Moie Uesugi
CSCI1300 A/B Testing and Eye Tracking

Heroku Site URL: csci1300-ab-testing.herokuapp.com

Part I: A/B Testing

Summary of Changes
Interface A: Button text is "Reserve" + company name; not in a grid; no titles
Interface B: Button text is "More Information"; in a grid layout; titles

Hypotheses[1]
1. Time to Click
Null Hypothesis: Changing the button text, layout, and title will not change the Time to Click.
Alternative Hypothesis: Changing the button text, layout, and title for B will make its Time to Click shorter.

The changed button title, "More Information," may appear to be a less committal choice than "reserve." The grid layout allows the user to see everything in one glance without scrolling, while titles allow the user to spend less time reading descriptions before clicking. For example, if the user already knows what Uber is, they may not need to read the description.

2. Click Through Rate
Null: Changing the button text, layout, and title will not change the Click Through Rate.
Alternative: Changing the button text, layout, and title will make more people click.

"More Information" is a less intimidating button title than "Reserve," so users will be more likely to click. The grid layout makes everything easier to see at once, so they may be more likely to actually pursue a service by clicking a button. Titles help give an easier identity to each button, so they are more likely to click on one.

3. Dwell Time
Null: Changing the button text, layout, and title will not change the Dwell Time.
Alternative: Interface B will have a lower Dwell Time.

"More Information" encourages finding out more about all the options, instead of clicking the button to make a "reserve" commitment. The grid layout and titles give a better summary of the options, so users will return faster, because they clearly saw there were multiple options to order a taxi.

4. Return Rate
Null: Changing the button text, layout, and title will not change the Return Rate.
Alternative: Changing the button text will increase the Return Rate, since B changes the user's mental model.

[1] All groupwork parts of this project were done with groupmates, Ivy Wong, Montana Fowler, and Harry He. Some written work may be edited for length and clarity.

The users will be persuaded to see “More Information,” exploring the different options instead of making the final reservation right away. The grid layout and addition of titles allows the user to see more options than in the listview (interface A). Because the user is exposed to more options, they are more likely to return to the page.

Analysis
My stats.py script (included in this handin) was used to compute and print all statistics. The following contains the output of each calculation as well as a short, high-level overview of the computation.

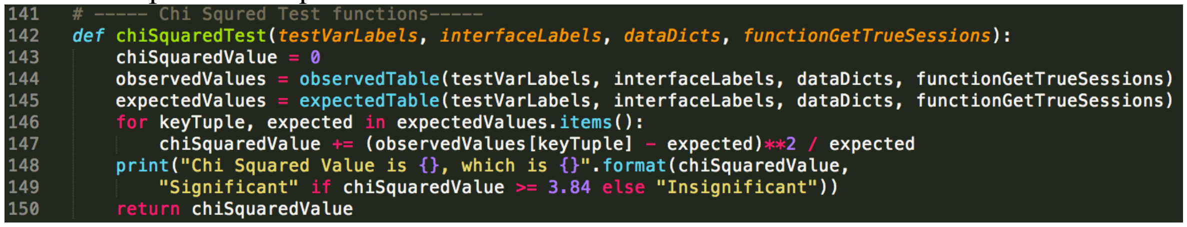
1. <u>Click Through Rate</u>
   Divide the number of sessions that had a click by the number of total sessions in the interface.
   Interface A: 54.83%
   Interface B: 63.64%
2. <u>Average Time to Click</u>
   Obtain a list of first clicks for each unique user, then compute the individual time to click for each (click time minus the page load time for that impression). Average across this list to get the average first time to click across all unique users that had at least one click.
   Interface A: 16505.24 milliseconds
   Interface B: 23992.67 milliseconds
3. <u>Return Rate</u>
   The percentage of unique sessions with clicks that also had a return. A session with a return is defined as any session in which there is at least one page load impression following the first click.
   Interface A: 64.71%
   Interface B: 52.38%
4. <u>Average Dwell Time</u>
   The average length of the first dwell time (ie; the amount of time between a user’s first click and their next page load) for all sessions with returns.
   Interface A: 33572.36
   Interface B: 6134.09

For each of these metrics, a statistical test was performed to determine the significance of the differences between interface A and B.

Click Through Rate

a. The Chi-Squared Test was chosen for this metric, as it is a categorical metric where the two categories are unique sessions that did click and those that did not.

