A/B Testing Project

- Conducting A/B tests for Maximum Bidding/Average Bidding systems
- Forming the potential basis for statistically significant data-driven decisions

S. Emre Kıyak
Data Science Researcher
s.emrekiyak@gmail.com

Outline

- What is A/B Testing?
- Brief Key points on Literature Review
- Business Problem / Case Study
- How to conduct A/B Testing?
- Data Exploration
- Testing Assumptions and Conducting A/B Tests
- Results and Conclusion
- References

What is A/B Testing?

- A statistical hypothesis testing and research methodology of comparing & testing two or more variants features in a service or product to determine which performs better

- Key Point: 🔍 Making inference whether a difference between two variants is statistically significant & to prove that when comparing two variants, the difference is obvious by leaving no room for the chance !

- For example:
  - Statistical comparison of the outcomes of a brand-new developed ML algorithm and an existing method

  - Should the purchase button be green or blue on the website?

- A widely used crucial data-driven approach for efficient, scientific decision making.

- Independent two-sample t-test, Two sample proportion hypothesis z-test

Brief Key points on Literature Review

- A/B tests are indispensable for identifying statistically significant results with respect to forming data-driven decisions rather than happened by chance. (King et al., 2017)
- The de facto standard for more efficient outcomes, improving customer experience, optimizing the design and decision making by leveraging data. (Kumar, 2019)
- Providing a "glimpse into the future" to observe how the regarding a product is likely to perform. (King et al., 2017)
- A/B testing has gone from a secret weapon within the purview of only a handful of tech companies to an increasingly ubiquitous and critical part of doing business online. (Siroker et al., 2013)
- The adoption and long-term success of hypothesis testing require considering how it may philosophically, culturally and organizationally fit the company. (Siroker et al., 2013)
- Mission-critical value of A/B testing in 2012 Presidential race in the U.S mentioned as the Presidential campaign fundraising machines by leading publications from TIME to Forbes. (Siroker et al., 2013)

Business Problem / Case Study

- Facebook has recently introduced a new type of bidding named "Average Bidding“ , as an alternative to the existing bidding system called "Maximum Bidding".
- Untitled_Company.com has decided to test this new feature.
- The company aims to measure whether the new bid type "Average Bidding" yields better results than current bid type "Maximum Bidding" using A/B testing.
- Research Question: Is "Average Bidding system" better/ beneficial than "Maximum Bidding system“ ?
- As the result of the first observations that have continued for about 40 days, the company expects the A/B test to be conducted and the results to be analysed.

Business Problem / Case Study

- "Untitled_Company.com" company's target audience has been randomly divided into two groups of equal size.

- A Facebook ad campaign with "Maximum Bidding" is served to "Control Group" and another campaign with "Average Bidding" is served to "Test Group".

  - Control Group : Maximum Bidding System

  - Test Group: Average Bidding System.

- The ultimate measure of success for "Untitled_Company.com" is "Purchase" metric.

Variables
- Impression
- Click
- Purchase
- Earning

Customer Journey

Impression: Customer sees an advertisement

Click: Customer clicks on the website link in the ad

Search: Customer searches in the website

View Content: Customer views the details of a product

Add to Chart: Customer adds product to cart

Purchase: Customer purchases the product

How to conduct A/B Testing?

- In order to conduct the A/B test, first of all, 2 statistical assumptions must be provided beforehand.
  1. The Assumption of Normality (Essential)
  2. The Assumption of Homogeneity of Variance

If assumptions
- are ensured → Parametric Test → Independent two-sample t-test
- are not ensured → Nonparametric Test → Mann–Whitney U test

The Assumption of Normality

- The first theoretical assumption of A/B testing to be tested. Most of the parametric tests require that the assumption of normality be met. It is necessary to test whether the distribution of a variable is the same as the theoretical normal distribution.
- Normality can be tested with graphs and several statistical tests such as The Shapiro-Wilk test, Chi-square normality test, Kolmogorov-Smirnov Goodness of Fit Test, Lilliefors Test etc.
- The Shapiro-Wilk test has been used in this research to examine whether the distribution of the data as a whole deviates from a comparable normal distribution.

The Assumption of Homogeneity of Variance

- ✓ The second theoretical assumption of A/B testing to be tested.
- ✓ Levene's test has been used in this research to test whether the variances of the distributions of two variables are similar. (whether variances are homogenous)
- ✓ Bartlett's Test, Brown-Forsythe Test could also be used to test the assumption of homogeneity of variance.

