A/B Testing and Eye Tracking

Heroku Site: fathomless-ocean-62704.herokuapp.com

Part I: A/B Testing

Hypotheses

Click through rate
Null: The click through rate of A and B will be the same.
Alternative: The click through rate of B will be higher than A because of the dark background in B makes the buttons stand out more, thus attract the user to it.

Average time to click
Null: The average time to click will be the same.
Alternative: The average click time will be lower on A because of the two column arrangement, which will let the user scan the information faster.

Dwell time
Null: The dwell time will be the same.
Alternative: The average dwell time will be longer on B because the user will have likely read through all the options, thus found the one they want given the four columns.

Return rate
Null: The return rate with be the same.
Alternative: The return rate will be higher on A because the images are larger, which will be memorable for the user.

Data Analysis

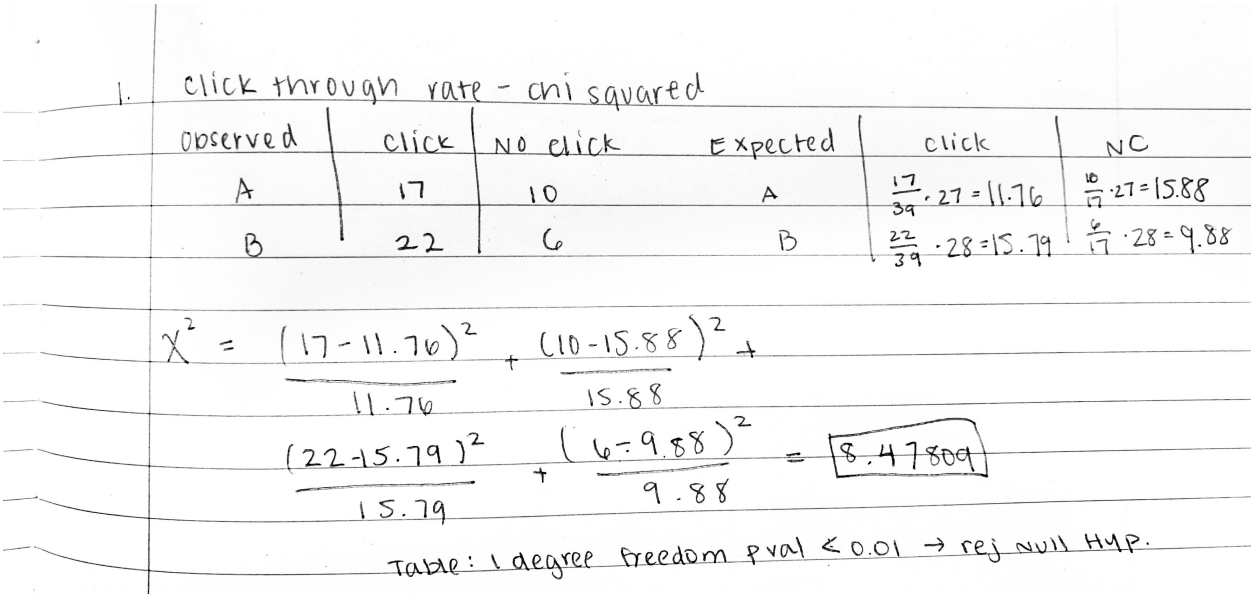
Click through rate

Computed Metrics:
A: 17/27 = .629
B: 22/28 = .786

Statistical Test:
We used chi-squared because it is a categorical test that determines if there is a relation between changes. This is relevant because we wanted to see if changing a feature, such as the background and button color, would change the number of clicks we would receive. This test was accomplished by summing (OF-EF)2 EF for each interface A and

B, where O is observed, E is expected, and F is the attribute/feature we are examining (in this case, click and no click).

Chi-squared = 8.47809
P value: 0.01 → We reject the null hypothesis and accept our alternative hypothesis since our p value is statistically significant.



