A/B Testing & Eye Tracking

Done With: Jun Yi Hu, Dani Smith, Shreeya Rajgarhia

A/B Testing and Eye Tracking for Memphis Taxi

We were posed with the task of using two data collection methods to determine the best design for a Taxi Listing page. By collecting A/B Testing and Eye Tracking data and performing a series of metric and statistical calculations, we attempted to distinguish the differences in usability and viability between two designs for the website.

A/B Testing: Data Collection

Our group spent some time designing two versions of the website. We then came up with null and alternative hypotheses about how each version would be used according to four metrics:

Click Through Rate:

- Null Hypothesis: Version A and B both have the same click through rate
- Alternative Hypothesis: Version A has a higher click through rate than Version B because the simplistic layout of A encourages clicking through every link for information before making a choice.

Time To Click:

- Null Hypothesis: Version A and B have the same time to click
- Alternative Hypothesis: Version B takes a longer time to click than Version A because there is more information to process on Version B, meaning the user would take more time to read before choosing.

Dwell Time:

- Null Hypothesis: Version A and B have the same dwell time
- Alternative Hypothesis: Version A has a longer dwell time than B because the user most likely needs to stay on the A reserve pages for longer to read the information about the options before returning, whereas Version B provides more information, meaning there will be less to read on the reserve page.

Return Rate:

- Null Hypothesis: Both versions have the same return rate
- Alternative Hypothesis: Version B has a higher return rate than A because it provides more information and has a more learnable design, meaning that, as long as the user has a satisfactory experience reserving, they would be more likely to trust it and come back again to reserve.

One we formed these hypotheses, we coded each version of the website, and put the site online using Heroku so that we could have users visit to collect usage data. The website can be found here: https://afternoon-shelf-26309.herokuapp.com/

A/B Testing: Data Analysis

After collecting our data logs, we split them up between our A and B sites into a spreadsheet.

Click Through Rate

To calculate our click through rate, we summed the total number of unique clicks for Version A and B, as well as the number of unique sessions. We then divided the number of unique clicks by the number of unique sessions.

A: 73.17%

B: 66.67%

Time-To-Click

To calculate the time-to-click, we found the number of unique sessions that clicked a link on the version they visited, then calculated the difference for their page load time and their click time. We then averaged all the resutls.

A: 26565.129 ms

B: 14006.8065 ms

Dwell Time

For calculating dwell time, we first summed the number of unique sessions that returned to the landing page after clicking another link. We then found the difference of their first click time and their 2nd page load time. We then found the average of all the results.

A: 55619.21 ms

B: 43545.31 ms

Return Rate

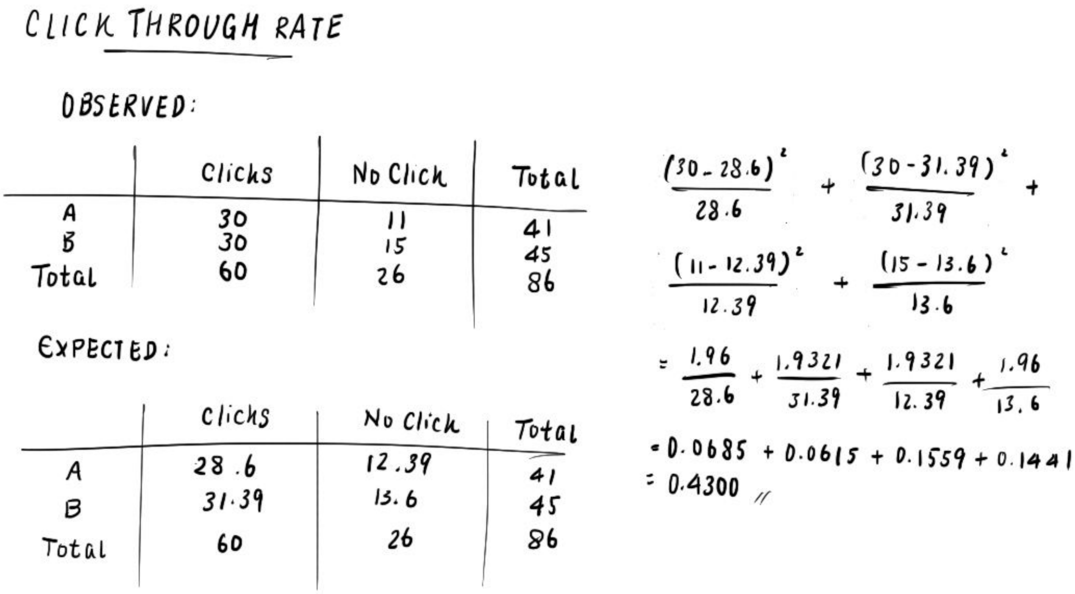
We calculated the return rate by taking the number of unique sessions that returned to the landing page, summing the number of unique sessions that left the page after their first click, and the number of unique sessions that visited the page and did not click anything at all. We then divided those 2 numbers together.

