University of Exeter coursework header sheet

Extension
211433

BEMM460J Statistics and Mathematics for Business Analytics (A, TERM2 2021/2)

1072885

720005065

Coursework: Individual report

Submission Deadline: Fri 13th May 2022 12:00

Personal tutor: Justin Tumlinson

Marker name: N/A

Word count: 3000

By submitting coursework you declare that you understand and consent to the University policies regarding plagiarism and mitigation (these can be seen online at www.exeter.ac.uk/plagiarism, and www.exeter.ac.uk/mitigation respectively), and that you have read your school's rules for submission of written coursework, for example rules on maximum and minimum number of words. Indicative/first marks are provisional only.

2022

Individual Report BEMM460J: Statistics and Mathematics for Business Analytics

MADE BY
ANAND PHADTARE

UNIVERSITY OF EXETER | ap1001@exeter.ac.uk

Contents

Executive Summary

This report provides an exploratory and descriptive analysis of the data obtained from an online tea retail store. The data of this store is obtained from a website named Kaggle.com which shares publicly available data for commercial and non-commercial purposes. The tools used for analysis of the data are: Microsoft Excel, Atom application (for Python programming) and Tableu application.
The report focuses on analyzing data exploratory level to understand the data and its limitations. To get statistically significant results from the dataset under consider, data preparation involved, cleaning the data, and manipulating it by treating Null values and outliers from the dataset.
Further analysis of this clean data was done using visualizations and statistical analysis, for comparing, summarizing, and interpreting patterns from the dataset. The report also focuses on email marketing, retention results and average number orders supplied to conclude final observations and provide recommendations for the business.

Introduction

The boom of internet has changed definition of shopping from its traditional methods. Online shopping had a slow and unsystematic start in India since access to internet in early years of rise was relatively slow and many potential customers of a business were not aware of use of internet shopping. Along with this, most customers are not ready to take risk of buying a product without examining it physically.

Traditionally, Indian people have always had a conservative approach when it comes to purchasing any product. Slowly but steadily Indian market started accepting the new way of shopping since it offered benefits in attractive fashion compared to traditional way of shopping. The shopping was made convenient, way faster and even cheaper in most cases. Being the second most populated country in the world and having an E-commerce market worth of about $3.9 billion in 2009(as per ‘India Goes Digital), the Indian market was a prime target for a lot of businesses running not just inside India, but also outside of it. According to report of an Internet and mobile association (IAMAI), e-commerce in India altogether has an astonishing record of Compound annual growth rate (CAGR) of 54.6%, which crossed $10.0 billion in 2007-2011. Earlier reach of a retail store would be limited to its neighboring areas or even cities in exceedingly rare cases. But online shopping helped in expanding this reach to such an extent, which no one would have ever considered trying to. Launch of Flipkart in 2007 amazed every business owner and introduced them to the most efficient way of shopping ever existed. Although it was convenient, the online shopping was a huge threat to numerous small-scale businesses in India. And most of these businesses had to enter the world of online shopping to start competing in the race against big online companies.

In this study, we will conduct exploratory data analysis for a dataset obtained from an online tea retail store, which sells tea of various flavors across four major cities in India. The dataset consists of data about the customers of business, average orders of each customer, their order frequency, favorite day of delivery, city where customers live in, etc. Along with these variables, the data set has a variable which stores data about, if a customer is retained or stopped buying products from the business. Email being one of the most advantageous tools of internet, the business used it to keep a close communication with their customers. Hence the dataset is also provided with data about number of emails sent, and open-rate and click-rate of those emails based on number of emails sent.

Objective of the study

- To perform exploratory data analysis on the dataset to analyze and identify any trends in the data
- Customer behavior analysis and how it is affecting the business under consideration
- Customer retention analysis to recommend improvement strategies

Data understanding and its limitations

About the dataset

The dataset was collected from kaggel.com which provides variety of publicly available dataset. Although this is a dataset of a business organization, any detailed or sensitive information of the business is not included in the data. The dataset is purely focused on various aspects of customer behavior without listing any personal details and limited demographics of the customers. The dataset consists of multiple categorical and quantitative variables. The timeline of sample data recorded is between 2008-2018. The dataset has 30801 rows with unique customer ids and 15 columns in the main datasheet, along with this there is another sheet named ‘data dictionary’ which provides the definition of variables such as following:

1. retained: 1, if customer is assumed to be active, 0 = otherwise
2. created: Date when the contact was created in the database - when the customer joined
3. firstorder: Date when the customer placed first order
4. lastorder: Date when the customer placed last order
5. esent: Number of emails sent
6. eopenrate: Number of emails opened divided by number of emails sent
7. eclickrate: Number of emails clicked divided by number of emails sent
8. avgorder: Average order size for the customer
9. ordfreq: Number of orders divided by customer tenure
10. favday: Customer's favorite delivery day
11. City: City where the customer resides in

Limitations

Even though the dataset is publicly available on Kaggle, there is no clear definition of the data if this is a sample data or the complete data of the business from which it is acquired. The definition of some variables is incomplete and insufficient when considering statistical analysis of the dataset. A proper definition would have enabled the study to provide with useful recommendations for improving overall business performance and/or retention strategies for the business. Ambiguity in the definition limits use of diagnostic analytics for the data under consideration.

Data Preparation

- The primary tool used to prepare both the dataset which was in '.xlsx' format: Microsoft Excel
- First step executed while preparing the datasets was cleaning the datasets to ensure smooth operations during data modelling phase of the report.
- Investigation was conducted to look for any missing or incorrect values to avoid any kind of obstructions for processes ahead in the report.

Data cleaning

- The dataset available from the source was already in a structured format along with well-defined column names and without presence of any extra space and/or incorrect characters.
- There were 20 missing values under columns: custid, created, firstorder, lastorder. Rows of these values had to be removed to avoid computational errors during data analysis.
- Columns associated with datetype data were untidy and had incorrect dates in some rows, which had to be fixed using filtering operations from Microsoft Excel

Data manipulation

- For ease of data analysis and visualization 1's and 0's from retained columns were converted to 'retained' and 'left' respectively by using 'if' function from Excel application

- A new column named: Active_days was created by calculating days between when customer id was created in the database system of the business and last order of that customer.
- After doing an exploratory analysis of the data using Atom (a python application), it was observed that there were many outliers in some of the columns.