Chi-squared test

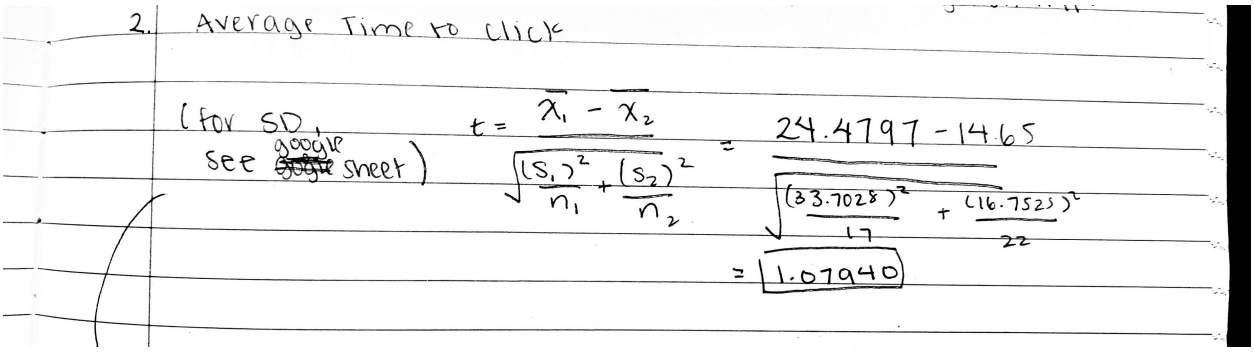
Average time to click

Computed Metrics:
A: 22.4797 sec
B: 14.65 sec

Statistical Test:
We used the T-test because the T-test is used to determine if there is a difference between two means. This is relevant since we are examining the average time before a user makes its first click between our two interfaces.

t = 0.9507
Degrees of freedom: 37
P value: 0.3479 → Our p value is not statistically significant, thus we accept the null hypothesis.

We computed the t-test through the equation shown below. We utilized Excel to compute the standard deviations and plugged them into the equation.



T-test equation

A mean	22.47976471			
Values	Distance from mean	Distance from mean squared		Sum of squared distance
3.666	-18.81376471	353.9577424		18174.06021
15.497	-6.982764706	48.75900294		
16.784	-5.695764706	32.44173558		Sum over n
7.118	-15.36176471	235.9838149		1135.878763
96.453	73.97323529	5472.03954		
8.401	-14.07876471	198.2116156		Sqrt of sum (final stdev)
4.267	-18.21276471	331.7047982		**33.70280052**
67.803	45.32323529	2054.195658		
10.46	-12.01976471	144.4747436		
2.953	-19.52676471	381.2945399		
4.34	-18.13976471	329.0510636		
3.022	-19.45776471	378.6046073		
9.738	-12.74176471	162.3525678		
5.699	-16.78076471	281.5940641		
16.379	-6.100764706	37.21933		
107.913	85.43323529	7298.837693		
1.663	-20.81676471	433.3376928		

B mean	14.65381818			
Values	Distance from mean	Distance from mean squared		Sum of squared distance
17.746	3.092181818	9.561588397		5893.611261
58.619	43.96518182	1932.937212		
5.146	-9.507818182	90.39860658		Sum over n
6.088	-8.565818182	73.37324112		280.6481553
6.23	-8.423818182	70.96071276		
1.889	-12.76481818	162.9405832		Sqrt of sum (final stdev)
11.765	-2.888818182	8.345270488		**16.75255668**
9.48	-5.173818182	26.76839458		
11.588	-3.065818182	9.399241124		
7.754	-6.899818182	47.60749094		
1.139	-13.51481818	182.6503105		
2.517	-12.13681818	147.3023556		
5.482	-9.171818182	84.12224876		
13.44	-1.213818182	1.473354579		
2.153	-12.50081818	156.2704552		
27.358	12.70418182	161.3962357		
57.021	42.36718182	1794.978095		
8.658	-5.995818182	35.94983567		
24.819	10.16518182	103.3309214		
3.092	-11.56181818	133.6756397		
2.899	-11.75481818	138.1757505		
37.501	22.84718182	521.993717		

