# Project_A_B_Test_MuccleHub

August 16, 2022

## 1 Capstone Project 1: MuscleHub AB Test


- Name: Petr Vlasak
- Email: petr.vlasakk@gmail.com
- GitHub: https://github.com/pvlasak


---

In this project a special Codecademy library is used which allows to type SQL Commands directly into Jupyter Notebook. Function is called sql_query and takes SQL query as an argument. Each query will return a Pandas DataFrame.

## 2 Initial Check of Data Sets

The sign-up process for MuscleHub has several steps: 1. Take a fitness test with a personal trainer (only Group A) 2. Fill out an application for the gym 3. Send in their payment for their first month's membership

### 2.1 Available Data sets

Suppose that two groups of customers are defined based on the fact, if they took a fitness test or not. - Group A: customers, who were given a fitness test - Group B: customer, who were not given a fitness test

SQLite database contains several tables: - `visits` contains information about potential gym customers who have visited MuscleHub - `fitness_tests` contains information about potential customers in "Group A", who were given a fitness test - `applications` contains information about any potential customers (both "Group A" and "Group B") who filled out an application. Not everyone in `visits` will have filled out an application. - `purchases` contains information about customers who purchased a membership to MuscleHub.

### 2.2 Visulalisation of "visits" data set:

| | index | first_name | last_name | email | gender |
|---|---|---|---|---|---|
| 0 | 0 | Karen | Manning | Karen.Manning@gmail.com | female |
| 1 | 1 | Annette | Boone | AB9982@gmail.com | female |
| 2 | 2 | Salvador | Merritt | SalvadorMerritt12@outlook.com | male |
| 3 | 3 | Martha | Maxwell | Martha.Maxwell@gmail.com | female |

```
4      4      Andre     Mayer         AndreMayer90@gmail.com    male

  visit_date
0     5-1-17
1     5-1-17
2     5-1-17
3     5-1-17
4     5-1-17
```

### 2.3 Visulalisation of "fitness_tests" data set:

```
   index first_name last_name                  email  gender  \
0      0        Kim    Walter   KimWalter58@gmail.com  female
1      1        Tom   Webster        TW3857@gmail.com    male
2      2     Marcus     Bauer  Marcus.Bauer@gmail.com    male
3      3    Roberta      Best      RB6305@hotmail.com  female
4      4     Carrie   Francis      CF1896@hotmail.com  female

  fitness_test_date
0        2017-07-03
1        2017-07-02
2        2017-07-01
3        2017-07-02
4        2017-07-05
```

### 2.4 Visulalisation of "applications" data set:

```
   index first_name last_name                    email  gender  \
0      0        Roy    Abbott    RoyAbbott32@gmail.com    male
1      1      Agnes   Acevedo  AgnesAcevedo1@gmail.com  female
2      2    Roberta   Acevedo         RA8063@gmail.com  female
3      3     Darren    Acosta  DAcosta1996@hotmail.com    male
4      4     Vernon    Acosta    VAcosta1975@gmail.com    male

  application_date
0       2017-08-12
1       2017-09-29
2       2017-09-15
3       2017-07-26
4       2017-07-14
```

### 2.5 Visulalisation of "purchases" data set:

```
   index first_name last_name                    email  gender purchase_date
0      0        Roy    Abbott    RoyAbbott32@gmail.com    male    2017-08-18
1      1    Roberta   Acevedo         RA8063@gmail.com  female    2017-09-16
2      2     Vernon    Acosta    VAcosta1975@gmail.com    male    2017-07-20
```

```
3      3     Darren    Acosta  DAcosta1996@hotmail.com    male    2017-07-27
4      4       Dawn    Adkins    Dawn.Adkins@gmail.com  female    2017-08-24
```

### 2.6 Preparation of Pandas Dataframe

Data have been combined by sequence of SQL commands, joined based on "first_name", "last_name" and "email" and only entries after date of 7-1-17 have been selected.

All data are save as Pandas Dataframe named "df".

```
  first_name last_name  gender                   email visit_date  \
0        Kim    Walter  female   KimWalter58@gmail.com     7-1-17
1        Tom   Webster    male        TW3857@gmail.com     7-1-17
2     Edward     Bowen    male  Edward.Bowen@gmail.com     7-1-17
3     Marcus     Bauer    male  Marcus.Bauer@gmail.com     7-1-17
4    Roberta      Best  female      RB6305@hotmail.com     7-1-17

  fitness_test_date application_date purchase_date
0        2017-07-03             None          None
1        2017-07-02             None          None
2              None       2017-07-04    2017-07-04
3        2017-07-01       2017-07-03    2017-07-05
4        2017-07-02             None          None
```

