Analyze_ab_test_results_notebook

March 23, 2022

1 Analyze A/B Test Results

This project will assure you have mastered the subjects covered in the statistics lessons. We have organized the current notebook into the following sections:

- Section **??**
- Section **??**
- Section **??**
- Section **??**
- Section **??**
- Section **??**

Specific programming tasks are marked with a **ToDo** tag.
## Introduction
A/B tests are very commonly performed by data analysts and data scientists. For this project, you will be working to understand the results of an A/B test run by an e-commerce website. Your goal is to work through this notebook to help the company understand if they should: - Implement the new webpage, - Keep the old webpage, or - Perhaps run the experiment longer to make their decision.

Each **ToDo** task below has an associated quiz present in the classroom. Though the classroom quizzes are **not necessary** to complete the project, they help ensure you are on the right track as you work through the project, and you can feel more confident in your final submission meeting the rubric specification.

> **Tip**: Though it's not a mandate, students can attempt the classroom quizzes to ensure statistical numeric values are calculated correctly in many cases.

## Part I - Probability
To get started, let's import our libraries.

```
In [1]: import pandas as pd
        import numpy as np
        import random
        import matplotlib.pyplot as plt
        %matplotlib inline
        #We are setting the seed to assure you get the same answers on quizzes as we set up
        random.seed(42)
```

1.0.1 ToDo 1.1

Now, read in the `ab_data.csv` data. Store it in `df`. Below is the description of the data, there are a total of 5 columns:

Data columns	Purpose	Valid values
user_id	Unique ID	Int64 values
timestamp	Time stamp when the user visited the webpage	-
group	In the current A/B experiment, the users are categorized into two broad groups. The `control` group users are expected to be served with `old_page`; and `treatment` group users are matched with the `new_page`. However, **some inaccurate rows** are present in the initial data, such as a `control` group user is matched with a `new_page`.	`['control', 'treatment']`
landing_page	It denotes whether the user visited the old or new webpage.	`['old_page', 'new_page']`
converted	It denotes whether the user decided to pay for the company's product. Here, `1` means yes, the user bought the product.	`[0, 1]`

Use your dataframe to answer the questions in Quiz 1 of the classroom.

Tip: Please save your work regularly.

a. Read in the dataset from the `ab_data.csv` file and take a look at the top few rows here:

```
In [2]: df = pd.read_csv('ab_data.csv')
        df.head(2)
```

```
Out[2]:    user_id                   timestamp    group landing_page  converted
        0   851104  2017-01-21 22:11:48.556739  control     old_page          0
        1   804228  2017-01-12 08:01:45.159739  control     old_page          0
```

b. Use the cell below to find the number of rows in the dataset.

```
In [3]: df.shape
```

```
Out[3]: (294478, 5)
```

c. The number of unique users in the dataset.

```
In [4]: df['user_id'].nunique()
```

```
Out[4]: 290584
```

d. The proportion of users converted.

```
In [5]: proportion_user_converted =df.query('converted == 1').user_id.nunique() / df.user_id.nun
        proportion_user_converted
```

```
Out[5]: 0.12104245244060237
```

e. The number of times when the "group" is `treatment` but "landing_page" is not a `new_page`.

```
In [6]: df.query('group == "treatment" and  landing_page == "new_page"').shape[0]
```

```
Out[6]: 145311
```

```
In [7]: df[(df['group']=='treatment') & (df['landing_page'] == 'old_page')].shape[0]
```

```
Out[7]: 1965
```

```
In [8]: df.landing_page.value_counts()
```

```
Out[8]: old_page    147239
        new_page    147239
        Name: landing_page, dtype: int64
```

f. Do any of the rows have missing values?

```
In [9]: df.isna().sum()
```

```
Out[9]: user_id         0
        timestamp       0
        group           0
        landing_page    0
        converted       0
        dtype: int64
```

1.0.2 ToDo 1.2

In a particular row, the **group** and **landing_page** columns should have either of the following acceptable values:

user_id	timestamp	group	landing_page	converted
XXXX	XXXX	`control`	`old_page`	X
XXXX	XXXX	`treatment`	`new_page`	X

It means, the `control` group users should match with `old_page`; and `treatment` group users should matched with the `new_page`.

However, for the rows where `treatment` does not match with `new_page` or `control` does not match with `old_page`, we cannot be sure if such rows truly received the new or old wepage.

