xyovdf

To apply this formula to a log entry from our Heroku logs you would subtract the first click time (green cell) time stamp by the initial page load time (blue cell). The result for this log is shown in the yellow cell.

$$\begin{aligned} \text{Time to Click} &= \text{click (leave time)} - \text{initial page load time} \\ &= 1571790971275 - 1571790963705 \\ &= 7570 \end{aligned}$$

*Note: These are just example calculations and explanations. The full 8 metric results can be found at the end of the metrics section.

For Excel Sheet with full calculations visit:

<https://docs.google.com/spreadsheets/d/11zf-nkDTcohHhiqQ4tz0sxgF08Qj1NtKmvEdRk6VDyY/edit?usp=sharing>

METRICS

DWELL TIME:

Dwell Time = return page load time – click time to leave page

<timestamp>	<ver. a/b>	<page load time>	<click time>	<id of clicked element>	<sessionID>	<Dwell Time>
2019-10-23T00:36:04.116267+00:00	app[web.1]: AB_TESTING: A	1571790963705	0	0	xyovdf	
2019-10-23T00:36:11.699169+00:00	app[web.1]: AB_TESTING: A	1571790963705	1571790971275	mp2	xyovdf	
2019-10-23T00:37:33.681566+00:00	app[web.1]: AB_TESTING: A	1571791053267	0	0	xyovdf	81992

In the example log above the click time is denoted by the green cell and the return page load time when the user returned is in the blue cell.

$$\begin{aligned}
 \text{Dwell Time} &= \text{return page load time} - \text{click time to leave page} \\
 &= 1571791053267 - 1571790971275 \\
 &= 81992
 \end{aligned}$$

RETURN RATE:

Return Rate = $\frac{\text{Number of sessions that left and returned}}{\text{Number of session that leave}}$

<timestamp>	<ver. a/b>	<page load time>	<click time>	<id of clicked element>	<sessionID>
2019-10-23T00:37:00.681492+00:00	app[web.1]: AB_TESTING: A	1571801821964	0		0 mkgywpvuvw
2019-10-23T00:37:04.610748+00:00	app[web.1]: AB_TESTING: A	1571801821964	1571801825888	mp3	mkgywpvuvw
2019-10-23T00:37:10.141291+00:00	app[web.1]: AB_TESTING: A	1571791030125	0		0 djngwkpn
2019-10-23T00:37:35.016472+00:00	app[web.1]: AB_TESTING: A	1571791030125	1571791055000	mp2	djngwkpn
2019-10-23T00:37:42.953965+00:00	app[web.1]: AB_TESTING: A	1571791062934	0		0 djngwkpn
2019-10-23T00:37:44.30985+00:00	app[web.1]: AB_TESTING: A	1571791062934	1571791064293	mp3	djngwkpn
2019-10-23T00:42:15.586811+00:00	app[web.1]: AB_TESTING: A	1571791335564	0		0 djngwkpn
2019-10-23T00:42:27.805718+00:00	app[web.1]: AB_TESTING: A	1571791335564	1571791347789	mp4	djngwkpn
2019-10-23T00:37:10.51026+00:00	app[web.1]: AB_TESTING: A	1571791030252	0		0 noadnrqd
2019-10-23T00:37:16.509513+00:00	app[web.1]: AB_TESTING: A	1571791030252	1571791036452	mp2	noadnrqd
2019-10-23T00:37:17.192134+00:00	app[web.1]: AB_TESTING: A	1571791037082	0		0 wagijhjp

In the example log above the yellow cells represents sessions that left and didn't return, and the blue cells are the session that did return.

*Note: The green cell is a session that never left the page, so it isn't considered in the calculation

$$\begin{aligned}
 \text{Return Rate} &= \frac{\text{Number of sessions that left and returned}}{\text{Number of sessions that leave}} \\
 &= \frac{1}{3} \\
 &= 0.3333 \dots \\
 &\approx 0.33
 \end{aligned}$$

*Note: These are just example calculations and explanations. The full 8 metric results can be found at the end of the metrics section.

