GloBox Data Report

Introduction:

This report contains important statistical analysis and A/B testing for GloBox data. The report contains the statistical analysis regarding both control and treatment group. GloBox is a e-commerce website looking for growth in sale and revenue in its food and drink category by launching a banner regarding this category on its home page to its mobile user. Before launching the new homepage Globox has done an experiment to ensure that if it really helping them in growth of their business. Globox has divided its users to Control and Treatment group that is A and B respectively. So, the purpose of this project is to do statistical analysis and Hypothesis testing on both groups and decide whether GloBox should launch the new homepage to all its users.

Statistical Analysis methods and tools used in this project:

SQL and Ms-excel for data download:

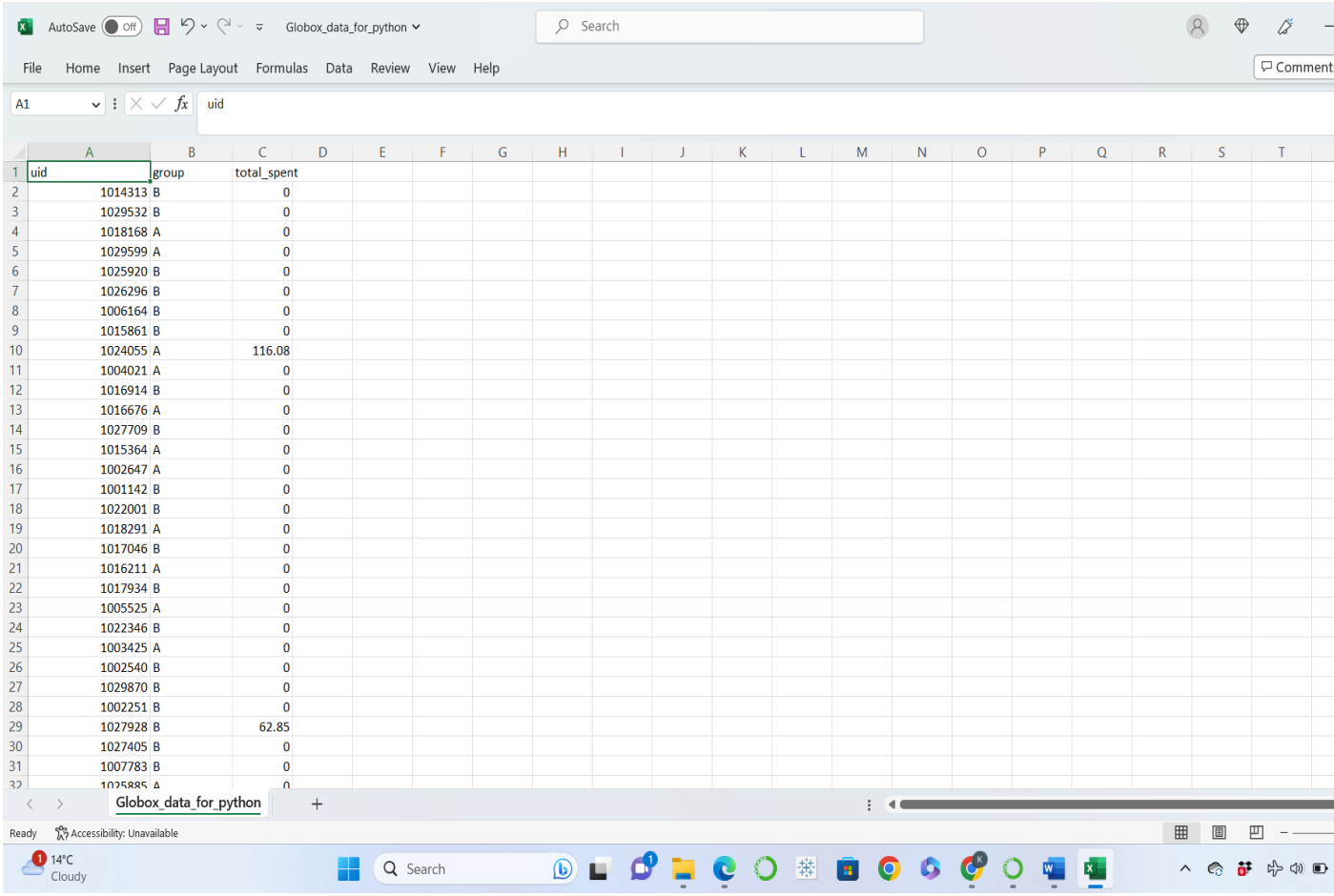
I have downloaded the Globox data using following SQL query and later save into excel worksheet as Globox data file for further analysis. I have only selected uid, spent and group column for our analysis. I also aggregated the spent column for retrieving all unique user in our experiment and replacing all the null values with 0.

```
WITH globox AS
 (
SELECT uid, "group", SUM(spent) AS total_spent
FROM groups
left join activity
using(uid)
GROUP BY uid,"group"
),

globox1 as
(select    uid,    "group",    (COALESCE(total_spent,    0))
total_spent from globox)

select * from globox1
```

