A/B Testing & EyeTracking

Introduction

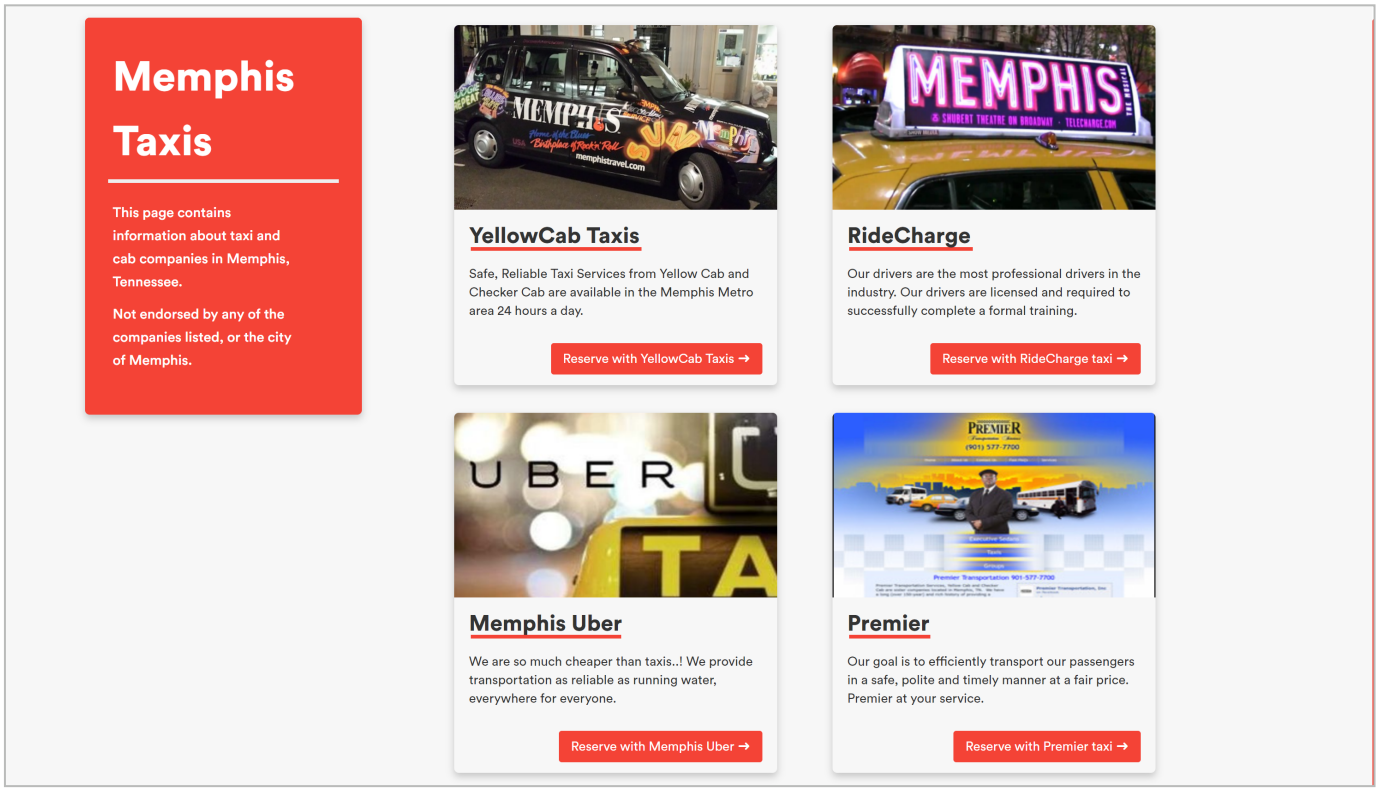
When validating your designs, it's important to observe real users and analyze their behavior to know if the changes you made have a real effect. In this assignment, we created two versions of a website listing several Memphis taxi services, collected and analyzed the data of over 40 users who visited the website, and used eye-tracking with two users to generate a heatmap and an animated replay of users' eye movement.