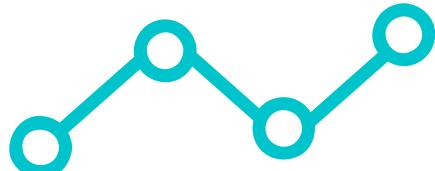
For Excel Sheet with full calculations visit:

<https://docs.google.com/spreadsheets/d/11zf-nkDTcohHhiqQ4tz0sxgF08Qj1NtKmvEdRk6VDyY/edit?usp=sharing>

METRIC RESULTS

Final Results		
Metrics	Version A	Version B
Click Through Rate	0.6667	0.4872
Average Dwell Time	311164.2	136432.9
Average Time- to Click	40074.6	8476.0526
Return Rate	0.500	0.526

As seen in the data above, the click through rate of Version A is slightly above Version B whereas the Dwell time and time-to-click of version A is a lot greater than Version B. Lastly, the Return Rate between Version A and Version B is similar but version B is slightly above.



STATISTICAL TESTS

CLICK-THROUGH RATE:

We used the chi-square test formula for the click through rate to test how likely it is that an observed distribution is due to chance. Also, we used this test because we are comparing categorical data of the observed clicks/no clicks and the expected clicks/no clicks rate. Because we can construct a table to organize the data, using this test will make finding the calculations much simpler.

CLICK-THROUGH DATA:

Click Through Rate

$$\chi^2 = \sum_{\text{data}} \frac{(O-E)^2}{E}$$

OBSERVED	#Sessions w/clicks	#Sessions w/o clicks	TOTAL
Interface A	20	10	30
Interface B	19	10	29
TOTAL	39	20	59

EXPECTED	#Sessions w/clicks	#Sessions w/o clicks	TOTAL
Interface A	$\frac{30 \times 39}{59} = 19.83050847 \dots$	$\frac{30 \times 20}{59} = 10.16949153 \dots$	30
Interface B	$\frac{29 \times 39}{59} = 19.16949153 \dots$	$\frac{29 \times 20}{59} = 9.830508475 \dots$	29
TOTAL	39	20	59

CHI-SQUARE CALCULATION:

$$\begin{aligned}
 & \sum_{\text{data}} \frac{(O-E)^2}{E} \\
 & \frac{(20-19.83050847)^2}{19.83050847} + \frac{(19-19.16949153)^2}{19.16949153} \\
 & + \frac{(10-10.16949153)^2}{10.16949153} + \frac{(10-9.830508475)^2}{9.830508475} \\
 & + \frac{(0.1694915254)^2}{19.83050847} + \frac{(-0.1694915254)^2}{19.16949153} \\
 & + \frac{(-0.1694915254)^2}{10.16949153} + \frac{(0.1694915254)^2}{9.830508475} \\
 & \frac{0.02872737719}{19.83050847} + \frac{0.02872737719}{19.16949153} \\
 & + \frac{0.02872737719}{10.16949153} + \frac{0.02872737719}{9.830508475} \\
 & 0.001448645516 + 0.00149859881 + 0.002824858757 + 0.00292226768 \\
 & 0.008694370763
 \end{aligned}$$

The degree of freedom is :

$$(row-1)*(col-1) = (2-1) * (2-1) = 1$$

STATISTICAL TESTS

RETURN RATE:

Similarly, we're dealing with categorical data, so we used the chi-squared test. We can compute the statistical data by comparing the observed return or no return data of site A and B to the expected value, which is the percentage of clicks or no clicks overall multiplied by the number of sessions.