Use **Quiz 2** in the classroom to figure out how should we handle the rows where the group and landing_page columns don't match?

a. Now use the answer to the quiz to create a new dataset that meets the specifications from the quiz. Store your new dataframe in **df2**.

```
In [10]: # Remove the inaccurate rows, and store the result in a new dataframe df2
         df_control =df.query('group == "control" and landing_page == "old_page"')
         df_treatment  =df.query('group == "treatment" and landing_page == "new_page"')
         frame = [df_control , df_treatment]
         df2= pd.concat(frame)
         df2.head()
```

```
Out[10]:    user_id                   timestamp    group landing_page  converted
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
         1   804228  2017-01-12 08:01:45.159739  control     old_page          0
         4   864975  2017-01-21 01:52:26.210827  control     old_page          1
         5   936923  2017-01-10 15:20:49.083499  control     old_page          0
         7   719014  2017-01-17 01:48:29.539573  control     old_page          0
```

```
In [11]: # Double Check all of the incorrect rows were removed from df2 -
         # Output of the statement below should be 0
         df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) == False].sh
```

```
Out[11]: 0
```

1.0.3 ToDo 1.3

Use **df2** and the cells below to answer questions for **Quiz 3** in the classroom.

a. How many unique **user_id**s are in **df2**?

```
In [12]: df2.user_id.nunique() , df2.shape[0]
```

```
Out[12]: (290584, 290585)
```

```
In [13]: df2.duplicated(subset=['user_id']).sum()
```

```
Out[13]: 1
```

b. There is one **user_id** repeated in **df2**. What is it?

```
In [14]: df2[df2.duplicated(subset=['user_id'])].user_id
```

```
Out[14]: 2893    773192
         Name: user_id, dtype: int64
```

c. Display the rows for the duplicate **user_id**?

```
In [15]: df2[df2.duplicated(subset=['user_id'])]
```

```
Out[15]:       user_id                   timestamp      group landing_page  converted
         2893   773192  2017-01-14 02:55:59.590927  treatment     new_page          0
```

d. Remove **one** of the rows with a duplicate **user_id**, from the **df2** dataframe.

```
In [16]: # Remove one of the rows with a duplicate user_id..
         # Hint: The dataframe.drop_duplicates() may not work in this case because the rows with
         df2 =  df2.drop_duplicates(subset=['user_id'])
         # Check again if the row with a duplicate user_id is deleted or not
         df2.duplicated(subset=['user_id']).sum()
```

```
Out[16]: 0
```

1.0.4 ToDo 1.4

Use **df2** in the cells below to answer the quiz questions related to **Quiz 4** in the classroom.
a. What is the probability of an individual converting regardless of the page they receive?

Tip: The probability you'll compute represents the overall "converted" success rate in the population and you may call it $p_{population}$.

```
In [17]: P_population = df2.query('converted == 1').shape[0] / df2['converted'].shape[0]
         P_population
```

```
Out[17]: 0.11959708724499628
```

b. Given that an individual was in the `control` group, what is the probability they converted?

```
In [18]: P_con_converted =df2.query('group =="control" and converted == 1').shape[0]/df2.query('
         P_con_converted
```

```
Out[18]: 0.1203863045004612
```

c. Given that an individual was in the `treatment` group, what is the probability they converted?

```
In [19]: P_treatment_converted =df2.query('group =="treatment" and converted == 1').shape[0]/df2
         P_treatment_converted
```

```
Out[19]: 0.11880806551510564
```

Tip: The probabilities you've computed in the points (b). and (c). above can also be treated as conversion rate. Calculate the actual difference (`obs_diff`) between the conversion rates for the two groups. You will need that later.

```
In [20]: # Calculate the actual difference (obs_diff) between the conversion rates for the two g
         obs_diff = P_con_converted - P_treatment_converted
         obs_diff

Out[20]: 0.0015782389853555567
```

d. What is the probability that an individual received the new page?

```
In [21]: P_new_page = df2.query('landing_page == "new_page"').shape[0]/df2.shape[0]
         P_new_page

Out[21]: 0.5000619442226688
```

e. Consider your results from parts (a) through (d) above, and explain below whether the new `treatment` group users lead to more conversions.