Outlier detection and treatment by using Inter Quartile Range (IQR):

- Exploratory analysis of the data revealed that, the dataset had a lot of outliers, which can be seen from following figure.

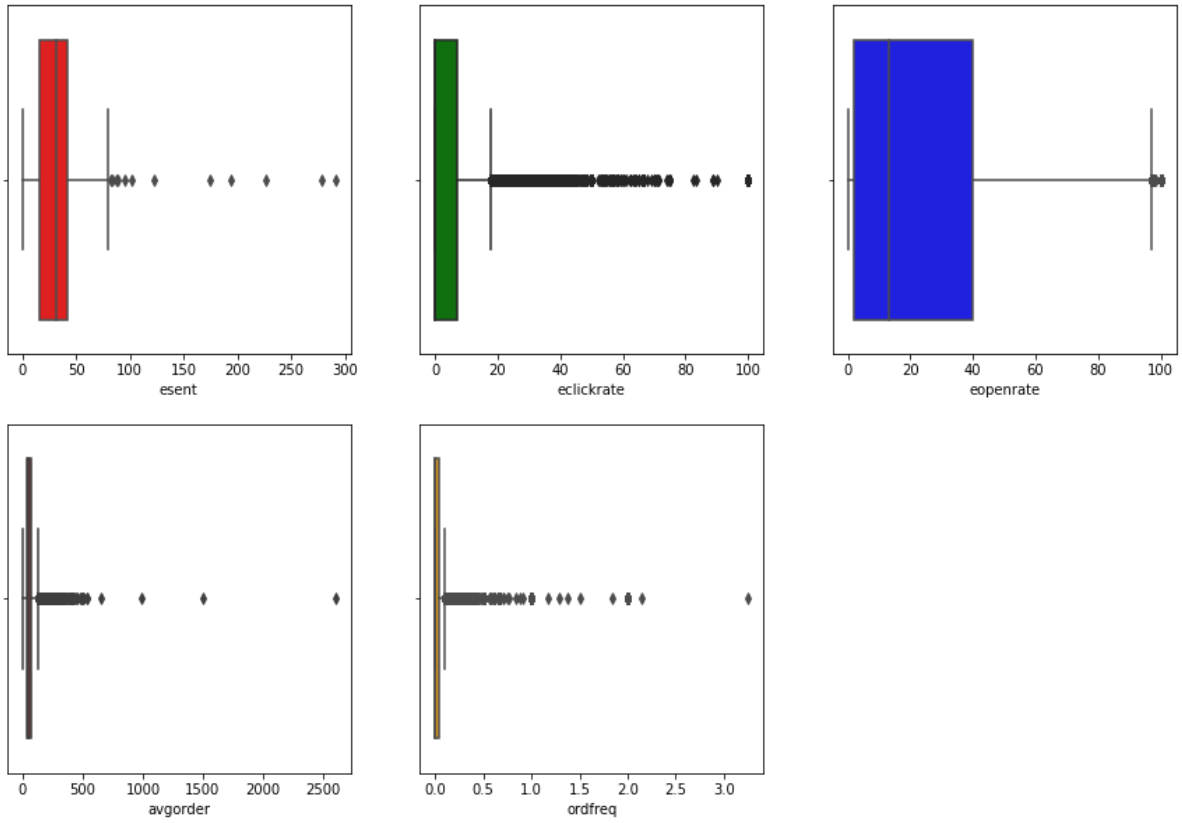


Figure No. (1)

- These many outliers in the data increases the variability in the data and diminished statistical power. To get statistically significant result in the study these outliers need to be treated. The outlier treatment was done using Inter-quartile range (IQR) method, which is a formula for measure of variability of datapoints in a dataset for a particular variable.
  Following diagram is a best example for representation of IQR using a boxplot():

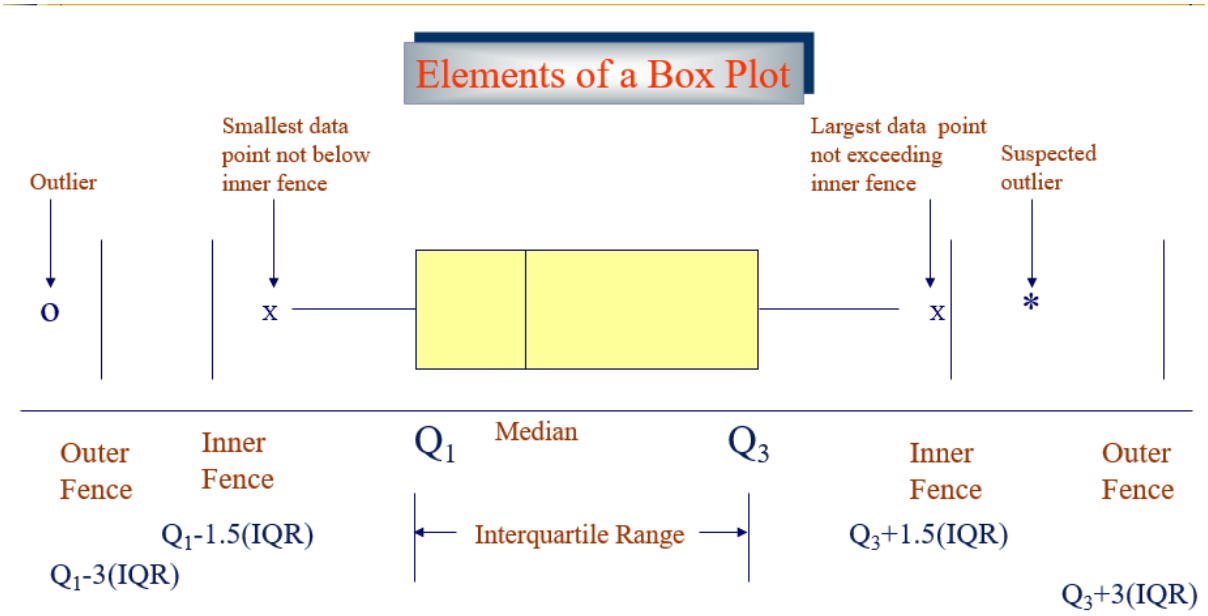


Figure No. (2)

To be able to do accurate statistical analysis of the data, these outliers were removed using python program which identifies outliers for variables from the figure 1 and removed them from the whole dataset along with rows associated for those values. (Code mentioned in the appendix)