A: 13.89%

B: 18.92%

After calculating each metric, we had to determine if the data and calculations were actually viable for use. So, we ran statistical tests - either a Chi Squared or T Test - on each to determine their statistical significance at a margin of 95%.

Click-Through Rate: Chi Squared Test



CLICK THROUGH RATE

OBSERVED:

	Clicks	No Click	Total
A	30	11	41
B	30	15	45
Total	60	26	86

EXPECTED:

	Clicks	No Click	Total
A	28.6	12.39	41
B	31.39	13.6	45
Total	60	26	86

$$\frac{(30-28.6)^2}{28.6} + \frac{(30-31.39)^2}{31.39} + \frac{(11-12.39)^2}{12.39} + \frac{(15-13.6)^2}{13.6}$$

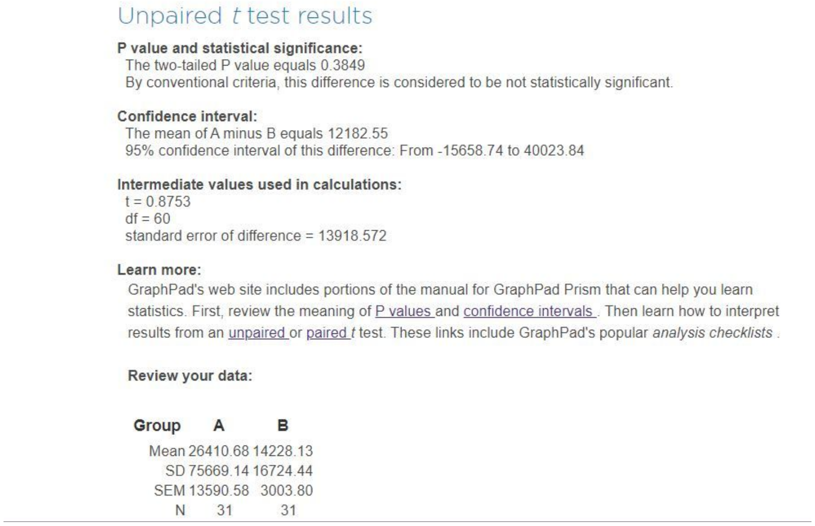
$$= \frac{1.96}{28.6} + \frac{1.9321}{31.39} + \frac{1.9321}{12.39} + \frac{1.96}{13.6}$$

$$= 0.0685 + 0.0615 + 0.1559 + 0.1441$$

$$= 0.4300 \; //$$

We chose the Chi Squared Test because it is meant to compare multiple variables, or categories of information. Here, we have two: whether or not the user clicked. Our resulting value, 0.4300, is far from the 95% value, which is 3.84, meaning that our results were not statistically significant. Therefore, we cannot say for certain that versions A and B do not have the same click-through rate.

Time-To-Click: T-Test



Unpaired *t* test results

P value and statistical significance:
The two-tailed P value equals 0.3849
By conventional criteria, this difference is considered to be not statistically significant.