As we can see from the previous code that the difference in conversions is in the favor of the control group with a probability of 0.12 against a probabilty of 0.118 for the treatment group

## Part II - A/B Test
Since a timestamp is associated with each event, you could run a hypothesis test continuously as long as you observe the events.
However, then the hard questions would be: - Do you stop as soon as one page is considered significantly better than another or does it need to happen consistently for a certain amount of time?
- How long do you run to render a decision that neither page is better than another?
These questions are the difficult parts associated with A/B tests in general.

1.0.5 ToDo 2.1

For now, consider you need to make the decision just based on all the data provided.

Recall that you just calculated that the "converted" probability (or rate) for the old page is *slightly* higher than that of the new page (ToDo 1.4.c).

If you want to assume that the old page is better unless the new page proves to be definitely better at a Type I error rate of 5%, what should be your null and alternative hypotheses (H_0 and H_1)?
You can state your hypothesis in terms of words or in terms of p_{old} and p_{new}, which are the "converted" probability (or rate) for the old and new pages respectively.

H0: P_old_page >= P_new_page , H1: P_new_page > P_old_page

1.0.6 ToDo 2.2 - Null Hypothesis H_0 Testing

Under the null hypothesis H_0, assume that p_{new} and p_{old} are equal. Furthermore, assume that p_{new} and p_{old} both are equal to the **converted** success rate in the `df2` data regardless of the page. So, our assumption is:

$p_{new} = p_{old} = p_{population}$

In this section, you will:

- Simulate (bootstrap) sample data set for both groups, and compute the "converted" probability p for those samples.
- Use a sample size for each group equal to the ones in the `df2` data.
- Compute the difference in the "converted" probability for the two samples above.
- Perform the sampling distribution for the "difference in the converted probability" between the two simulated-samples over 10,000 iterations; and calculate an estimate.

Use the cells below to provide the necessary parts of this simulation. You can use **Quiz 5** in the classroom to make sure you are on the right track.

a. What is the **conversion rate** for p_{new} under the null hypothesis?

```
In [22]: P_new = df2.query('converted ==1').shape[0]/df2.shape[0]
         P_new
```

```
Out[22]: 0.11959708724499628
```

b. What is the **conversion rate** for p_{old} under the null hypothesis?

```
In [23]: P_old = df2.query('converted ==1').shape[0]/df2.shape[0]
         P_old
```

```
Out[23]: 0.11959708724499628
```

c. What is n_{new}, the number of individuals in the treatment group? *Hint*: The treatment group users are shown the new page.

```
In [24]: n_new = df2.query('group == "treatment" and landing_page == "new_page"').shape[0]
         n_new
```

```
Out[24]: 145310
```

d. What is n_{old}, the number of individuals in the control group?

```
In [25]: n_old =df2.query('group == "control" and landing_page == "old_page"').shape[0]
         n_old
```

```
Out[25]: 145274
```

```
In [26]: P_new - P_old
```

```
Out[26]: 0.0
```

```
In [27]: df2.dtypes

Out[27]: user_id          int64
         timestamp       object
         group           object
         landing_page    object
         converted        int64
         dtype: object
```

e. Simulate Sample for the `treatment` Group Simulate n_{new} transactions with a conversion rate of p_{new} under the null hypothesis. *Hint*: Use `numpy.random.choice()` method to randomly generate n_{new} number of values. Store these n_{new} 1's and 0's in the `new_page_converted` numpy array.

```
In [28]: # Simulate a Sample for the treatment Group
         treatment_g = df2[df2['group']=='treatment'].converted
         new_page_converted = np.random.choice(treatment_g, size=df2.shape[0])
         new_page_converted = np.array(new_page_converted)
```

f. Simulate Sample for the `control` Group Simulate n_{old} transactions with a conversion rate of p_{old} under the null hypothesis. Store these n_{old} 1's and 0's in the `old_page_converted` numpy array.

```
In [29]: # Simulate a Sample for the control Group
         control_g = df2[df2['group']=='control'].converted
         old_page_converted = np.random.choice(control_g, size=df2.shape[0])
         old_page_converted = np.array(old_page_converted)
```

g. Find the difference in the "converted" probability $(p'_{new} - p'_{old})$ for your simulated samples from the parts (e) and (f) above.