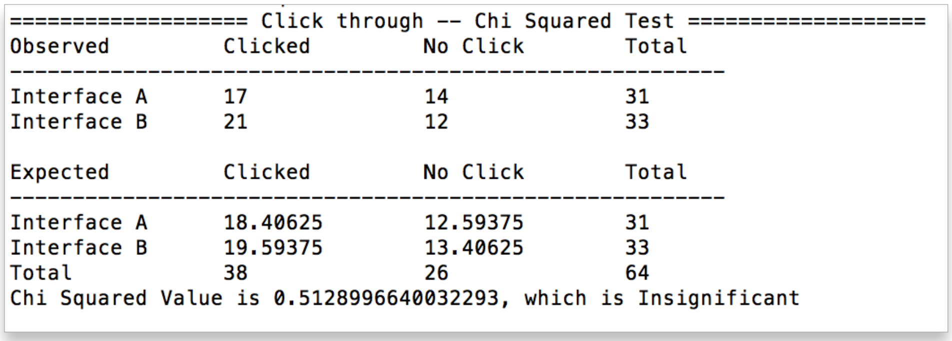
b. The computed chi-squared value was 0.51.



```
# ----- Chi Squred Test functions-----
def chiSquaredTest(testVarLabels, interfaceLabels, dataDicts, functionGetTrueSessions):
    chiSquaredValue = 0
    observedValues = observedTable(testVarLabels, interfaceLabels, dataDicts, functionGetTrueSessions)
    expectedValues = expectedTable(testVarLabels, interfaceLabels, dataDicts, functionGetTrueSessions)
    for keyTuple, expected in expectedValues.items():
        chiSquaredValue += (observedValues[keyTuple] - expected)**2 / expected
    print("Chi Squared Value is {}, which is {}".format(chiSquaredValue,
        "Significant" if chiSquaredValue >= 3.84 else "Insignificant"))
    return chiSquaredValue
```

Figure 1.1. Code snippet for chi squared value calculation.

As visible in Figure 1.1, computation simply followed the formula provided, where the "observedValues" and "expectedValues" are maps from (row, col) tuples to the observed or expected value respectively in that location of the table.
The function passed in as the last argument to chiSquaredTest is a function that takes a dataDict and returns a set representative of all sessions for which a condition is true. For the click through rate, this was *clickSessions*; which returns a set of unique sessions that contained a click.



```
=================== Click through -- Chi Squared Test ===================
Observed          Clicked          No Click          Total
-------------------------------------------------------------
Interface A       17               14                31
Interface B       21               12                33

Expected          Clicked          No Click          Total
-------------------------------------------------------------
Interface A       18.40625         12.59375          31
Interface B       19.59375         13.40625          33
Total             38               26                64
Chi Squared Value is 0.5128996640032293, which is Insignificant
```

Figure 1.2. Observed and Expected Chi Squared Tables for Click Through Rate

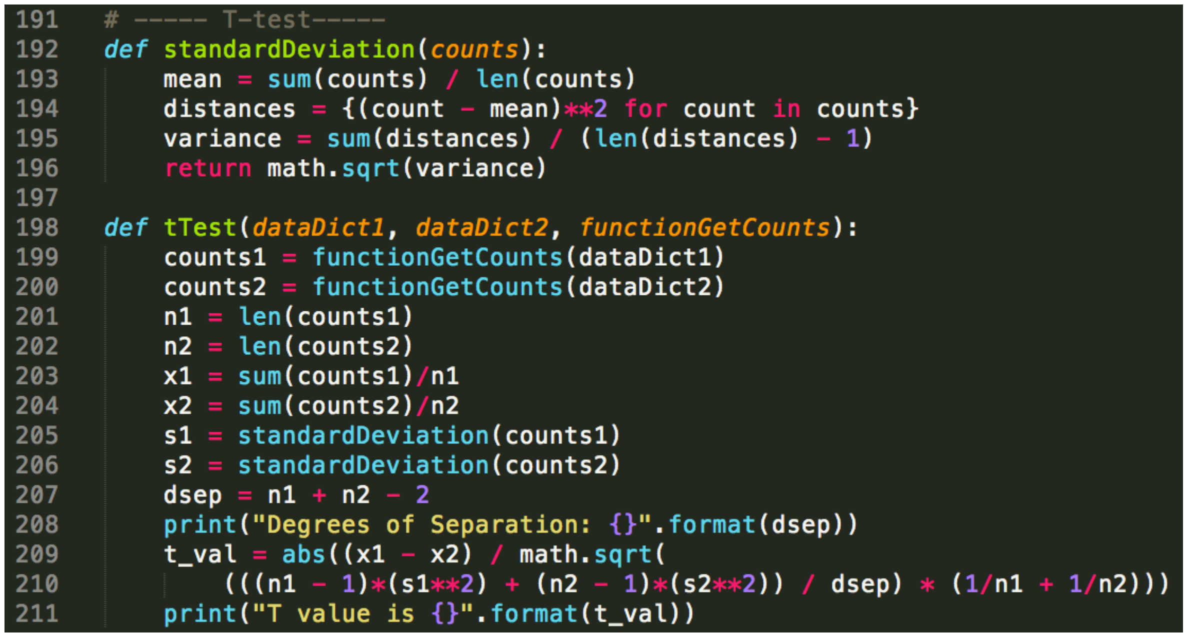
The "expected" table was calculated by multiplying each "Clicked"/"No Click" cell of the observed table by the total ratio of unique sessions with clicks to unique sessions, 38/64.