Confidence interval:
The mean of A minus B equals 12182.55
95% confidence interval of this difference: From -15658.74 to 40023.84

Intermediate values used in calculations:
t = 0.8753
df = 60
standard error of difference = 13918.572

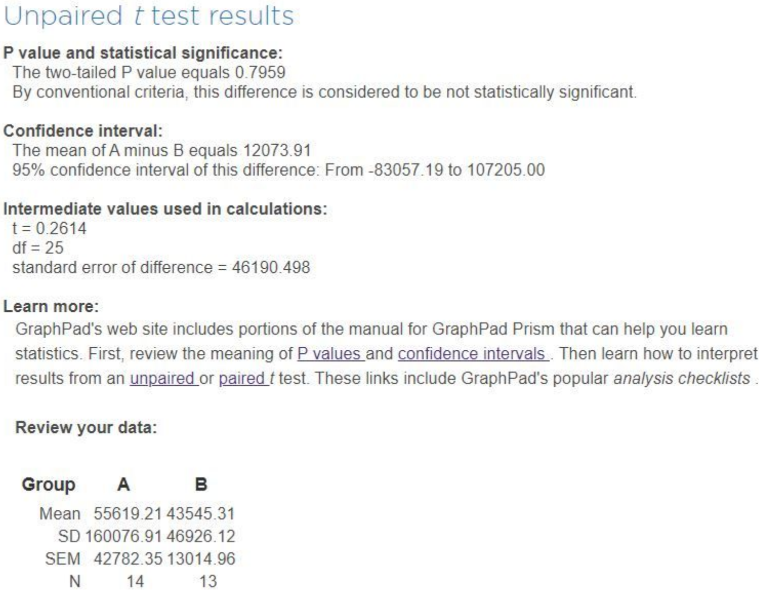
Learn more:
GraphPad's web site includes portions of the manual for GraphPad Prism that can help you learn statistics. First, review the meaning of P values and confidence intervals . Then learn how to interpret results from an unpaired or paired *t* test. These links include GraphPad's popular *analysis checklists* .

Review your data:

Group	A	B
Mean	26410.68	14228.13
SD	75669.14	16724.44
SEM	13590.58	3003.80
N	31	31

We used a T-Test for our Time-To-Click because it is meant to compare means between different samples. Our p-value, 0.3849, falls well below the required range for 95%, which is 6.313752. So, we cannot conclude whether our null hypothesis is false, as our data was not statistically significant here.

Dwell Time: T-Test



Unpaired *t* test results

P value and statistical significance:
The two-tailed P value equals 0.7959
By conventional criteria, this difference is considered to be not statistically significant.

Confidence interval:
The mean of A minus B equals 12073.91
95% confidence interval of this difference: From -83057.19 to 107205.00

Intermediate values used in calculations:
t = 0.2614
df = 25
standard error of difference = 46190.498

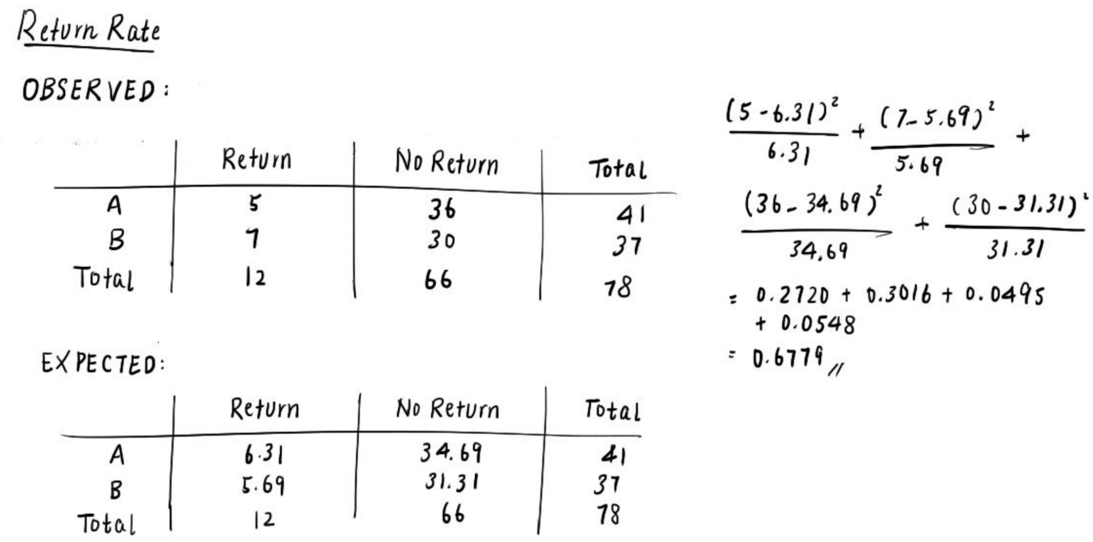
Learn more:
GraphPad's web site includes portions of the manual for GraphPad Prism that can help you learn statistics. First, review the meaning of P values and confidence intervals . Then learn how to interpret results from an unpaired or paired *t* test. These links include GraphPad's popular *analysis checklists* .