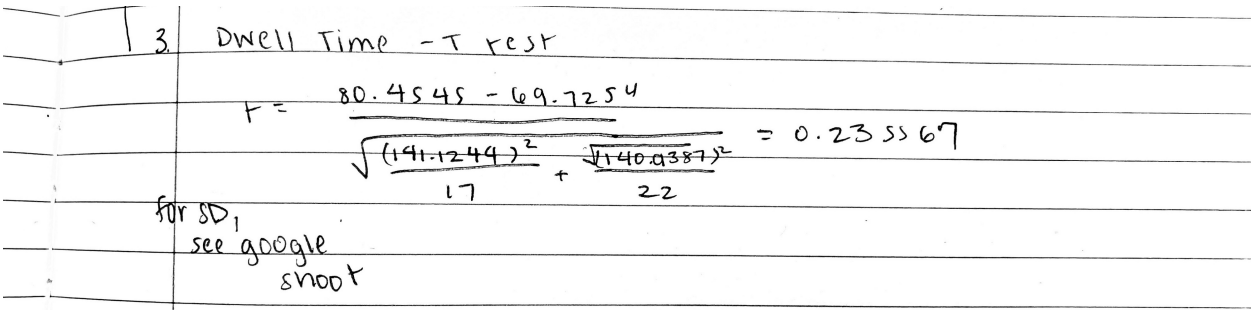
Calculating standard deviations

Dwell time

Computed Metrics:
A: 80.4545
B:69.7254

Statistical Test:
Given that we are comparing the average dwell times between the two interfaces, the t-test is the appropriate statistical test.

t = 0.1812
Degrees of freedom = 21
P value: 0.8580 → Our p value is not statistically significant. As a result, we accept the null hypothesis.



T-test equation

A mean	80.45454545			
Value	Dist to average	Dist squared		Sum of dists
428.517	348.0624545	121147.4723		199161.0989
270.04	189.5854545	35942.64458		
106.199	25.74445455	662.7769398		Sum over n
48.733	-31.72154545	1006.256446		19916.10989
7.004	-73.45054545	5394.982628		
6.557	-73.89754545	5460.847224		Sqrt
5.405	-75.04954545	5632.434273		**141.1244483**
4.019	-76.43554545	5842.392609		
3.073	-77.38154545	5987.903577		
2.939	-77.51554545	6008.659787		
2.514	-77.94054545	6074.728626		

B mean	69.78808333			
Value	Dist to average	Dist squared		Sum of dists
403.838	334.0499167	111589.3468		218500.807
334.587	264.7989167	70118.46627		
35.373	-34.41508333	1184.397961		Sum over n
19.518	-50.27008333	2527.081278		19863.70973
9.333	-60.45508333	3654.817101		
7.848	-61.94008333	3836.573923		Sqrt
7.015	-62.77308333	3940.459991		**140.9386737**
4.958	-64.83008333	4202.939705		
4.54	-65.24808333	4257.312379		
3.661	-66.12708333	4372.79115		
3.449	-66.33908333	4400.873978		
3.337	-66.45108333	4415.746476		