c. Since there are two rows and two columns (Figure 1.2), there is 1 degree of freedom; we compare 0.51 to the critical value in the table, 3.84. Since $0.51 < 3.84$, the click through rate is not statistically significant.
The corresponding p value, which represents the probability that the chi-squared value would be 0.51 or greater under the null hypothesis, is 0.473929. Since p is greater than 0.05, we see again that the difference in click through rate between interfaces A and B is not statistically significant.

Time to Click

a. The Independent Samples T-Test ("T-Test") is used, as the time to click is a quantitative difference for which each session has a different value.

b. The computed T-value is 0.645.



```
# ----- T-test-----
def standardDeviation(counts):
    mean = sum(counts) / len(counts)
    distances = {(count - mean)**2 for count in counts}
    variance = sum(distances) / (len(distances) - 1)
    return math.sqrt(variance)

def tTest(dataDict1, dataDict2, functionGetCounts):
    counts1 = functionGetCounts(dataDict1)
    counts2 = functionGetCounts(dataDict2)
    n1 = len(counts1)
    n2 = len(counts2)
    x1 = sum(counts1)/n1
    x2 = sum(counts2)/n2
    s1 = standardDeviation(counts1)
    s2 = standardDeviation(counts2)
    dsep = n1 + n2 - 2
    print("Degrees of Separation: {}".format(dsep))
    t_val = abs((x1 - x2) / math.sqrt(
        (((n1 - 1)*(s1**2) + (n2 - 1)*(s2**2)) / dsep) * (1/n1 + 1/n2)))
    print("T value is {}".format(t_val))
```

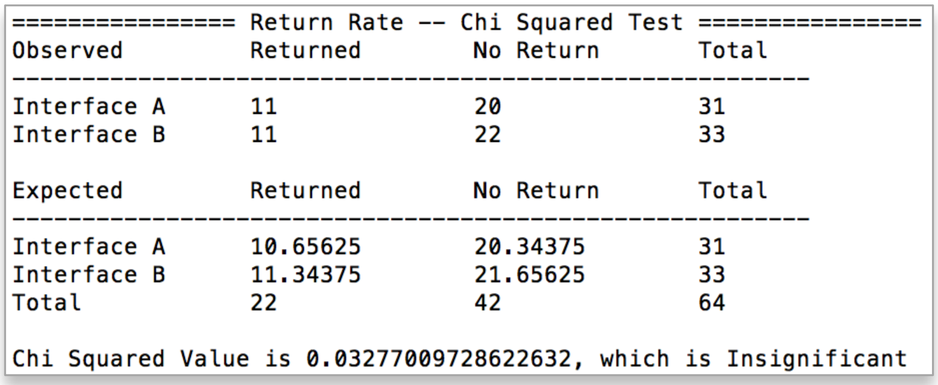
Figure 2.1. Code snippet for T-Test calculation

Again, the computation followed the formula provided, where an additional method was defined to calculate the standard deviation. *tTest*, like *chiSquaredTest* (Figure 1.1), takes a function as a parameter. For *tTest*, this function is one that outputs the counts for every unique session. For calculating the Time to Click, this function is *allTimeToClicks*, which returns the time to click for each relevant session with a click.

c. Given $N_1 = 17$ and $N_2 = 21$, there are 36 degrees of separation. Using the T-Test chart, the T-value must be >= 2.028 for the difference in Time to Click between interfaces to be statistically significant, but it is not.
The corresponding p-value, which represents the probability of the given results under the null hypothesis, is 0.51. Since 0.51 > 0.05, the results are not statistically significant.