Review your data:

Group	A	B
Mean	55619.21	43545.31
SD	160076.91	46926.12
SEM	42782.35	13014.96
N	14	13

We also use a T-Test for our Dwell time for the same reason, as we are working with averages of two independent samples. Our p-value, 0.7959, also falls well below the required range for 95%, which is 6.313752, meaning we cannot conclude whether our null hypothesis is false, as our results were not statistically significant.

Return Rate: Chi Squared Test



Return Rate

OBSERVED:

	Return	No Return	Total
A	5	36	41
B	7	30	37
Total	12	66	78

EXPECTED:

	Return	No Return	Total
A	6.31	34.69	41
B	5.69	31.31	37
Total	12	66	78

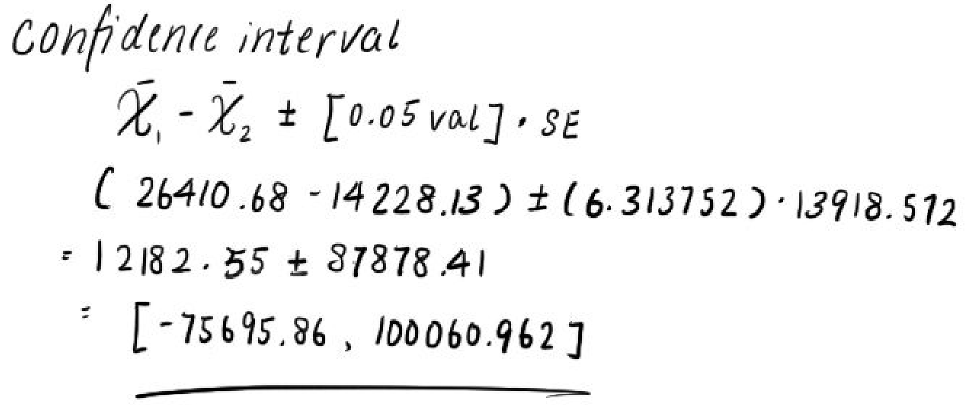
$$\frac{(5-6.31)^2}{6.31} + \frac{(7-5.69)^2}{5.69} + \frac{(36-34.69)^2}{34.69} + \frac{(30-31.31)^2}{31.31}$$

$$= 0.2720 + 0.3016 + 0.0495 + 0.0548$$

$$= 0.6779 \; //$$

We chose the Chi-Squared Test for our Return Rate as we are again comparing multiple variables, in this case the number of users who did vs didn’t return. Our calculated value, 0.6779, is lower than our target value of 3.84, meaning that our results were not statistically significant. We, again, cannot say for certain that versions A and B do not have the same click-through rate.