After fetching all the data from GlobBox database I saved into excel csv file as:



uid	group	total_spent
1014313	B	0
1029532	B	0
1018168	A	0
1029599	A	0
1025920	B	0
1026296	B	0
1006164	B	0
1015861	B	0
1024055	A	116.08
1004021	A	0
1016914	B	0
1016676	A	0
1027709	B	0
1015364	A	0
1002647	A	0
1001142	B	0
1022001	B	0
1018291	A	0
1017046	B	0
1016211	A	0
1017934	B	0
1005525	A	0
1022346	B	0
1003425	A	0
1002540	B	0
1029870	B	0
1002251	B	0
1027928	B	62.85
1027405	B	0
1007783	B	0
1025885	A	0

Statistical Analysis and A/B testing Using python.

I used the following python code for Statical analysis on Total Spent and Conversion rate for both control and treatment groups.

The First step is to import all the required library and methods for our code. Then store the globox data csv file into pandas data frame and named as globox_data.

Home Page - Select o | Globox_Data_File - Ju | about:blank | Globox_Data_File | about:blank | Globox_Data_File - Ju

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

GloBox Data analysis and A/B testing For Control and Treatment Group

In [137]:
```
#importing libraries
```

In [138]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as st
from scipy.stats import ttest_ind
```

In [139]:
```
# Downloading Globox
```

In [140]:
```
globox_data = pd.read_csv('Globox_data_for_python.csv')
```

In [141]:
```
globox_data
```

Out[141]:

	uid	group	total_spent
0	1014313	B	0.00
1	1029532	B	0.00
2	1018168	A	0.00
3	1029599	A	0.00
4	1025920	B	0.00
...	...	...	...
48938	1042801	A	0.00
48939	1045885	A	102.81
48940	1045538	A	129.67
48941	1046699	B	0.00
48942	1030280	B	0.00

48943 rows × 3 columns

14°C Cloudy | Search

The next step is to start analysis on total_Spent column of dataframe and storing total_spent separately by two groups that is A and B.

Home Page - Select or | Globox_Data_File - Ju | about:blank | Globox_Data_File | about:blank | Globox_Data_File - Ju

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

48943 rows × 3 columns

Total Spent Analysis

In [142]:
```
Spent_A = globox_data.query("group=='A'")["total_spent"]
```

In [143]:
```
Spent_A
```

Out[143]:
```
2          0.00
3          0.00
8        116.08
9          0.00
11         0.00
           ...
48932      0.00
48933      0.00
48938      0.00
48939    102.81
48940    129.67
Name: total_spent, Length: 24343, dtype: float64
```

In [144]:
```
Spent_B = globox_data.query("group=='B'")["total_spent"]
```

In [145]:
```
Spent_B
```

Out[145]:
```
0        0.0
1        0.0
4        0.0
5        0.0
6        0.0
         ...
48935    0.0
48936    0.0
48937    0.0
48941    0.0
48942    0.0
Name: total_spent, Length: 24600, dtype: float64
```

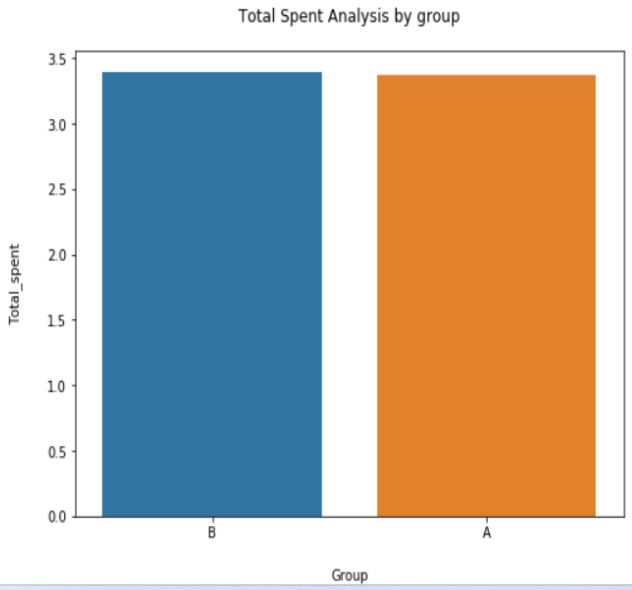
14°C Cloudy | Search

Plotting total spent by both control and treatment group.