Result of this outlier treatment method helped in getting fairely normal distribution of variables compared to their earlier version. Following figures shows the distribution of variables before and after the outlier removal:

Distribution with outliers:

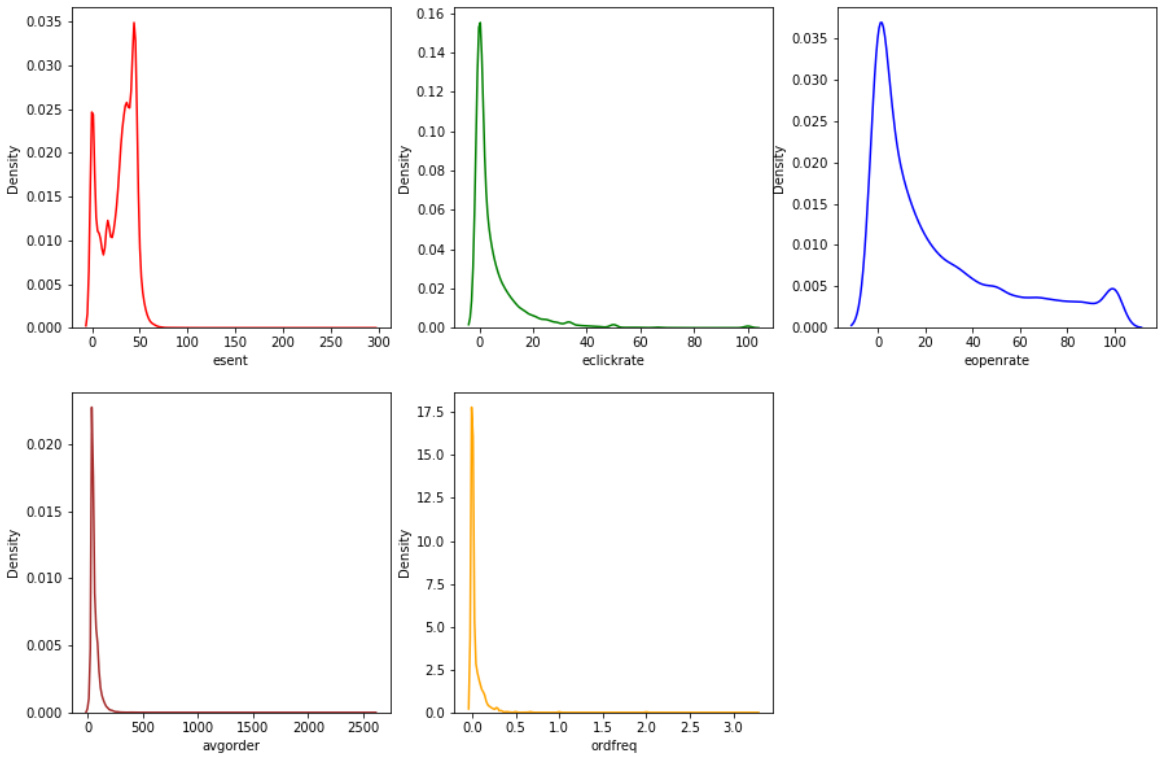


Figure No. (3)

Distribution after outlier removed:

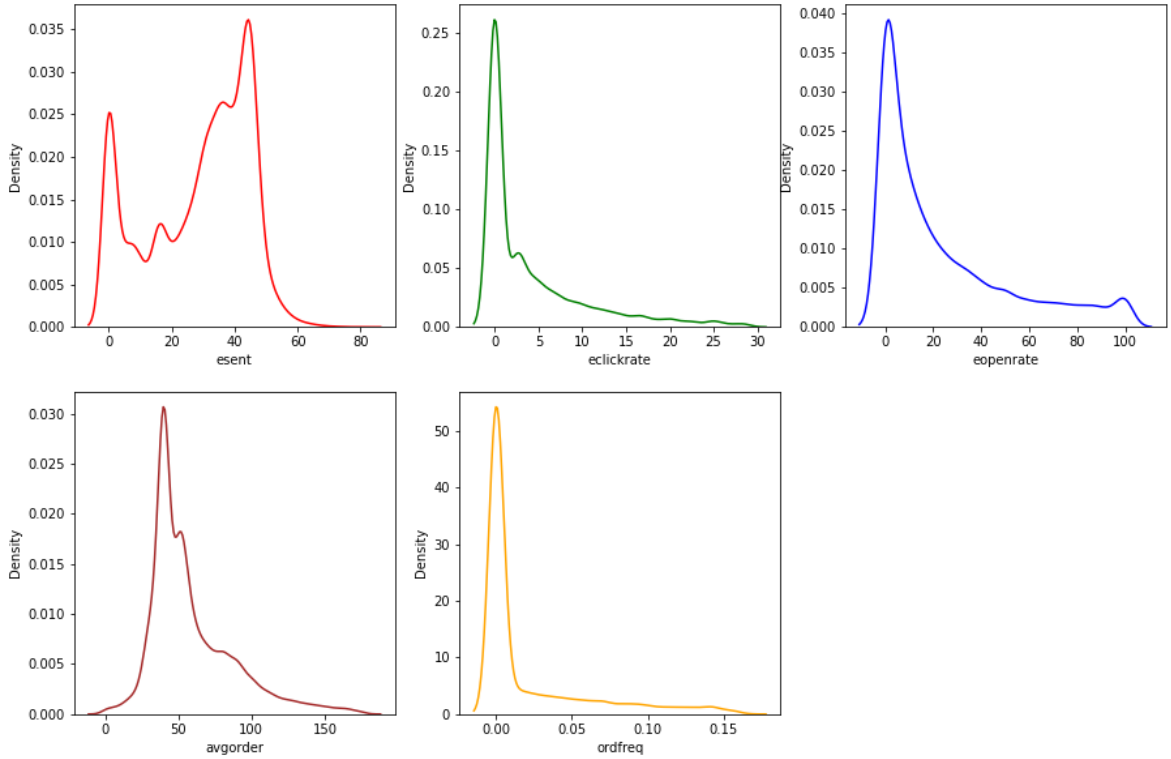


Figure No. (4)

- From above graphs we can see that all these variables have a non- normal distribution with skewness towards left, except for ‘esent’ variable.