Click-Through Rate: Confidence Interval



confidence interval

$$\bar{x}_1 - \bar{x}_2 \pm [0.05 \text{ val}] \cdot SE$$

$$(26410.68 - 14228.13) \pm (6.313752) \cdot 13918.572$$

$$= 12182.55 \pm 87878.41$$

$$= [-75695.86, 100060.962]$$

The confidence interval that we calculated has 0 within its range. This indicates that the difference between our Click-Through Rates is not statistically significant. So, we cannot definitely confirm or deny either our null or alternative hypothesis.

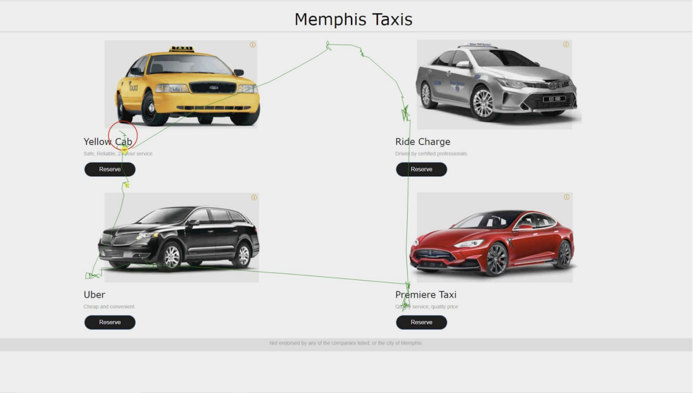
EYE TRACKING

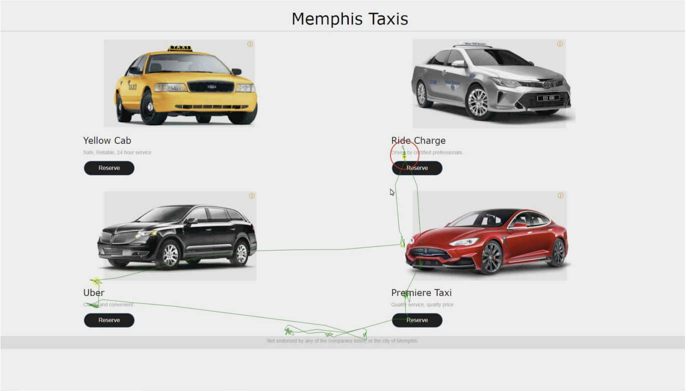
While A/B Testing is an effective way to test an interface design, it does not necessarily provide the full picture of the design’s strengths and flaws. So, we also performed some eye tracking tests on each of our designs. Before running the testing, we hypothesized how the web pages may be interacted with differently.