Part 1: A/B Testing

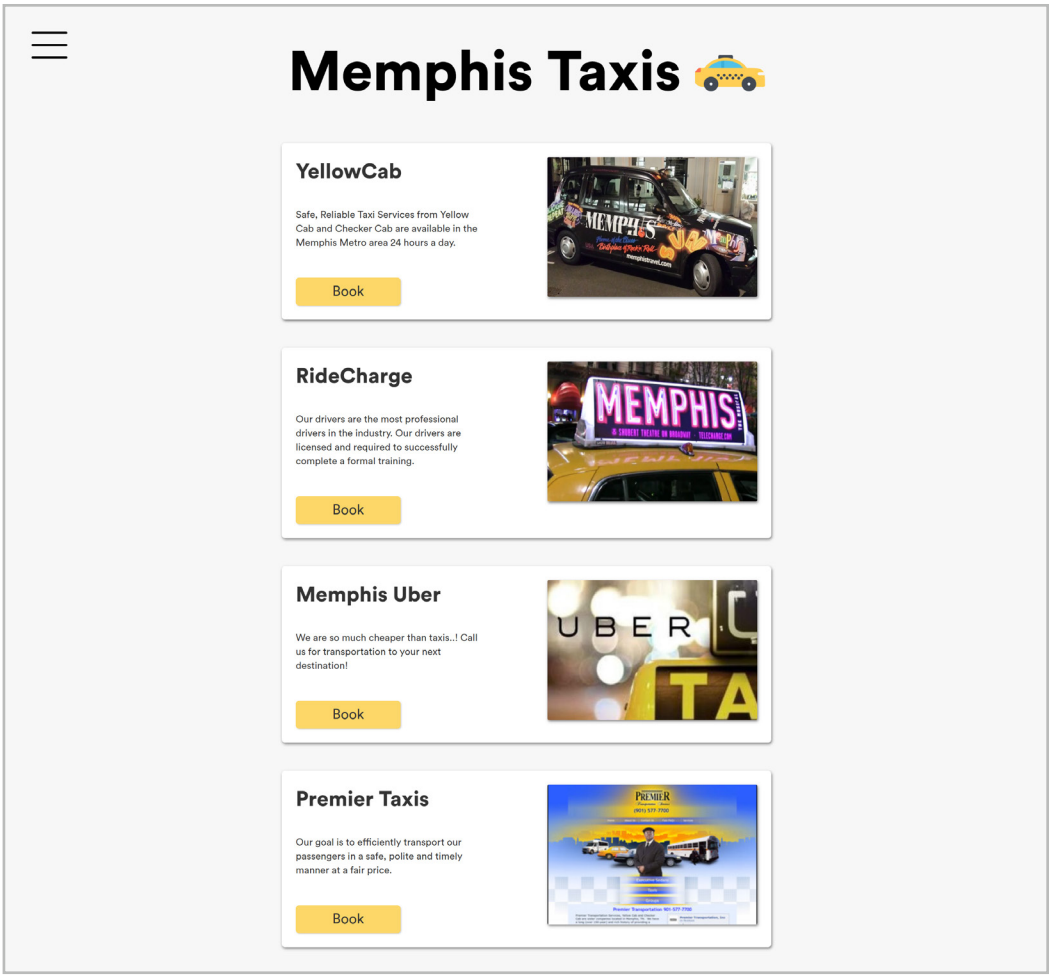
I. Design

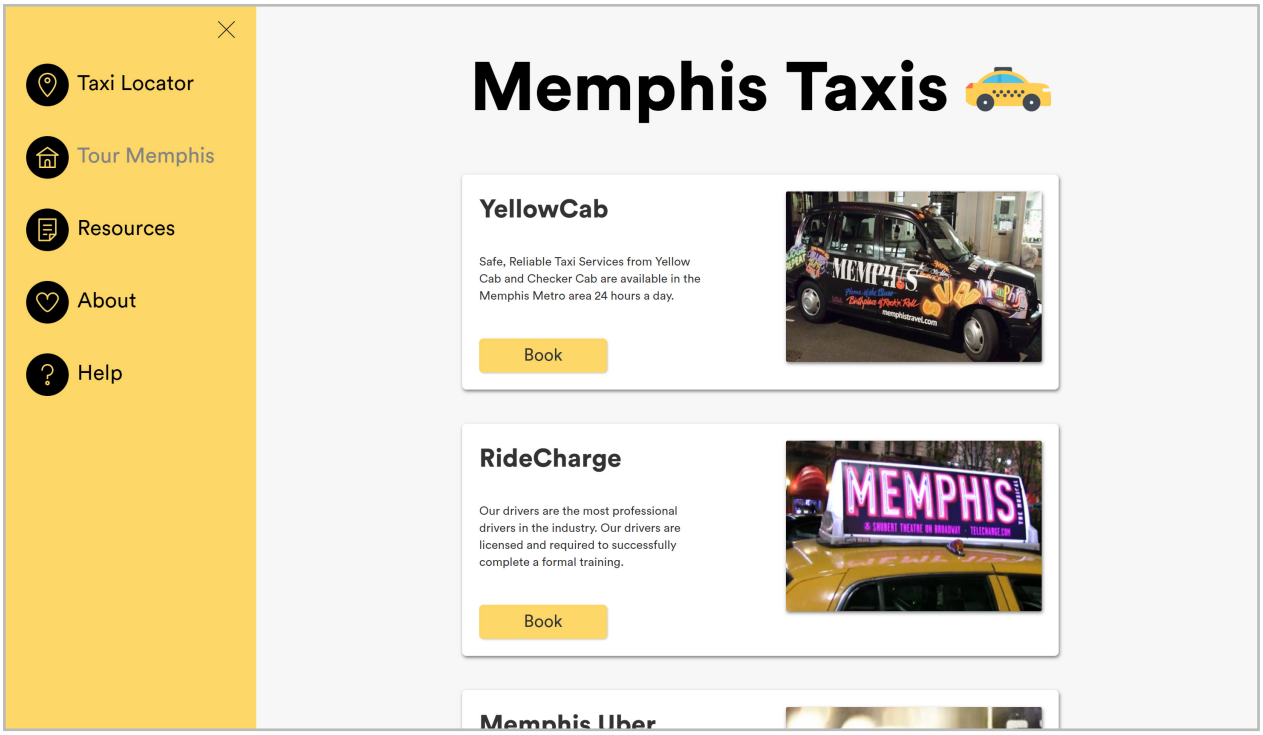
You can find the versions hosted at https://shrouded-beyond-84576.herokuapp.com/.

