A/B Testing

Introduction

The number of visitors on our website equals the number of opportunities we have to expand our business by acquiring new customers and build relationships by catering to existing ones. And it is our conversion funnel that decides whether our website gets good traffic and if it converts more visitors. Businesses want visitors to take action (also called a conversion) on their website, and the rate at which a site can drive this is called its "conversion rate." The more optimized our funnel, the higher is the visitors' chance to convert.

One way to optimize our website's funnel is **A/B testing**. A/B testing (also sometimes referred to as split testing) is the practice of showing two variants of the same web page to different segments of visitors at the same time and comparing which variant drives more conversions. Typically, the one that gives higher conversions is the winning variant, applying, which can help us optimize our site for better results.

Our Project and Concern

For this project, we will be working to understand the results of an A/B test run by an e-commerce website. Our goal is to work through this notebook to help the company understand if they should implement the new page, keep the old page, or perhaps run the experiment longer to make their decision.

Importing Data Sets

Let us first import the given data sets and have a view about different columns and how the data is enumerated.

[Code]

```
> data_set_1 = read.csv(file = "E:/R workshop/ab_data.csv" , header = T)
> head(data_set_1)
  user_id timestamp     group landing_page converted
1  851104   11:48.6   control     old_page         0
2  804228   01:45.2   control     old_page         0
3  661590   55:06.2 treatment     new_page         0
4  853541   28:03.1 treatment     new_page         0
5  864975   52:26.2   control     old_page         1
6  936923   20:49.1   control     old_page         0
```

```
> data_set_2 = read.csv(file = "E:/R workshop/countries.csv" , header = T)
> head(data_set_2)
```

```
  user_id country
1  834778      UK
2  928468      US
3  822059      UK
4  711597      UK
5  710616      UK
6  909908      UK
```

Understanding and exploring the datasets

Now as we have imported the data let's try and understand the data sets individually for different columns

Data set 1:

[Code]

```
> nrow(data_set_1)
[1] 294478
> ncol(data_set_1)
[1] 5
> colnames(data_set_1)
[1] "user_id"      "timestamp"    "group"        "landing_page" "converted"
```

1. user_id: this is the unique identity of each customer who have taken part in the experiment and performed some action.

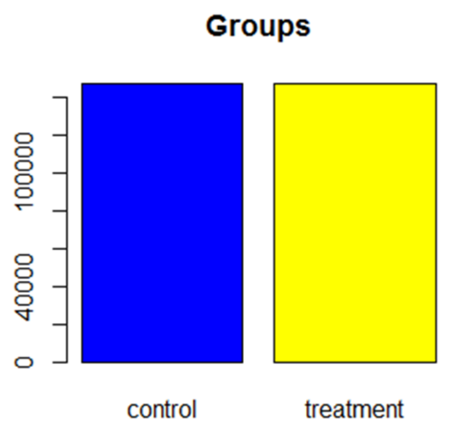
2. timestamp: this denotes the time on which the customers have performed some action in the e-commerce website.

3. group :

[Code]

```
> table(data_set_1$group)

  control treatment
   147202    147276
> barplot(table(data_set_1$group), main = 'Groups', col = c('Blue','Yellow'))
```



We observe that the entire data set is almost equally divided among control and treatment.Now let us understand what control and treatment are.

Control: this is the unchanged or original version of the web page.

Treatment: this is the changed or modified version of the web page.

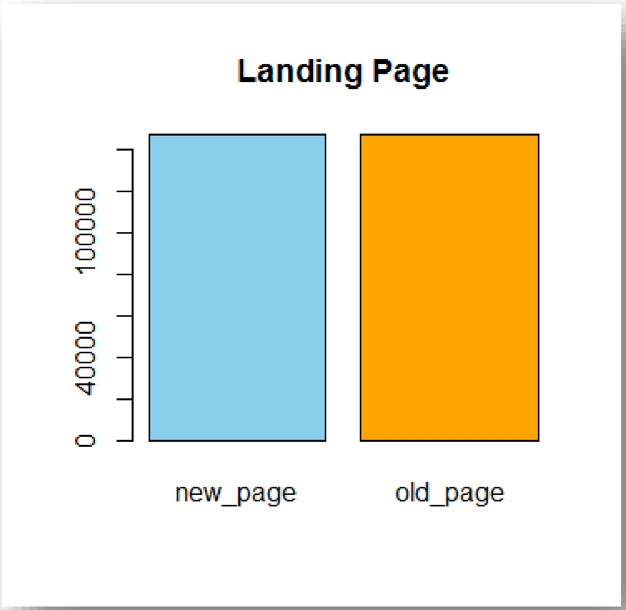
4. Landing_page:

[Code]

```
> table(data_set_1$landing_page)

new_page old_page
  147239   147239

> barplot(table(data_set_1$landing_page), main = 'Landing Page', col =
c('SkyBlue','Orange'))
```



We observe that there are two landing pages that the customers drive into when they open the e-commerce website. Old page and New page. And for the given data sets we have equal number of responses for both the lading pages.

5. Converted: As we have seen have, this column is mainly divided into two categories.
    - ‘0’: the customers did not get converted or did not show satisfactory concern about the changes
    - ‘1’: the customers got converted or liked the changes

We will later discuss about the number of conversions in details when we begin our analysis on our data sets.

Data set 2:

[Code]

```
> nrow(data_set_2)
[1] 290584
> ncol(data_set_2)
[1] 2
> colnames(data_set_2)
[1] "user_id" "country"
```

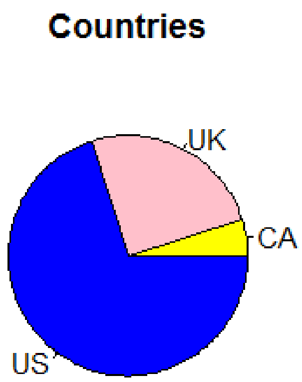
We observe that even the second data set has the same number of columns as data set 1. And both the data sets are linked to each other through the primary key user_id.

We have another column ‘countries’,that depicts the country from which response has been made by the customers in the e-commerce website.

[Code]

```
> table(data_set_2$country)

    CA     UK     US 
 14499  72466 203619 
> pie(table(data_set_2$country), main= 'Countries', col = c('yellow','pink','blue'))
```



We can clearly observe that the respondents are maximum from United States, followed by United Kingdom and then from Canada.

Analysis of the Data set

We have seen above that the two data sets are linked to each other through the primary key ‘user_id’. Therefore it would be justified if we combine the data sets before performing our analysis.

Joining the data sets

[Code]

```
> data_set = merge.data.frame(data_set_2,data_set_1, by = 'user_id')
```

```
> head(data_set)
  user_id country timestamp     group landing_page converted
1  630000      US   26:06.5 treatment     new_page         0
2  630001      US   16:42.6 treatment     new_page         1
3  630002      US   20:56.4   control     old_page         0
4  630003      US   09:31.5 treatment     new_page         0
5  630004      US   23:58.8 treatment     new_page         0
6  630005      US   22:25.9 treatment     new_page         1
```

Now that we have merged the data sets, we will be working on the final merged data set as we proceed further

Checking missing values

Let us now check if there is any missing value or unreported data our data set.

[Code]

```
> sapply(data_set, function(x){sum(is.na(x))})
     user_id      country    timestamp        group landing_page    converted
           0            0            0            0            0            0
```

We can observe that we do not have any missing or unreported value in our data set.

Checking Redundancies

If need to be aware about the duplicate data entries in the data set. If there is any then it has to be removed or acted upon accordingly.

[Code]

```
> data_duplicate <- sqldf("select user_id, timestamp, count(*) from data_set
group by user_id, timestamp")
> head(data_duplicate)

  user_id timestamp count(*)
1  630000   26:06.5        1
2  630001   16:42.6        1
3  630002   20:56.4        1
4  630003   09:31.5        1
5  630004   23:58.8        1
6  630005   22:25.9        1

> View(data_duplicate)
> nrow(data_duplicate)

[1] 294478
```

We see that we get the same number of rows as the original data set. So primarily there is no redundancy in the data set.