RETURN RATE DATA:

Return Rate

$$\chi^2 = \sum_{\text{data}} \frac{(O-E)^2}{E}$$

OBSERVED	#Sessions that returned	#Sessions that did not return	TOTAL
Interface A	10	10	20
Interface B	10	9	19
TOTAL	20	19	39

EXPECTED	#Sessions that returned	#Sessions that did not return	TOTAL
Interface A	$\frac{20 \times 20}{39} = 10.25641026\dots$	$\frac{20 \times 19}{39} = 9.743589744\dots$	20
Interface B	$\frac{19 \times 20}{39} = 9.743589744\dots$	$\frac{19 \times 19}{39} = 9.256410256\dots$	19
TOTAL	20	19	39

CHI-SQUARE CALCULATION:

$$\begin{aligned}
 & \sum_{\text{data}} \frac{(O-E)^2}{E} \\
 & + \frac{(10-10.25641026)^2}{10.25641026} + \frac{(10-9.743589744)^2}{9.743589744} \\
 & + \frac{(10-9.743589744)^2}{9.743589744} + \frac{(9-9.256410256)^2}{9.256410256} \\
 & + \frac{(-0.2564102564)^2}{10.25641026} + \frac{(0.2564102564)^2}{9.743589744} \\
 & + \frac{(0.2564102564)^2}{9.743589744} + \frac{(-0.2564102564)^2}{9.256410256} \\
 & + \frac{0.06574621959}{10.25641026} + \frac{0.06574621959}{9.743589744} \\
 & + \frac{0.06574621959}{9.743589744} + \frac{0.06574621959}{9.256410256} \\
 & 0.00641025641 + 0.006747638327 + 0.006747638327 + 0.007102777186 \\
 & = 0.02700831025
 \end{aligned}$$

The degree of freedom is :

$$(\text{row}-1) * (\text{col}-1) = (2-1) * (2-1) = 1$$

STATISTICAL TESTS

DWELL-TIME

Since we're dealing with continuous data, we used the T-Test for calculating the dwell time because it is used to compare the mean of two given samples, in this case, it would be site A and site B. It will help us determine whether there's a significant difference of dwell time between site A and site B.

DWELL-TIME DATA:

DWELL TIME	
Version A X_1	Version B X_2
81992	19143
398749	9794
7934	1041122
:	:
2092	4929
3094	7601
18666	4389
ΣX	3111642
N	10
ΣX^2	6818170180724
\bar{X}	311164.2

T-TEST CALCULATIONS:

DWELL TIME T-TEST

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2} \right) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}} \\ &= \frac{311164.2 - 124834.4545}{\sqrt{\left[\left(6818170180724 - \frac{(3111642)^2}{10} \right) + \left(114060908563 - \frac{(1373179)^2}{11} \right) \right] \cdot \left[\frac{1}{10} + \frac{1}{11} \right]}} \\ &= \frac{186329.7455}{261758.5725} \\ &= 0.7118381786 \end{aligned}$$

STATISTICAL TESTS

AVERAGE TIME TO CLICK

Similarly to dwell time, we are dealing with continuous data, and we used the T-Test for calculating the average time to click because it is used to compare the mean of two given groups.

DWELL-TIME DATA:

TIME TO CLICK

	Version A X_1	Version B X_2
1570	1888	
424802	6834	
3924	3303	
:	:	
2428	6821	
83499	2816	
46313	60174	
ΣX	801000	161045
N	19	18
ΣX^2	206210154374	4608160409
\bar{X}	42157.89474	8946.944444

T-TEST CALCULATIONS:

TIME TO CLICK T-TEST

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2} \right) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}} \\ &= \frac{42157.89474 - 8946.944444}{\sqrt{\left[\left(206210154374 - \frac{(801000)^2}{19} \right) + \left(4608160409 - \frac{(161045)^2}{18} \right) \right] \cdot \left[\frac{1}{19} + \frac{1}{18} \right]}} \\ &= \frac{33210.95029}{23298.4563473} \\ &= 1.42545711162 \end{aligned}$$

EVALUATION OF METRICS

CHI-TEST RESULTS:

Metrics	Chi-squared value	Percentage point chi-squared distribution at p > 0.05 at df=1
Click through rate	0.00869	3.84
Return rate	0.02708	3.84

For click through rate, since 0.00869 is not greater than 3.84, the click through rate difference between version A and version B is not statistically significant.

Similarly, since 0.02708 is not greater than 3.84, the return rate difference between version A and version B is not statistically significant.