Return Rate

a. The Chi-Squared Test is used for this metric, as, again, the metric is a binary true/false (does the unique session return to the page?) which naturally creates two categories.
b. The computed chi-squared value is 0.0328.



```
================ Return Rate -- Chi Squared Test ================
Observed         Returned        No Return        Total
-----------------------------------------------------------
Interface A      11              20               31
Interface B      11              22               33

Expected         Returned        No Return        Total
-----------------------------------------------------------
Interface A      10.65625        20.34375         31
Interface B      11.34375        21.65625         33
Total            22              42               64

Chi Squared Value is 0.03277009728622632, which is Insignificant
```

Figure 3.1. Observed and Expected Chi Squared Tables for Return Rate

The same code as above (see Figure 1.1) was used, where the function was supplied with *allReturnedSessions*, which again takes a representation of the log data and returns the sessions that have a return page load impression.

c. Similar to the Click Rate calculation, the degree of separation is 1. We again compare 0.0328 to the critical value in the table, 3.84. Since $0.0328 < 3.84$, the click through rate is not statistically significant.
The corresponding p value is 0.856283. Since p is greater than 0.05, it again follows that the difference in return rate between interfaces A and B is not statistically significant.

Dwell Time

a. The T-Test is used for this metric, as the time it takes to return to the page after a click is a quantitative difference for which each session has a different value.
b. The T-value for the dwell time is 1.148. See Figure 2.1 for the method to calculate the T-value; for the dwell time, the function *allDwellTime* is passed into *tTest*, which returns the first dwell time for each of the sessions that have a dwell time.
c. There are 20 degrees of separation for this t-test calculation, as $N_1 = N_2 = 11$. Looking at the t-value table of critical values, the t-value must be greater than 1.724718 to have a p-value of 0.05 or smaller. Since $1.148 < 1.724718$, the difference in dwell time between the two interfaces is not statistically significant.

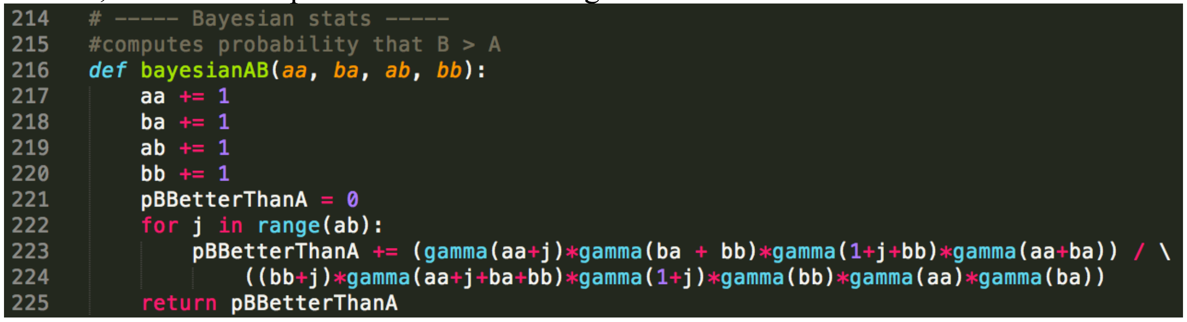
Bayesian Probability

By the Bayesian A/B test for click through, the probability that interface B is better than interface A for the click through rate is 0.76, while the probability that interface A is better than interface B by the same metric is 0.24.
These probabilities were calculated based on the alternate format of the Bayesian probability test in Evan Miller's implementations of A/B testing, such that the equation could be expressed with the gamma function, as below.[2]

$$\Pr(p_B > p_A) = \sum_{i=0}^{\alpha_B - 1} \frac{\Gamma(\alpha_A + i)\Gamma(\beta_A + \beta_B)\Gamma(1 + i + \beta_B)\Gamma(\alpha_A + \beta_A)}{(\beta_B + i)\Gamma(\alpha_A + i + \beta_A + \beta_B)\Gamma(1 + i)\Gamma(\beta_B)\Gamma(\alpha_A)\Gamma(\beta_A)}$$

Figure 4.1. Bayesian Probability Test expressed in terms of gamma[3]

In code, this can be expressed as the following:



```
# ----- Bayesian stats -----
#computes probability that B > A
def bayesianAB(aa, ba, ab, bb):
    aa += 1
    ba += 1
    ab += 1
    bb += 1
    pBBetterThanA = 0
    for j in range(ab):
        pBBetterThanA += (gamma(aa+j)*gamma(ba + bb)*gamma(1+j+bb)*gamma(aa+ba)) / \
            ((bb+j)*gamma(aa+j+ba+bb)*gamma(1+j)*gamma(bb)*gamma(aa)*gamma(ba))
    return pBBetterThanA
```

