Analyze A/B Test Results

You may either submit your notebook through the workspace here, or you may work from your local machine and submit through the next page. Either way assure that your code passes the project RUBRIC (https://review.udacity.com/#!/projects/37e27304-ad47-4eb0-a1ab-8c12f60e43d0/rubric). **Please save regularly

This project will assure you have mastered the subjects covered in the statistics lessons. The hope is to have this project be as comprehensive of these topics as possible. Good luck!

Table of Contents

- Introduction
- Part I - Probability
- Part II - A/B Test
- Part III - Regression

Introduction

A/B tests are very commonly performed by data analysts and data scientists. It is important that you get some practice working with the difficulties of these

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 page, keep the old page, or perhaps run the experiment longer to make their decision.

As you work through this notebook, follow along in the classroom and answer the corresponding quiz questions associated with each question. The labels for each classroom concept are provided for each question. This will assure you are on the right track as you work through the project, and you can feel more confident in your final submission meeting the criteria. As a final check, assure you meet all the criteria on the RUBRIC (https://review.udacity.com/#!/projects/37e27304-ad47-4eb0-a1ab-8c12f60e43d0/rubric).

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.` Now, read in the `ab_data.csv` data. Store it in `df`. **Use your dataframe to answer the questions in Quiz 1 of the classroom.**

a. Read in the dataset and take a look at the top few rows here:

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

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
2	661590	2017-01-11 16:55:06.154213	treatment	new_page	0
3	853541	2017-01-08 18:28:03.143765	treatment	new_page	0
4	864975	2017-01-21 01:52:26.210827	control	old_page	1

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

In [3]:
```
df.shape[0]
```

Out[3]:
```
294478
```

c. The number of unique users in the dataset.

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

Out[4]:
```
user_id         290584
timestamp       294478
group                2
landing_page         2
converted            2
dtype: int64
```

d. The proportion of users converted.

In [5]:
```
df['converted'].mean()
```

Out[5]:
```
0.11965919355605512
```

e. The number of times the `new_page` and `treatment` don't line up.

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

Out[6]: 3893

f. Do any of the rows have missing values?

In [7]:
```
df.isnull().any().any()
```

Out[7]: False

2. For the rows where **treatment** is not aligned with **new_page** or **control** is not aligned with **old_page**, we cannot be sure if this row truly received the new or old page. Use **Quiz 2** in the classroom to provide how we should handle these rows.

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 [8]:
```
df2 = df.query('~((group == "treatment" and landing_page != "new_pag
e") or (group != "treatment" and landing_page == "new_page"))')
df2.shape[0]
```

Out[8]: 290585

In [9]:
```
# Double Check all of the correct rows were removed - this should be 0
df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_pag
e')) == False].shape[0]
```

Out[9]: 0

3. Use **df2** and the cells below to answer questions for **Quiz3** in the classroom.

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

In [10]:
```
df2.user_id.nunique()
```

Out[10]: 290584

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

In [11]:
```
df2[df2['user_id'].duplicated()]
```

Out[11]:

	user_id	timestamp	group	landing_page	converted
2893	773192	2017-01-14 02:55:59.590927	treatment	new_page	0

c. What is the row information for the repeat **user_id**?

In [12]:
```
df2[df2['user_id'].duplicated()].index
```

Out[12]:
```
Int64Index([2893], dtype='int64')
```

d. Remove **one** of the rows with a duplicate **user_id**, but keep your dataframe as **df2**.

In [13]:
```
df2 = df2.drop([2893])
```

4. Use **df2** in the below cells 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?

In [14]:
```
df2['converted'].mean()
```

Out[14]:
```
0.11959708724499628
```

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

In [15]:
```
df2[df2['group'] == 'control']['converted'].mean()
```

Out[15]:
```
0.1203863045004612
```

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

In [16]:
```
df2[df2['group'] == 'treatment']['converted'].mean()
```

Out[16]:
```
0.11880806551510564
```

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

In [17]:
```
df2[df2['landing_page'] == 'new_page']['user_id'].count() / df2.shape[
0]
```

Out[17]:
```
0.50006194422266881
```

e. Use the results in the previous two portions of this question to suggest if you think there is evidence that one page leads to more conversions? Write your response below.

No, there is no suficient evidence to conclude that the new landing page for the treatment group produces more conversions than the current page for the control group. The conversion rate for the former is, in fact, lower than that for the latter.

Part II - A/B Test

Notice that because of the time stamp associated with each event, you could technically run a hypothesis test continuously as each observation was observed.