Version A represented the options in a grid-format, where users could see all four options at one time (depicted below):



Version B represented the options in a scrolling format, where the user must scroll to see all options. Additionally, version B included an expandable side menu (depicted below):





II. Data Analyses

Hypotheses:

Click through rate:

- Null hypothesis: Versions A and B will receive the same amount of clicks over a session
- Alternative hypothesis: The click through rate for Version A will be greater than that of Version B, because the information in Version A is presented in a grid which is easier to read.

Dwell time:

- Null hypothesis: Versions A and B will have the same average time for each unique session that leaves the page and returns.
- Alternative hypothesis: The dwell time for Version A will be less than that of Version B, because users will be able to more quickly complete the desired task with the more consolidated layout.

Return rate:

- Null hypothesis: Versions A and B will have the same proportion of unique sessions that left the landing page and returned.
- Alternative hypothesis: The return rate for Version A will be less than that of Version B, because users will not need to backtrack due to the more intuitive grid design.

Time to click:

- Null hypothesis: The average time it took a session to do the first click with be the same across versions A and B.
- Alternative hypothesis: The time to click for Version A will be less than that of Version B, because more buttons are immediately visible on the opening screen due to the grid layout.

Metric Calculations

Calculations for Version A:

Click through rate:

# of unique users: 38
# of unique clicks (# of users who made a click): 20
Click rate = 20/38 = 52.6%

Dwell time:

Average of (2nd page load time - click time) for a user's click (if they returned)
= 21311 ms

Return rate:

# of unique sessions that return after leaving: 12
# of sessions that leave: 20
12/20 = 60%

Time to click:

Average of (click time - page load time) for each user's first click (if they made one)
= 15071 ms

Calculations for Version B:

Click through rate:

# of unique users: 32
# of unique clicks (# of users who made a click): 17
Click rate = 17/32 = 53.1%

Dwell time:

Average of (2nd page load time - click time) for a user's click (if they returned)
= 70763 ms

Return rate:

# of unique sessions that return after leaving: 8
# of sessions that leave: 32
8/17 = 47%

Time to click:

Average of (click time - page load time) for each user's first click (if they made one)
= 11347 ms

In summary, the calculations are as follows:

	Version A	**Version B**
Click through rate	52.6%	53.1%
Dwell time	21311 ms	62397 ms
Return rate	60%	47%
Time to click	15071 ms	11347 ms

Statistical Tests

1. Click through rate

To test if the click through rate of Version A was different from Version B, we will use a chi-squared test because click through rate is categorical data (either the user clicks through or doesn't).

Observed	Click	No click	Total
Version A	20	18	38
Version B	17	15	32
Total	37	33	70

Expected Click Through Rate: 37/70 * 100 = 52.9%

Expected	Click	No click	Total
Version A	20.1	17.9	38
Version B	16.9	15.1	32
Total	37	33	70

$$\chi^2 = \frac{(20-20.1)^2}{20.1} + \frac{(18-17.9)^2}{17.9} + \frac{(17-16.9)^2}{16.9} + \frac{(15-15.1)^2}{15.1} = .0023$$

In a χ^2 table, the χ^2 value for 1 degree of freedom at a significance level of .05 is 3.84.

