**Apoorva Kadam**

Dataset: Each row in this dataset contains information about a person who signed up for a demo of a software product. The objective is to build a model to predict whether or not a given user will "convert" and purchase the software.

Data Feature Descriptions:

 num\_visits - the number of distinct sessions the user spent with the software

 total\_minutes\_on\_demo - the number of total minutes the user spent using the software

 user\_age - the age of the user in years

 demo\_source - how the user initially discovered the software

 data\_rows\_input - how many rows of their own data the user uploaded to the software

 users\_in\_network - how many people the user knows who already use the software

 outreach\_emails\_sent - the number of emails sent to the user by the software company while the user is in the demo period

 support\_tickets\_filed - the number of support requests sent by the user to the software company while using the demo

 converted - boolean label of whether the user converted and purchased the software.

**Research question:**

1. Which demo source is better? Email or website?
2. Total minutes spent on email demo is more or on website?
3. User age of email demo source v/s website demo source user age
4. For user who checked demo software sent through email weather total minutes spent on demo software is a significant factor in predicting whether or not that user converted the software?
5. For user who checked demo software on website weather total minutes spent on demo software is a significant factor in predicting whether or not that user converted the software?
6. For user who checked demo software sent through email weather age group is a significant factor in predicting whether or not that user converted the software?
7. For user who checked demo software sent through email weather age group is a significant factor in predicting whether or not that user converted the software?
8. Which

Table function on the converted variable of subset demo source email shows that 693 people purchased software and for website demo source 879 people converted software. This clearly shows that people who check demo on website are more likely to purchase the software than the one whom the demo source was sent though email.

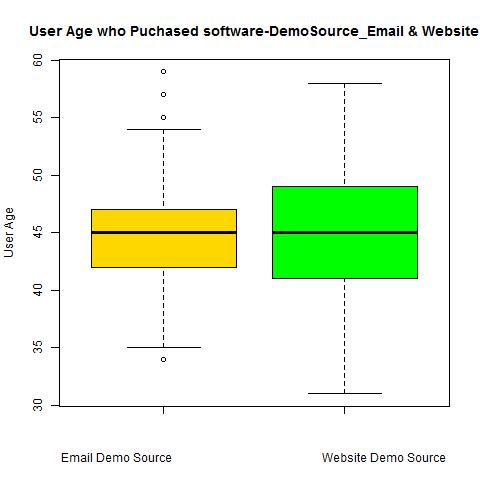
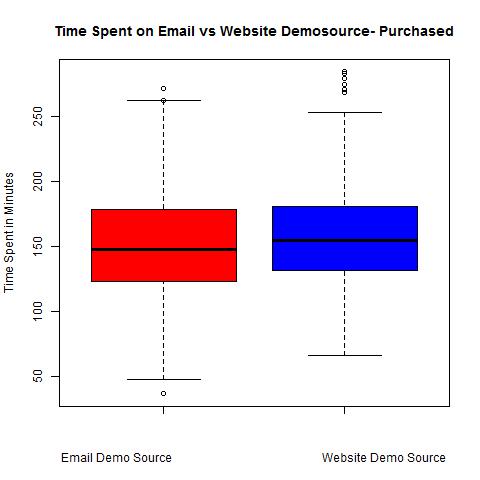
**Descriptive Statistics:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Demo\_Source - Email** | Min | 1st Qu. | Median | Mean | 3rd Qu | Max |
| total\_minutes\_on\_demo | 11.46 | 126.70 | 149.50 | 148.60 | 170.70 | 277.20 |
| user\_age | 25.0 | 42.0 | 45.0 | 45.1 | 48.0 | 65.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Demo\_Source - Website** | Min | 1st Qu. | Median | Mean | 3rd Qu | Max |
| total\_minutes\_on\_demo | 27.97 | 126.00 | 146.40 | 148.10 | 167.40 | 284.30 |
| user\_age | 23.00 | 40.00 | 44.00 | 44.06 | 48.00 | 59.00 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Demo\_Source – Email-Purchased** | Min | 1st Qu. | Median | Mean | 3rd Qu | Max |
| total\_minutes\_on\_demo | 36.75 | 122.90 | 147.90 | 151.70 | 178.70 | 271.90 |
| user\_age | 34.0 | 42.0 | 45.0 | 44.5 | 47.0 | 59.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Demo\_Source – Website-Purchased** | Min | 1st Qu. | Median | Mean | 3rd Qu | Max |
| total\_minutes\_on\_demo | 65.92 | 131.20 | 154.30 | 157.80 | 180.90 | 284.30 |
| user\_age | 31.00 | 41.00 | 45.00 | 45.03 | 49.00 | 58.00 |



**Observations:**

* For demo source email, when we observe the total time spent on demo min is 11.46 minutes average time spent is around 148 minutes. When compared this with statistics for purchased user’s min time spent is 36 minutes and average time spent is 150 minutes. This clearly shows that people who are interested in software and who wanted to purchase software did spend more time on demo.
* When we consider user age for any demo source, we can see that even though user below 30yrs spent time on demo source, the people who purchased software is above 30 years only.