Calculating standard deviations

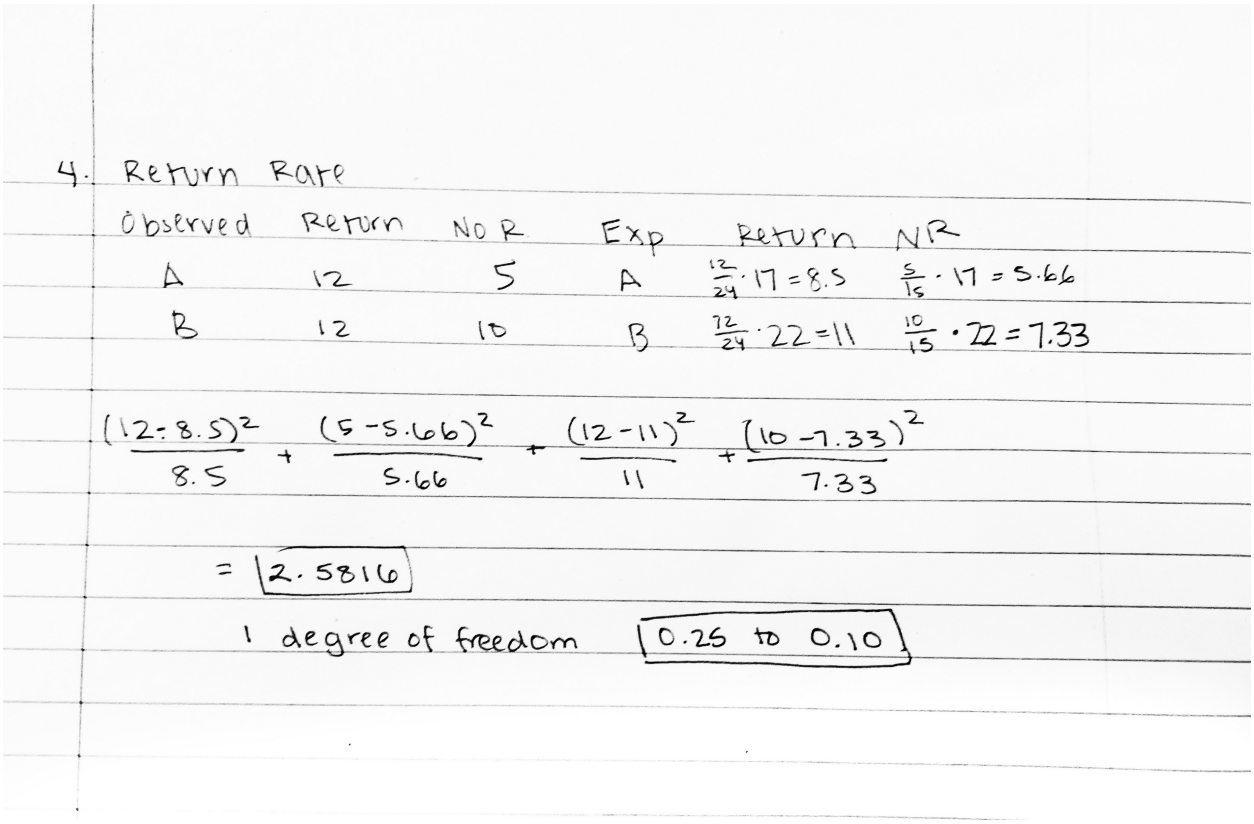
Return rate

Computed Metrics:
A: 12/17
B:12/22

Statistical Test:
Since we are using the return rate to compare the proportion of users that return to the site after a click, we are testing to see if there are any features in how our website is laid out (2 vs 4 columns) that may result in a higher return rate. Thus, the chi-squared test is appropriate.

We compute the chi squared test the same way that we computed our click through rate (see that explanation for more detail).

Chi-squared = 2.5816
P value: 0.10–0.25 → As a result, we accept our null hypothesis.



Chi-squared test

95% Confidence Interval Calculation

With a 95% confidence interval, we conclude that the difference between the average time to click interface A and B is from -8.31 to 23.97. Note that our lower bound is negative, which is technically not a real time, thus it can be stated that the difference between the click time average of A and B is from 0 to 23.97. The negative result can be attributed to our large standard deviation. As a result, to have a high confidence interval of 95%, our computed difference in the average time is also large.

$$\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N+2-2}(\frac{1}{N_1}+\frac{1}{N_2})}$$

$$=\sqrt{(\frac{1}{37}(16\cdot 33.7028^2+21\cdot 16.75255^2)(\frac{1}{17}+\frac{1}{22})}$$