Data Exploration

- After reading the data for both groups, shape, data types, head and the existence of null values have been summarized for the first look.

- Number of Observations = 40

- Shape (40, 4)

- No missing values

- No outliers

```
In[3]: check_df(control_group)
##################### Shape #####################
(40, 4)
##################### Types #####################
Impression    float64
Click         float64
Purchase      float64
Earning       float64
dtype: object
##################### Head #####################
   Impression   Click  Purchase  Earning
0    82529.46 6090.08    665.21  2311.28
1    98050.45 3382.86    315.08  1742.81
2    82696.02 4167.97    458.08  1797.83
##################### Tail #####################
    Impression   Click  Purchase  Earning
37   123678.93 3649.07    476.17  2187.72
38   101997.49 4736.35    474.61  2254.56
39   121085.88 4285.18    590.41  1289.31
##################### NA #####################
Impression    0
Click         0
Purchase      0
Earning       0
dtype: int64
##################### Quantiles #####################
               0.00     0.05     0.50      0.95      0.99      1.00
Impression 45475.94 79412.02 99790.70 132950.53 143105.79 147539.34
Click       2189.75  3367.48  5001.22   7374.36   7761.80   7959.13
Purchase     267.03   328.66   531.21    748.27    790.19    801.80
Earning     1253.99  1329.58  1975.16   2318.53   2481.31   2497.30
```

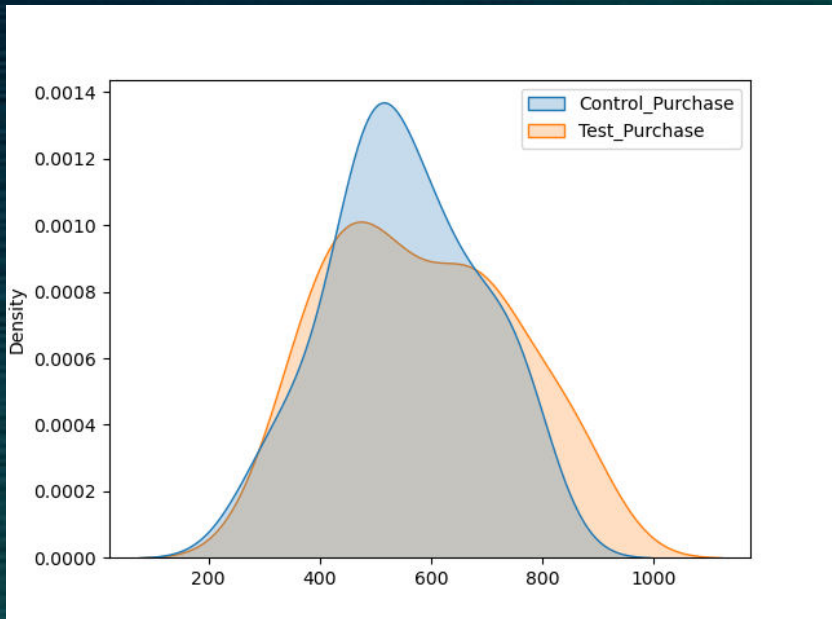
```
In[4]: check_df(test_group)
##################### Shape #####################
(40, 4)
##################### Types #####################
Impression    float64
Click         float64
Purchase      float64
Earning       float64
dtype: object
##################### Head #####################
   Impression   Click  Purchase  Earning
0   120103.50 3216.55    702.16  1939.61
1   134775.94 3635.08    834.05  2929.41
2   107806.62 3057.14    422.93  2526.24
##################### Tail #####################
    Impression   Click  Purchase  Earning
37   116481.87 4702.78    472.45  2597.92
38    79033.83 4495.43    425.36  2595.86
39   102257.45 4800.07    521.31  2967.52
##################### NA #####################
Impression    0
Click         0
Purchase      0
Earning       0
dtype: int64
##################### Quantiles #####################
               0.00     0.05      0.50      0.95      0.99      1.00
Impression 79033.83 83150.50 119291.30 153178.69 158245.26 158605.92
Click       1836.63  2600.36   3931.36   5271.19   6012.88   6019.70
Purchase     311.63   356.70    551.36    854.21    876.58    889.91
Earning     1939.61  2080.98   2544.67   2931.31   3091.94   3171.49
```

Data Exploration

Let's take a closer look at 'Purchase' Variable in terms of the confidence interval (% 95) , mean and displaying the distribution.

```
In[5]: sms.DescrStatsW(control_group["Purchase"]).tconfint_mean()
Out[5]: (508.0041754264924, 593.7839421139709)
In[6]: sms.DescrStatsW(test_group["Purchase"]).tconfint_mean()
Out[6]: (530.5670226990063, 633.645170597929)
```

Confidence Intervals:
If the number of observations is conducted for 100 days, purchases for control group in 95 out of 100 days would be between 508 and 594 values and purchases for test group in 95 out of 100 days would be between 530 and 633 values.