Since .0023 < 3.84, we fail to reject the null hypothesis and conclude that the click through rates of Versions A and B do not differ.

2. Dwell time

To test if the dwell time of Version A was different from the dwell time of Version B, we will use a t-test because the data is quantitative (we measure dwell time in ms).

Dwell Time A	(Dwell Time A)^2	(Dwell Time A - Avg Dwell Time A)^2	Dwell Time B	(Dwell Time B)^2	(Dwell Time B - Avg Dwell Time B)^2
2398	5750404	357701569	118428	14025191184	3139472961
3553	12623809	315346564	3869	14969161	3425526784
4227	17867529	291863056	13974	195272676	2344786929
3629	13169641	312653124	3836	14714896	3429390721
12190	148596100	83192641	347195	120544368025	81109900804
8545	73017025	162970756	3307	10936249	3491628100
2601	6765201	350064100	4125	17015625	3395625984
148680	22105742400	16222862161	4443	19740249	3358666116
5976	35712576	235162225			
Sum = 191799		Sum = 18331816196	Sum = 62397		Sum = 103694998399
N = 9		s^2 = 2291477024.5	N = 8		s^2 = 14813571199.9

$$t = \frac{21311 - 62397}{\sqrt{\left(\frac{(9-1)2291477024.5 + (8-1)14813571199.9}{9+8-2}\right)\left(\frac{1}{9}+\frac{1}{8}\right)}} = -.94$$

In a t-table, with 15 degrees of freedom and a significance level of .05, the critical value is 2.13.

Since abs(-.94) < 2.13, we fail to reject the null and conclude that the dwell time does not differ between Version A and Version B.

3. Return rate

To test if the return rate of Version A was different from Version B, we will use a chi-squared test because return rate is categorical data (either the user clicks through or doesn't).

Observed	**Return**	**No return**	**Total**
Version A	12	8	20
Version B	8	9	17
Total	20	17	37

Expected Return Rate: 20/37 * 100 = 54.1%

Expected	**Return**	**No return**	**Total**
Version A	10.8	9.2	20
Version B	9.2	7.8	17
Total	20	17	37

$$\chi^2 = \frac{(12-10.8)^2}{10.8} + \frac{(8-9.2)^2}{9.2} + \frac{(8-9.2)^2}{9.2} + \frac{(9-7.8)^2}{7.8} = .63$$

In a χ^2 table, the χ^2 value for 1 degree of freedom at a significance level of .05 is 3.84.

Since .63 < 3.84, we fail to reject the null hypothesis and conclude that the return rates of Versions A and B do not differ.

4. Time to click

To test if the time to click of Version A was different from the time to click of Version B, we will use a t-test because the data is quantitative (we measure time to click in ms).

Time to Click A	(Time to Click A)²	(Time to Click A - Avg Time to Click A)²	Time to Click B	(Time to Click B)²	(Time to Click B - Avg Time to Click B)²
4414	19483396	113571649	17546	307862116	38427601
15667	245454889	355216	3295	10857025	64834704
3989	15912121	122810724	7912	62599744	11799225
1153	1329409	193710724	10027	100540729	1742400
13480	181710400	2531281	9790	95844100	2424249
28866	833245956	190302025	8614	74200996	7469289
12445	154878025	6895876	10155	103124025	1420864
10386	107868996	21949225	24083	579990889	162205696
16676	278088976	2576025	5416	29333056	35176761
68571	4701982041	2862250000	6489	42107121	23600164
12411	154032921	7075600	6814	46430596	20548089
20034	401361156	24631369	2173	4721929	84162276
11375	129390625	13660416	12222	149377284	765625
29109	847333881	197065444	9089	82609921	5098564
3113	9690769	142993764	14015	196420225	7118224
1522	2316484	183575401	39055	1525293025	767733264
18733	350925289	13410244	6207	38526849	26419600
7934	62948356	50936769			
10561	111534721	20340100			
10989	120758121	16662724			
Sum = 301428		Sum = 4187304576	Sum = 192902		Sum = 1260946595
N = 20		s^2 = 220384451.4	N = 17		s^2 = 78809162.2

$$t = \frac{15071 - 11347}{\sqrt{\left(\frac{(20-1)220384451.4 + (17-1)78809162.2}{20+17-2}\right)\left(\frac{1}{20}+\frac{1}{17}\right)}} = .9048$$

In a t-table, with 35 degrees of freedom and a significance level of .05, the critical value is 2.03.