Descriptive statistics:

Following table shows the values of Skewness and Kurtosis for all these variables:

	Email sent	Email open-rate	Email click-rate	Average Order	Order Frequency
Sknewss	-0.47	1.29	1.82	1.35	1.92
Kurtosis	-0.96	0.66	2.87	1.89	2.69

Table No. (1)

Based on above graph and table we can say that the data under consideration is not normally distributed. Left-leaning curve indicates a positive skewness while right-leaning curve indicates negative skewness

In order to reject the Null hypothesis that, the data is non-normally distributed, we have also calculated p-value for these variables using Shaprio-wilk test, by using a function in python, following figure shows the result of this test:

```
#Normality Check

st.shapiro(df_cleaned['esent']) ShapiroResult(statistic=0.9164579510688782, pvalue=0.0)
st.shapiro(df_cleaned['eclickrate']) ShapiroResult(statistic=0.7110521793365479, pvalue=0.0)
st.shapiro(df_cleaned['eopenrate']) ShapiroResult(statistic=0.804276704788208, pvalue=0.0)
st.shapiro(df_cleaned['avgorder']) ShapiroResult(statistic=0.8792672753334045, pvalue=0.0)
st.shapiro(df_cleaned['ordfreq']) ShapiroResult(statistic=0.6160522103309631, pvalue=0.0)
```

Table of measures of variability

Measures	esent	eopenrate	eclickrate	avgorder	ordfreq	Active_days
Median	33	11.9047619	0	50.73	0	16
Mean	28.54971163	23.45084989	4.067205297	58.81848344	0.020738311	132.631942
Standard Deviation	16.37660303	27.91086059	6.174244525	28.98549259	0.038112775	268.009896
Variance	268.1931268	779.016139	38.12129545	840.1587809	0.001452584	71829.3043
First quartile	16	0	0	40.02	0	0
Second quartile	33	11.9047619	0	50.73	0	16
Third quartile	42	35.71428571	6.060606061	72.22	0.025735294	109
Inter Quartile Range	26	35.71428571	6.060606061	32.2	0.025735294	109
Minimum	0	0	0	0	0	0
Maximum	80	100	28.57142857	176.84	0.163265306	1998
Range	80	100	28.57142857	176.84	0.163265306	1998

Table No. (2)
Note: The python is showing p-value is 0.0 since the actual p-value is extremely insignificant compared to alpha value: 0.05, but python rounds up the value to 0.0

Since p-value is 0.0 for all these variables, it means the tests are significant and the null hypothesis can be rejected.

Above table provides measures of variability of all quantitative variables from the dataset, which provides numeric information of distribution of the datapoints. From the table we can see minimum value of all the parameters are zero, same applies to median value of ‘eclickrate’ and ‘ordfreq’

Correlations between variables:

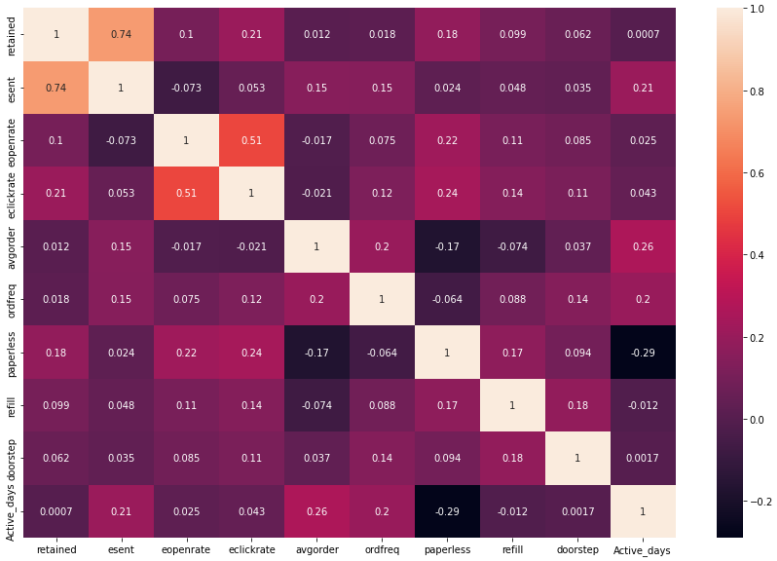


Figure No. (5)

Modelling and Analysis

Retention performance of the business

From this pie chart we can see that the business is able to retain 79.965 % of its customers, whereas 20.04% of customer chose to leave and stop purchasing tea from the business. After cleaning the customer count was reduced to 275690. The graph represents retention data of same count of customers.

The report will further study the behavior of customers who left as well as those who chose to continue placing orders from the business.

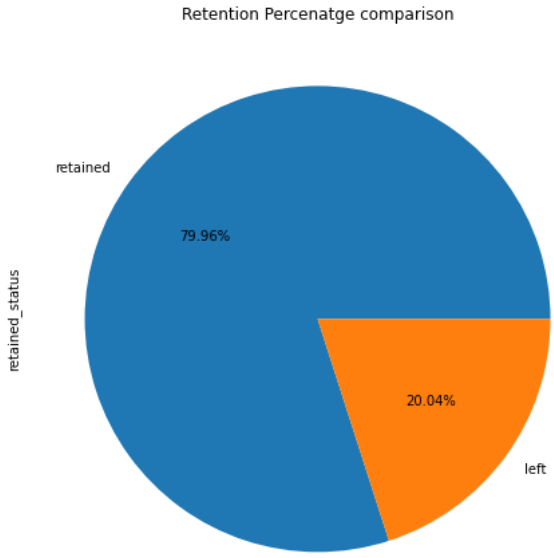


Figure No. (5)

Which days customers prefer to place an order for delivery:

From the bar graph it is clear that Monday, Tuesday are most preferred days having delivery requests around 6000 on each of these days, whereas Saturday and Sunday receive only very request around 1500 combining both days.