Mean 'Purchase' control group: 550.89
Mean 'Purchase' test group: 582.11

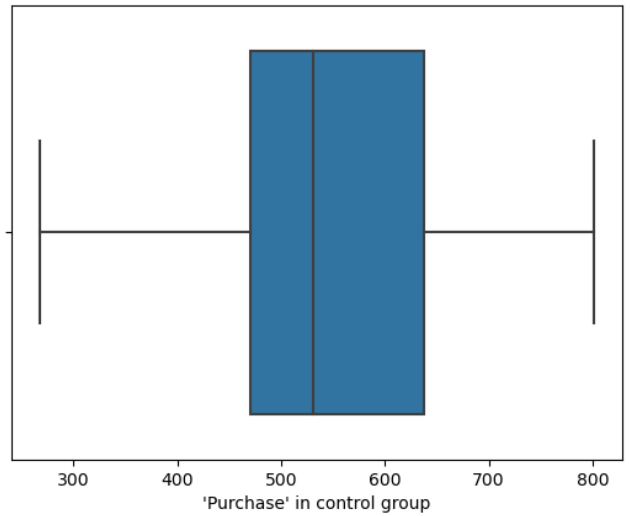
Median 'Purchase' control group: 531.21
Median 'Purchase' test group: 551.36

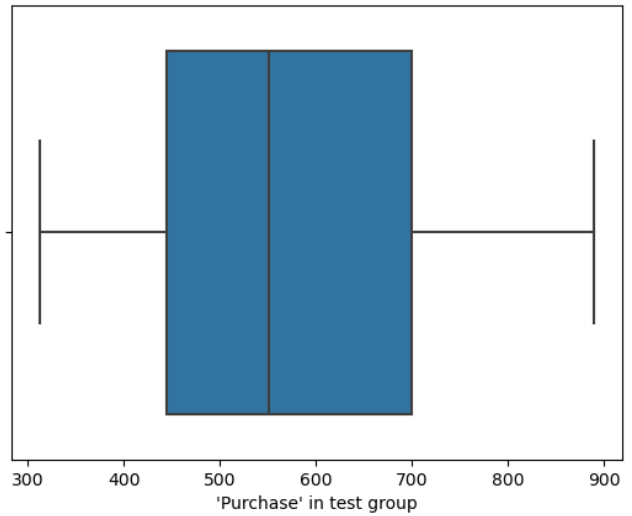
The close proximity of mean and median is a strong indication that the distribution is uniform, homogeneous, and symmetrical.
- However, it will be tested.

The mean of Purchase in Test group(Average Bid) is higher than control group(Maximum Bid).
However we cannot only rely on this information.

Is there statistically significant difference between the means of 'Purchase' of the Control and Test Groups?

This question brings out the necessity of A/B Testing!





Before conducting the A/B Testing, functions have been created for multiple uses

```python
def normality_test(dataframe, col_name, plot=False):
    """The Assumptions of normality: The first theoretical assumption of AB testing to be tested.
    Most of the parametric tests require that the assumption of normality be met.
    The Shapiro-Wilk test is used to examine whether the distribution of the data as a whole
    deviates from a comparable normal distribution.

    H0: The assumption of the normal distribution is provided.
    H1: The assumption of the normal distribution is NOT provided."""

    # from scipy.stats import shapiro
    test_stats, p_value = shapiro(dataframe[col_name])
    print("Test Statistic= %.4f, p-value = %.4f" % (test_stats, p_value))

    if plot:
        stats.probplot(dataframe[col_name], dist="norm", plot=pylab)
        pylab.title(f"Q-Q Plot")
        pylab.show()
```

Normality test function has been created to examine the assumption of normality with Shapiro-Wilks statistical test that also enables to visualize Q-Q plot.

```python
def testing_variance_homogeneity(arg1, arg2):
    """
    The Assumption of Homogeneity of Variance: The second theoretical assumption of AB testing to be tested.
    Levene's test is used to test whether the variances of the distributions of two variables are similar.
    (whether variances are homogenous)

    H0: Variances are homogenous
    H1: Variances are not homogenous
    """
    # from scipy import stats
    test_stats, p_value = stats.levene(arg1, arg2)
    print("Test Statistic= %.4f, p-value = %.4f" % (test_stats, p_value))
```