T-TEST RESULTS:

Metrics	Degree of Freedom	t-value	t-distribution at p > 0.05 at given df
Dwell Time	19	0.7118	1.729133
Time-to-Click	35	1.4254	2.0301

For Dwell time, since 0.7118 is not greater than 1.729133, the Dwell time difference between version A and version B is not statistically significant.

For time to click, since 1.4254 is not greater than 2.0301, the time-to-click difference between version A and version B is not statistically significant.

CONFIDENCE INTERVAL:

With 95% confidence we know the difference in means of the time-to-click between Version A and Version B is between -14.087 and 80.509 seconds. The standard error of the difference is 23.298 Seconds, and the margin of error is 47.297 Seconds.

$$\text{Confidence Interval} = 33.210 \pm 47.297$$

$$\text{Margin of error} = (\text{t-distribution value at } p = 0.05 \text{ and } df = 35) * \text{Standard Error}$$

If a 95% confidence interval includes zero, then there is no statistically meaningful or statistically significant difference between the groups. Our 95% confidence interval lies between -14.087 and 80.509 seconds which includes 0 seconds as part of our confidence interval. Therefore, we can conclude that the difference between the time-to-click between Version A and Version B is not statistically meaningful or statistically significant.

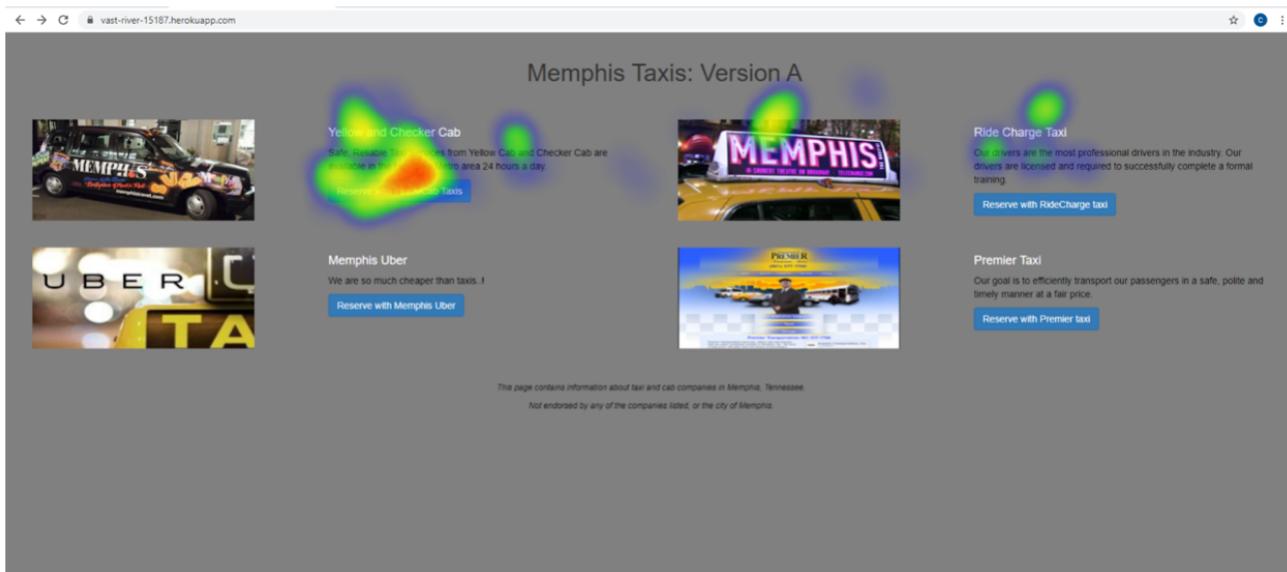
EYE TRACKING DATA

QUALITATIVE HYPOTHESIS:

