A/B Testing

Hypotheses

1. Click Through Rate:
    a. Null Hypothesis: Version A and Version B will have the same click through rate.
    b. Alternative Hypothesis: The click through rate for version A will be greater than that for version B because version A has clearly defined buttons that help with learnability. Also, version A has clear affordances because the button changes colors when the user hovers over it.
2. Time to Click:
    a. Null Hypothesis: The time to click will be equal for version A and version B.
    b. Alternative Hypothesis: The time to click will be quicker for version A than that for version B because the buttons in version A are easily identifiable whereas version B has each taxi company in a separate blue box that appears to be clickable but actually isn't. This leads to confusion and means users take a longer time to identify the small "Reserve" label in the corner of the box that is the clickable element.
3. Dwell Time:
    a. Null Hypothesis: Both versions A and B will have the same dwell time.
    b. Alternative Hypothesis: The dwell time for version A will be higher than that of version B. This is because version A has condensed, vague descriptions of the cab companies compared to more detailed descriptions in version B. This means that the user will spend more time gathering information on each cab company's

website when they navigate there from version A compared to version B that presents all of that information at a glance.

4. Return Rate:
    a. Null Hypothesis: Version A and version B will have the same return rate.
    b. Alternative Hypothesis: The return rate for version A will be higher than that for version B because in terms of usability, version A requires scrolling through the page to look at all the options which will lead to the users clicking on the first available button, reading more about the taxi company, and then returning in order to fully explore other options. However, in version B, the user is able to see all the taxi companies at once without scrolling, meaning that they can make an informed choice and not return to our page as often.

Data Analysis

1. The Four Metrics
    a. Click Through Rate:
        i. Version A: 39/39 = 100%
        ii. Version B: 20/20 = 100%
    b. Time to Click:
        i. Version A: 11200.05128 ms
        ii. Version B: 14061.9 ms
    c. Dwell Time:
        i. Version A: 14990.24 ms

ii. Version B: 16362.4 ms

d. Return Rate:

i. Version A: 25/39 = 64.1%

ii. Version B: 10/20 = 50%

2. Statistical Tests

a. Click Through Rate:

i. We used the Chi-squared test for the click through rate because it is categorical data. There are only two possibilities: clicked or no click.

Observed	**Clicked**	**No Click**	**Total**
A	39	0	39
B	20	0	20
Total	59	0	59

<u>Calculating the Expected Values</u>

$$(59 \cdot 39) \div 59 = 39$$

$$(59 \cdot 20) \div 59 = 20$$

$$(0 \cdot 39) \div 59 = 0$$

$$(0 \cdot 20) \div 59 = 0$$

Expected	**Clicked**	**No Click**	**Total**
A	39	0	39
B	20	0	20
Total	59	0	59

$$x^2 = \sum_{all\ cases} \frac{(observed - expected)^2}{expected}$$

$$\frac{(39-39)^2}{39} + \frac{(20-20)^2}{20} + \frac{(0-0)^2}{0} + \frac{(0-0)^2}{0} = 0$$

Degrees of Freedom: $(2-1)(2-1) = 1$

Value from Table: 3.84

$0 < 3.84$ so it is not statistically significant

ii. It makes sense that it is not statistically significant because the click through rate is 100% for both interfaces so there is no difference. Because it is not statistically significant, we cannot reject our null hypothesis. It also means that we cannot accept our alternative hypothesis. This is because there is no statistically significant difference in click through rate for the two versions, meaning that we cannot determine if the difference in the clicks was because of the different designs.

b. Time to Click:

i. We used the Independent Samples t-test because we are calculating the difference in the means of the two interfaces. We are working on the difference in the means of the time to click for the two interfaces.

$$\overline{X_1} = 11200.5128 \quad \overline{X_2} = 14061.9$$

$$N_1 = 39 \quad N_2 = 20$$

$$s_1 = \sqrt{\frac{1}{39}\sum_{i=1}^{39}(x_i - 11200.5128)^2}$$

$$s_1 = 15446.11468$$

$$s_2 = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(x_i - 14061.9)^2}$$

$$s_1 = 18586.2398$$

$$t = \frac{11200.5128 - 14061.9}{\sqrt{\frac{38(15446.11468)^2+19(18586.2398)^2}{57}(\frac{1}{39}+\frac{1}{20})}}$$

$$t = -0.628291 \quad |t| = 0.628291$$

Degrees of freedom: 39 + 20 - 2 = 57

Value from table: 2.00

$|t| < 2.00$ so it is not statistically significant

ii. Because it is not statistically significant, we cannot reject our null hypothesis. It also means that we cannot accept our alternative hypothesis.

This is because there is no statistically significant difference in time to click for the two versions, meaning that we cannot determine if the difference in the time to click was because of the different designs.

c. Dwell Time:

i. We used the Independent Samples t-test because we are calculating the difference in the means of the two interfaces. We are working on the difference in the means of the dwell time for the two interfaces.

$$\overline{X_1} = 14990.24 \quad \overline{X_2} = 16362.4$$

$$N_1 = 39 \quad N_2 = 20$$

$$s_1 = \sqrt{\frac{1}{39}\sum_{i=1}^{39}(x_i - 14990.24)^2}$$

$$s_1 = 15756.47478$$

$$s_2 = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(x_i - 16362.4)^2}$$

$$s_1 = 8160.65078$$

$$t = \frac{14990.24 - 16362.4}{\sqrt{\frac{38(15756.47478)^2+19(8160.65078)^2}{57}(\frac{1}{39}+\frac{1}{20})}}$$

$$t = -0.364152 \quad |t| = 0.364152$$

Degrees of freedom: 39 + 20 - 2 = 57

Value from table: 2.00

$|t| < 2.00$ so it is not statistically significant

ii. Because it is not statistically significant, we cannot reject our null hypothesis. It also means that we cannot accept our alternative hypothesis. This is because there is no statistically significant difference in dwell time for the two versions, meaning that we cannot determine if the difference in dwell time was because of the different designs.

d. Return Rate:

i. We used the Chi-squared test for the return rate because it is categorical data. The user either returned to the page or didn't.