Testing variance homogeneity function has been created to examine the assumption of homogeneity of variance

Continued.

```python
def ab_testing(arg1, arg2):
    """
    If the assumptions of normality & variance homogeneity are met=> AB TESTING(Independent two-sample t-test- parametric)
    If the assumptions are not met => MannWhitneyU (nonparametric)"""
    # H0: M1 = M2  (There is no statistically significant difference between the groups with % 95 confidence)
    # H1: M1 != M2 (There is a statistically significant difference between the groups with % 95 confidence)

    test_stats, p_value = stats.ttest_ind(arg1, arg2, equal_var=True)
    print("Test Statistic= %.4f, p-value = %.4f" % (test_stats, p_value))
```

```python
def mann_whitney_u_test(arg1, arg2):
    """
    If the assumptions of normality & variance homogeneity are not met=> MannWhitneyU (nonparametric)"""
    # H0: M1 = M2  (There is no statistically significant difference between the groups with % 95 confidence)
    # H1: M1 != M2 (There is a statistically significant difference between the groups with % 95 confidence)

    test_stats, p_value = stats.mannwhitneyu(arg1, arg2)
    print("Test Statistic= %.4f, p-value = %.4f" % (test_stats, p_value))
```

According to the results of testing assumptions prior to A/B testing, either parametric or nonparametric tests must be used.

Defining Main Hypotheses

H0: M1 = M2 (There is no statistically significant difference between the means of 'Purchase' of the Control and Test Groups)

H1: M1 != M2 (There is a statistically significant difference between the means of 'Purchase' of the Control and Test Groups)

1. The Assumption of Normality

First of all, it is necessary to check whether the assumption of normality is met within the two groups.
The Shapiro-Wilk test is conducted to examine if the variable has normal distribution.

HYPOTHESES
H0: The assumption of the normal distribution is provided.
H1: The assumption of the normal distribution is NOT provided.

According to Shapiro Wilk's test result;

```
In[20]: normality_test(control_group, "Purchase", True)
Test Statistic= 0.9773, p-value = 0.5891
```

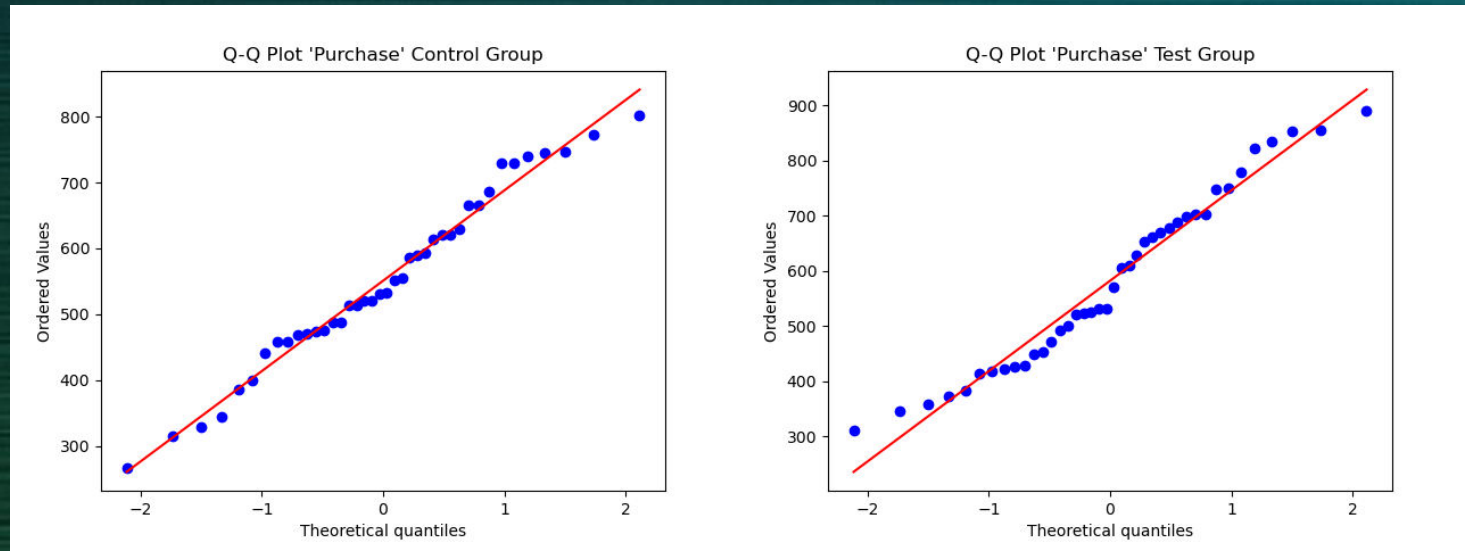
p-value = 0.5891 > 0.05 → H0 cannot be rejected
The assumption of the normality is provided for Purchase variable in Control Group

```
In[21]: normality_test(test_group, "Purchase", True)
Test Statistic= 0.9589, p-value = 0.1541
```