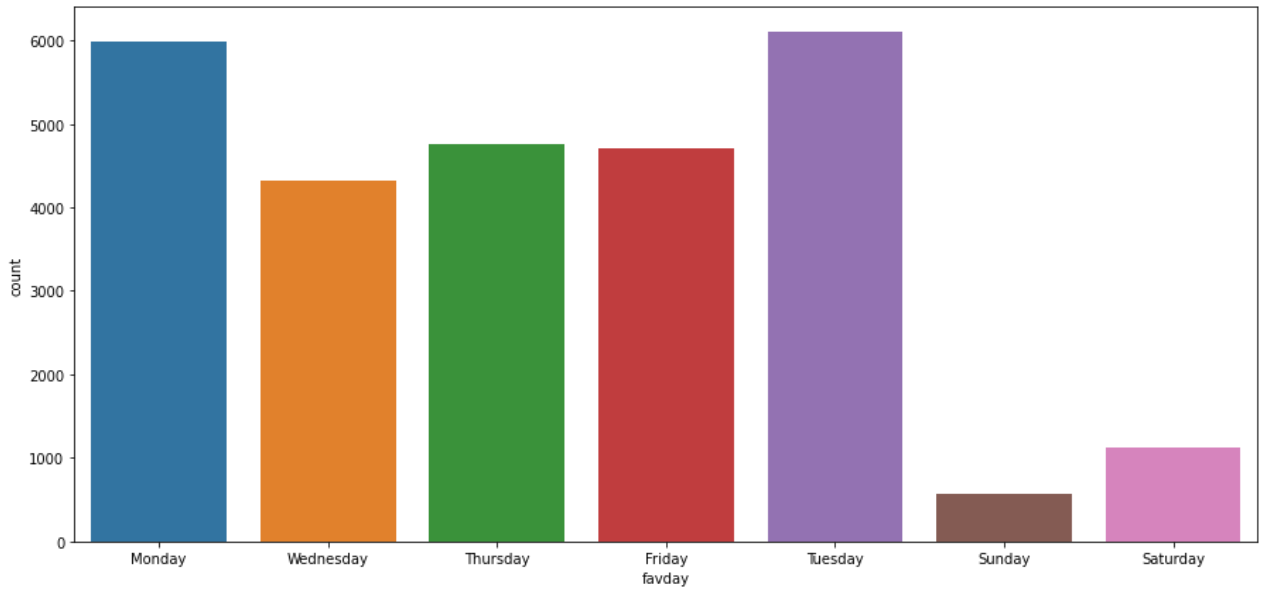


Figure No. (6)

Comparison of loyal vs customers who left based on the cities:

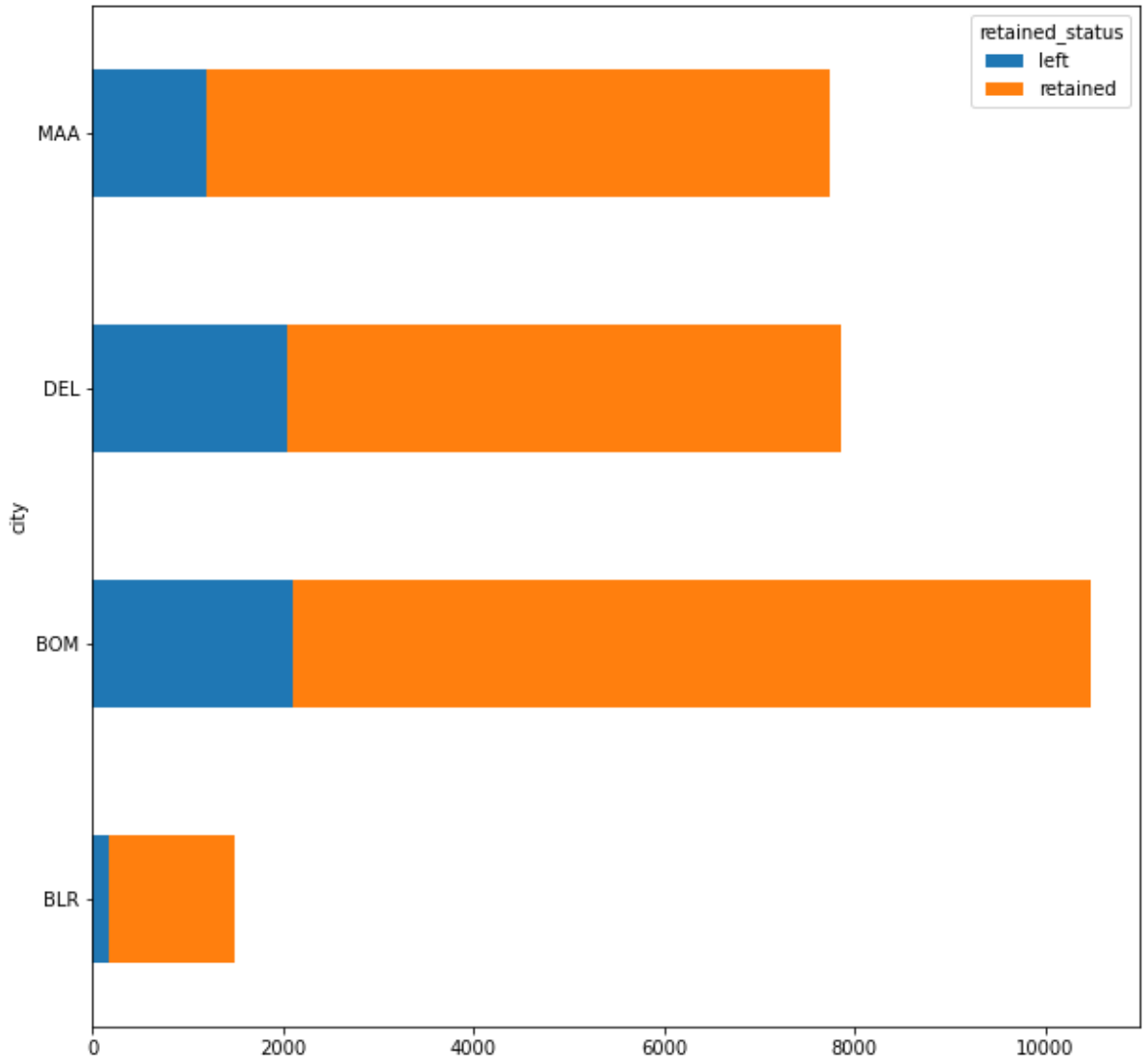
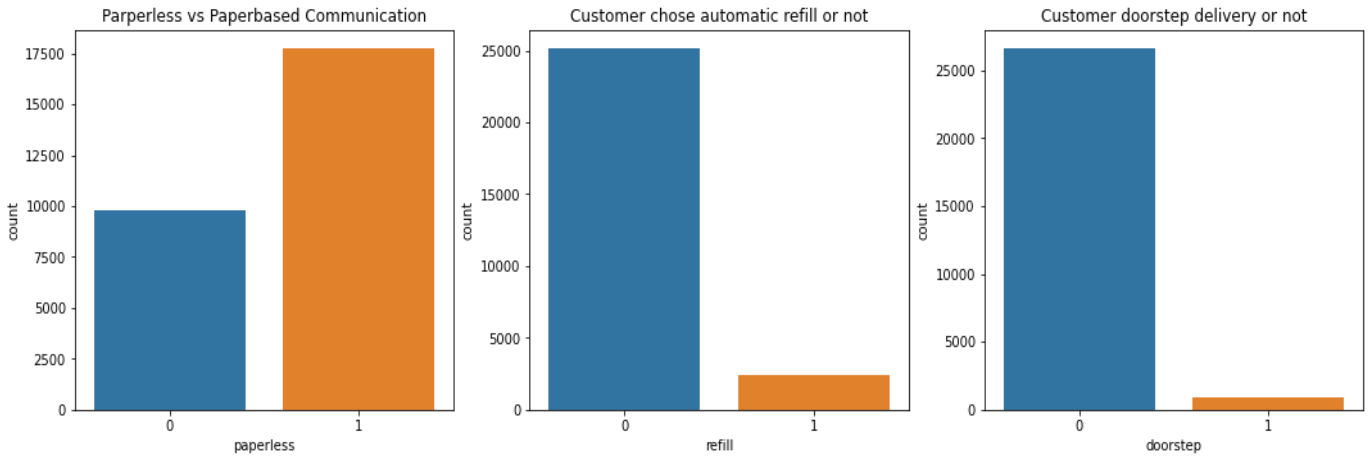