$$=8.23593$$

Calculating standard error

$$\bar{X_1}-\bar{X_2}\pm z\cdot se$$

$$=22.4797-14.65\pm 1.96\cdot 8.23593$$

$$=23.97,-8.31$$

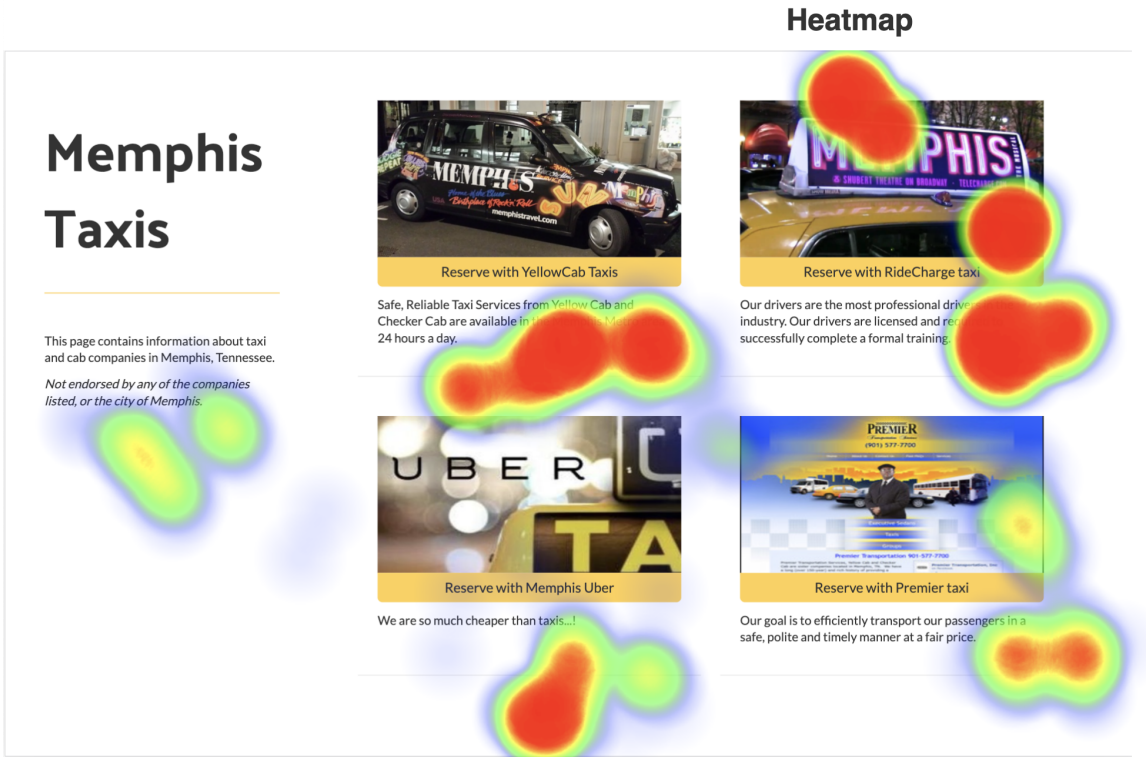
Confidence error

Part II: Eye Tracking

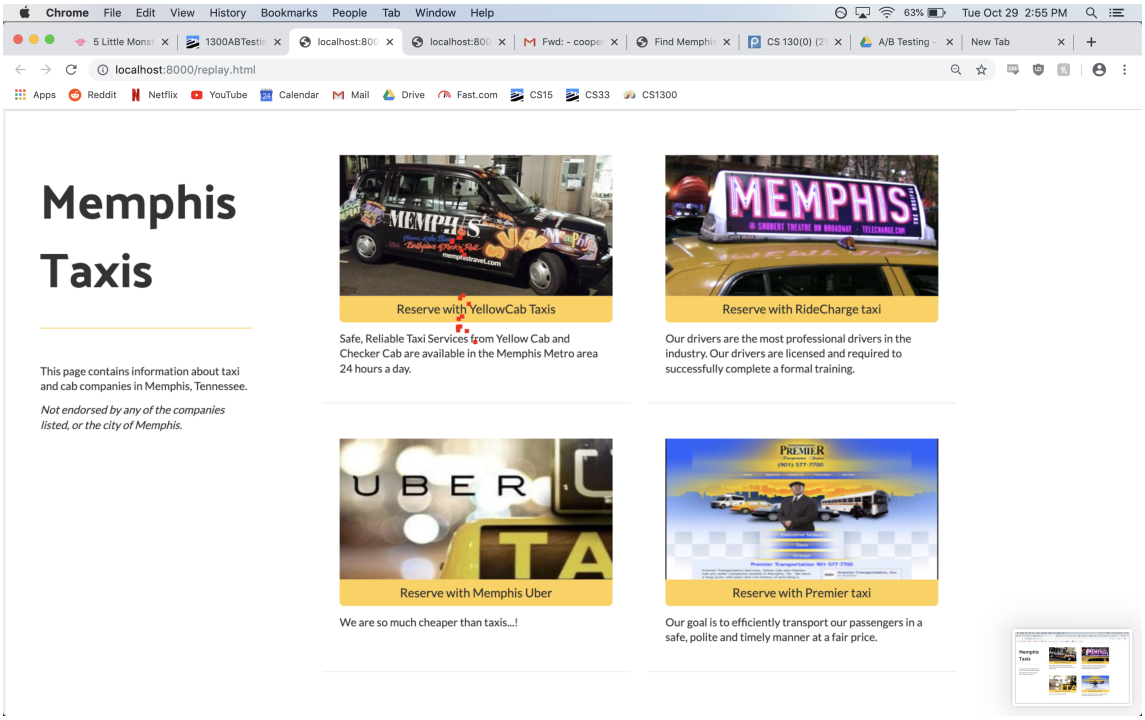
Hypothesis: Version A will have more evenly spread out eye gazes because the four images are better balanced in a grid, whereas Version B is stacked in a row, making the center more heavily weighted.