```
...
48935    0.0
48936    0.0
48937    0.0
48941    0.0
48942    0.0
Name: total_spent, Length: 24600, dtype: float64
```

Plotting total spent by group

In [210]:
```
plt.figure(figsize=(8,6))
x=globox_data['group']
y=globox_data['total_spent']
sns.barplot(x, y, ci=False)
plt.title('Total Spent Analysis by group', pad=20)
plt.xlabel('Group', labelpad=20)
plt.ylabel('Total_spent', labelpad=20);
```



Calculating average spent per user and standard deviation by both groups.

Average spent by control group is 3.374518

Average spent by treatment group is 3.390667

Standard deviation for A is 25.936391

Standard deviation for B is 25.414110

Also starting A/B testing or hypothesis testing by stating our null and alternate hypothesis.

Hypothesis Test, to test that there is difference between mean total spent per user in Control and Treatment groups.

Null Hypothesis: H0 : There is no significant difference between mean total spent by per user in Control and Treatment groups.

Alternate Hypothesis: H1 : There is a significant difference between mean total spent by per user in Control and Treatment groups.

Home Page - Select or | Globox_Data_File - Jup | about:blank | Globox_Data_File | about:blank | Globox_Data_File - Jup

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

```
In [147]: avg_spent_per_user_A_B = globox_data.groupby('group')['total_spent'].mean()

In [148]: avg_spent_per_user_A_B

Out[148]: group
          A    3.374518
          B    3.390867
          Name: total_spent, dtype: float64

In [111]: stdev_spent_per_user_A_B = globox_data.groupby('group')['total_spent'].std()

In [112]: stdev_spent_per_user_A_B

Out[112]: group
          A    25.936391
          B    25.414110
          Name: total_spent, dtype: float64

In [113]: # A/B testing on total_spent for control and treatment group
```

Hypothesis Test, to test that there is diffrence between mean total spent per user and Treatment groups

Null Hypothesis: H0 : There is no significant diffrence between mean total spent b Control and Treatment groups

Alternate Hypothesis: H1 : There is a significant diffrence between mean total spe user in Control and Treatment groups

```
In [114]: #Seperating data into two groups

In [115]: Group_A = globox_data[globox_data.group=='A']
          Group_B = globox_data[globox_data.group=='B']

