King - Per Marcus

Per Marcus

November 7, 2018

Assignment

Setup

The aim of this assignment has been to show if a diffence exists between two testgroups, `a` and `b`. The groups consists of users of a game where 80 % of the users belong to group `a` and 20 % belong to group `b`.

Two data sets have been used for the assignment, `assignment` and `activity`. Both have been called from Googles cloud services using R and the package `BigRQuery`. Since the amount of data was large, a sample of 1000 users have been used. These 1000 users was randomly selected from the `assignment` table and then all data connected to these users have been called from the `activity` table. These 1000 unique users had a total of 24646 different observations.

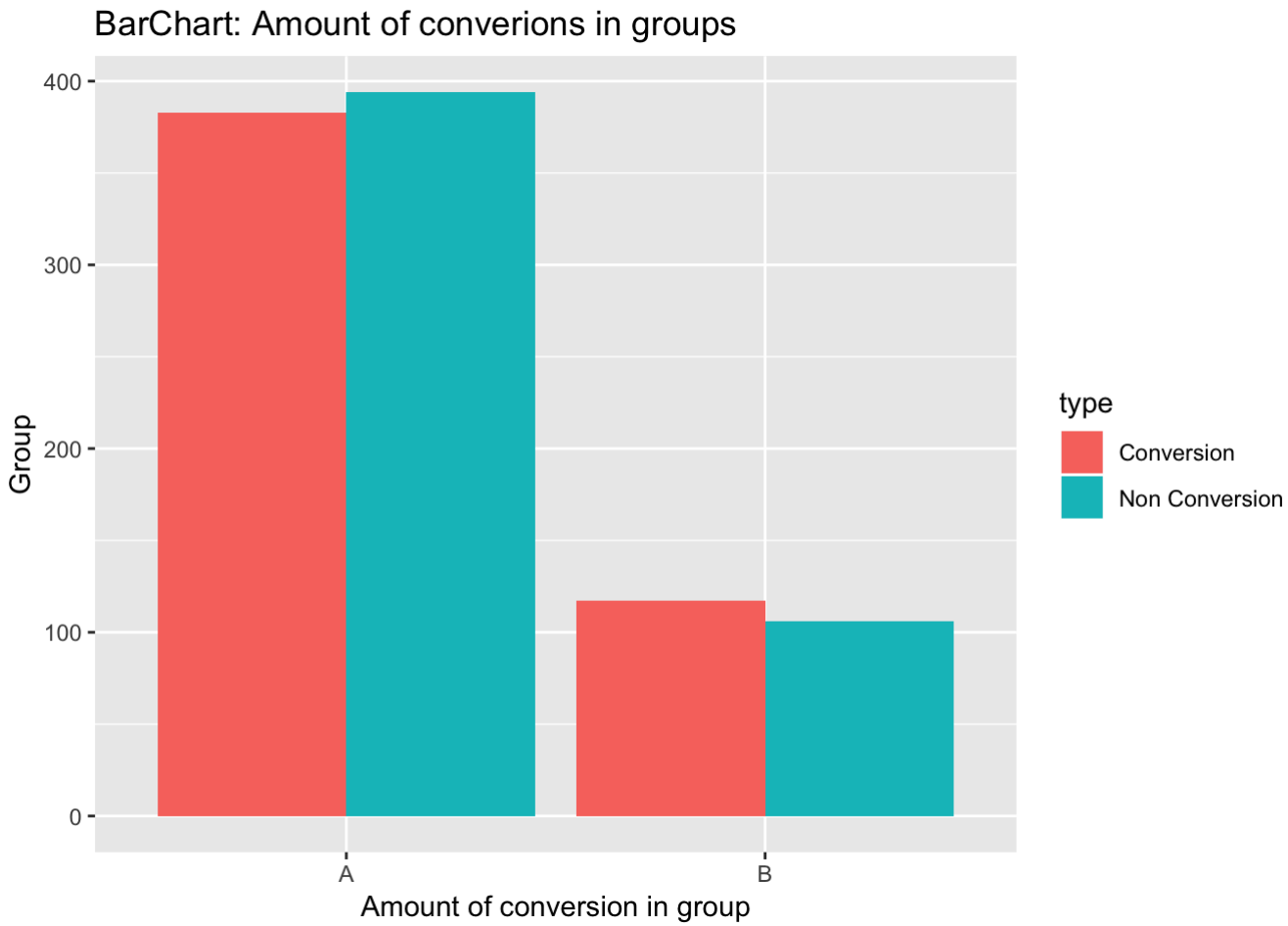
The assignment will be split in two parts. The first part will handle simple statistics to understand if conversion and puchases are larger in group A or B, both in proportion and in frequency. I will also present plots on the distribution of `gameends`, `purchases` and `total_puchases` of set time period.