p-value = 0.1541 > 0.05 → H0 cannot be rejected
The assumption of the normality is provided for Purchase variable in Test Group.

- Purchase variable in both groups have normal distribution.

As can be seen in the Q-Q Plot charts, the purchase variable is normally distributed in both groups.



--> First Assumption has ensured!

-Since the most important criterion for the application of parametric tests is providing the assumption of normality, it is of great importance.

2. The Assumption of Homogeneity of Variance

Levene's test is used to test whether the variances of two variables are homogenous.

HYPOTHESES
H0: Variances are homogenous
H1: Variances are not homogenous

```
In[23]: testing_variance_homogeneity(control_group["Purchase"], test_group["Purchase"])
Test Statistic= 2.6393, p-value = 0.1083
```

According to Levene's test result;

p-value = 0.10 > 0.05 → H0 cannot be rejected

The assumption of homogeneity of variance is provided which means that variances are homogenous, the variances of the distributions of two variables are similar.

- Second Assumption has also ensured!
- Since 2 of the theoretical assumptions are met, the Parametric t-test will be conducted!

A/B Testing (Independent two-sample t-test)

Independent two-sample t-test (parametric) will be applied since two assumptions are ensured beforehand.

HYPOTHESES
H0: M1 = M2 (There is no statistically significant difference between the means of 'Purchase' of the Control and Test Groups)
H1: M1 != M2 (There is a statistically significant difference between the means of 'Purchase' of the Control and Test Groups)

```
In[25]: ab_testing(control_group["Purchase"], test_group["Purchase"])
Test Statistic= -0.9416, p-value = 0.3493
```

According to the result of A/B testing;

Since p-value = 0.34 > 0.05 → H0 cannot be rejected!

- When evaluated statistically according to the Purchase variable, there is no statistically significant difference between the means of 'Purchase' of the Control and Test Groups.

- All in all, it turns out that there is no statistically significant difference between the Maximum Bidding and Average Bidding system with % 95 confidence in terms of Purchase variable

- In terms of the means of Purchase variable, Average Bidding promises higher purchases. In order to get more precise results, the number of observations should be increased. Recommendations will be treated in the final section.

- Let's also have a closer look at 'Earning' variable and add "Conversion Rate" variable by doing a little Feature Engineering!

Let's take a look at Earning variable!

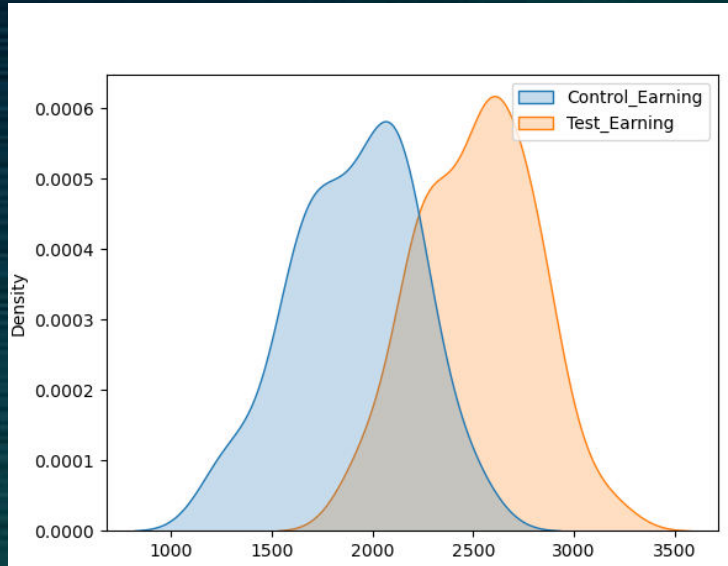
```
In[7]: sms.DescrStatsW(control_group["Earning"]).tconfint_mean()
Out[7]: (1811.6904932901248, 2005.4461063153722)
In[8]: sms.DescrStatsW(test_group["Earning"]).tconfint_mean()
Out[8]: (2424.4690197727855, 2605.312445528449)
```

Confidence Intervals (% 95)
If the number of observations is conducted for 100 days, Earning for control group in 95 out of 100 days would be between 1811 and 2005 values and Earning for test group in 95 out of 100 days would be between 2424 and 2605 values.