Figure No. (7)

From the above horizontal bar graph, we can clearly see that, highest number loyal or customers who chose to continue the purchase tea from the business resides in Mumbai city, but the most customer who stopped purchasing also stay in the same city. While the retention rate is good in both Mumbai and Chennai, same could not be said about Delhi; out of 7853 customers, 2037 customer stopped purchasing tea from the business. While Banglore has the smallest customer base compared to other cities, only 180 customers out of 1489 from the city of Banglore stopped business with the retail store.

Customers choices for additional services:



Note: 1, if customer subscribed, 0 if otherwise, *Figure No. (8)*

From above bar graphs we can see that, around 10000 customers have still not opted for paperless communication, which is about 58% of overall unique customers. The refill and doorstep service are least subscribed by only near 1000-2000 customers out of 275690.

11 years of business journey based on average of orders supplied

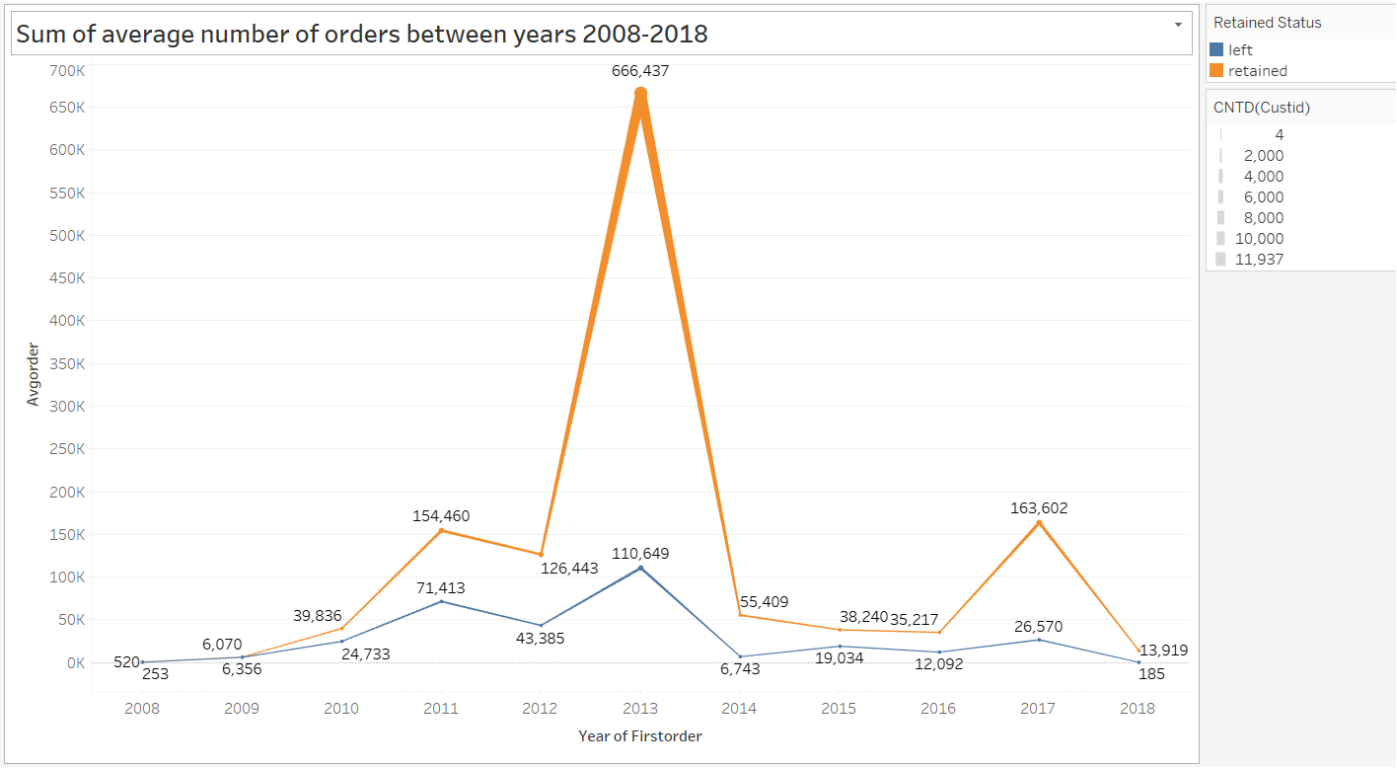


Figure No. (9)

From the timeline graph above we can see that there has been a dramatic increase the average number orders completed in the year 2013, and sell of that year must have contribute highest gross value compared to any other year or even combining of rest of the years from 2008-

2018.Morever, 2013 year had the highest number of active customers, but least number of customer left when considering proportion of overall customers. Conversely the drop in number orders in next year was equally dramatic as it falls by 83%, and the number keeps dropping till 2016, until the business made a come-back in 2017 by selling average number of orders which only slightly higher than that of year 2011. From year 2008 till 2011, the business had a steady increase the number of orders till the number drop in year 2012.