Figure 4.2. Calculation for the Bayesian probability that B is greater than A

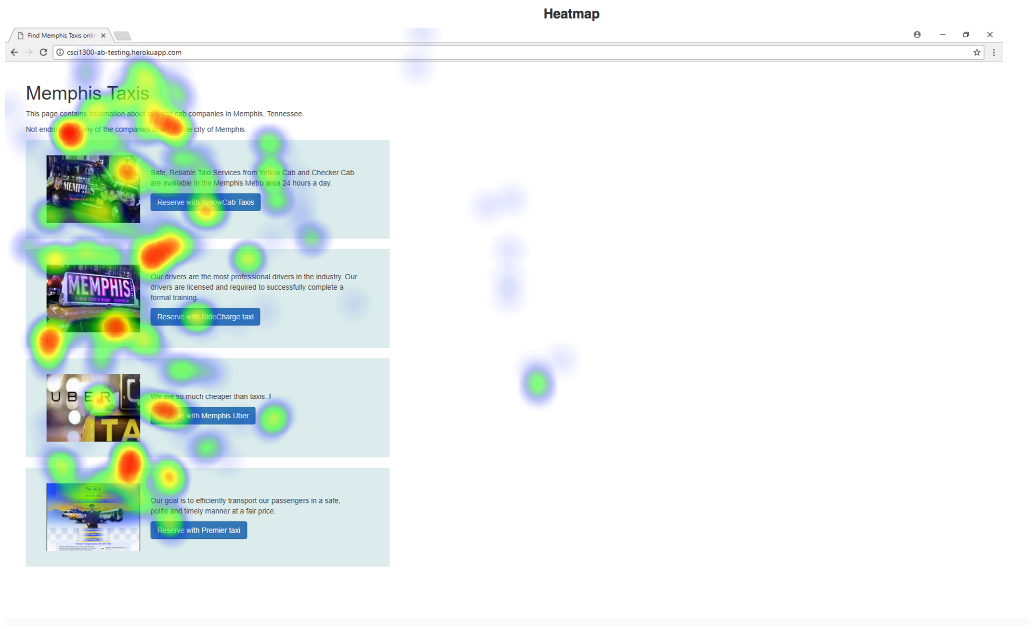
[2] This can also be done equivalently with scipy.special.beta, using the original formula provided
[3] Screenshot from www.evanmiller.org/bayesian-ab-testing.html

Part II: Eye Tracking

Hypotheses

- In version A, the map will look denser on the left half of the screen, whereas it will be sparser in version B, because all of the content is on the left half of the screen in version A.
- In version B, users will look repeatedly at the same part of the page multiple times, whereas users will only look at each entry once in version A because version B facilitates more comparison between entries.

Results – Heatmap (heatmap-a.html, heatmap-b.html)


Heatmap
Find Memphis Taxis onli
csci1300-ab-testing.herokuapp.com
Memphis Taxis
Reserve with YellowCab Taxis
Reserve with RideCharge taxi
Reserve with Memphis Uber
Reserve with Premier taxi

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.

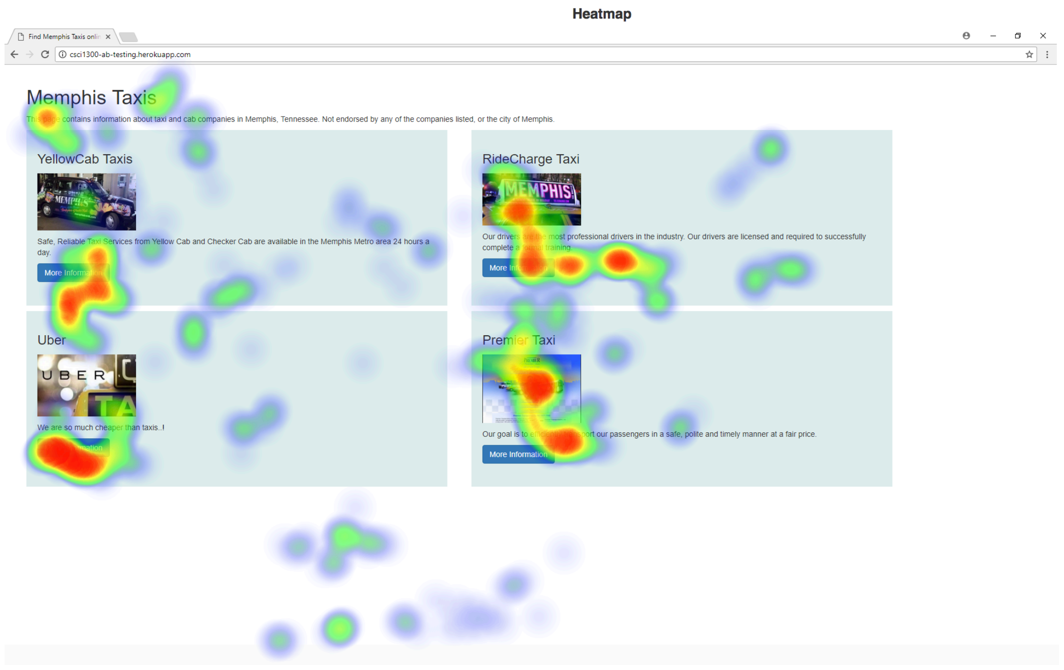
Safe, Reliable Taxi Services from Yellow Cab and Checker Cab are available in the Memphis Metro area 24 hours a day.