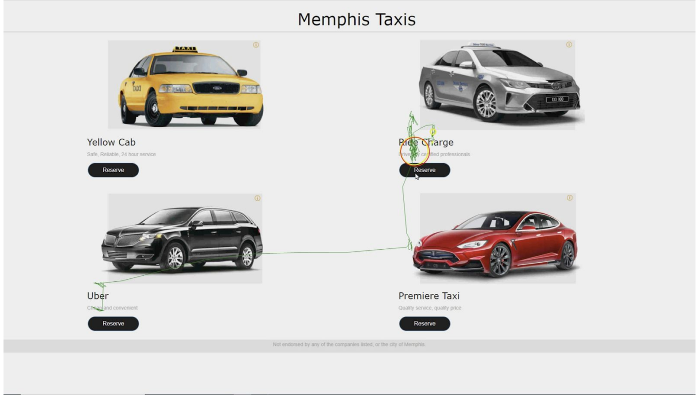
Qualitative Hypothesis

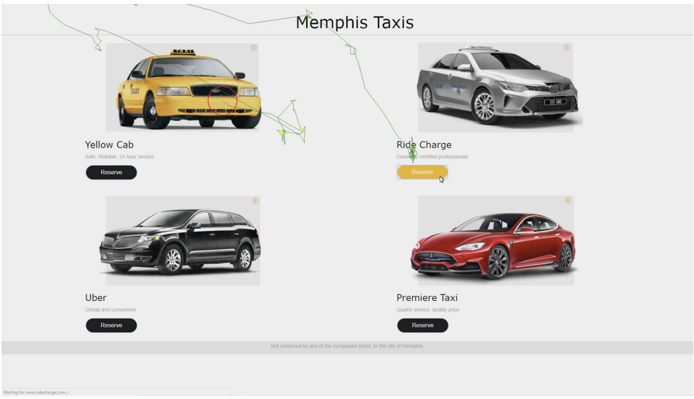
We expect Version A to be read with a more circular eye motion that covers a larger portion of the screen, whereas Version B will be read with sweeping eye movements left and right across the top portion of the screen. This is because Version A has the information arranged in a large, centered grid, while Version B has it set up in a single row near the top.

Version A: Video Link


Memphis Taxis
Yellow Cab
Safe, Reliable, 24 hour service
Reserve
Ride Charge
Driven by certified professionals
Reserve
Uber
Cheap and convenient
Reserve
Premiere Taxi
Quality service, quality price
Reserve
Not endorsed by any of the companies listed, or the city of Memphis.