To understanding the incomparable surge of orders received in 2013, external research was conducted. According to an economic article from India-Times, tea production in India rose by 6.5% in year 2013 producing1200 million which was very compare all the previous years of tea production. Hypothetically if the overall production of tea in India could have been the reason for the huge surge in the average order supply in the business, a test need to be carried to determine the result of the hypothesis.

Regression Model to determine relationship between gross tea production in India versus average orders supplied by the business

To carry out this test, tea production data from year 2008-2017 was obtained from official website of : India Tea association. And by calculating sum of average number orders year wise following table was formulated for further testing:

	Y Dependent	X Independent
	Average Orders	Tea Production in KGs
2008	773.96	980,818
2009	12425.44	979,000
2010	64569.5	966,400
2011	225873.01	1,115,720
2012	169828.51	1,126,330
2013	777086.11	1,200,040
2014	62152.01	1,207,310
2015	57273.79	1,208,660
2016	47309.03	1,267,360
2017	190171.5	1,321,760

Table No. (3)

To check for relation between two variables a Regression model is suitable choice to evaluate the hypothesis. By using Regression function from the Microsoft Excel following results were obtained for the test:

Summary Report:

Regression Statistics	
Multiple R	0.299872718
R Square	0.089923647
Adjusted R Square	-0.023835897
Standard Error	232504.9615
Observations	10

Table No. (4)

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	42731734224	42731734224	0.790471232	0.399901312
Residual	8	4.32468E+11	54058557120		
Total	9	4.752E+11			

Table No. (5)

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 80.0%*	*Upper 80.0%*
Intercept	-458365.0072	700217.7958	-0.654603482	0.531088464	-2073070.14	1156340.125	-1436440	519709.9301
Tea Production in KGs	0.544350328	0.61225939	0.889084491	0.399901312	-0.867522357	1.956223013	-0.31086	1.399563618

Table No. (6)

And following are the regression graphs of same test:

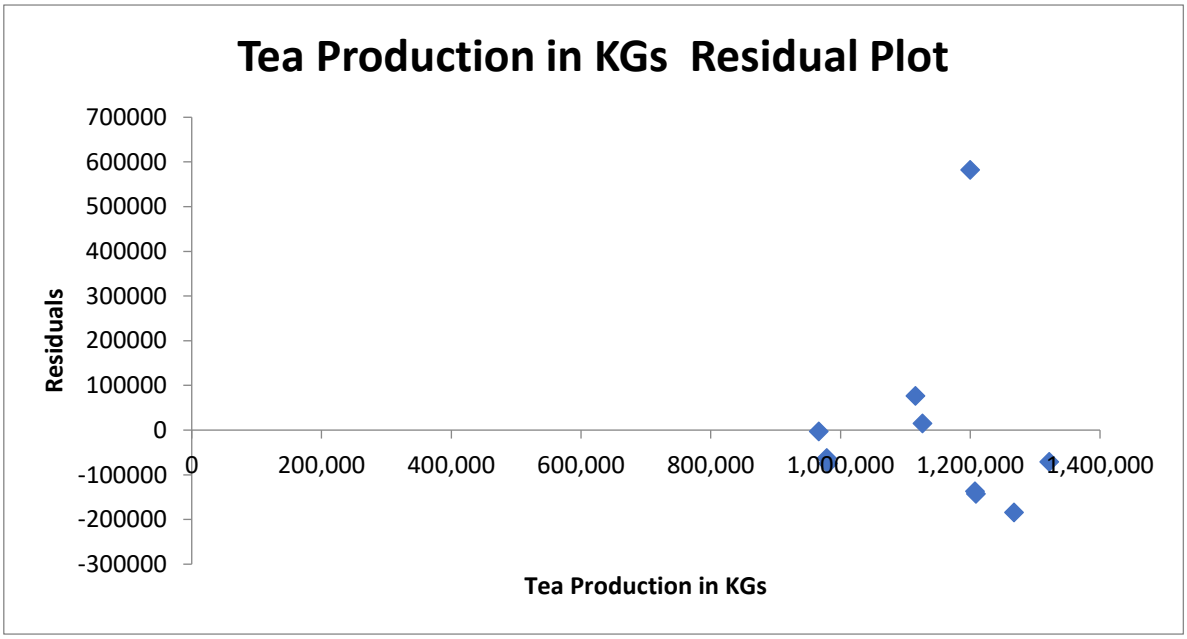


Figure No. (10)