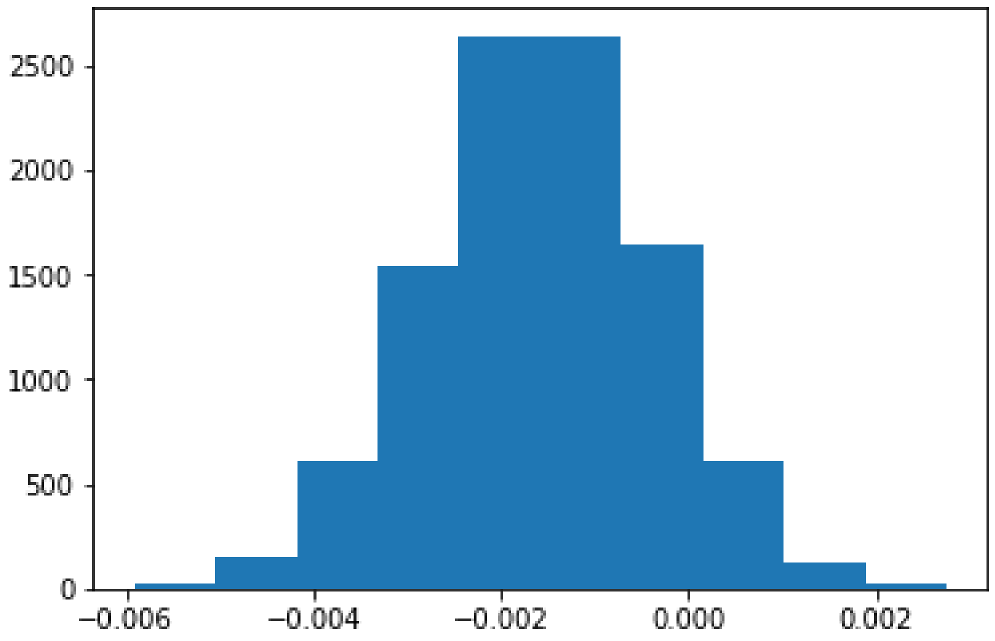
```
In [30]: new_page_converted.mean() - old_page_converted.mean()

Out[30]: -0.00027186631060209521
```

h. Sampling distribution Re-create `new_page_converted` and `old_page_converted` and find the $(p'_{new} - p'_{old})$ value 10,000 times using the same simulation process you used in parts (a) through (g) above.

Store all $(p'_{new} - p'_{old})$ values in a NumPy array called `p_diffs`.

```
In [31]: # Sampling distribution
         p_diffs = []
         for _ in range(10000):
             boot=df2.sample(df2.shape[0],replace = True)
             new_page_converted = boot.query('group =="treatment"')
             old_page_converted = boot.query('group == "control"')
             P_neww = new_page_converted.query('converted == 1').shape[0]/new_page_converted.sha
             P_oldd = old_page_converted.query('converted == 1').shape[0]/old_page_converted.sha
             p_diffs.append(P_neww - P_oldd)
```