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

We are so much cheaper than taxis..!

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

Figure 5.1. Heat Map for Interface A



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.

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

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

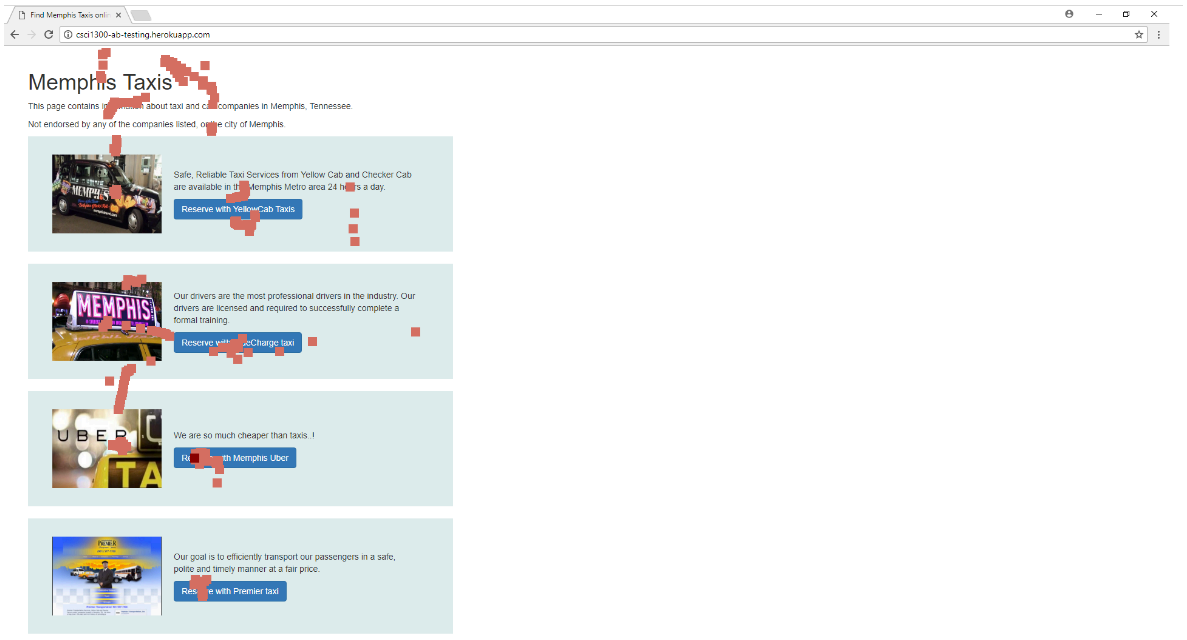
We are so much cheaper than taxis..!

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

Figure 5.2. Heat Map for Interface B

Results – Replay (replay-a.html, replay-b.html)

Interface A[4]


Find Memphis Taxis onli
csci1300-ab-testing.herokuapp.com
Reserve with YellowCab Taxis
Reserve with RideCharge taxi
Reserve with Memphis Uber
Reserve with Premier taxi

Memphis Taxis

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.

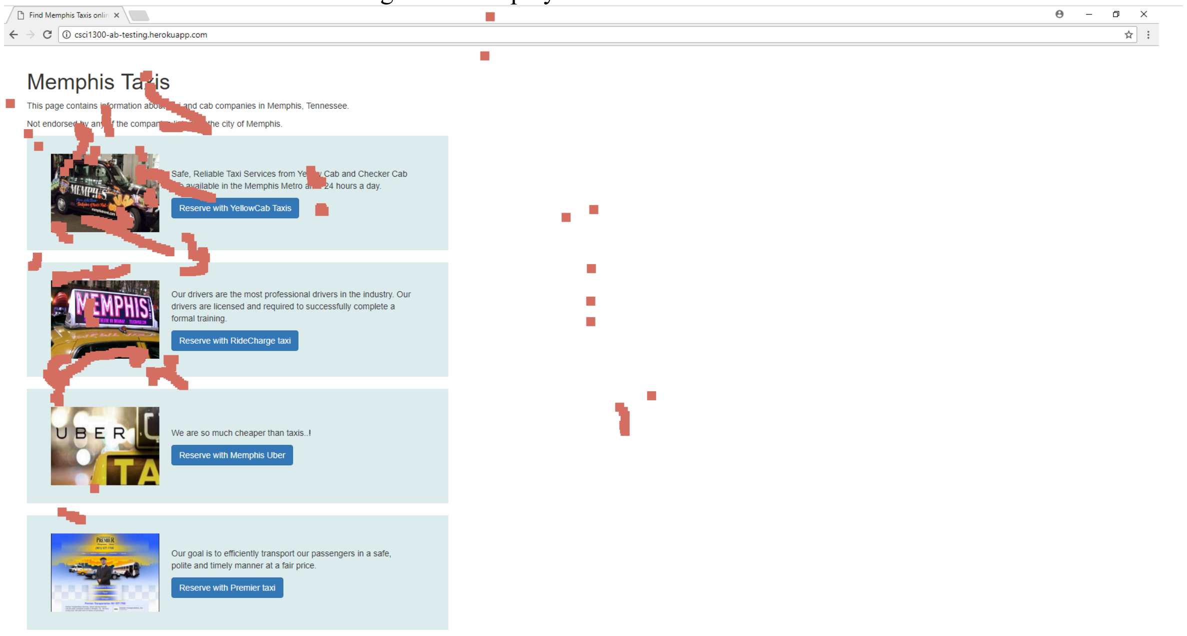
Safe, Reliable Taxi Services from Yellow Cab and Checker Cab are available in the Memphis Metro area 24 hours a day.