We think this makes sense because the horizontal layout of Version B's images puts the center two images more prominently on display, and the first and last images become visually marginal. Version A, in the meantime, has images that are more equal and viewers aren't given a subconscious bias.

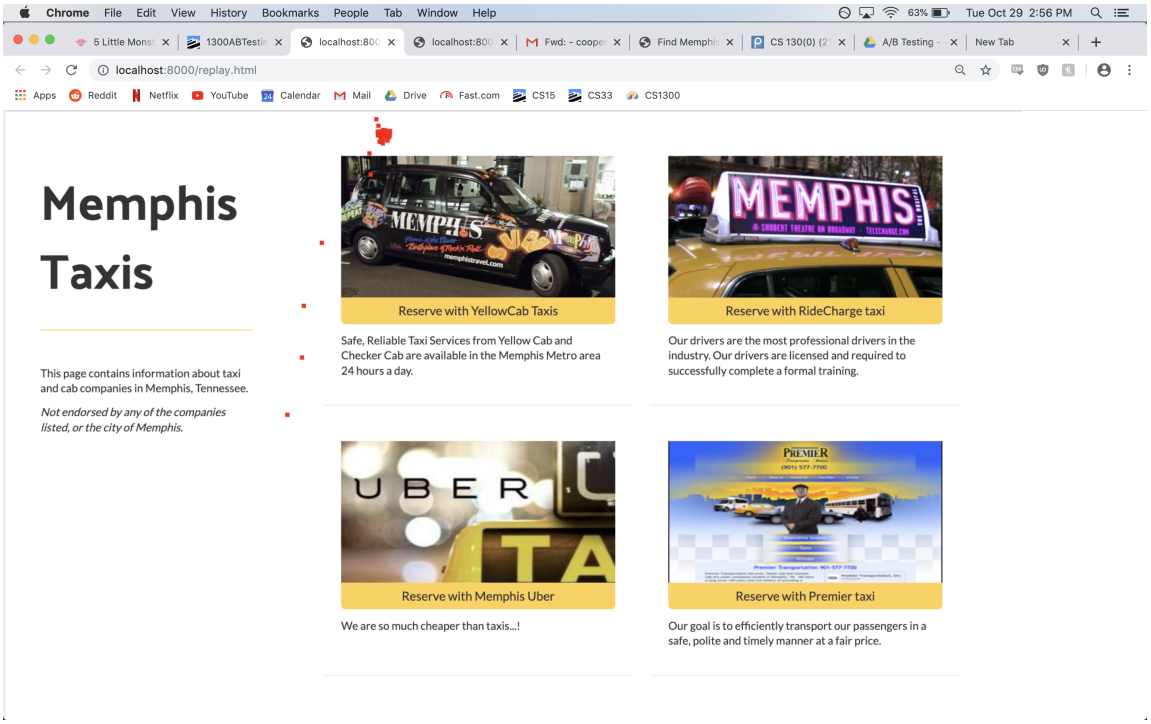
Session 1



Version A Heat map



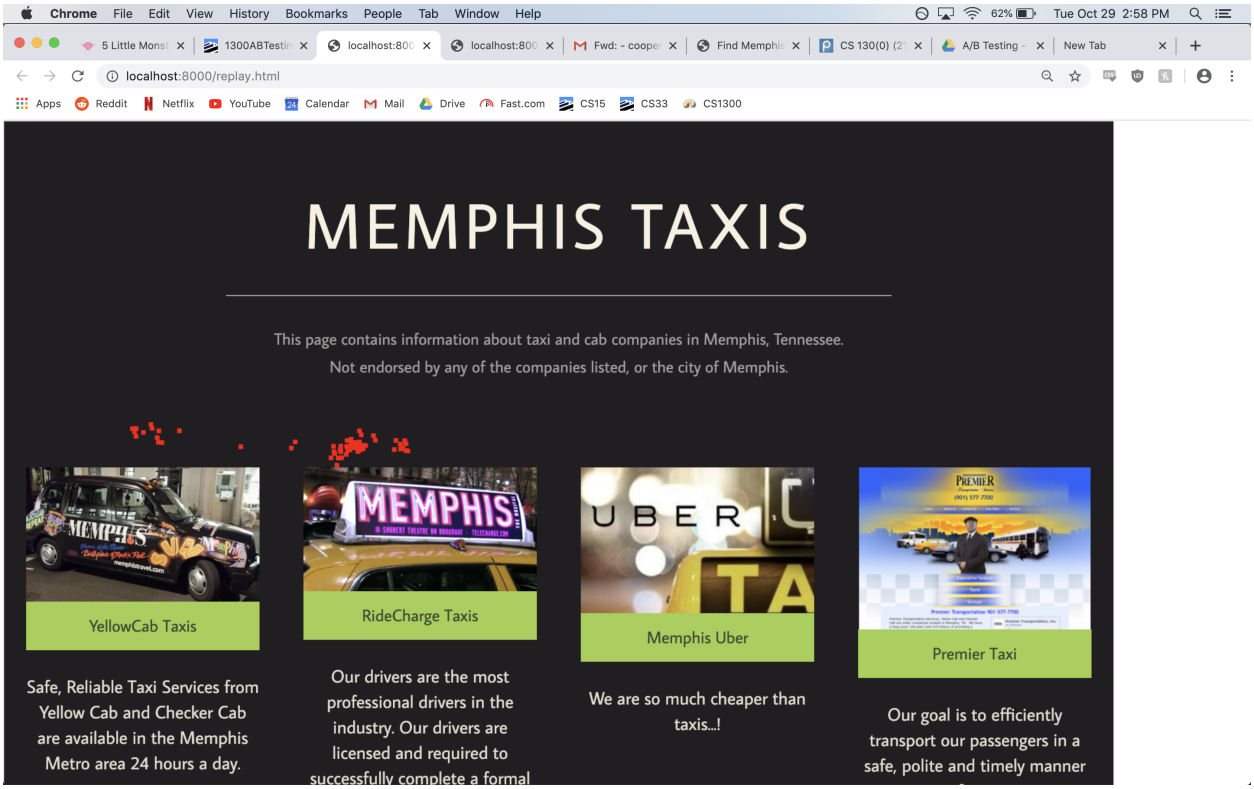
Version A Eye tracking — in progress



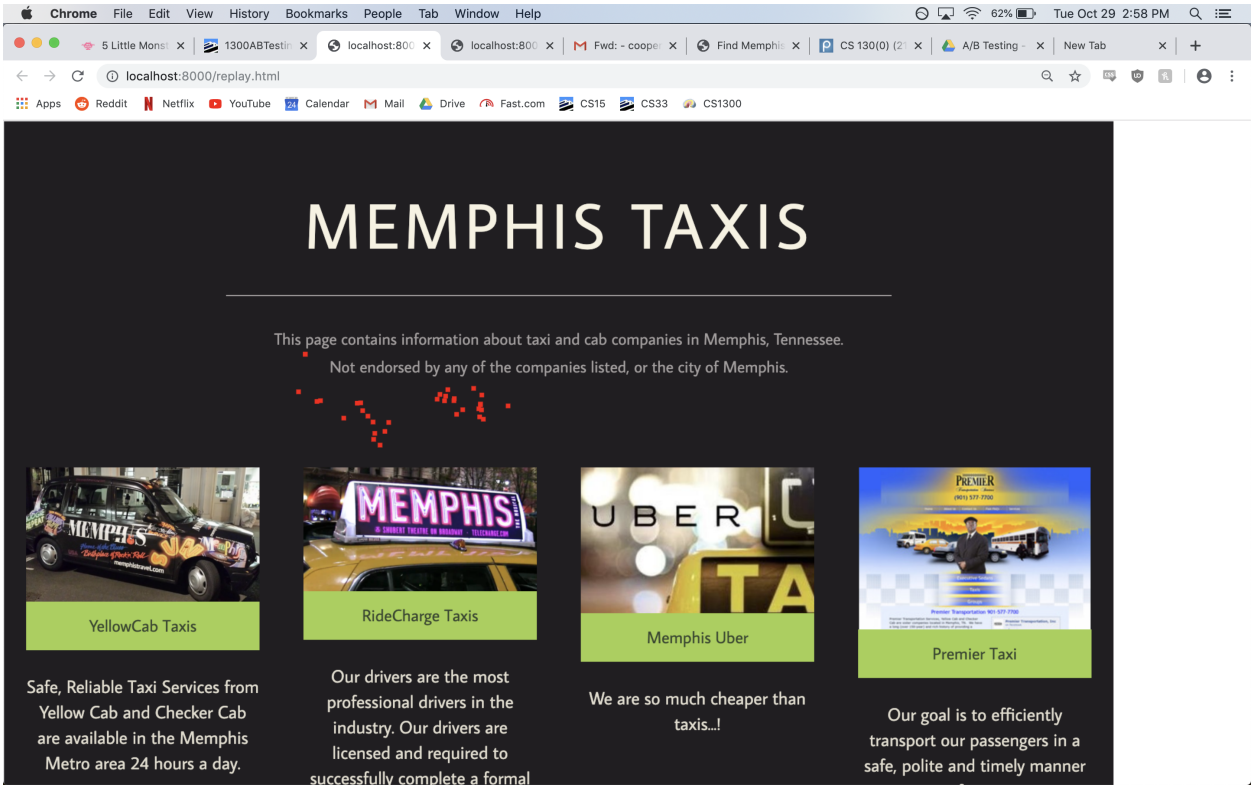
Version A Eye tracking — final

Session 2

Version B Heat map



Version B Eye tracking — in progress



Version A Eye tracking —final

Results
Our hypothesis was correct in that with Version A, we saw a wide, evenly-spaced distribution in our heat map. Version B had the hotspots of the map mostly around the center (and a little to the left), virtually ignoring the image to the right.

Part III: Comparison