In [116]: Group_A
```

14°C Cloudy | Search

Separating data into two groups.

Home Page - Select or | Globox_Data_File - Jup | about:blank | Globox_Data_File | about:blank | Globox_Data_File - Jup

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

In [114]:
```
#Seperating data into two groups
```

In [115]:
```
Group_A = globox_data[globox_data.group=='A']
Group_B = globox_data[globox_data.group=='B']
```

In [116]:
```
Group_A
```

Out[116]:

	uid	group	total_spent
2	1018168	A	0.00
3	1029599	A	0.00
8	1024055	A	116.08
9	1004021	A	0.00
11	1016676	A	0.00
...	...	...	...
48932	1038187	A	0.00
48933	1048845	A	0.00
48938	1042801	A	0.00
48939	1045885	A	102.81
48940	1045538	A	129.67

24343 rows × 3 columns

In [117]:
```
Group_B
```

Out[117]:

	uid	group	total_spent
0	1014313	B	0.0
1	1029532	B	0.0
4	1025920	B	0.0
5	1026296	B	0.0
6	1006164	B	0.0
...	...	...	...

14°C Cloudy | Search

T-test done for determining our pvalue for the test which is 0.944 approx.

A/B testing Result:

For 5% significance level pvalue is greater than 0.05 that is 0.944 approximately, we fail to reject the null hypothesis that there is no difference in mean amount spent per user between the control and treatment group.

Home Page - Select o | Globox_Data_File - Ju | about:blank | Globox_Data_File | about:blank | Globox_Data_File - Ju

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

24343 rows × 3 columns

In [117]:
```
Group_B
```

Out[117]:

	uid	group	total_spent
0	1014313	B	0.0
1	1029532	B	0.0
4	1025920	B	0.0
5	1026296	B	0.0
6	1006164	B	0.0
...	...	...	...
48935	1033944	B	0.0
48936	1044676	B	0.0
48937	1044260	B	0.0
48941	1046699	B	0.0
48942	1030280	B	0.0

24600 rows × 3 columns

In [118]:
```
# T-test for t-statistics and determining pvalue
```

In [119]:
```
ttest_ind(Group_A.total_spent,Group_B.total_spent,equal_var=False)
```

Out[119]:
```
Ttest_indResult(statistic=-0.07042491002232797, pvalue=0.9438557531531265)
```

A/B testing Result

pvalue is greater than 0.05 that is 0.944 approximately, We fail to reject the null h there is no diffrence in mean amount spent per user between the control and tre group.

14°C Cloudy | Search

Calculating confidence interval for A and B group with 95% confidence interval for average amount spent.

Confidence interval for A = (3.049,3.7)

Confidence interval for B = (3.07,3.7)

Home Page - Select or | Globox_Data_File - Jup | about:blank | Globox_Data_File | about:blank | Globox_Data_File - Jup

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

48935	1033944	B	0.0
48936	1044676	B	0.0
48937	1044260	B	0.0
48941	1046699	B	0.0
48942	1030280	B	0.0

24600 rows × 3 columns

In [118]:
```
# T-test for t-statistics and determining pvalue
```

In [119]:
```
ttest_ind(Group_A.total_spent,Group_B.total_spent,equal_var=False)
```

Out[119]:
```
Ttest_indResult(statistic=-0.07042491002232797, pvalue=0.9438557531531265)
```

A/B testing Result

pvalue is greater than 0.05 that is 0.944 approximately, We fail to reject the null hy there is no diffrence in mean amount spent per user between the control and trea group.

Confidence interval Calculation

In [120]:
```
Confidence_interval_A = st.t.interval(alpha=0.95, df=len(Spent_A)-1, loc=np.mean(Spent_A), scale=st.sem(Spent_
```

In [121]:
```
Confidence_interval_A
```

Out[121]:
```
(3.0486876385889445, 3.7003492972709022)
```

In [122]:
```
Confidence_interval_B = st.t.interval(alpha=0.95, df=len(Spent_B)-1, loc=np.mean(Spent_B), scale=st.sem(Spent_
```

In [123]:
```
Confidence_interval_B
```

Out[123]:
```
(3.073269643161592, 3.708464248623692)
```

14°C Cloudy | Search

Now the second statistical testing about another important metric that is conversion rate for both control and treatment group.

First, we count all the users in control and treatment group.

```
group
A      24343
B      24600
```

Then we count all the unique user from both the groups who converted, which means they spent at least once for purchasing required product from the website.