However, then the hard question is 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.` For now, consider you need to make the decision just based on all the data provided. 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 your null and alternative hypotheses be? You can state your hypothesis in terms of words or in terms of p_{old} and p_{new}, which are the converted rates for the old and new pages.

$$H_0 : p_{new} - p_{old} \leq 0$$
$$H_1 : p_{new} - p_{old} > 0$$

`2.` Assume under the null hypothesis, p_{new} and p_{old} both have "true" success rates equal to the **converted** success rate regardless of page - that is p_{new} and p_{old} are equal. Furthermore, assume they are equal to the **converted** rate in **ab_data.csv** regardless of the page.

Use a sample size for each page equal to the ones in **ab_data.csv**.

Perform the sampling distribution for the difference in **converted** between the two pages over 10,000 iterations of calculating an estimate from the null.

Use the cells below to provide the necessary parts of this simulation. If this doesn't make complete sense right now, don't worry - you are going to work through the problems below to complete this problem. You can use **Quiz 5** in the classroom to make sure you are on the right track.

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

In [18]:
```
p_new = df2['converted'].mean()
p_new
```

Out[18]: 0.11959708724499628

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

In [19]:
```
p_old = df2['converted'].mean()
p_old
```

Out[19]: 0.11959708724499628

c. What is n_{new}?

In [20]:
```
n_new = df2[df2.landing_page == 'new_page']['user_id'].count()
n_new
```

Out[20]: 145310

d. What is n_{old}?

In [21]:
```
n_old = df2[df2.landing_page == 'old_page']['user_id'].count()
n_old
```

Out[21]: 145274

e. Simulate n_{new} transactions with a convert rate of p_{new} under the null. Store these n_{new} 1's and 0's in **new_page_converted**.

In [22]:
```
 new_page_converted = np.random.choice([0,1], n_new, [(1-p_new), p_new
])
```

f. Simulate n_{old} transactions with a convert rate of p_{old} under the null. Store these n_{old} 1's and 0's in **old_page_converted**.

In [23]:
```
old_page_converted = np.random.choice([0,1], n_old, [(1-p_old), p_old
])
```

g. Find p_{new} - p_{old} for your simulated values from part (e) and (f).

In [24]:
```
new_page_converted.mean() - old_page_converted.mean()
```

Out[24]:
```
0.00088101521039490871
```

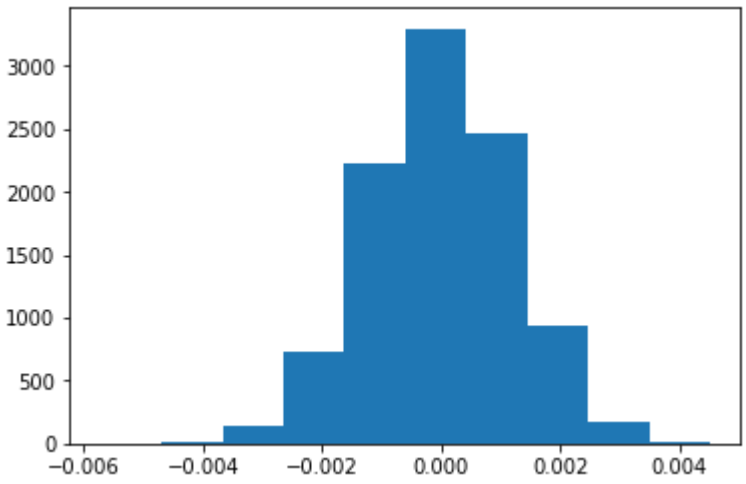
h. Simulate 10,000 p_{new} - p_{old} values using this same process similarly to the one you calculated in parts **a. through g.** above. Store all 10,000 values in **p_diffs**.

In [25]:
```
new_page_converted_sim = np.random.binomial(n_new, p_new, 10000) / n_n
ew
old_page_converted_sim = np.random.binomial(n_old, p_old, 10000) / n_o
ld
p_diffs = new_page_converted_sim - old_page_converted_sim
```

i. 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.