## 3 Hypotesis Testing

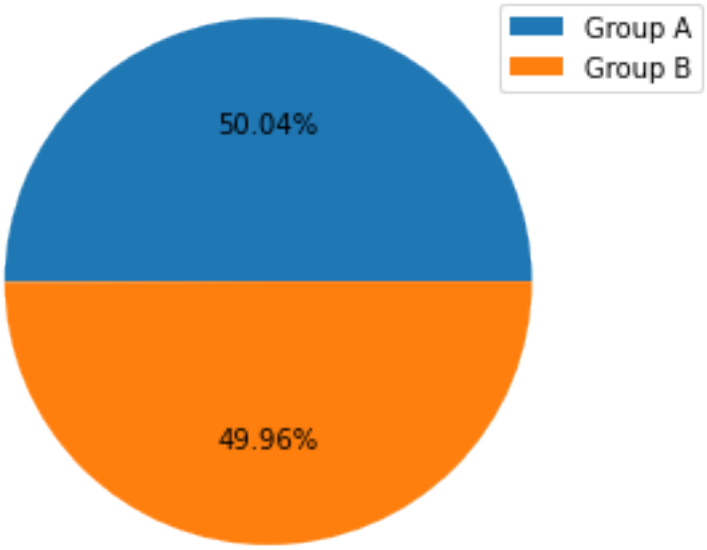
### 3.1 Investigation of the A and B groups

New column "ab_test_group" is created that contains value `A` if value in `fitness_test_date` is not `None` and `B` if is `None`.

Number of users for both groups A and B is counted:

```
  ab_test_group  count_of_users
0             A            2504
1             B            2500
```

Number of user for each group is visualised like a pie chart:




**Number of users in each group is almost equal**

### 3.2 Who picks up an application?

New column in `df` called `is_application` which is `Application` if `application_date` is not `None` and `No Application`, otherwise.

New DataFrame named as `app_counts` is saved that contains the amount of users that either do or don't pick up application in each group A and B.

A new column called `Total`, which is the sum of `Application` and `No Application` is defined.

A new column called `Percent with Application`, which is equal to `Application` divided by `Total` is defined.

```
is_application ab_test_group  Application  No Application  Total  \
0                           A          250            2254   2504
1                           B          325            2175   2500

is_application  Percent with Application
0                                0.09984
1                                0.13000
```

Higher percentage of customers in group B turned in an application.

#### 3.2.1 Chi Square Test

This type of test can help us to identify if the probabily of picking up an application depends on categorical variable - Group A or B.

`P-value for conducted Chi Square test is: 0.00096`

P-value is below a threshold of 0.05 and null hypotesis can be rejeted. There is a statistically significant association between the given fitness test and filled out application. E.g. two groups A and B appear to have different probabilities of filling out an application.

**I can conclude that the customers from B-group are more likely to pick up an application**

### 3.3 Who purchases a membership from those who picked up application?

A column to `df` called `is_member` which is `Member` if `purchase_date` is not `None`, and `Not Member` otherwise.

A new DataFrame called `just_apps` the contains only people who picked up an application is created.

```
   first_name last_name  gender                    email visit_date  \
2      Edward     Bowen    male   Edward.Bowen@gmail.com     7-1-17
3      Marcus     Bauer    male   Marcus.Bauer@gmail.com     7-1-17
9    Salvador  Cardenas    male  SCardenas1980@gmail.com     7-1-17
11    Valerie     Munoz  female     VMunoz1998@gmail.com     7-1-17
35    Michael     Burks    male         MB9820@gmail.com     7-1-17

   fitness_test_date application_date purchase_date ab_test_group  \
2               None       2017-07-04    2017-07-04             B
3         2017-07-01       2017-07-03    2017-07-05             A
9         2017-07-07       2017-07-06          None             A
11        2017-07-03       2017-07-05    2017-07-06             A
35              None       2017-07-07    2017-07-13             B

   is_application   is_member
2     Application      Member
3     Application      Member
9     Application  Not Member
11    Application      Member
35    Application      Member
```

New dataframe includes number of customers that purchased or not purchased membership after they picked up an application. It shows total number of customer and percentage of membership purchase for both groups.

```
is_member ab_test_group  Member  Not Member  Total  Percent Purchase
0                     A     200          50    250          0.800000
1                     B     250          75    325          0.769231
```

Percentage value shows that the customers from Group A are more likely to purchase membership if they pick up an application. Another Chi-Square Test is important to be conducted to check if the difference in purchase for those two groups is statistically significant.

### 3.3.1 Chi Square Test

```
P-value for conducted Chi Square test is: 0.43259
```

The p-value is higher than 0.05 and the null hypotesis therefore can't be rejected. It means there is statistically no significant association between the fitness test taken (group A and group B) and the purchase of membership **if application is picked up**.

Therefore the difference found previously is not significant and there is not high probability that people from group A will purchase membership than people from group B if application is already picked up.