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

We are so much cheaper than taxis..!

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

Figure 6.1. Replay in Action for Interface A


Find Memphis Taxis onli
csci1300-ab-testing.herokuapp.com
Reserve with YellowCab Taxis
Reserve with RideCharge taxi
Reserve with Memphis Uber
Reserve with Premier taxi

Memphis Taxis

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.

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

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

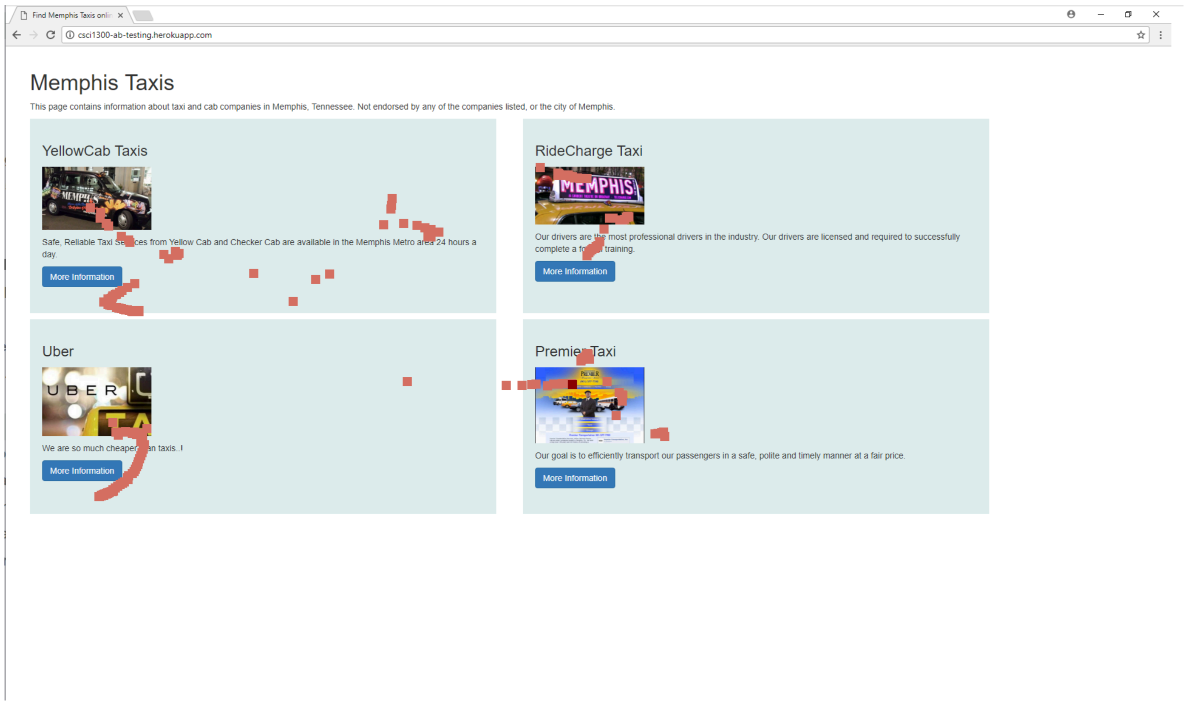
We are so much cheaper than taxis..!

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

Figure 6.2 Replay at End for Interface A

[4] Eye tracking logs for interface A were cleaned to prevent eye tracking data while the user was on a different site from inter

Interface B



Memphis Taxis

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.