```
group
```

```
Converted_A     955
Converted_B     1139
```

Now we can Start our hypothesis testing on conversion rate.

Home Page - Select o | Globox_Data_File - Jup | about:blank | Globox_Data_File | about:blank | Globox_Data_File -

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

Conversion Rate Analysis

In [124]:
```
count_of_user_A_B = globox_data.groupby('group')['uid'].count()
```

In [125]:
```
count_of_user_A_B
```

Out[125]:
```
group
A    24343
B    24600
Name: uid, dtype: int64
```

In [126]:
```
Converted_A = globox_data.query('group=="A" & total_spent!= 0 ')
```

In [127]:
```
Converted_A
```

Out[127]:

	uid	group	total_spent
8	1024055	A	116.08
72	1004755	A	77.39
86	1027437	A	31.97
178	1017801	A	40.89
190	1004202	A	67.12
...	...	...	...
48848	1043070	A	89.09
48890	1033644	A	41.30
48926	1045701	A	29.12
48939	1045885	A	102.81
48940	1045538	A	129.67

955 rows × 3 columns

In [128]:
```
Converted_B = globox_data.query('group=="B" & total_spent!= 0 ')
```

In [129]:
```
Converted_B
```

14°C Cloudy | Search

First, we will determine our null and alternate hypothesis for conversion rate.

Hypothesis Test to determine that there is difference between conversion_rate by in Control and Treatment groups.

Null Hypothesis: H0 : There is no significant difference between conversion_rate in Control and Treatment groups.

Alternate Hypothesis: H1 : There is significant difference between conversion_rate in Control and Treatment groups.

Home Page - Select or | Globox_Data_File - Jup | about:blank | Globox_Data_File | about:blank | Globox_Data_File -

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

In [129]:
```
Converted_B
```

Out[129]:

	uid	group	total_spent
27	1027928	B	62.85
79	1017307	B	8.66
188	1001685	B	83.37
217	1007932	B	36.34
342	1028100	B	157.15
...	...	...	...
48774	1047062	B	60.40
48777	1042411	B	162.86
48805	1045330	B	114.48
48818	1041236	B	30.62
48861	1044992	B	139.33

1139 rows × 3 columns

Hypothesis Test to determine that there is diffrence between conversion_rate and Treatment groups

Null Hypothesis: H0 : There is no significant diffrence between conversion_ra Treatment groups

Alternate Hypothesis: H1 : There is significant diffrence between conversion_ and Treatment groups

In [130]:
```
Conversion_rate_A = (955/24343)*100
Conversion_rate_A
```

Out[130]: 3.9230990428459926

In [131]: Conversion_rate_B = (1139/24600)*100

14°C Cloudy | Search

Adding a new column 'converted' into our dataframe and given the value 1 and 0 to converted and non-converted users.

Now calculating conversion_rate, standard deviation, and standard error for both the control and treatment groups.

Home Page - Select or | Globox_Data_File - Jup | about:blank | Globox_Data_File | about:blank | Globox_Data_File

File | C:/Users/meena/Downloads/Globox_Data_File%20(1).html

In [131]:
```
Conversion_rate_B = (1139/24600)*100
Conversion_rate_B
```

Out[131]:
```
4.630081300813008
```

In [132]:
```
globox_data['converted'] = np.where(globox_data['total_spent'] != 0, 1, 0)
```

In [133]:
```
globox_data
```

Out[133]:

	uid	group	total_spent	converted
0	1014313	B	0.00	0
1	1029532	B	0.00	0
2	1018168	A	0.00	0
3	1029599	A	0.00	0
4	1025920	B	0.00	0
...	...	...	...	...
48938	1042801	A	0.00	0
48939	1045885	A	102.81	1
48940	1045538	A	129.67	1
48941	1046699	B	0.00	0
48942	1030280	B	0.00	0

48943 rows × 4 columns

In [134]:
```
conversion_rate = globox_data.groupby('group')['converted']

std_p = lambda x: np.std(x, ddof=0)              # Std. deviation of the proportion
se_p = lambda x: st.sem(x, ddof=0)            # Std. error of the proportion (std / sqrt(n))

conversion_rate = conversion_rate.agg([np.mean, std_p, se_p])
conversion_rate.columns = ['conversion_rate', 'std_deviation', 'std_error']