Checking random allocation of treatment and control groups

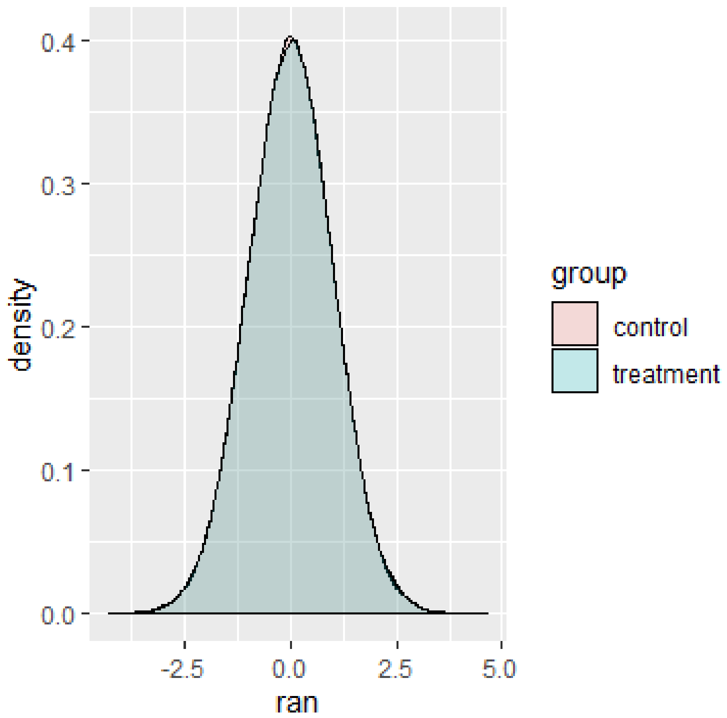
In A/B testing procedure we are aware that the treatment and control groups are generally randomly allocated to the response of the customers. Let us check if this is applicable in our given data set.

[Code]

```
> library("ggplot2", lib.loc="~/R/win-library/3.5")
> data_set = cbind(data_set,  ran = rnorm(294478))
> head(data_set)

  user_id country timestamp     group landing_page converted        ran
1  630000      US   26:06.5 treatment     new_page         0 -0.1259390
2  630001      US   16:42.6 treatment     new_page         1 -1.7083307
3  630002      US   20:56.4   control     old_page         0  0.8088963
4  630003      US   09:31.5 treatment     new_page         0  0.2308176
5  630004      US   23:58.8 treatment     new_page         0  0.5418345
6  630005      US   22:25.9 treatment     new_page         1  0.8489695

> ggplot(data_set, aes(x=ran, fill= group)) + geom_density(alpha=.2) +
xlab("ran")
```



From the plot we observe that the treatment and control groups are randomly allocated to the customer responses.

Testing

To check which page (old page or new page) is performing better we first test for their conversions in the three different countries, United States (US), United Kingdom (UK), and Canada (CA).

United States

Old Page

Null Hypothesis: Conversions of control and treatment are same.

[Code]

```
> library(dplyr)
> US_old= filter(data_set, country == 'US', landing_page == 'old_page')
> table(US_old$group,US_old$converted)

                0     1
  control   89446 12270
  treatment  1216   185
> fisher.test(matrix(c(12270,89446,185,1216),ncol=2) , alternative= 'less')

        Fisher's Exact Test for Count Data

data:  matrix(c(12270, 89446, 185, 1216), ncol = 2)
p-value = 0.1046
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.00000 1.03235
sample estimates:
odds ratio
 0.9016943
```

New Page

Null Hypothesis: Conversions of control and treatment are

[Code]

```
> US_new= filter(data_set, country == 'US', landing_page == 'new_page')
> table(US_new$group,US_new$converted)

                0     1
  control    1189   154
  treatment 89832 12072
> fisher.test(matrix(c(154,1189,12072,89832),ncol=2) , alternative= 'less')

        Fisher's Exact Test for Count Data

data:  matrix(c(154, 1189, 12072, 89832), ncol = 2)
p-value = 0.3534
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.00000 1.11272
sample estimates:
odds ratio
 0.9638063
```

Conclusion:

Observing both the p-values for old page and new page, we accept the null hypothesis for both the cases at 5% level of significance.