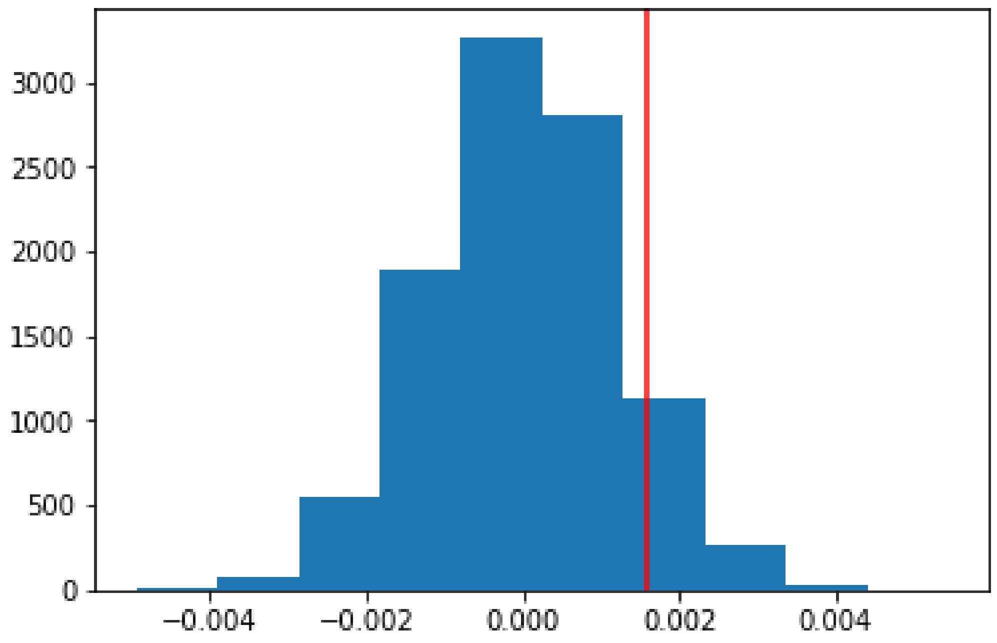
```
In [32]: plt.hist(p_diffs);
```



i. Histogram Plot a histogram of the **p_diffs**. Does this plot look like what you expected? Use the matching problem in the classroom to assure you fully understand what was computed here.
Also, use `plt.axvline()` method to mark the actual difference observed in the `df2` data (recall `obs_diff`), in the chart.

Tip: Display title, x-label, and y-label in the chart.

```
In [33]: p_diffs = np.array(p_diffs)
         null_vals =np.random.normal(0,p_diffs.std(),p_diffs.size)
         plt.hist(null_vals);
         plt.axvline(x=obs_diff,c='r');
```



j. What proportion of the **p_diffs** are greater than the actual difference observed in the `df2` data?

```
In [34]: (null_vals > obs_diff).mean()

Out[34]: 0.09410000000000003
```

k. Please explain in words what you have just computed in part **j** above.
- What is this value called in scientific studies?
- What does this value signify in terms of whether or not there is a difference between the new and old pages? *Hint*: Compare the value above with the "Type I error rate (0.05)".

> **It's called the P-value which is the probability of observing our statistics, We compare it with the alpha which is the Type one error rate, if the p-value is small we choose H1 and if its large we choose H0**

l. Using Built-in Methods for Hypothesis Testing We could also use a built-in to achieve similar results. Though using the built-in might be easier to code, the above portions are a walk-through of the ideas that are critical to correctly thinking about statistical significance.

Fill in the statements below to calculate the: - `convert_old`: number of conversions with the old_page - `convert_new`: number of conversions with the new_page - `n_old`: number of individuals who were shown the old_page - `n_new`: number of individuals who were shown the new_page

```
In [35]: df2.head()

Out[35]:    user_id                   timestamp    group landing_page  converted
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
```

```
1   804228  2017-01-12 08:01:45.159739  control     old_page          0
4   864975  2017-01-21 01:52:26.210827  control     old_page          1
5   936923  2017-01-10 15:20:49.083499  control     old_page          0
7   719014  2017-01-17 01:48:29.539573  control     old_page          0
```

```
In [36]: import statsmodels.api as sm

         # number of conversions with the old_page
         convert_old = df2.query('converted =="0"')

         # number of conversions with the new_page
         convert_new =df2.query('converted ==1')

         # number of individuals who were shown the old_page
         n_old = df2.query('converted =="0"').shape[0]

         # number of individuals who received new_page
         n_new = df2.query('converted ==1').shape[0]
```

```
/opt/conda/lib/python3.6/site-packages/statsmodels/compat/pandas.py:56: FutureWarning: The panda
  from pandas.core import datetools
```

m. Now use `sm.stats.proportions_ztest()` to compute your test statistic and p-value. Here is a helpful link on using the built in.
The syntax is:

```
proportions_ztest(count_array, nobs_array, alternative='larger')
```

where, - `count_array` = represents the number of "converted" for each group - `nobs_array` = represents the total number of observations (rows) in each group - `alternative` = choose one of the values from `[two-sided, smaller, larger]` depending upon two-tailed, left-tailed, or right-tailed respectively. >**Hint**: It's a two-tailed if you defined H_1 as $(p_{new} = p_{old})$. It's a left-tailed if you defined H_1 as $(p_{new} < p_{old})$. It's a right-tailed if you defined H_1 as $(p_{new} > p_{old})$.
The built-in function above will return the z_score, p_value.