Since .9048 < 2.03, we reject the null and conclude that the time to click does not differ between Version A and Version B.

5. Confidence Interval

Average Time to Click A	Average Time to Click B	Average Time to Click A − Average Time to Click B = $\bar{x}$
Sum = 301428	Sum = 192902	
N = 20	N = 17	
Average = Sum/N = 15071	Average = Sum/N = 11347	3724

Given:
N = 37
$\bar{x}$ = 3724
s^2_B = 78809162.2
s^2_A = 220384451.4

Standard error = $\sqrt{(\frac{(20-1)220384451.4 + (17-1)78809162.2}{20+17-2})(\frac{1}{20} + \frac{1}{17})}$ = 4115.8

$t_{df = 35,\ \alpha = .05}$ = 2.03 (from t-table)

95% confidence interval = $\bar{x} \pm t_{df = 35,\ \alpha = .05}$ * (standard error)

= 3724 ± 2.03(4115.8) = (-4631.1, 12079.1)

Because the confidence interval includes zero, we can say with 95% confidence that the time to click for Version A and Version B does not differ.

Part 2: Eye Tracking

Next, we analyzed how participants interacted with the interfaces using an eye tracker. We had one participant view Version A and the other view Version B. Our hypotheses about what the eye tracking data would look like as well as the actual results are discussed below.

Hypothesis

We hypothesized that Version A will have a greater proportion of eye gazes at the sides of the screen than Version B. This is because the content is organized in a grid, which spans more of the screen horizontally. Version B will likely have a greater proportion of eye gazes going down the center of the screen because the information is organized in a scrolling format.