## 3.4 Who purchases a membership from all visitors?

New dataframe takes into account all customers and counts the amount of membership purchase for both group, including percentage value.

| is_member | ab_test_group | Member | Not Member | Total | Percent Purchase |
|---|---|---|---|---|---|
| 0 | A | 200 | 2304 | 2504 | 0.079872 |
| 1 | B | 250 | 2250 | 2500 | 0.100000 |

Previously, only people who had **already picked up an application** were considered, it shows there was no significant difference in membership between Group A and Group B.

Now, all people who **visit MuscleHub** are considered, there might be a significant difference in purchase of membership between Group A and Group B.

### 3.4.1 Chi Square Test

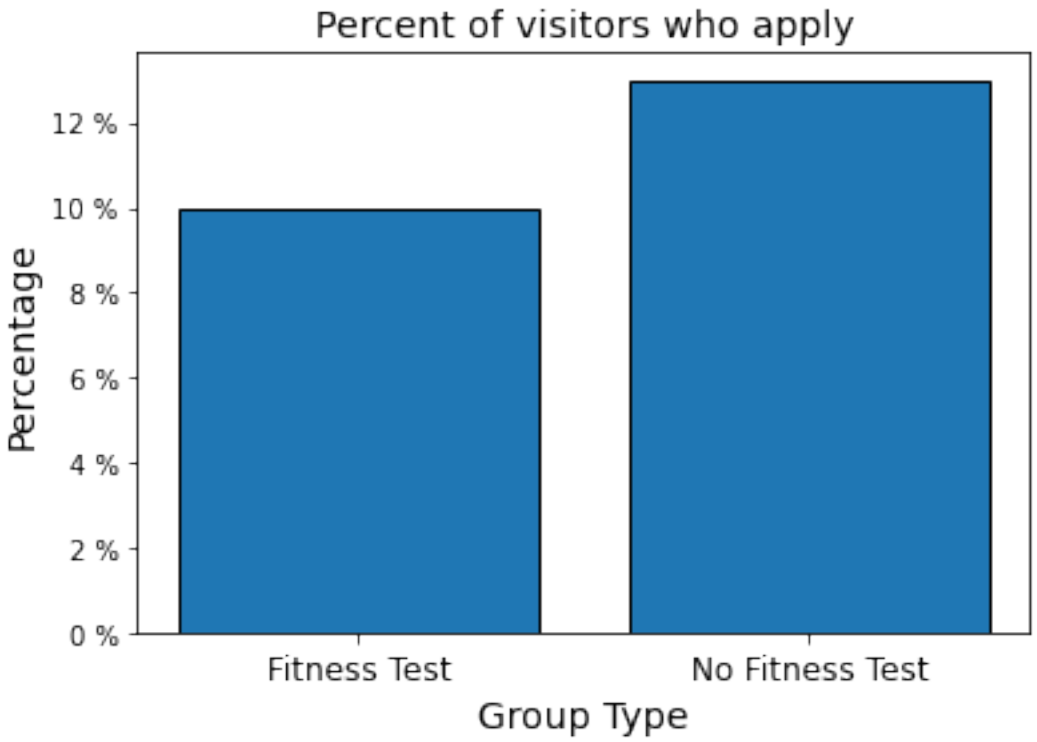
```
P-value for conducted Chi Square test is: 0.01472
```

Based on resultant p-value the null hypotesis has to be rejected because there is a statistically significant difference in purchased membership for those two groups A and B (there is a significant association between taken fitness test and membership purchase)