The second part will present modelling and test if one can estimate the probability if players in each group will `convert` or not.

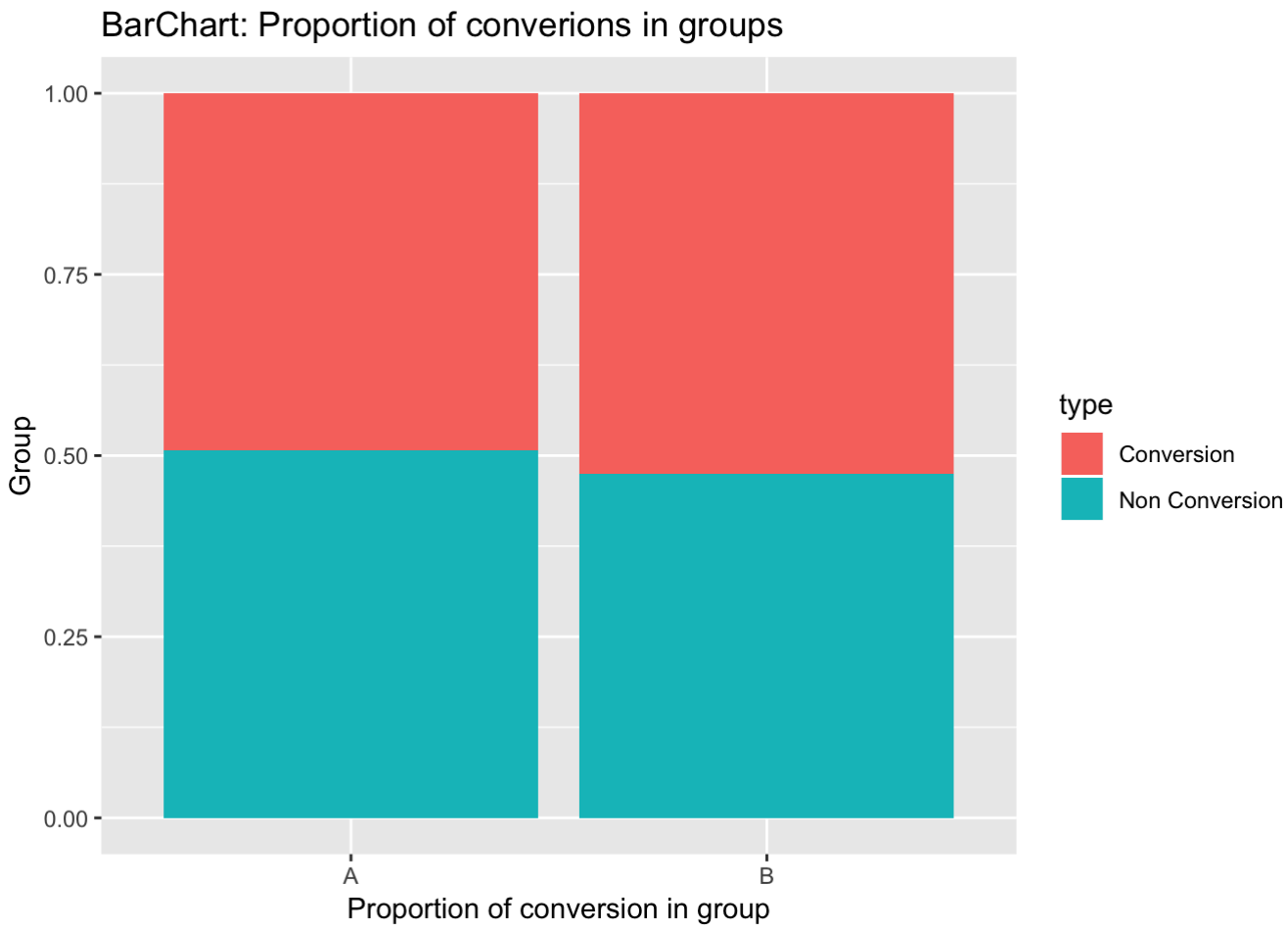
Simple statistcs

Following section will present the plots and statistics. Plots will be presented to visualize the results and statistics will be presented to give inference.