In [26]:
```
p_diffs = np.array(p_diffs)
plt.hist(p_diffs)
```

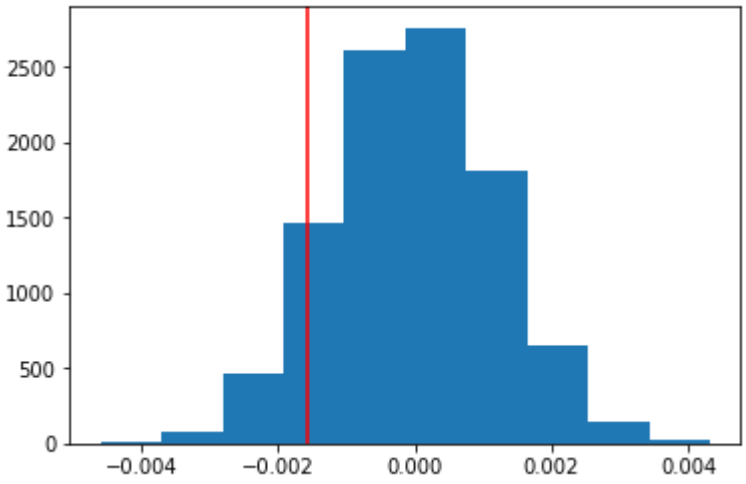
Out[26]:
```
(array([  1.00000000e+00,   5.00000000e+00,   1.34000000e+02,
          7.35000000e+02,   2.22700000e+03,   3.29800000e+03,
          2.47200000e+03,   9.42000000e+02,   1.66000000e+02,
          2.00000000e+01]),
 array([-0.0057216 , -0.00469817, -0.00367473, -0.0026513 , -0.001627
86,
        -0.00060443,  0.00041901,  0.00144244,  0.00246588,  0.003489
31,
         0.00451275]),
 <a list of 10 Patch objects>)
```



j. What proportion of the **p_diffs** are greater than the actual difference observed in **ab_data.csv**?

In [27]:
```
conv_rate_new = df2.query('converted == 1 and landing_page == "new_pag
e"').user_id.nunique() / n_new
conv_rate_old = df2.query('converted == 1 and landing_page == "old_pag
e"').user_id.nunique() / n_old
obs_diff = conv_rate_new - conv_rate_old
obs_diff
null_vals = np.random.normal(0, p_diffs.std(), p_diffs.size)
plt.hist(null_vals)
plt.axvline(obs_diff, c='red')
```

Out[27]: <matplotlib.lines.Line2D at 0x7f6aa5d11278>



In [28]:
```
(null_vals > obs_diff).mean()
```

Out[28]: 0.90580000000000005

k. In words, explain what you just computed in part **j.**. What is this value called in scientific studies? What does this value mean in terms of whether or not there is a difference between the new and old pages?

The calculation in part j produces the percentage of the simulated conversion rate differences between new and old landing pages greated than the observed conversion rate differences. In statistical studies, the result is called the p-value.

A low p-value allows us to reject the null hypothesis. In this case, our alpha is 0.05 as we have a type I error rate of 5%. The p-value is quite large at 0.91; therefore, we fail to reject the null hypothesis. In practical terms, we can conclude that it is better off keeping the current page.

l. 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 walkthrough of the ideas that are critical to correctly thinking about statistical significance. Fill in the below to calculate the number of conversions for each page, as well as the number of individuals who received each page. Let `n_old` and `n_new` refer the the number of rows associated with the old page and new pages, respectively.

In [29]:
```
convert_old = df[df['group'] == 'control']['converted'].sum()
convert_new = df[df['group'] == 'treatment']['converted'].sum()
n_old = df[df['group'] == 'control']['converted'].shape[0]
n_new = df[df['group'] == 'treatment']['converted'].shape[0]

print('convert_old: ', convert_old)
print('convert_new: ', convert_new)
print('n_old: ', n_old)
print('n_new: ', n_new)
```

```
convert_old:  17723
convert_new:  17514
n_old:  147202
n_new:  147276
```

m. Now use `stats.proportions_ztest` to compute your test statistic and p-value. Here (http://knowledgetack.com/python/statsmodels/proportions_ztest/) is a helpful link on using the built in.

In [30]:
```
import statsmodels.api as sm

z_score, p_value = sm.stats.proportions_ztest([convert_new, convert_ol
d], [n_new, n_old], alternative='larger')
print('z-score: {}'.format(z_score))
print('p-value: {}'.format(p_value))
```

```
/opt/conda/lib/python3.6/site-packages/statsmodels/compat/pandas.py:5
6: FutureWarning: The pandas.core.datetools module is deprecated and
will be removed in a future version. Please use the pandas.tseries mo
dule instead.
  from pandas.core import datetools

z-score: -1.2369217547321678
p-value: 0.8919419336512124
```

In [31]:
```
from scipy.stats import norm

z_critical_value = norm.ppf(1-0.05)
print('critcal value of z-score at 5% alpha is: {}'.format(z_critical_
value))
```

```
critcal value of z-score at 5% alpha is: 1.6448536269514722
```

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.**?