**T-test: Difference in means of the two samples**

* Email User\_age vs website user\_age. In this P-value = 0.01898235, and Conf Int =

- 0.97059710 -0.08711803, x mean = 44.50072, y mean = 45.02958. P-value is less than 0.05 and confidence interval does not include zero hence it is statistically significant. Hence test is significant and the means of 2 groups do not overlap.

* Email\_ total\_minutes\_on\_demo vs website total\_minutes\_on\_demo. In this P-value = 0.002301062, and Conf Int = -10.038365 -2.186099, x mean = 151.6688, y mean = 157.7810. P-value is less than 0.05 and confidence interval does not include zero hence it is statistically significant. Hence test is significant and the means of 2 groups do not overlap.

**Effect size: d = M1 - M2 / pooled   
    where pooled =√ [( 12+  22) / 2]**

* **pooled = sqrt((4.031174^2+4.895867^2)/2)= 4.484411**
* **pooled = sqrt( (40.99933^2+37.26926^2)/2) =** **39.17871**
* Email User\_age vs website user\_age 44.50072-45.02958/ 4.484411= 0.117933
* Email\_ total\_minutes\_on\_demo vs website total\_minutes\_on\_demo

151.6688-157.7810/ 39.17871= 0.1560082

**Regression**: Here our goal is to estimate the probability of binary response (Converted/ Not) for demo source email and website based on one or more feature.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Demo source | Dependent Variable | Independent Variables | Intercept | Co-efficients | P-vale | Equ |
| Email | Converted | total\_minutes\_on\_demo  +num\_visits | -7.75868 | 0.009711  0.729734 | 3.97e-08  2e-16 | Converted = -7.75868+0.009711  \*total minutes+ 0.729734 \*num\_vists |
| Website | Converted | total\_minutes\_on\_demo  +user\_age+num\_vists | -8.79649 | 0.018686  0.066521  0.331702 | 2e-16  3.86e-11  2e-16 | Converted = -8.79649+0.018686\* total\_minutes +0.066521  \*user\_age+0.331702\* num\_visits |

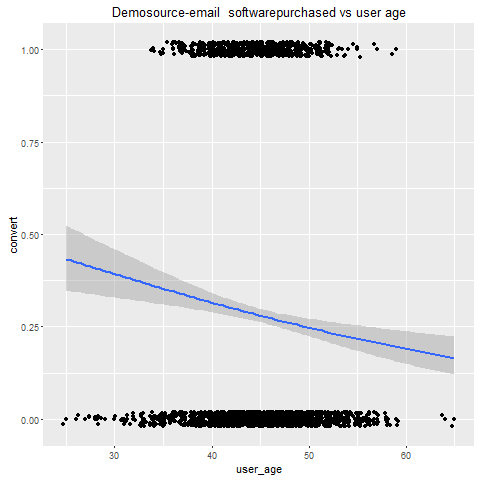
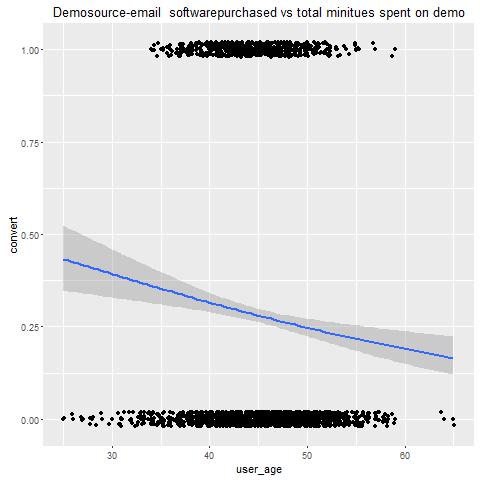
**Confidence Interval: 95% confidence that population mean lie between the two interval lower and upper**

**Regression1:**

|  |  |  |
| --- | --- | --- |
|  | Upper CI | Lower CI |
| (Intercept) | 0.001076469 | 0.0001693897 |
| total\_minutes\_on\_demo | 1.013263195 | 1.0062648842 |
| num\_visits | 2.251100081 | 1.9118090973 |

**Regression 2**:

|  |  |  |
| --- | --- | --- |
|  | Upper CI | Lower CI |
| (Intercept) | 0.0005619739 | 4.071425e-05 |
| total\_minutes\_on\_demo | 1.0223324686 | 1.015403e+00 |
| user\_age | 1.0900755632 | 1.047907e+00 |
| num\_visits | 1.4947801050 | 1.298780e+00 |

****