Tip: You don't have to dive deeper into z-test for this exercise. **Try having an overview of what does z-score signify in general.**

```
In [37]: n_control =df2.query('group == "control" and converted ==1').shape[0]
         n_treatment =df2.query('group == "treatment" and converted ==1').shape[0]
         count_array = [n_control,n_treatment]
         count_array
         nobs_array =[df2.query('group=="control"').shape[0],df2.query('group =="treatment"').sh

         nobs_array , count_array

Out[37]: ([145274, 145310], [17489, 17264])
```

```
In [38]: import statsmodels.api as sm
         # ToDo: Complete the sm.stats.proportions_ztest() method arguments
         z_score, p_value = sm.stats.proportions_ztest(count_array[::-1],nobs_array[::-1],altern
         print(z_score, p_value)
```

```
-1.31092419842 0.905058312759
```

n. What do the z-score and p-value you computed in the previous question mean for the conversion rates of the old and new pages? Do they agree with the findings in parts **j.** and **k.**?

> **Tip**: Notice whether the p-value is similar to the one computed earlier. Accordingly, can you reject/fail to reject the null hypothesis? It is important to correctly interpret the test statistic and p-value.
>
> **we will fail to reject the null hypothesis because the p-value is large > 0.05**

### Part III - A regression approach

1.0.7 ToDo 3.1

In this final part, you will see that the result you achieved in the A/B test in Part II above can also be achieved by performing regression.

a. Since each row in the df2 data is either a conversion or no conversion, what type of regression should you be performing in this case?

> **logistic regression**

b. The goal is to use **statsmodels** library to fit the regression model you specified in part **a.** above to see if there is a significant difference in conversion based on the page-type a customer receives. However, you first need to create the following two columns in the df2 dataframe: 1. intercept - It should be 1 in the entire column. 2. ab_page - It's a dummy variable column, having a value 1 when an individual receives the **treatment**, otherwise 0.

```
In [39]: df2['intercept'] =1
         df2[['control','treatment']] = pd.get_dummies(df2['group'])
         df2.head()
```

```
Out[39]:    user_id                   timestamp    group landing_page  converted  \
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
         1   804228  2017-01-12 08:01:45.159739  control     old_page          0
         4   864975  2017-01-21 01:52:26.210827  control     old_page          1
         5   936923  2017-01-10 15:20:49.083499  control     old_page          0
         7   719014  2017-01-17 01:48:29.539573  control     old_page          0

            intercept  control  treatment
         0          1        1          0
         1          1        1          0
         4          1        1          0
         5          1        1          0
         7          1        1          0
```

c. Use **statsmodels** to instantiate your regression model on the two columns you created in part (b). above, then fit the model to predict whether or not an individual converts.

```
In [40]: lm = sm.Logit(df2['converted'],df2[['intercept','treatment']])
         results = lm.fit()

Optimization terminated successfully.
         Current function value: 0.366118
         Iterations 6
```

d. Provide the summary of your model below, and use it as necessary to answer the following questions.

```
In [41]: results.summary2()

Out[41]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                  Results: Logit
         =================================================================
         Model:              Logit            No. Iterations:   6.0000
         Dependent Variable: converted        Pseudo R-squared: 0.000
         Date:               2022-02-23 13:01 AIC:              212780.3502
         No. Observations:   290584           BIC:              212801.5095
         Df Model:           1                Log-Likelihood:   -1.0639e+05
         Df Residuals:       290582           LL-Null:          -1.0639e+05
         Converged:          1.0000           Scale:            1.0000
         ------------------------------------------------------------------
                       Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
         ------------------------------------------------------------------
         intercept    -1.9888    0.0081  -246.6690  0.0000  -2.0046  -1.9730
         treatment    -0.0150    0.0114    -1.3109  0.1899  -0.0374   0.0074
         =================================================================

         """
```

e. What is the p-value associated with **ab_page**? Why does it differ from the value you found in **Part II**?

Hints: - What are the null and alternative hypotheses associated with your regression model, and how do they compare to the null and alternative hypotheses in **Part II**? - You may comment on if these hypothesis (Part II vs. Part III) are one-sided or two-sided. - You may also compare the current p-value with the Type I error rate (0.05).

As we can see in Part II the p-value is equal to 0.9 and in Part III it's equal to 0.18, p-value_partII > p-value_partIII, but both of them is larger than the alpha (type one error which is equal to 0.05) p-value_partII > p-value_partIII > alpha so in that case we will fail to reject the H0 (null hypothesis).

The p-value is large so we fail to reject the null hypothesis

```
In [42]: results.pvalues
```

```
Out[42]: intercept    0.000000
         treatment    0.189886
         dtype: float64
```

The p-value is large (p-value > 0.05(alpha)) so we will fail to reject the null hypothesis

f. Now, you are considering other things that might influence whether or not an individual converts. Discuss why it is a good idea to consider other factors to add into your regression model. Are there any disadvantages to adding additional terms into your regression model?