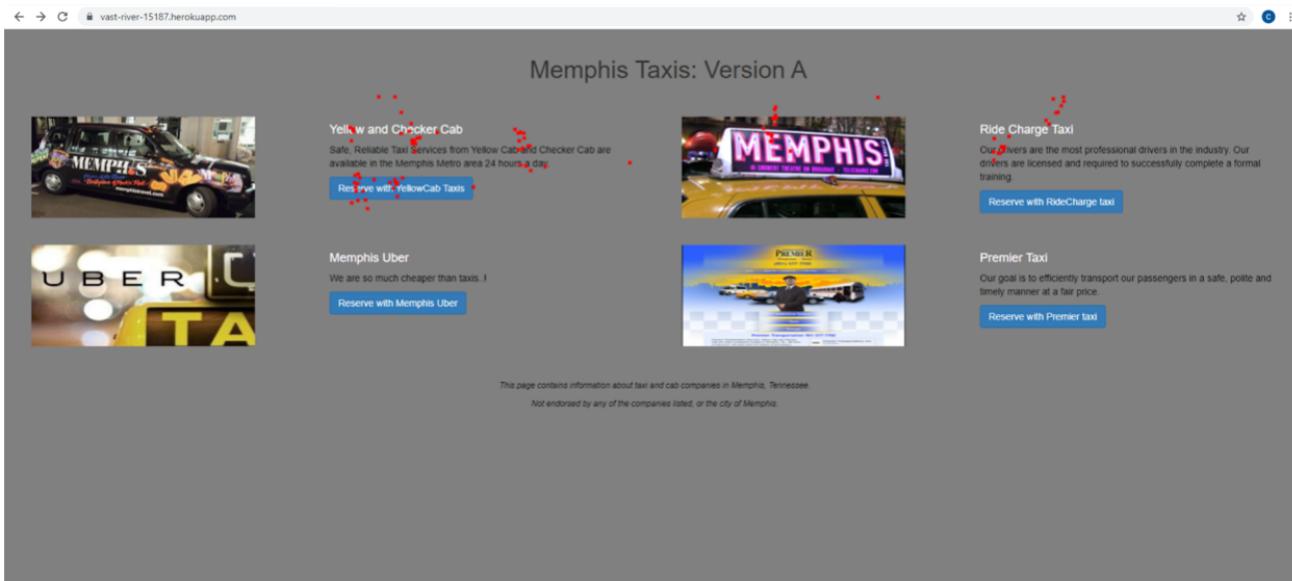
Version B will have a greater proportion of eye gazes on the images on the screen than Version A because there is a greater hierarchy on the images than the buttons and text on the site.

Version A will have a greater proportion of eye gazes at the top of the screen because the text and buttons are aligned left to the images.

Heatmap Site A



Replay: Action Shot Site A



EYE TRACKING DATA

Replay: Final Shot Site A

This screenshot shows the Memphis Taxis: Version A website. The page title is "Memphis Taxis: Version A". It features four main sections: "Yellow Cab and Checker Cab" (with a car image), "Memphis Uber" (with a car image), "Ride Charge Taxi" (with a car image), and "Premier Taxi" (with a car image). Each section includes a brief description and a "Reserve with [Company Name] Taxi" button. A small note at the bottom states: "This page contains information about taxi and cab companies in Memphis, Tennessee. Not endorsed by any of the companies listed, or the city of Memphis." A heatmap overlay is present, showing high engagement with the Yellow Cab and Checker Cab section.

Heat Map Site B

This screenshot shows the Memphis Taxis Version B website. The page title is "Memphis Taxis Version B". It features four main sections: "Yellow Cab and Checker Cab" (with a car image), "Memphis Uber" (with a car image), "Ride Charge Taxi" (with a car image), and "Premier Taxi" (with a car image). Each section includes a brief description and a "Reserve with [Company Name] Taxi" button. A small note at the bottom states: "This page contains information about taxi and cab companies in Memphis, Tennessee. Not endorsed by any of the companies listed, or the city of Memphis." A heatmap overlay is present, showing high engagement with the Uber section and the Ride Charge Taxi section.

EYE TRACKING DATA

Replay: Action Shot Site B

This screenshot shows a web page titled "Memphis Taxis Version B". The page features four taxi service options: Yellow Cab Taxi, RideCharge Taxi, Memphis Uber, and Premier Taxi. Each service is represented by a small image and a "Reserve with [Service Name] Taxi" button. Below each button is a brief description of the service's offerings. Red dots and lines indicate eye tracking data, showing where users are looking across the page. A footer note states: "This page contains information about taxi and cab companies in Memphis, Tennessee. Not endorsed by any of the companies listed, or the city of Memphis."