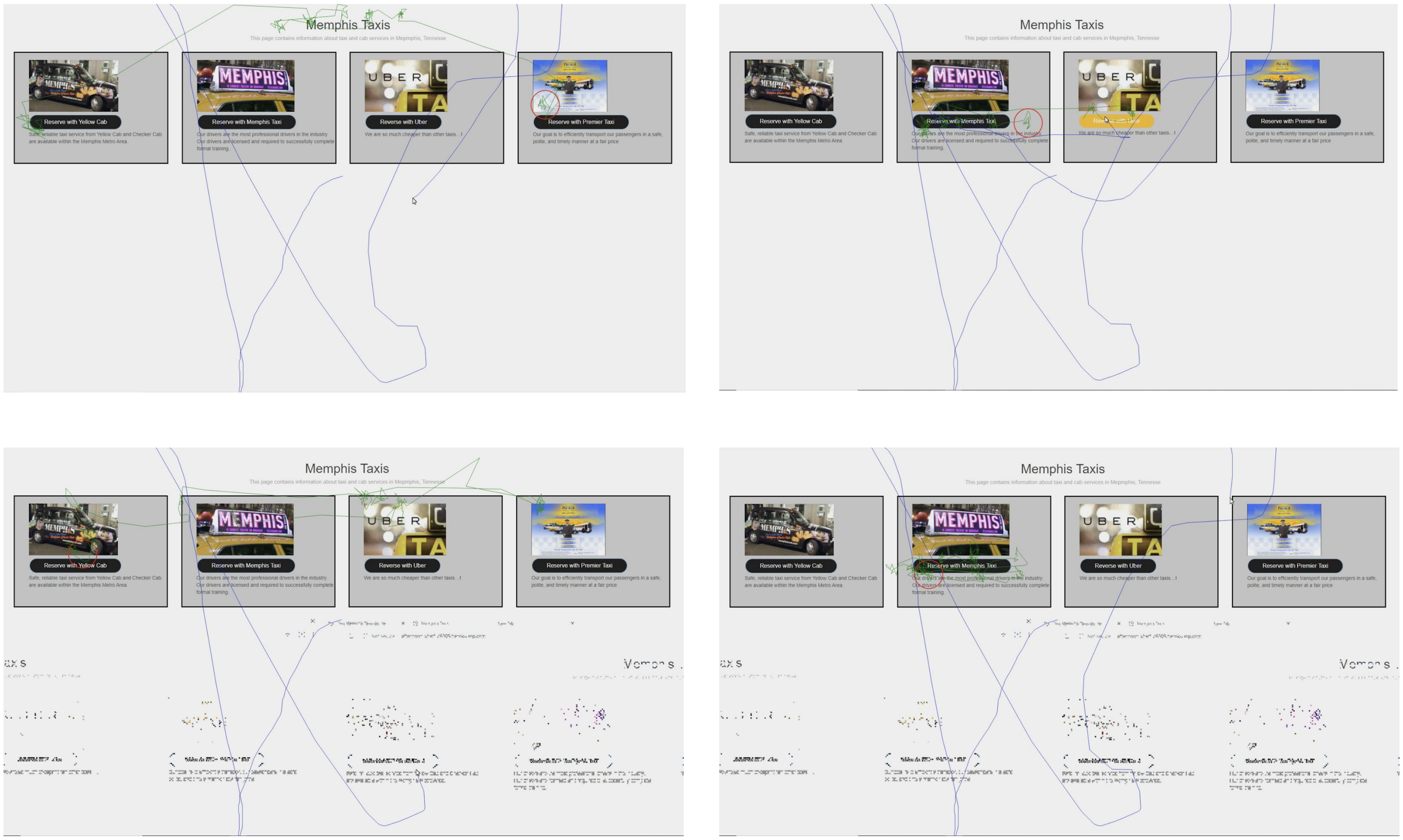

Memphis Taxis
Yellow Cab
Safe, Reliable, 24 hour service
Reserve
Ride Charge
Driven by certified professionals
Reserve
Uber
Cheap and convenient
Reserve
Premiere Taxi
Quality service, quality price
Reserve
Not endorsed by any of the companies listed, or the city of Memphis.


Memphis Taxis
Yellow Cab
Safe, Reliable, 24 hour service
Reserve
Ride Charge
Driven by certified professionals
Reserve
Uber
Cheap and convenient
Reserve
Premiere Taxi
Quality service, quality price
Reserve
Not endorsed by any of the companies listed, or the city of Memphis.


Memphis Taxis
Yellow Cab
Safe, Reliable, 24 hour service
Reserve
Ride Charge
Driven by certified professionals
Reserve
Uber
Cheap and convenient
Reserve
Premiere Taxi
Quality service, quality price
Reserve
Not endorsed by any of the companies listed, or the city of Memphis.
Waiting for www.ridecharge.com...

As we expected, the visitor to this site started off by looking through all the options in a circular motion before spending some time looking at the details given for Ride Charge and clicking on the button to reserve it.

Version B: Video Link



Memphis Taxis

This page contains information about taxi and cab services in Mepmphis, Tennesse

Reserve with Yellow Cab

Safe, reliable taxi service from Yellow Cab and Checker Cab are available within the Memphis Metro Area.

Reserve with Memphis Taxi

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

Reverse with Uber

We are so much cheaper than other taxis...!

Reserve with Premier Taxi

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

The user started by looking through all the options from the right to left. Then, they spent some time comparing the details given for Memphis Taxi and Uber before finally clicking the button to reserve with Uber.

COMPARISON:

While our metrics seemed accurate to our alternative hypotheses, our statistical tests showed everything to be statistically insignificant, so we cannot give the company any certain recommendations based on our A/B testing data. However, our eye tracking data gave us valuable insight on the current designs. If the company's priority is the customer's time, we recommend using Version A, as we found our user made a decision faster on that page. However, if the priority is serving an audience of people new to the area, we recommend Version B, as it provides more information, and its more condensed layout is more accommodating to future additions of companies and information to the page.

The data from the eye tracking was collected from only one user each and hence was not as comprehensive as the data collected from the A/B testing. The eye tracking was more objective since we did not say anything that would've affected how they interacted with the website, and is a good way to see how specific elements on the page affect user interaction. However, users in A/B testing may have only used the website the way they did because of how they were asked to use the site for testing.

While thinking about our own metrics, we also considered the possible misuse of data for unethical design practices in other applications.

Click-Through Rate:

If a high click-through rate is observed for a page, a company could take advantage of this by adding advertisements that users need to see or click to continue to the page they're

Conclusion

Despite the issues with out A/B Testing data, this still proved to be a valuable exercise in data collection and analysis practices for interface testing. In the future, we could perhaps conduct more extensive testing, and givemore consistent, unbiased instructions to users testing the interface.

Sources

Most of the sources we referenced while coding our web pages are sourced within comments in the code. Aside from those, we also referred to https://www.w3schools.com/css/css_grid.asp for help with setting up our CSS Grid on each page.