As shown above, I have computed a z-score of -1.23 and a p-value of 0.89, the latter of which is similar to my previously calculated p-value of 0.91. Because the critical value of z-score is 1.64 at a 5% alpha, we can proceed to reject our null hypothesis.

This confirms our prior conclusion that there is no statistical significance to indicate any difference in conversation rates between the old and new pages. Said differently, we should stay with the current landing page.

Yes, my results here are in agreement with my findings in parts j and k.

Part III - A regression approach

1. In this final part, you will see that the result you acheived in the previous A/B test can also be acheived by performing regression.

a. Since each row 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** to fit the regression model you specified in part **a.** to see if there is a significant difference in conversion based on which page a customer receives. However, you first need to create a colun for the intercept, and create a dummy variable column for which page each user received. Add an **intercept** column, as well as an **ab_page** column, which is 1 when an individual receives the **treatment** and 0 if **control**.

In [32]:
```
df2['intercept'] = 1
df2[['to_drop', 'ab_page']] = pd.get_dummies(df2['group'])
df2.drop(['to_drop'], axis=1, inplace=True)
df2.head()
```

Out[32]:

	user_id	timestamp	group	landing_page	converted	intercept	ab_page
0	851104	2017-01-21 22:11:48.556739	control	old_page	0	1	0
1	804228	2017-01-12 08:01:45.159739	control	old_page	0	1	0
2	661590	2017-01-11 16:55:06.154213	treatment	new_page	0	1	1
3	853541	2017-01-08 18:28:03.143765	treatment	new_page	0	1	1
4	864975	2017-01-21 01:52:26.210827	control	old_page	1	1	0

c. Use **statsmodels** to import your regression model. Instantiate the model, and fit the model using the two columns you created in part **b.** to predict whether or not an individual converts.

In [33]:
```
log_mod = sm.Logit(df2['converted'], df2[['intercept', 'ab_page']])
```

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

In [34]:
```
results = log_mod.fit()
results.summary()
```

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

Out[34]:

Logit Regression Results

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290582
Method:	MLE	**Df Model:**	1
Date:	Sat, 21 Apr 2018	**Pseudo R-squ.:**	8.077e-06
Time:	13:24:21	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
		LLR p-value:	0.1899

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-1.9888	0.008	-246.669	0.000	-2.005	-1.973
ab_page	-0.0150	0.011	-1.311	0.190	-0.037	0.007

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

Hint: What are the null and alternative hypotheses associated with your regression model, and how do they compare to the null and alternative hypotheses in the **Part II**?

The p-value associated with ab_page is 0.19, which is much lower than the value of 0.91 found in Part II. There is a discrepancy in p-values because Part II and III do not test the same hypotheses.

In Part III d, the null hypothesis is concered with the lack of relationship between the conversion rate and the landing page. The alternative states the opposite: correlation exists between the two terms. With a p-value of 0.19, we fail to reject the null hypothesis since there is no correlation between the dependent and independent variables.

In Part II, the null hypothesis is different in that it is based on whether the conversion rate for the new page is less than or equal to the conversion rate for the old page. The alternative states the opposite: the conversion rate for the new page is larger than the conversion rate for the old page.

Because the hypotheses being tested in Part II and III are not the same, their associated p-values carry different meanings. Therefore, the difference in p-values from Part II and III makes sense.

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 more terms into my regression model could help us gain deeper insights into our findings. For example, it might be helpful to look at where users come from before arriving at the landing page. The sources of users could include email campaigns, referal partners, paid marketing sites, etc. Different sources might generate different conversion rates.

On the other hand, there are potential disadvantages for doing so. The existence of multiple factors gives rise to multicollinearity. It refers to a situation in which predictors are correlated with each other, and it could make my regression model unstable and my results difficult to interpret.

g. Now along with testing if the conversion rate changes for different pages, also add an effect based on which country a user lives. You will need to read in the **countries.csv** dataset and merge together your datasets on the approporiate rows. Here (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.join.html) are the docs for joining tables.

Does it appear that country had an impact on conversion? Don't forget to create dummy variables for these country columns - **Hint: You will need two columns for the three dummy varaibles.** Provide the statistical output as well as a written response to answer this question.