Adding other factor will help with the precision of the model

g. Adding countries Now along with testing if the conversion rate changes for different pages, also add an effect based on which country a user lives in.

1. You will need to read in the **countries.csv** dataset and merge together your `df2` datasets on the appropriate rows. You call the resulting dataframe `df_merged`. Here are the docs for joining tables.
2. Does it appear that country had an impact on conversion? To answer this question, consider the three unique values, `['UK', 'US', 'CA']`, in the `country` column. Create dummy variables for these country columns. >**Hint:** Use `pandas.get_dummies()` to create dummy variables. **You will utilize two columns for the three dummy variables.**

Provide the statistical output as well as a written response to answer this question.

```
In [43]: # Read the countries.csv
         df3 = pd.read_csv('countries.csv')
         df3.head()
         df3['country'].unique()
```

```
Out[43]: array(['UK', 'US', 'CA'], dtype=object)
```

```
In [44]: df_merge = pd.merge(df2,df3)
         df_merge.head()
```

```
Out[44]:    user_id                   timestamp    group landing_page  converted  \
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
         1   804228  2017-01-12 08:01:45.159739  control     old_page          0
         2   864975  2017-01-21 01:52:26.210827  control     old_page          1
         3   936923  2017-01-10 15:20:49.083499  control     old_page          0
         4   719014  2017-01-17 01:48:29.539573  control     old_page          0

            intercept  control  treatment country
         0          1        1          0      US
         1          1        1          0      US
         2          1        1          0      US
         3          1        1          0      US
         4          1        1          0      US
```

```
In [45]: df_merge.head()

Out[45]:    user_id                   timestamp    group landing_page  converted  \
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
         1   804228  2017-01-12 08:01:45.159739  control     old_page          0
         2   864975  2017-01-21 01:52:26.210827  control     old_page          1
         3   936923  2017-01-10 15:20:49.083499  control     old_page          0
         4   719014  2017-01-17 01:48:29.539573  control     old_page          0

            intercept  control  treatment country
         0          1        1          0      US
         1          1        1          0      US
         2          1        1          0      US
         3          1        1          0      US
         4          1        1          0      US
```

```
In [46]: # Create the necessary dummy variables
         df_merge[['UK', 'US','CA']] = pd.get_dummies(df_merge['country'])
         df_merge = df_merge.drop(columns=['CA'])
         df_merge.head()

Out[46]:    user_id                   timestamp    group landing_page  converted  \
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
         1   804228  2017-01-12 08:01:45.159739  control     old_page          0
         2   864975  2017-01-21 01:52:26.210827  control     old_page          1
         3   936923  2017-01-10 15:20:49.083499  control     old_page          0
         4   719014  2017-01-17 01:48:29.539573  control     old_page          0

            intercept  control  treatment country  UK  US
         0          1        1          0      US   0   0
         1          1        1          0      US   0   0
         2          1        1          0      US   0   0
         3          1        1          0      US   0   0
         4          1        1          0      US   0   0
```

```
In [47]: df_merge['intercept']=1
         lo = sm.Logit(df_merge['converted'],df_merge[['intercept','treatment','UK','US']])
         results = lo.fit()
         results.summary2()

Optimization terminated successfully.
         Current function value: 0.366113
         Iterations 6


Out[47]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                   Results: Logit
         ==================================================================
Model:                Logit            No. Iterations:   6.0000
Dependent Variable:   converted        Pseudo R-squared: 0.000
Date:                 2022-02-23 13:01 AIC:              212781.1253
No. Observations:     290584           BIC:              212823.4439
Df Model:             3                Log-Likelihood:   -1.0639e+05
Df Residuals:         290580           LL-Null:          -1.0639e+05
Converged:            1.0000           Scale:            1.0000
-------------------------------------------------------------------
              Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
-------------------------------------------------------------------
intercept    -1.9893    0.0089  -223.7628  0.0000  -2.0067  -1.9718
treatment    -0.0149    0.0114    -1.3069  0.1912  -0.0374   0.0075
UK           -0.0408    0.0269    -1.5161  0.1295  -0.0934   0.0119
US            0.0099    0.0133     0.7433  0.4573  -0.0162   0.0359
==================================================================

"""
```

```
In [48]: results.pvalues
```

```
Out[48]: intercept    0.000000
         treatment    0.191245
         UK           0.129501
         US           0.457282
         dtype: float64
```

As we can see the values of the p-values are large so in this case we will fail to reject the null hypothesis. Hypothesis testing allow us to use only sample data to draw conclusions about an entire population

h. Fit your model and obtain the results Though you have now looked at the individual factors of country and page on conversion, we would now like to look at an interaction between page and country to see if are there significant effects on conversion. **Create the necessary additional columns, and fit the new model.**

Provide the summary results (statistical output), and your conclusions (written response) based on the results.