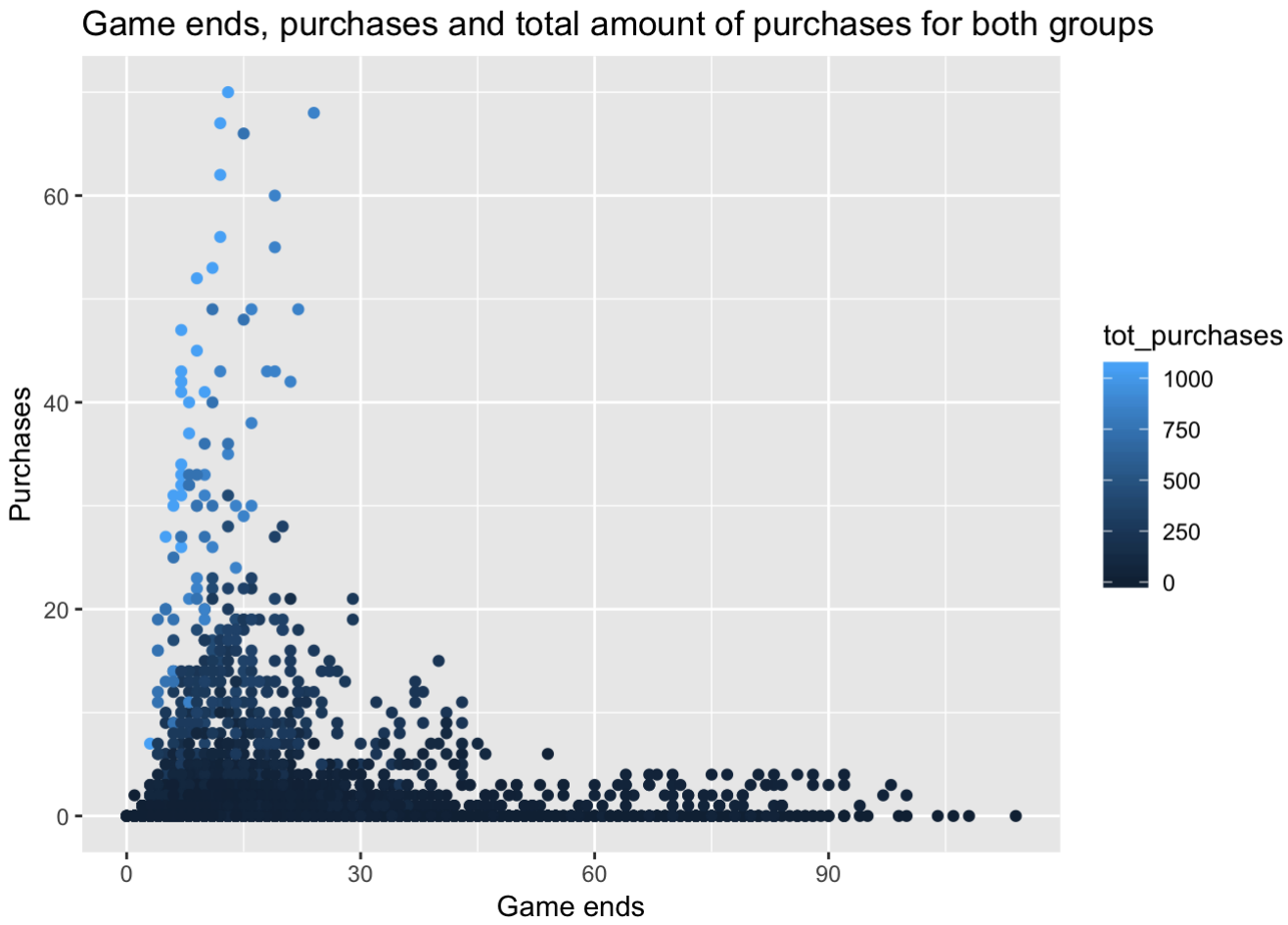
Plots/Visualisation



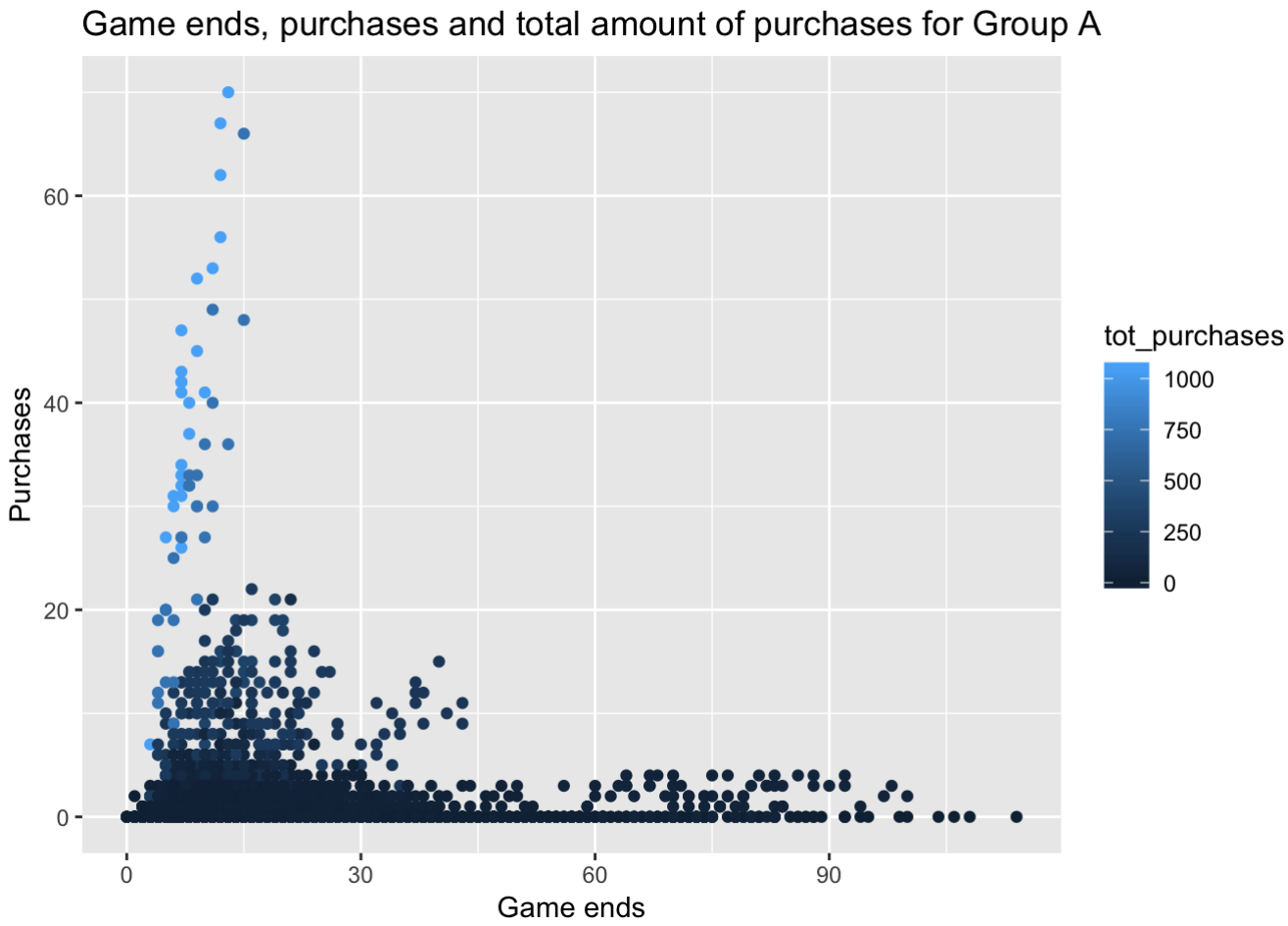
Amount of conversion in groups implies that it is more common to convert in group B than in group A.



Proportion of conversion in groups show that it is more common to convert in group B than in group A.