Observed	**Returned**	**No Return**	**Total**
A	25	14	39
B	10	10	20
Total	35	24	59

Calculating the Expected Values

$(35 \cdot 39) \div 59 = 23.1$

$(35 \cdot 20) \div 59 = 11.9$

$(24 \cdot 39) \div 59 = 15.9$

$(24 \cdot 20) \div 59 = 8.1$

Expected	**Returned**	**No Returned**	**Total**
A	23.1	15.9	39
B	11.9	8.1	20
Total	35	24	59

$$x^2 = \sum_{all\ cases} \frac{(observed - expected)^2}{expected}$$

$$\frac{(25 - 23.1)^2}{23.1} + \frac{(14-15.9)^2}{15.9} + \frac{(10 - 11.9)^2}{11.9} + \frac{(10 - 8.1)^2}{8.1} = 1.13236$$

Degrees of Freedom: $(2 - 1)(2 - 1) = 1$

Value from Table: 3.84

$1.13236 < 3.84$ so it is not statistically significant

ii. Because it is not statistically significant, we cannot reject our null hypothesis. It also means that we cannot accept our alternative hypothesis. This is because there is no statistically significant difference in return rate for the two versions, meaning that we cannot determine if the difference in return rate was because of the different designs.

3. 95% Confidence Interval

$$\overline{X_1} = 11200.5128 \quad \overline{X_2} = 14061.9$$

$$N_1 = 39 \quad N_2 = 20$$

$$s_1 = \sqrt{\frac{1}{39}\sum_{i=1}^{39}(x_i - 11200.5128)^2}$$

$$s_1 = 15446.11468$$

$$s_2 = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(x_i - 14061.9)^2}$$

$$s_1 = 18586.2398$$

$$se = \sqrt{\frac{38(15446.11468)^2+19(18586.2398)^2}{57}\left(\frac{1}{39} + \frac{1}{20}\right)}$$

$$se = 4554.24008$$

Degrees of freedom: 39 + 20 - 2 = 57

Value from table: 2.00

$$2.00 \cdot 4554.24008 = 9108.48016$$

$$\overline{X_1} - \overline{X_2} = 11200.5128 - 14061.9$$

$$\overline{X_1} - \overline{X_2} = -2861.3872$$

$$-2861.3872 - 9108.48016 = -11969.86736$$

$$-2861.3872 + 9108.48016 = 6247.09296$$

95% Confidence Interval is -11969.86736 to 6247.09296

This interval contains 0 so there is not a statistically significant difference.

Eye Tracking

Hypothesis

Version A will have a greater proportion of eye gazes in the center than version B because all the elements in version A are in a column down the center of the screen. Version A has eye catching images and colors mainly in the middle of the screen whereas version B is more spread out across the width of the screen.

Data Analysis

We included all 6 of our screenshots in our handin. These screenshots support our hypothesis. Version A's eye gazes are much more concentrated in the center of the screen whereas version B's eye gazes are spread out throughout the entire width of the page. These observations make sense because the elements of version A are exclusively arranged in the middle of the screen in only one column, while version B makes use of the entire width of the page by utilizing a 2x2 grid and showing multiple entries in each row.

Comparison

1. What to do

   Because there is no statistical significance in our A/B tests, we can assume that there are no overall differences between Version A and Version B. However, according to the Eye Tracking tests, we can say that both versions are the opposite of each other: with one design *being concentrated in the center* while the other is *completely spread out*. Version A's elements are too crowded in the middle while Version B's elements blended in too much and therefore, does not catch the user's attention. The best recommendation is to

combine both versions and redesign the website page. The new design should have its elements spread out similar to what is on Version B, but there should be certain elements such as images or the details buttons that stand out to the users like Version A to make *navigation easier*.

2. A/B vs eye tracking

   The A/B tests are mostly about gathering statistics based on the number of clicks and time spent doing each action while eye tracking gives us more information on the user's behavior.

   Some of the advantages of A/B testing are:

   - easy to get *clear evidence* when comparing different ideas
   - *quantitative data* can be obtained
   - *reliable data* in short amount of time

   The advantages of eye tracking compared to A/B testing:

   - more detailed data on user's behaviors
     - shows the elements that people tend to ignore/focus or those that distract the users away from the key components
     - shows clearly which components are important or irrelevant
     - good for identifying how users are interacting with the interface

3. Unethical metrics

   - *Dwell Time* can be used *unethically* if it is maximized through applying *addictive tools and patterns* that intentionally lure users to spend longer time on the webpage without providing more information and functionalities. For example,

many websites that provide entertaining reading experience have unrelated and distractive ads on the side of their contents that often prevent the users from focusing on what they want and in the end spend longer time on the website.

- *Number of Clicks* can be used *unethically* if it is maximized by *deceiving users and creating functionalities* that does not contribute to the website's functionalities. When used positively, it can measure the level of user engagement with the webpage. It is commonly used in e-commerce platforms. However, if used unethically, designers can put nested links on the webpage, making users click on newly popped up links after they click on the initial link they see. This way, click through rate can be maximized, but no real tasks gets done.