In [35]:
```
df_countries = pd.read_csv('countries.csv')
df_joined = df_countries.set_index('user_id').join(df2.set_index('user
_id'), how='inner')
df_joined.head()
```

Out[35]:

	country	timestamp	group	landing_page	converted	intercept	ab_pa
user_id							
834778	UK	2017-01-14 23:08:43.304998	control	old_page	0	1	0
928468	US	2017-01-23 14:44:16.387854	treatment	new_page	0	1	1
822059	UK	2017-01-16 14:04:14.719771	treatment	new_page	1	1	1
711597	UK	2017-01-22 03:14:24.763511	control	old_page	0	1	0
710616	UK	2017-01-16 13:14:44.000513	treatment	new_page	0	1	1

In [36]:
```
df_joined['country'].unique()
```

Out[36]:
```
array(['UK', 'US', 'CA'], dtype=object)
```

In [37]:
```
df_joined[['UK', 'US', 'CA']] = pd.get_dummies(df_joined['country'])
log_mod_2 = sm.Logit(df_joined['converted'], df_joined[['intercept',
'US', 'CA']])
results = log_mod_2.fit()
results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.366116
         Iterations 6
```

Out[37]:

Logit Regression Results

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290581
Method:	MLE	**Df Model:**	2
Date:	Sat, 21 Apr 2018	**Pseudo R-squ.:**	1.521e-05
Time:	13:24:22	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
		LLR p-value:	0.1984

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-2.0375	0.026	-78.364	0.000	-2.088	-1.987
US	0.0507	0.028	1.786	0.074	-0.005	0.106
CA	0.0408	0.027	1.518	0.129	-0.012	0.093

In [38]:
```
np.exp(results.params)
```

Out[38]:
```
intercept    0.130350
US           1.052027
CA           1.041647
dtype: float64
```

UK is the baseline category for its omission from the logistic regression model. My results indicate that US and Canadian users are 1.05 and 1.04 times, respectively, as likely to convert as UK users.

The p-values of 0.074 and 0.129 for US and CA terms, respectively, are larger than an alpha of 0.05. Thus, both are not statistically significant. Practically speaking, differences of 4 - 5% among three countries do not seem significant to suggest the adoption of new landing pages based on countries.

h. 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 there significant effects on conversion. Create the necessary additional columns, and fit the new model.

Provide the summary results, and your conclusions based on the results.

In [39]:
```
log_mod_3 = sm.Logit(df_joined['converted'], df_joined[['intercept',
'ab_page', 'US', 'CA']])
results = log_mod_3.fit()
results.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.366113
         Iterations 6
```

Out[39]:

Logit Regression Results

Dep. Variable:	converted	**No. Observations:**	290584
Model:	Logit	**Df Residuals:**	290580
Method:	MLE	**Df Model:**	3
Date:	Sat, 21 Apr 2018	**Pseudo R-squ.:**	2.323e-05
Time:	13:24:23	**Log-Likelihood:**	-1.0639e+05
converged:	True	**LL-Null:**	-1.0639e+05
		LLR p-value:	0.1760

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
intercept	-2.0300	0.027	-76.249	0.000	-2.082	-1.978
ab_page	-0.0149	0.011	-1.307	0.191	-0.037	0.007
US	0.0506	0.028	1.784	0.074	-0.005	0.106
CA	0.0408	0.027	1.516	0.130	-0.012	0.093

In [40]:
```
np.exp(results.params)
```

Out[40]:
```
intercept    0.131332
ab_page      0.985168
US           1.051944
CA           1.041599
dtype: float64
```

With all three factors present in this model, the p-values have either stayed the same or increased. According to the results, this new model has a worse fit than previous ones in which there is no interaction between a/b test landing pages and countires.

Conclusions

Congratulations on completing the project!

Gather Submission Materials

Once you are satisfied with the status of your Notebook, you should save it in a format that will make it easy for others to read. You can use the **File -> Download as -> HTML (.html)** menu to save your notebook as an .html file. If you are working locally and get an error about "No module name", then open a terminal and try installing the missing module using `pip install <module_name>` (don't include the "<" or ">" or any words following a period in the module name).

You will submit both your original Notebook and an HTML or PDF copy of the Notebook for review. There is no need for you to include any data files with your submission. If you made reference to other websites, books, and other resources to help you in solving tasks in the project, make sure that you document them. It is recommended that you either add a "Resources" section in a Markdown cell at the end of the Notebook report, or you can include a `readme.txt` file documenting your sources.

Submit the Project

When you're ready, click on the "Submit Project" button to go to the project submission page. You can submit your files as a .zip archive or you can link to a GitHub repository containing your project files. If you go with GitHub, note that your submission will be a snapshot of the linked repository at time of submission. It is recommended that you keep each project in a separate repository to avoid any potential confusion: if a reviewer gets multiple folders representing multiple projects, there might be confusion regarding what project is to be evaluated.

It can take us up to a week to grade the project, but in most cases it is much faster. You will get an email once your submission has been reviewed. If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com. In the meantime, you should feel free to continue on with your learning journey by continuing on to the next module in the program.