Memphis Taxis Version B

Yellow Cab Taxi

Safe, Reliable Taxi Services from Yellow Cab and Checker Cab are available in the Memphis Metro area 24 hours a day.

RideCharge Taxi

Our drivers are the most professional drivers in the industry. Our drivers are licensed and required to successfully complete a formal training.

Memphis Uber

We are so much cheaper than taxis..!

Premier Taxi

Our goal is to efficiently transport our passengers in a safe, polite and timely manner at a fair price.

This page contains information about taxi and cab companies in Memphis, Tennessee. Not endorsed by any of the companies listed, or the city of Memphis.

Replay: Final Shot Site B

This screenshot shows the same "Memphis Taxis Version B" website as the previous one, but with more extensive and dense red dot and line annotations. These annotations cover a larger portion of the page, particularly around the taxi images and the "Reserve with [Service Name] Taxi" buttons, indicating a more focused user interaction with these specific elements compared to the initial action shot.

Memphis Taxis Version B

Yellow Cab Taxi

Safe, Reliable Taxi Services from Yellow Cab and Checker Cab are available in the Memphis Metro area 24 hours a day.

RideCharge Taxi

Our drivers are the most professional drivers in the industry. Our drivers are licensed and required to successfully complete a formal training.

Memphis Uber

We are so much cheaper than taxis..!

Premier Taxi

Our goal is to efficiently transport our passengers in a safe, polite and timely manner at a fair price.

This page contains information about taxi and cab companies in Memphis, Tennessee. Not endorsed by any of the companies listed, or the city of Memphis.

EYE TRACKING ANALYSIS



OUR GROUP ANALYSIS:

- According to the eye tracking data, in Version B there is more scanning of information but it is mainly targeted towards looking at the images and buttons instead of the company descriptions as we have hypothesized. It seems that the user of Version B clicked on "Reserve with Memphis Uber" not because they read the description of the company but because of the image and the familiarity to Uber as a form of cheap transportation.
- Moreover, the user made an effort to read the description of the first company however the user did not do the same for the other three and based their decision mainly on the image.
- According to the heatmap and replay for Version A, the user has taken the time to read the description of the company and make an informed decision as we have hypothesized. However, the user only looked at the first two options instead of all four. This could mean that the user read the description for the first company and found it to be satisfying or the layout of the options doesn't encourage users to explore all the options available.

COMPARISON & RECOMMENDATIONS

RECOMMENDATION BASED ON RESULTS

Based on the evidence from the A/B testing and Eye Tracking, we would recommend the company stakeholders to use site B. However, It greatly depends on the company's goal and their popularity. There was a greater proportion of eye gaze on the images on the webpage so we would suggest the addition of eye-catching images to the website to prompt the user to click

the buttons underneath the images. In Version A, the eye tracking results and the heatmap showed that the users eyes were concentrated at the top of the page but it is more ideal if

the user scans the entire page. In addition, we would recommend the stakeholders to conduct more tests with a larger sample to make a better informed decision.



A/B TESTING VS. EYE TRACKING DATA

In comparison, our A/B testing and eye tracking data provided similar behavior trends. From the eye tracking data, you can observe that the users tracking site A carefully read the text to try to make an informed decision instead of scanning the whole page at first glance. The data in the A/B testing supports this claim because the average dwell time for site A was greater than Version B's dwell time. The disadvantages of using A/B testing versus eye tracking is that you can't visualize the eye gazing of certain elements on the site.

UNETHICAL USE OF METRICS

Two metrics that could be used unethically are dwell time and return rate. A metric measuring dwell time can be used unethically if maximizing dwell time means that the website relies on increasing engagement and keeping track of the users activity. Also, to increase dwell time, the company can unethically input ads to the page based off of their past searches to bring significance to elements on the website. In addition, a metric measuring return rate can be used unethically if increasing how often the user leave the landing page and returns relies on adding more clickable elements that provide sources to irrelevant information.

