Captone - MuscleHub fun_with_a/b_tests

Patrik Liba

Initial environment, data manipulations with SQL

- Preparing the data for general A/B testing
- Df have 5004 rows and 6 columns

1. Not all visits in `visits` occurred during the A/B test. You'll only want to pull data where `visit_date` is on or after `7-1-17`.
2. You'll want to perform a series of `LEFT JOIN` commands to combine the four tables that we care about. You'll need to perform the joins on `first_name`, `last_name`, and `email`. Pull the following columns:

- `visits.first_name`
- `visits.last_name`
- `visits.gender`
- `visits.email`
- `visits.visit_date`
- `fitness_tests.fitness_test_date`
- `applications.application_date`
- `purchases.purchase_date`

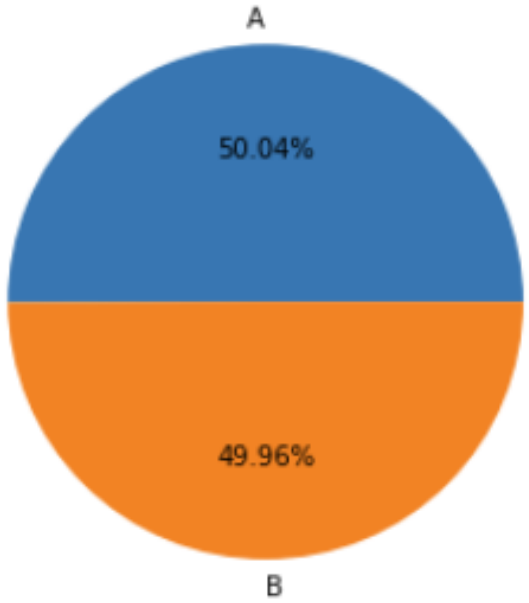
Save the result of this query to a variable called `df`.

Hint: your result should have 5004 rows. Does it? YES

```
df = sql_query('''
SELECT visits.first_name,
       visits.last_name,
       visits.visit_date,
       fitness_tests.fitness_test_date,
       applications.application_date,
       purchases.purchase_date
FROM visits
LEFT JOIN fitness_tests
    ON fitness_tests.first_name = visits.first_name
    AND fitness_tests.last_name = visits.last_name
    AND fitness_tests.email = visits.email
LEFT JOIN applications
    ON applications.first_name = visits.first_name
    AND applications.last_name = visits.last_name
    AND applications.email = visits.email
LEFT JOIN purchases
    ON purchases.first_name = visits.first_name
    AND purchases.last_name = visits.last_name
    AND purchases.email = visits.email
WHERE visits.visit_date >= '7-1-17'
''')
df
```

	first_name	last_name	visit_date	fitness_test_date	application_date	purchase_date
0	Kim	Walter	7-1-17	2017-07-03	None	None
1	Tom	Webster	7-1-17	2017-07-02	None	None

```
plt.pie(ab_counts.first_name.values, labels=['A', 'B'], autopct='%0.2f%%'
plt.axis('equal')
plt.show()
plt.savefig('ab_test_pie_chart.png')
```



Group A and Group B; 50:50 plot

- The initial grouping of the people into two groups which appears to be 50:50 as Jennet likes.
- Throughout this presentation, we will hit the specifically tests with added code
- Presentation ends with managerial critical point of view

More people turned it into application from Group B

- Tendency of Group B to turn it into application, why:
- Trainer at 1st approach did motivate them better
- More courageous individuals
- Habits of people from Group B

```
# civ
app_pivot = app_counts.pivot(columns='is_application',
                             index='ab_test_group',
                             values='first_name').reset_index()
print(app_pivot)
```

```
is_application ab_test_group  Application  No Application
0                           A          250            2254
1                           B          325            2175
```

Define a new column called `Total`, which is the sum of `Application` and `No Application`.

```
# assuming in pivot
app_pivot['Total'] = app_pivot.Application + app_pivot['No Application']
```

Calculate another column called `Percent with Application`, which is equal to `Application` divided by `Total`.

```
app_pivot['Percent with Application'] = app_pivot.Application / app_pivot.Total
print(app_pivot)
```

```
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
```

It looks like more people from Group B turned in an application. Why might that be?

We need to know if this difference is statistically significant.