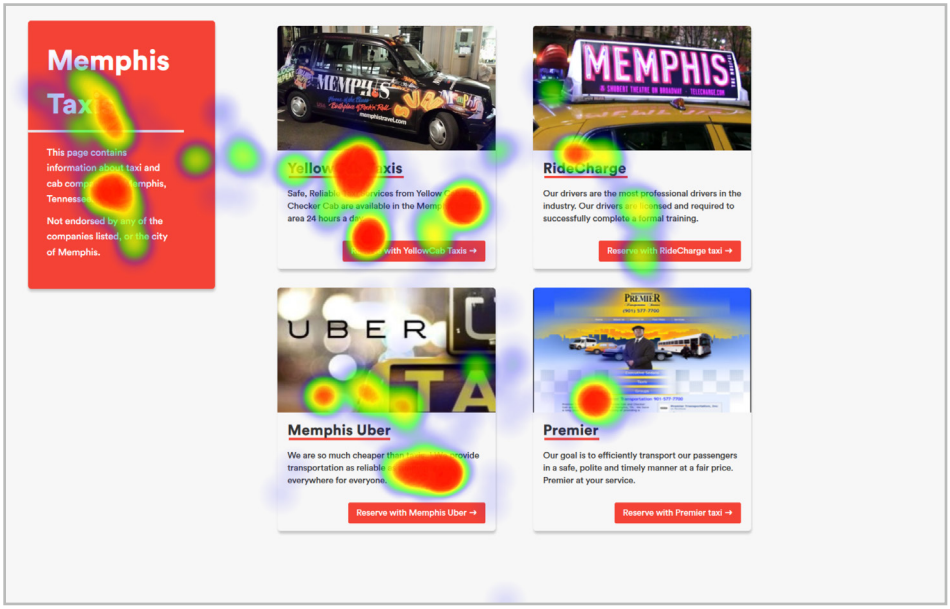


Fig 1:
Heatmap for Version A

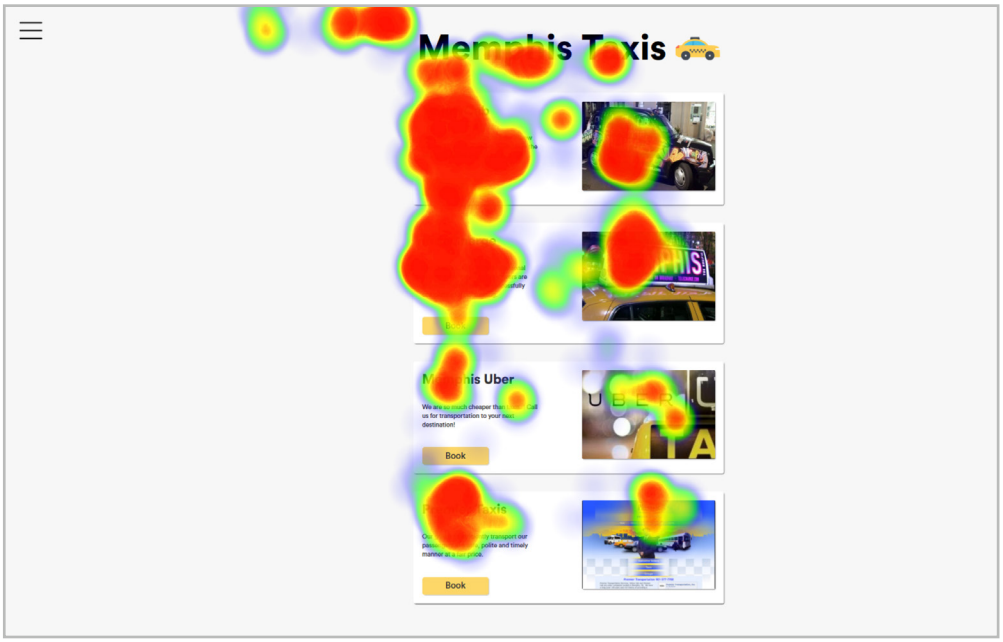


Fig 2:
Heatmap for Version B

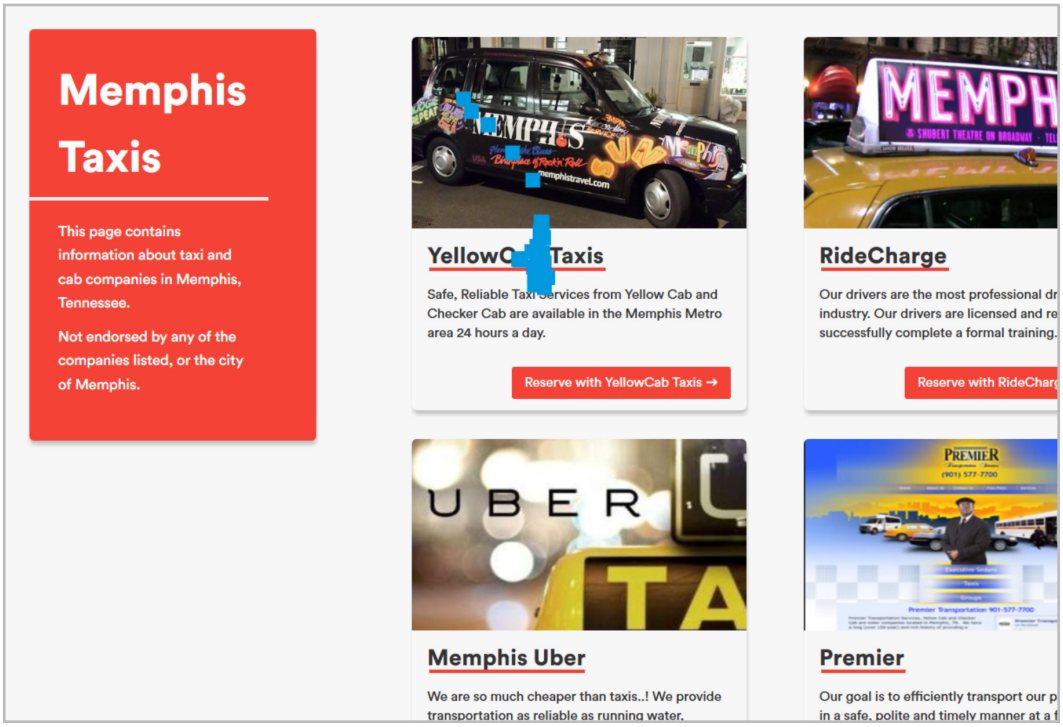


Fig 3:
In-progress shot for Version A

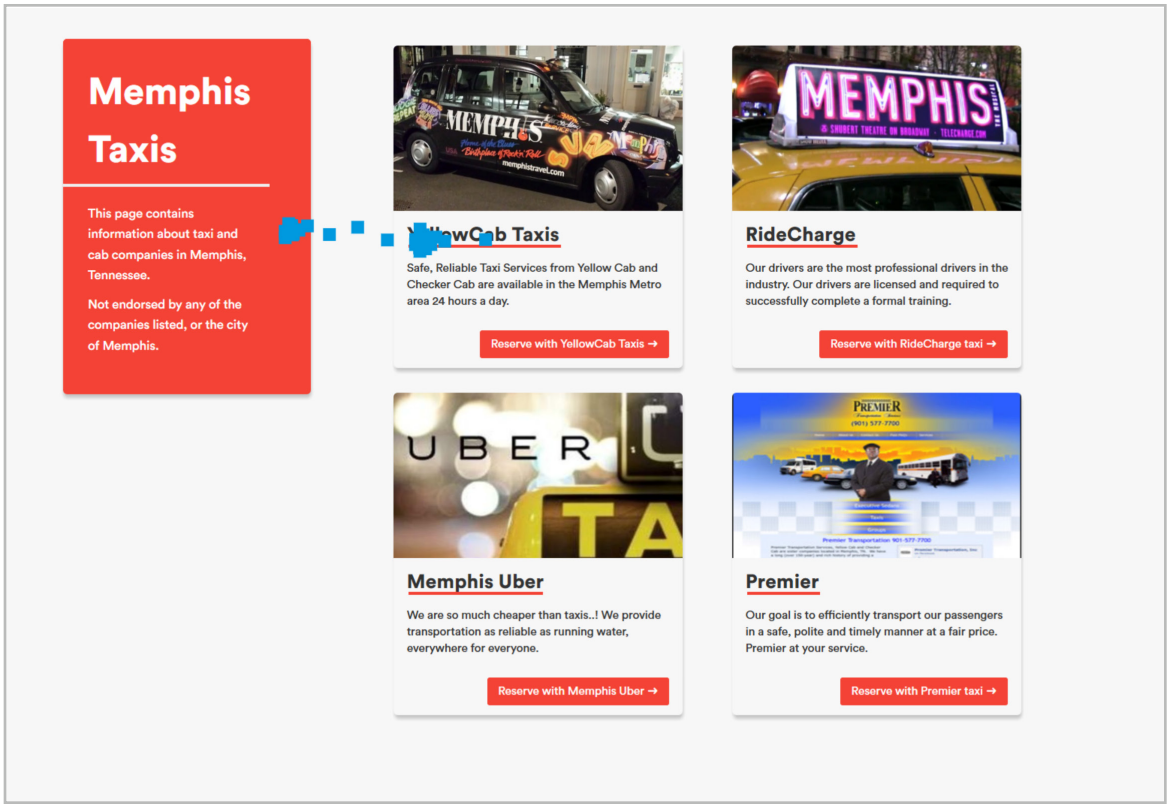


Fig 4:
End shot for Version A

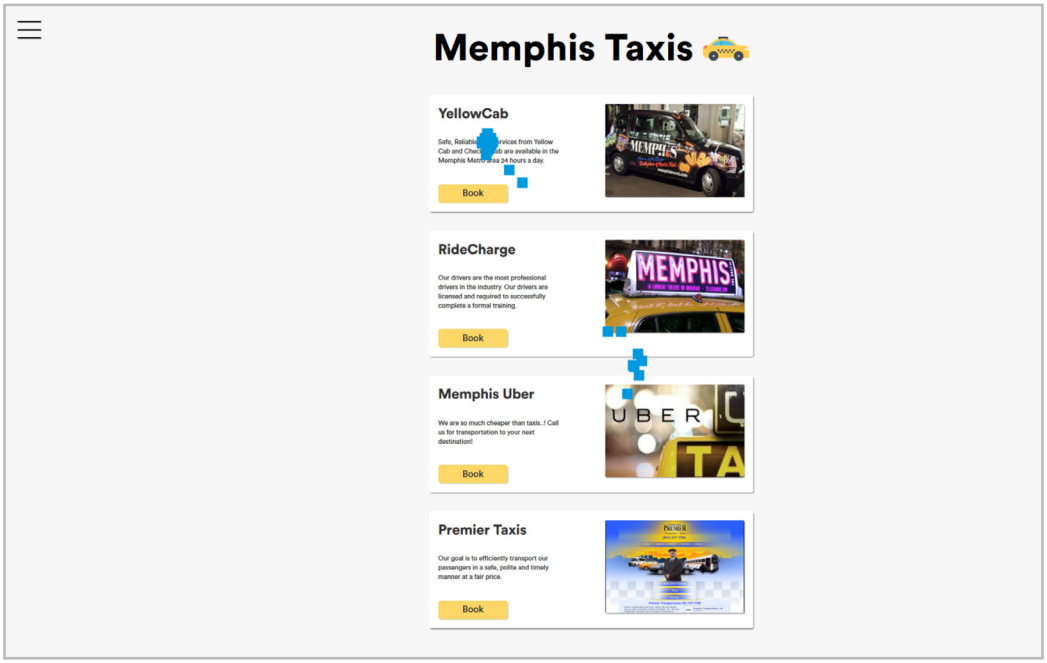


Fig 5:
In-progress shot for Version B

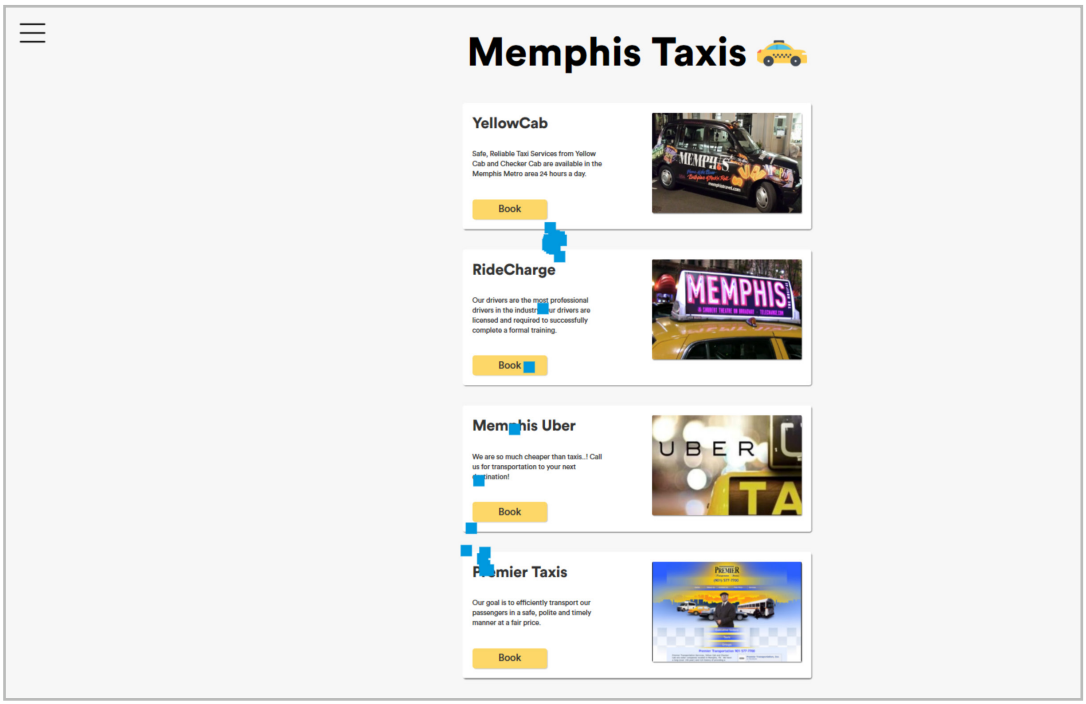


Fig 6:
End shot for Version A

Analysis

The eye tracking data from both versions of the interface support our hypothesis about the range of eye gazes we would see from Version A versus Version B. It is clear that the areas the user interacted with the most in Version A consume a greater horizontal scope of the page. On the contrary, the user who viewed Version B had eye gazes within a much narrower range of the page.

Part 3: Comparison

Question 1

Even though none of the metrics were different enough between Versions A and B to be statistically significant, I would recommend to a committee of stakeholders that