The plot above shows that purchases are more common for players whom play between 0 and 30 `gameends` per day. It is also these players whom use the most `purchases` in total. It also seems as if the linear relationship in this data, `gameends` between 0 and 30, is stronger than if one would look at the data as a whole. This group would be interesting to take a closer look at.



The behaviour of group A is similar to that of the total, which is to be expected since group A contains 80 % of the data.

Behaviour of group B seems to slightly differ in linear relationship from group A, where `gameends` 0 to 25 seems so give slightly more purchases.

Statistics

To check if the groups differ simple t.tests have been performed. Group A is defined by X and group B is defined by Y.

Difference in `conversion` :

```
##
##  Welch Two Sample t-test
##
## data:  group_a$conversion and group_b$conversion
## t = -5.1619, df = 8908.5, p-value = 2.498e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05440771 -0.02445824
## sample estimates:
## mean of x mean of y
## 0.4914874 0.5309203
```

Difference in `purchases` :

```
##
##  Welch Two Sample t-test
##
## data:  group_a$purchases and group_b$purchases
## t = -5.3735, df = 7218.7, p-value = 7.963e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3436951 -0.1599589
## sample estimates:
## mean of x mean of y
## 0.4957176 0.7475446
```

With statistical certainty, we can say that group B has a higher conversion rate and a higher purchase rate.

Results from statistics

The results tell us that, using the gameversion which was used for group B will be better than using the one for group A.

Modelling

The aim of my study in the modelling part will be to see if we can estimte `conversion` based on the data that exists within the activity data table. `Conversion` is set if a user has ever done a purchase within game.

For modelling purposes the data table was split in two parts, training and validation/test.

Concerns about data for the modelling

Before I started modelling the data I realized that some of the analysis which I considered to do was not possible. I thought it would be interesting to see if the total number of played games had impact on `conversion` , but the table does not contain all data from `installation_date` to first `activity_date` . This means that this analysis is not possible.

To get around this problem I choose to look at the total amount of `gameends` per player during the set time frame to see if that had any impact on `conversion` .

Modelling groups

The modelling was done for both of the groups, one modell for group A and one modell for group B. I applied a simple GLM-logit model to estimate the probability of `conversion` for players based on `total_gameends`, where `total_gameends` is the total amount of `gameends` during the time of the study.

The aim of the models is to see if players in group A or group B have a higher probability to `convert`.

Validation of models

The validation on modelling has been done by RMSE, accuracy and confussion matrix. Results from the two models are as follows.

Results of model for Group A

RMSE:

```
## [1] NaN
```

Confusion matrix:

```
##           Reference
## Prediction  0  1
##          0 54 70
##          1 23 15
```

Error estimation:

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   4.259259e-01  -1.188178e-01   3.486889e-01   5.059237e-01   5.246914e-01
## AccuracyPValue  McnemarPValue
##   9.952860e-01   1.842462e-06
```

Results of model for group B

RMSE:

```
## [1] NaN
```

Confusion matrix:

```
##           Reference
## Prediction  0  1
##          0  4  4
##          1 14 16
```

Error estimation:

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##     0.52631579     0.02285714     0.35818344     0.69019270     0.52631579
## AccuracyPValue  McnemarPValue
##     0.56548616     0.03389485
```

It is clear that the models are not accurat. The RMSE for both models are high, meaning that the volatility in the models are high. The accuracy, which is based on the confussion matrix is ok, but the threshold for the confussion matrix is manually set to 0.5, why both accuracy and confussion matrix may not give a true picture.

Becasue of this it is my recommendation to add data to the model for a more accurat results in regards to estimation of `conversion`.

Overall

Game version for group B shows better results. For this reason, it is recommended to use version B.

For further analysis it would be interesting to check if the average number of `gameends` per player differed between the two groups. I would also like to check if the number of `purchases` was different depending on how many `gameends` the players did in each group. This was shortly mentioned in the plots section. In that section we can see visually that players with 0 to 30 `gameends` have a higher tendency for `conversion` and `purchases`. Splitting the data in different groups based on `gameends` and creating inference on these individual groups could give an understanding in what groups would be better to target in regards to `converion` and `purchases`.

Regarding modelling more data or a different setup, such as the previously mentioned data split, is needed to better estimate the probability of `conversion`.

Reagaring what was tested, I quess it is the number of games a player can play for free per day.

//Per Marcus