Based on the data analyses from A/B testing, we proved two hypotheses: 1) the click through rate of B is higher than A because the dark background in B makes the buttons stand out more and 2) the return rate is higher on B. However, the average time to click and dwell time was the same for Version A and B because the results of the statistical tests were not significant enough to reject our null hypotheses. Memphis Taxis Co. should use Version B because it has a higher click through rate and return rate. Version B should also be further redesigned to improve its average time to click and dwell time.

Both our A/B testing data and eye tracking data support the idea that the horizontal layout in version B is more readable, and encourages users to click on a button. For AB testing, this led to a higher click through rate and return rate for version B. For eye tracking, this led to a concentration of hotspots in the center of the page.

A/B testing is more advantageous than eye tracking because one can easily analyze larger data sets (larger user testing) and calculate whether the results are statistically significant. This may provide more concrete information than eye tracking, which can vary from user to user and is more difficult to test on a large number of users. On the other hand, eye tracking provides unique information (such as which areas on the site are hotspots) that cannot be provided through A/B testing.

One metric that can be used unethically is click-through rate. For example, if a certain button style or placement leads to more clicks, companies may use this knowledge to advertise products that may damage users' health, such as cigarettes. Another metric that can be used unethically is the dwell time. If companies know a user will stay on a certain page longer, they can exploit this to show a greater number of ads to the user. Additionally, they may repeat features that cause the user to stay longer on a page on other pages, making the website more addictive.