using Version A would lead to more profitable results. Based on the eye tracking results, it is clear that in Version A, the user examines information in all quadrants, as well as the text area explaining the purpose of the page. This does not occur with Version B - rather, the users seem to look mainly at the navigation menu and the two taxi options in the first column.

Question 2

Our A/B Testing data differs from our eye tracking data in that our data from A/B testing is not statistically significant so it doesn't tell us much about which interface is better, while the eye tracking data shows us clearer differences between the way users behave on each website. Specifically, from the eyetracking, we can tell that Version A allows a user to assess more information on the page in one scan.

The advantages of eye tracking over A/B testing is that the results are more visual and allow you to localize exactly where on the interface a user is focusing, while the advantages of A/B testing over eye tracking is that you are able to view the exact clicks the user performs on each version of the interface (significant data).

Question 3

Two metrics that could used be unethically in modern websites are **shopping cart abandonment** (a measure of users who add products to their shopping cart but do not go through with the purchase) and **bounce rate** (a measure of users who go onto the landing page of your website, do nothing, and leave).

To minimize **shopping cart abandonment**, websites can advertise false sales and display a countdown clock to trick users into thinking that a purchase has to be made urgently– for example, clothing websites like boohoo.com advertise "limited time" sales all year long.

To minimize **bounce rate,** a developer could code the website in such a way that pressing the back button returns to a landing page that automatically redirects to the current page (this a tactic commonly used among modern websites). This makes users unable to actually use their back button efficiently, and takes away their freedom to navigate away from the page.