conversion_rate.style.format('{:.3f}')
```

14°C Cloudy | Search

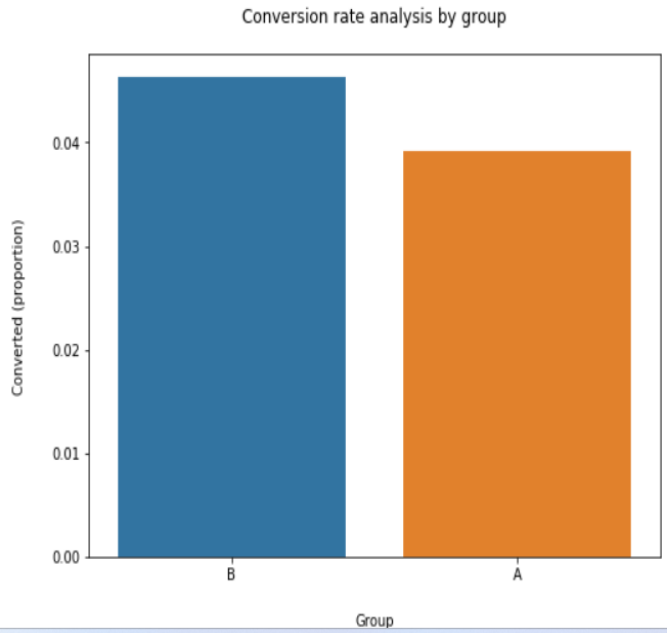
Plotting conversion rate for both control and treatment groups on bar-graph.

Out[134]:

group	conversion_rate	std_deviation	std_error
A	0.039	0.194	0.001
B	0.046	0.210	0.001

Plotting conversion rate for group A and B

In [135]:
```
plt.figure(figsize=(8,6))
x=globox_data['group']
y=globox_data['converted']
sns.barplot(x, y, ci=False)
plt.title('Conversion rate analysis by group', pad=20)
plt.xlabel('Group', labelpad=20)
plt.ylabel('Converted (proportion)', labelpad=20);
```



For hypothesis testing for conversion-rate we used z-test because if sample size is more than 30, we can use z-test.

Now we will calculate the pvalue, 95% confidence interval for both the group, and z-statistics for our A/B testing.

```
z statistic: -3.864
p-value: 0.0001
ci 95% for control group: [0.037, 0.042]
ci 95% for treatment group: [0.044, 0.049]
```

For 5% significance level, p = 0.0001, statistically significant which is less than 0.05. So, we can reject the null hypothesis that there is no difference in the user conversion rate between the control and treatment.

In [188]:
```
#Z-test for z-statistics and determining p-value and confidence-interval
```

In [189]:
```
from statsmodels.stats.proportion import proportions_ztest, proportion_confint
A_converted = globox_data[globox_data['group'] == 'A']['converted']
B_converted = globox_data[globox_data['group'] == 'B']['converted']
n_A = A_converted.count()
n_B = B_converted.count()
total = [A_converted.sum(), B_converted.sum()]
n_AB = [n_A, n_B]

z_stat, pval = proportions_ztest(total, nobs = n_AB)
(lower_con, lower_treat), (upper_con, upper_treat) = proportion_confint(total, nobs=n_AB, alpha=0.05)

print(f'z statistic: {z_stat:.3f}')
print(f'p-value: {pval:.4f}')
print(f'ci 95% for control group: [{lower_con:.3f}, {upper_con:.3f}]')
print(f'ci 95% for treatment group: [{lower_treat:.3f}, {upper_treat:.3f}]')
```

```
z statistic: -3.864
p-value: 0.0001
ci 95% for control group: [0.037, 0.042]
ci 95% for treatment group: [0.044, 0.049]
```

p = 0.0001, statistically significant. We can reject the null hypothesis that there i in the user conversion rate between the control and treatment.

END

In [ ]:

Summary of the report:

For all the above statistical analysis we can say that in respect of mean total_spent by both the control and treatment groups there is no significance difference between both the groups. Although we can see that the conversion rate in treatment group is significantly higher that control group. So, we can conclude that treatment group are more responding to our GloBox new launch banner on homepage for food and drinks category. And growth in conversion rate really effect in our revenue and sales positively and future growth of business. So, I am very positive about that we can launch the new webpage to all the GloBox customer for future growth of business.