YellowCab Taxis

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

More Information

RideCharge Taxi

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

More Information

Uber

We are so much cheaper than taxis..!

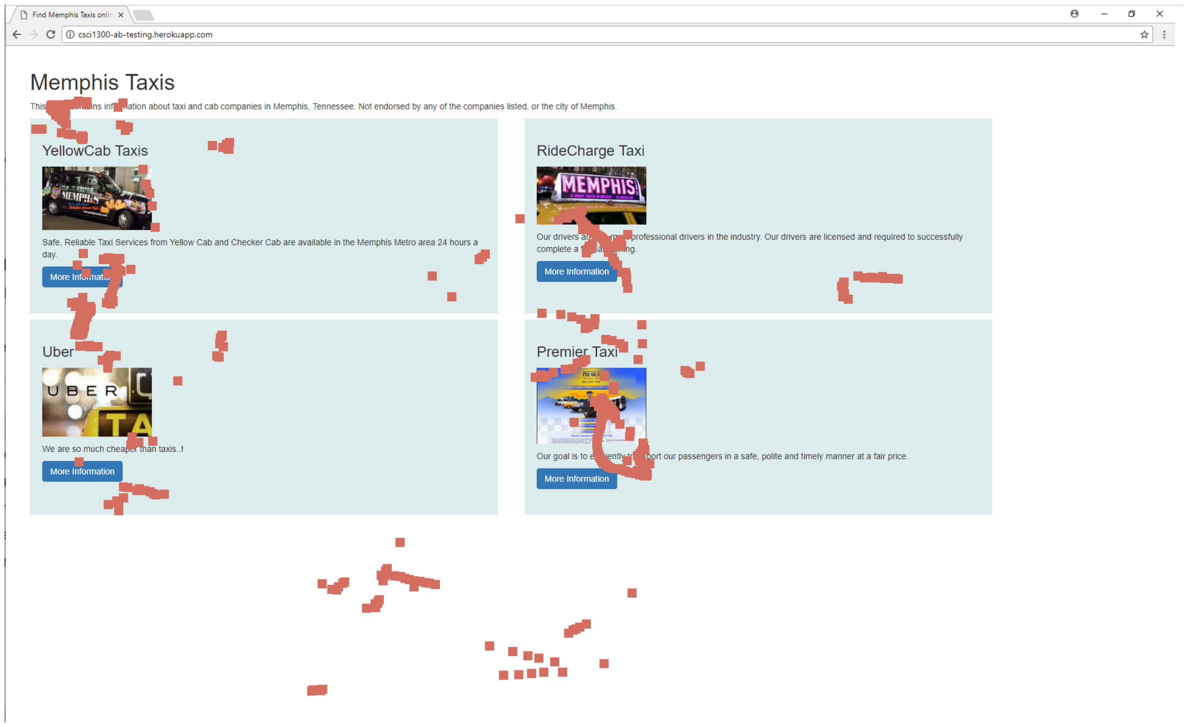
More Information

Premier Taxi

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

More Information

Figure 6.3. Replay in Action for Interface B



Memphis Taxis

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.

YellowCab Taxis

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

More Information

RideCharge Taxi

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

More Information

Uber

We are so much cheaper than taxis..!

More Information

Premier Taxi

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

More Information

Figure 6.4. Replay at end of Interface B

Analysis
Our first hypothesis appeared to be reflected in the results, with the majority of results centered on the left half of the page in interface A and on the four distinct sections in interface B. Of note is that within the content area, interface A's heatmap is much more spread out; in interface B, most of the focus points are concentrated to the image or button in the section. The second hypothesis did not appear to be correct; the replay for interface A shows the user scanning up and down the page visually multiple times.

Part III: Comparison

1. Given that the results of this test were inconclusive, further tests are necessary to determine which, if any, of the differences between interfaces made any change to the metrics that you (Memphis Taxi Co.) value. For the three changes (layout, button text, and titles), it appears from the eye-tracking data that car titles may not be as of interest for users as was initially hypothesized. The user focused visually on the image and button, as is most visible in Figure 5.2; I would therefore recommend that future tests explore further changes to the button text, as that change may be statistically significant where the combination of the three factors was not.

2. The A/B testing data, at least in the context of this experiment, allows for more accurate aggregation of data as well as insight into *what* the user did on the site; in turn, the eye tracking test reflects *why* the user (in this case, the single user for each interface) may have clicked. Had the difference in click through rate between the two interfaces been statistically significant, the eye tracking test would have been able to show us exactly which changes motivated the increased rate. With the current eye tracking test results, for example, we might be able to assert that the change in layout or button text motivated this increase, but A/B testing data is necessary to show that this focus and effect holds true in a statistically significant way.