There is close proximity of mean and median in both group, distributions seem symmetrical.

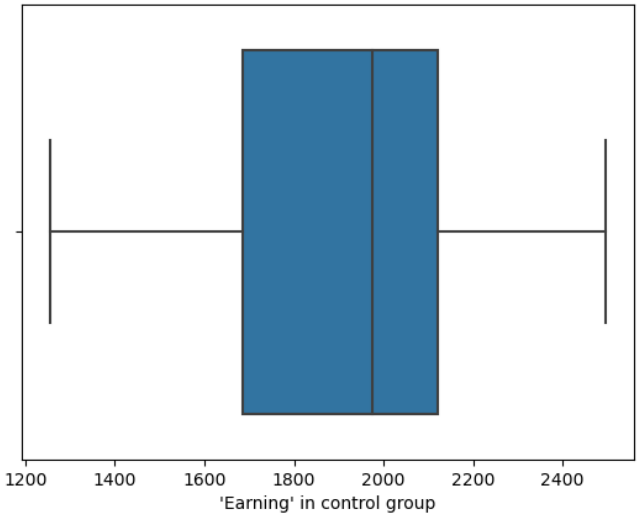
Mean 'Earning' in control group: 1908.57
Mean 'Earning' in test group: 2514.89

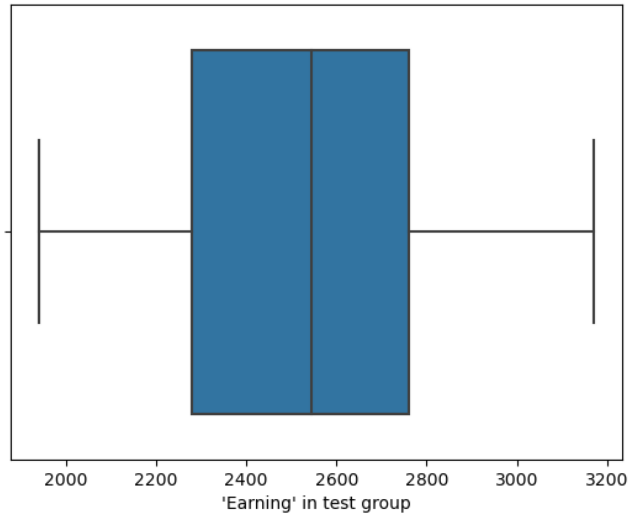
Median 'Earning' in control group: 1975.16
Median 'Earning' in test group: 2544.67



The mean of Earning in Test group(Average Bid) is higher than control group(Maximum Bid).

Let's conduct a test if this is statistically meaningful and significant?