Choose a hypothesis tests, import it from `scipy` and perform it. Be sure to note the p-value. Is this result significant?

```
from scipy.stats import chi2_contingency
# create matrix for it
contingency = [[250, 2254], [325, 2175]]
_, pval,_, _ = chi2_contingency(contingency) # chi2 is the right one, because we have cols of Interface A vs. Interface B
print(pval) # significant difference
```

```
0.0009647827600722304
```

Percent of people who picked up applications purchased memberships

- When people took fit_test they were inclined to purchase a membership. Due to ->
- Motivation by the trainers
- Courage of an individual
- Habits
- Body weight
- High blood pressure

Great! Now, let's do a `groupby` to find out how many people in `just_apps` are and aren't members from each group. Follow the same process that we did in Step 4, including pivoting the data. You should end up with a DataFrame that looks like this:

is_member	ab_test_group	Member	Not Member	Total	Percent Purchase
0	A	?	?	?	?
1	B	?	?	?	?

Save your final DataFrame as `member_pivot`.

```
]: member_count = just_apps.groupby(['ab_test_group', 'is_member']).first_name.count().reset_index()
member_pivot = member_count.pivot(columns='is_member',
                                  index='ab_test_group',
                                  values='first_name').reset_index()

member_pivot['Total'] = member_pivot.Member + member_pivot['Not Member']
member_pivot['Percent Purchase'] = member_pivot.Member / member_pivot.Total
member_pivot
```

]:

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

It looks like people who took the fitness test were more likely to purchase a membership **if** they picked up an application. Why might that be?

Just like before, we need to know if this difference is statistically significant. Choose a hypothesis tests, import it from `scipy` and perform it. Be sure to note the p-value. Is this result significant?

```
]: contingency = [[200, 50], [250, 75]]
_, pval,_, _ = chi2_contingency(contingency)
print(pval) # the value is NOT significant > 0.05
```

```
0.43258646051083327
```

Previously, we looked at what percent of people **who picked up applications** purchased memberships. What we really care about is what percentage of **all visitors** purchased memberships. Return to `df` and do a `groupby` to find out how many people in `df` are and aren't members from each group. Follow the same process that we did in Step 4, including pivoting the data. You should end up with a DataFrame that looks like this:

Test for significant difference between Group A and Group B

- P-value of 0.014 = 1.5 %

It is less than 0.05 meaning the data didn't occour due to chance.

Previously, we looked at what percent of people **who picked up applications** purchased memberships. What we really care about is what percentage of **all visitors** purchased memberships. Return to `df` and do a `groupby` to find out how many people in `df` are and aren't members from each group. Follow the same process that we did in Step 4, including pivoting the data. You should end up with a DataFrame that looks like this:

is_member	ab_test_group	Member	Not Member	Total	Percent Purchase
0	A	?	?	?	?
1	B	?	?	?	?

Save your final DataFrame as `final_member_pivot`.

```
final_member_count = df.groupby(['ab_test_group', 'is_member']).first_name.count().reset_index()
final_member_pivot = final_member_count.pivot(columns='is_member',
                                              index='ab_test_group',
                                              values='first_name').reset_index()

final_member_pivot['Total'] = final_member_pivot.Member + final_member_pivot['Not Member']
final_member_pivot['Percent Purchase'] = final_member_pivot.Member / final_member_pivot.Total
final_member_pivot
```

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, when we only considered people who had **already picked up an application**, we saw that there was no significant difference in membership between Group A and Group B.

Now, when we consider all people who **visit MuscleHub**, we see that there might be a significant different in memberships between Group A and Group B. Perform a significance test and check.

```
contingency = [[200, 2304], [250, 2250]]
_, pval,_, _ = chi2_contingency(contingency)
print(pval) # indicates that the data didn't occour by a chance.
```

```
0.014724114645783203
```

- # Summary of interviews.
- Attracting people with no atheltic bakground is the case of a successful gym, quantity! The people who are working and they train 3 times a week to stay healthy, who doesn't push their limits, knowing they won't become next Uzain Bold. Are precious! And you need to aim the attention to.
- Designing appropriate training plans, not starving diets, but motivating them with their personal bests every trainig so they can see the progress and will stay motivated, share experiences, and live a healthy life!
- For you as a gym owner it means, you gotta work buddy!

- Let’s make you rich buddy!

- **Thanks for your attention**