United Kingdom

Old page

Null Hypothesis: Conversions of control and treatment are same

[Code]

```
> UK_old= filter(data_set, country == 'UK', landing_page == 'old_page')
> table(UK_old$group,UK_old$converted)

                0     1
  control   31996  4364
  treatment   422    50
> fisher.test(matrix(c(4364,31996,50,422),ncol=2) , alternative= 'less')

        Fisher's Exact Test for Count Data

data:  matrix(c(4364, 31996, 50, 422), ncol = 2)
p-value = 0.8434
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.00000 1.50001
sample estimates:
odds ratio
  1.151182
```

New page

Null Hypothesis: Conversions of control and treatment are same

[Code]

```
> UK_new= filter(data_set, country == 'UK', landing_page == 'new_page')
> table(UK_new$group,UK_new$converted)

                0     1
  control     417    64
  treatment 31731  4375
> fisher.test(matrix(c(64,417,4375,31731),ncol=2) , alternative= 'less')

        Fisher's Exact Test for Count Data

data:  matrix(c(64, 417, 4375, 31731), ncol = 2)
p-value = 0.8073
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.00000 1.39688
sample estimates:
odds ratio
  1.113104
```

Conclusion

Observing both the p-values for old page and new page, we accept the null hypothesis for both the cases at 5% level of significance.

Canada

Old page

Null Hypothesis: Conversions of control and treatment are same

[Code]

```
> CA_old= filter(data_set, country == 'CA', landing_page == 'old_page')
> table(CA_old$group,CA_old$converted)
```

```
              0    1
  control   6343  855
  treatment   77   15
> fisher.test(matrix(c(855,6343,15,77),ncol=2), alternative = 'less')

        Fisher's Exact Test for Count Data

data:  matrix(c(855, 6343, 15, 77), ncol = 2)
p-value = 0.1289
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.175385
sample estimates:
odds ratio
 0.6919866
```

New Page

Null Hypothesis: Conversions of control and treatment are same

[Code]

```
> CA_new= filter(data_set, country == 'CA', landing_page == 'new_page')
> table(CA_new$group,CA_new$converted)

              0    1
  control     88   16
  treatment 6484  817
> fisher.test(matrix(c(16,88,817,6484),ncol=2) , alternative= 'less')

        Fisher's Exact Test for Count Data

data:  matrix(c(16, 88, 817, 6484), ncol = 2)
p-value = 0.9281
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 2.305677
sample estimates:
odds ratio
  1.442885
```

Conclusion:

Observing both the p-values for old page and new page, we accept the null hypothesis for both the cases at 5% level of significance.

For the entire data set

It is conclusive from the above sections that the conversions of the respondent customers from the New Page and Old Page are the same. Let us look at the same conversion for the entire data set.

Old Page

Null Hypothesis: Conversions of control and treatment are same for all the countries at different timestamp.

[Code]

```
> old_page= filter(data_set, landing_page == 'old_page')
> table(old_page$group,old_page$converted)
```

```
                0      1
  control   127785  17489
  treatment   1715    250

> fisher.test(matrix(c(17489,127785,250,1715),ncol=2) , alternative= 'less')


Fisher's Exact Test for Count Data

data:  matrix(c(17489, 127785, 250, 1715), ncol = 2)
p-value = 0.1861
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.053837
sample estimates:
odds ratio
 0.9388623
```

New Page

Null Hypothesis: Conversions of control and treatment are same for all the countries at different timestamp.

[Code]

```
> new_page= filter(data_set, landing_page == 'new_page')
> table(new_page$group,new_page$converted)

                0      1
  control     1694    234
  treatment 128047  17264
> fisher.test(matrix(c(234,1649,17264,128047),ncol=2) , alternative= 'less')


Fisher's Exact Test for Count Data

data:  matrix(c(234, 1649, 17264, 128047), ncol = 2)
p-value = 0.7785
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.183106
sample estimates:
odds ratio
  1.052515
```

Conclusion:

From the above p-values we accept the null hypothesis at 5% level of significance or in other word the conversions of new page and old page for all the countries combined at different timestamps are the same.

Note: **In the above cases Fisher's Exact test has been used instead of chi-square test. In case of large samples (sample size greater than 5 or 10) both the test gives almost the same result.**

Measures that can likely be taken by the e-commerce website

1. The company may run the experiment longer and try to get more responses if that helps the outcome results.

2. The company may try with a different alternative treatment or improve upon their selection of treatments to test the response accordingly in the new page as well as the old page.

3. If at all the company wants to take decisions based on the above experimental analysis then it must rely upon factors like cost and development functionalities of the landing page. The company should take a decision that minimizes the technical aspect.

4. As we observed the respondents are maximum from a particular country. In this type of cases the company can figure out some way to stratify the population of customers attending the experiment so that it can get a better response about conversion.