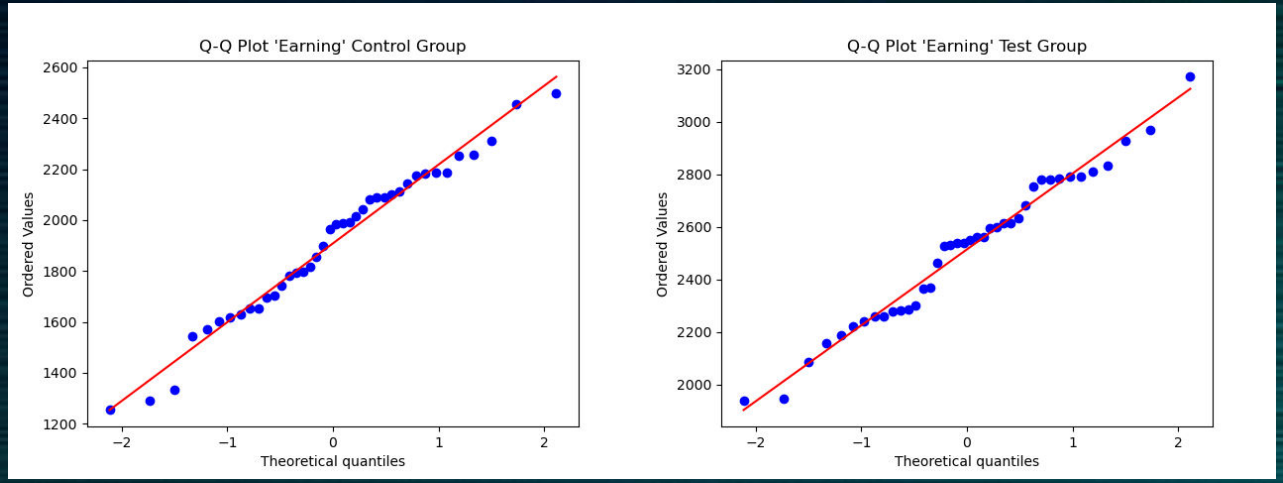




1. The Assumption of Normality (Earning)

HYPOTHESES
H0: The assumption of the normal distribution is provided.
H1: The assumption of the normal distribution is NOT provided.

```
In[41]: normality_test(control_group, "Earning", True)
Test Statistic= 0.9756, p-value = 0.5306
In[42]: normality_test(test_group, "Earning", True)
Test Statistic= 0.9780, p-value = 0.6163
```



According to Shapiro Wilk's test result;

p-value = 0.5306 > 0.05 → H0 cannot be rejected
The assumption of the normality is provided for Earning variable in Control Group

p-value = 0.6163 > 0.05 → H0 cannot be rejected
The assumption of the normality is provided for Earning variable in Test Group.

- Earning variable in both groups have normal distribution.

--> First Assumption has ensured!

2. The Assumption of Homogeneity of Variance (Earning)

HYPOTHESES
H0: Variances are homogenous
H1: Variances are not homogenous

```
In[43]: testing_variance_homogeneity(control_group["Earning"], test_group["Earning"])
Test Statistic= 0.3532, p-value = 0.5540
```

According to Levene's test result;

p-value = 0.5540 > 0.05 => H0 cannot be rejected
The assumption of the homogeneity of variance is provided for Earning variable.
Variances are homogenous.

The assumptions are ensured for A/B Testing, so parametric test will be conducted!

A/B Testing for Earning variable (Independent two-sample t-test)

Independent two-sample t-test (parametric) will be applied since two assumptions are ensured beforehand.

HYPOTHESES
H0: M1 = M2 (There is no statistically significant difference between the means of Earning' of the Control and Test Groups)
H1: M1 != M2 (There is a statistically significant difference between the means of 'Earning' of the Control and Test Groups)

```
In[44]: ab_testing(control_group["Earning"], test_group["Earning"])
Test Statistic= -9.2545, p-value = 0.0000
```

According to the result of A/B testing;

Since p-value = 0.00 < 0.05 → H0 is rejected!

- When evaluated statistically according to the Earning variable, there is a statistically significant difference between the means of Earning' of the Control and Test Groups.
- All in all, it turns out that there is a statistically significant difference between the Maximum Bidding and Average Bidding system with % 95 confidence in terms of Earning variable
- In terms of the means of Earning variable, Average Bidding promises higher yields.

Let's calculate Conversion Rate and conduct the test!

```
# Conversion Rate
control_group["Conversion Rate"] = (control_group["Purchase"] / control_group["Click"]) * 100
test_group["Conversion Rate"] = (test_group["Purchase"] / test_group["Click"]) * 100
```

```
In[46]: control_group.head()
Out[46]:
   Impression   Click  Purchase  Earning  Conversion Rate
0    82529.46 6090.08    665.21  2311.28            10.92
1    98050.45 3382.86    315.08  1742.81             9.31
2    82696.02 4167.97    458.08  1797.83            10.99
3   109914.40 4910.88    487.09  1696.23             9.92
4   108457.76 5987.66    441.03  1543.72             7.37
```

```
In[47]: test_group.head()
Out[47]:
   Impression   Click  Purchase  Earning  Conversion Rate
0   120103.50 3216.55    702.16  1939.61            21.83
1   134775.94 3635.08    834.05  2929.41            22.94
2   107806.62 3057.14    422.93  2526.24            13.83
3   116445.28 4650.47    429.03  2281.43             9.23
4   145082.52 5201.39    749.86  2781.70            14.42
```

Mean 'Conversion rate' in control group: 11.59
Mean 'Conversion rate' in test group: 15.65

Median 'Conversion rate' in control group: 10.95
Median 'Conversion rate' in test group: 14.61

1. The Assumption of Normality (Conversion Rate)

HYPOTHESES
H0: The assumption of the normal distribution is provided.
H1: The assumption of the normal distribution is NOT provided.