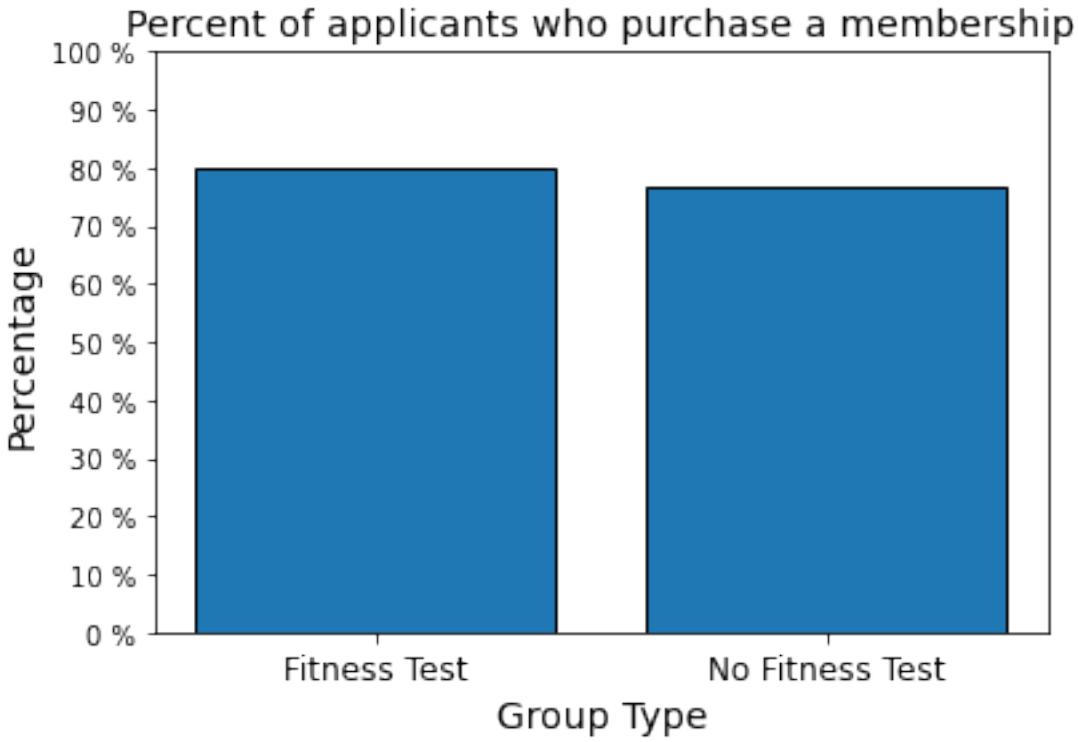
**Customers from Group B are more likely to purchase membership**

# 4 Visualisation of difference between Group A and B

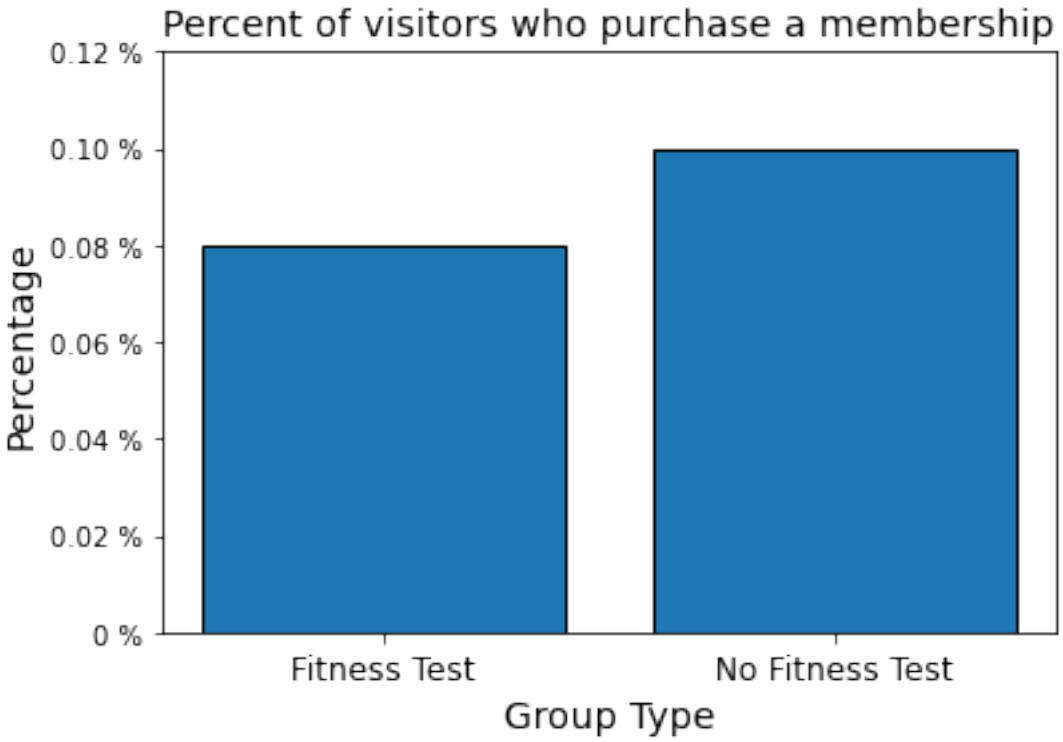
## 4.1 Percent of visitors who apply



### 4.2 Percent of applicants who purchase a membership



### 4.3 Percent of visitors who purchase a membership




## 5 Conclusion

1. Bar chart plots show that the people from category B are more likely to fill out an application and purchase membership.
2. The difference described in previous point is found to be statistically significant, which means there is a significant association between categorical value and percentage of picked up applications and purchased memberships.
3. The difference between percentage of purchased memberships is not statistically significant if only customers with filled out application are taken into account.

**4. Taking a fitness test with a personal trainer can reduce the amount of customers who finally purchase a membership. Therefore is recommended to skip the first step in the application process and offer a fitness test as an additional service to MuscleHub customers**