Tip: Conclusions should include both statistical reasoning, and practical reasoning for the situation.

Hints: - Look at all of p-values in the summary, and compare against the Type I error rate (0.05). - Can you reject/fail to reject the null hypotheses (regression model)? - Comment on the effect of page and country to predict the conversion.

```
In [49]: df_merge[['old_page','new_page']] = pd.get_dummies(df_merge['landing_page'])
         df_merge.head()
```

```
Out[49]:    user_id                   timestamp    group landing_page  converted  \
         0   851104  2017-01-21 22:11:48.556739  control     old_page          0
```

```
1   804228  2017-01-12 08:01:45.159739  control     old_page          0
2   864975  2017-01-21 01:52:26.210827  control     old_page          1
3   936923  2017-01-10 15:20:49.083499  control     old_page          0
4   719014  2017-01-17 01:48:29.539573  control     old_page          0

   intercept  control  treatment country  UK  US  old_page  new_page
0          1        1          0      US   0   0         0         1
1          1        1          0      US   0   0         0         1
2          1        1          0      US   0   0         0         1
3          1        1          0      US   0   0         0         1
4          1        1          0      US   0   0         0         1
```

```
In [50]: df_merge['UK_ab']= df_merge['UK']*df_merge['treatment']
         df_merge['US_ab']=df_merge['US']*df_merge['treatment']
```

```
In [51]: # Fit your model, and summarize the results
         df_merge['intercept']=1
         lo = sm.Logit(df_merge['converted'],df_merge[['intercept','treatment','UK','US','UK_ab'
         results = lo.fit()
         #results.pvalues
         results.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.366109
         Iterations 6
```

```
Out[51]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                  Results: Logit
         ==================================================================
         Model:              Logit            No. Iterations:   6.0000
         Dependent Variable: converted        Pseudo R-squared: 0.000
         Date:               2022-02-23 13:01 AIC:              212782.6602
         No. Observations:   290584           BIC:              212846.1381
         Df Model:           5                Log-Likelihood:   -1.0639e+05
         Df Residuals:       290578           LL-Null:          -1.0639e+05
         Converged:          1.0000           Scale:            1.0000
         -------------------------------------------------------------------
                       Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
         -------------------------------------------------------------------
         intercept    -1.9865    0.0096  -206.3440  0.0000  -2.0053  -1.9676
         treatment    -0.0206    0.0137    -1.5052  0.1323  -0.0473   0.0062
         UK           -0.0175    0.0377    -0.4652  0.6418  -0.0914   0.0563
         US           -0.0057    0.0188    -0.3057  0.7598  -0.0426   0.0311
         UK_ab        -0.0469    0.0538    -0.8718  0.3833  -0.1523   0.0585
         US_ab         0.0314    0.0266     1.1807  0.2377  -0.0207   0.0835
         ==================================================================
```

"""

> **Hypothesis testing allow us to use only sample data to draw conclusions about an entire population, H0 (null hypothesis) is true before we collect data, we calculate the p-value to know if we fail to reject or reject the null hypothesis, In our case we will fail reject the null hypothesis. Finally, from all the results we observed from the previous calculation, I think the company should stick to the old page.**

## Final Check!
Congratulations! You have reached the end of the A/B Test Results project! You should be very proud of all you have accomplished!

> **Tip**: Once you are satisfied with your work here, check over your notebook to make sure that it satisfies all the specifications mentioned in the rubric. You should also probably remove all of the "Hints" and "Tips" like this one so that the presentation is as polished as possible.

## Submission You may either submit your notebook through the "SUBMIT PROJECT" button at the bottom of this workspace, or you may work from your local machine and submit on the last page of this project lesson.

1. Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

2. Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

3. Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```
In [52]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Analyze_ab_test_results_notebook.ipynb'])

Out[52]: 0
```