According to Shapiro Wilk's test result;

```
In[48]: normality_test(control_group, "Conversion Rate", True)
Test Statistic= 0.8720, p-value = 0.0003
```

p-value for control group is: 0.0003 < 0.05 → H0 is rejected!
Conversion Rate variable in control group does not seem to have a normal distribution.

```
In[49]: normality_test(test_group, "Conversion Rate", True)
Test Statistic= 0.8381, p-value = 0.0000
```

p-value for test group is: 0.0000 < 0.05 → H0 is rejected!
Conversion Rate variable in test group does not seem to have a normal distribution.

THE ASSUMPTION OF NORMALITY IS NOT PASSED! ===> NONPARAMETRIC

2. The Assumption of Homogeneity of Variance (Conversion Rate)

HYPOTHESES
H0: Variances are homogenous
H1: Variances are not homogenous

```
In[50]: testing_variance_homogeneity(control_group["Conversion Rate"], test_group["Conversion Rate"])
Test Statistic= 2.0759, p-value = 0.1536
```

p value for homogenous variances: 0.1536 > 0.05 → H0 cannot be rejected!
Variances are homogenous. The second assumption is ensured.

Mann–Whitney U test (non-parametric)

HYPOTHESES
H0: M1 = M2 (There is no statistically significant difference between the means of "Conversion Rate" variable of the Control and Test Groups)
H1: M1 != M2 (There is a statistically significant difference between the means of "Conversion Rate" variable of the Control and Test Groups)

```
In[51]: mann_whitney_u_test(control_group["Conversion Rate"], test_group["Conversion Rate"])
Test Statistic= 459.0000, p-value = 0.0005
```

p-value for Mann-Whitney-U non-parametric test = 0.0005 < 0.05 → H0 is rejected!

- There is a statistically significant difference between the means of "Conversion Rate" variable in the Control and Test Groups with 95 % confidence.

- Test group's Conversion Rate is higher than the control group.

Results & Conclusion

'Purchase' Evaluation Metric

- Even though the mean of Purchase is higher in Average Bidding, it turned out that "there is no statistically significant difference between two bidding systems as the result of A/B testing. So, the current Maximum Bidding system can be continued and long-term changes can be evaluated by increasing the number of observations (more than 40).

- Also, since there is no significant difference between the two methods, customers can be asked which method they would prefer by conducting surveys.

'Earning' Evaluation Metric

- When focusing on the 'Earning' metric, as it is proven that there is both mathematically and statistically significant difference between the means of Earning variable between both bidding system. Thus, the Average Bidding system ends up increasing the metric of 'Earning'.

- For this reason, choosing Average Bidding method can be more profitable and beneficial for "Untitled_Company.com"

'Conversion Rate' Evaluation Metric

- The initial version of the data did not contain 'Conversion Rate' variable, out of curiosity, I aimed to calculate and investigate the possible impact of 'Conversion Rate'. {Conversion Rate = Purchase / Click}

- Results indicate that from the point of 'Conversion Rate', there is both mathematically and statistically significant difference between the means of 'Conversion Rate' variable between both bidding system. 'Average Bidding' system seems to have higher 'Conversion Rate' than Maximum Bidding.

References

- Kumar, R. (2019, March). Data-Driven Design: Beyond A/B Testing. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (pp. 1-2).
- King, R., Churchill, E. F., & Tan, C. (2017). Designing with data: Improving the user experience with A/B testing. " O'Reilly Media, Inc.".
- Siroker, D., & Koomen, P. (2013). A/B testing: The most powerful way to turn clicks into customers. John Wiley & Sons.
- Oskarsdottir, E. M. (2016). Towards a Data-Driven Pricing Decision With the Help of A/B Testing.
- Deng, A., & Shi, X. (2016, August). Data-driven metric development for online controlled experiments: Seven lessons learned. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 77-86).
- https://www.veribilimiokulu.com/bootcamp-programlari/veri-bilimci-yetistirme-programi/
- A/B Testing Udacity
- https://www.statisticshowto.com/assumption-of-normality-test/


thank you
merci
gracias
danke
謝謝
ngiyabonga
teşekkür ederim
obrigado
obrigada
dziękuję
bedankt
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
dakujem
arigatô
grazie
tapadh leat
спасибо
рахмат
Баярлалаа
faafetai lava
kiitos
dankie
dhanyavad
hvala
mauruuru
köszönöm
gràcie
bayarlaa
enkosi
nanni
nandri
vinaka
спасибі
blagodaram
mersi
barka
welalin
tack
spas
kia ora
dank je
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
akun
dankon
ačiū
chnorakaloutioun
gratias ago
gràcies
sulpáy
tack
takk
trugarez
sagolun
sobodi
děkuji
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
rahmet
rahmat
kam sah hamnida
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercè
мерси