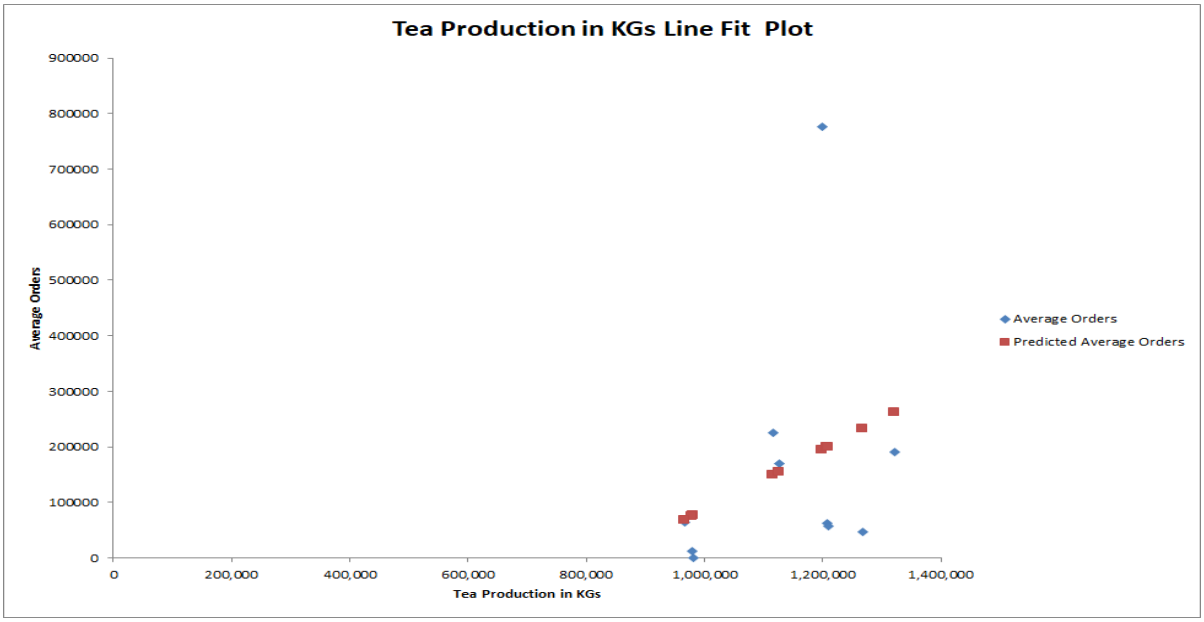
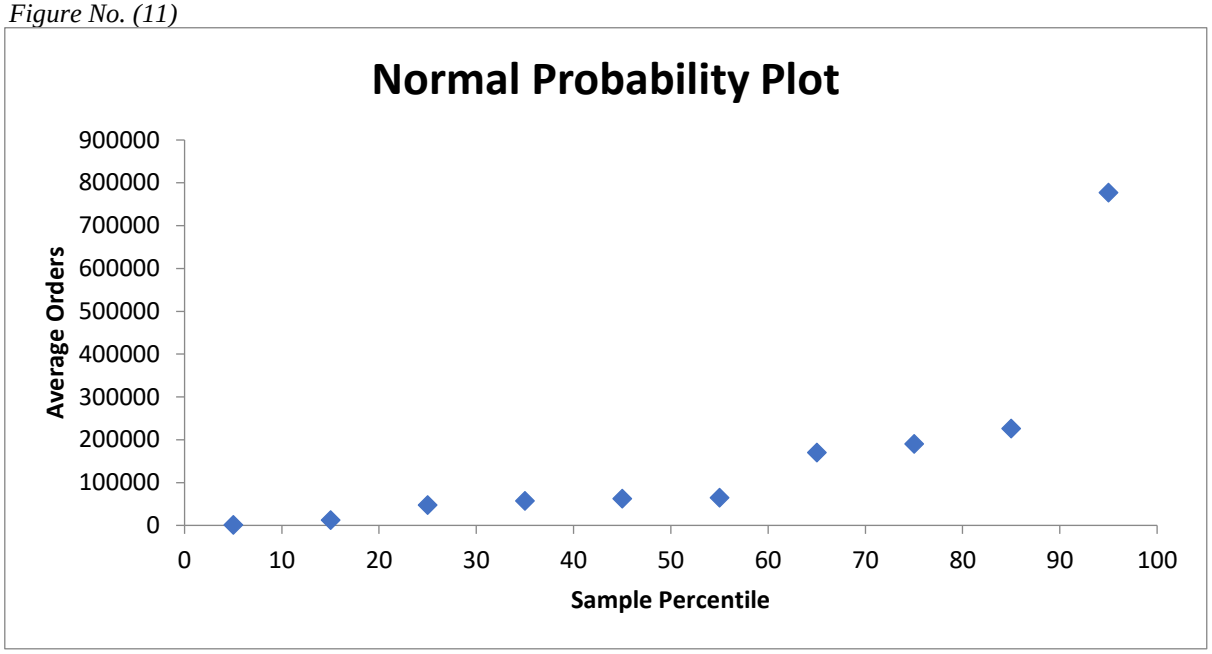


Figure No. (11)



Based on above results from Table No. 4 ,5 and 6 and graphs the hypothesis can be rejected, since the test is not statistically significant. The p-value (0.399) of the regression test is extremely high to consider any relationship between the two variables. Hence the conclusion of the test is that the overall amount of tea production in India has not affected the average number of orders supplied by the business.

Since a plausible external parameter is not affecting the average orders of the business, it is safe to assume that the surge and drop in orders could be caused due to an internal factor from the business. To further investigate multiple regression tests were conducted between combination of variables available in the datasets, but none of the test were able to produce any statistically significant results.

Analysis of Email marketing and retention of selective years:

	2011			2012			2013		
	Sent	**Open-rate**	**Click-rate**	**Sent**	**Open-rate**	**Click-rate**	**Sent**	**Open-rate**	**Click-rate**
count	3399	3399	3399	2394	2394	2394	14017	14017	14017
mean	**25.35628**	**13.841281**	**1.971215**	**31.165414**	**16.886204**	**2.335468**	**30.520297**	**26.26398**	**4.662552**
std	20.16284	23.7543	4.35558	19.381477	25.384762	4.747361	13.557901	27.81704	6.244199
min	0	0	0	0	0	0	0	0	0
25%	0	0	0	10	0	0	23	4.651163	0
50%	30	0	0	40	4.444444	0	33	15.90909	2.5
75%	45	17.021277	2.173913	45	23.313953	2.380952	41	39.47368	7.142857
max	70	100	28.571429	69	100	28.571429	80	100	28.57143

	2014			2017		
	Sent	**Open-rate**	**Click-rate**	**Sent**	**Open-rate**	**Click-rate**
count	1215	1215	1215	3442	3442	3442
mean	**14.77531**	**40.869791**	**8.579629**	**30.67606**	**25.638579**	**4.549036**
std	13.54034	32.543566	9.129919	13.51598	27.634355	6.215776
min	0	0	0	0	0	0
25%	7	12.5	0	24	4.444444	0
50%	9	36.111111	6.451613	33	15.384615	2.439024
75%	16	66.666667	16.666667	41	37.5	6.666667
max	58	100	28.571429	66	100	28.571429

Table No. (7)

Another probable reason for variations in the average number of orders could be the email marketing and/or the rate of retention, but since there is no strong evidence on the context of email marketing used to retain the customers, the prediction cannot be tested for significance due to ambiguity in the available data. But from above table we can see that, for year 2014, the mean value of email sent id lowest compared to other years but also has highest mean value of both 'Open-rate' and 'Click-rate'. Although the retention bar graph has a compelling similarity with the average orders graph, but that is since the proportion of customers is strongly related to average number of orders.

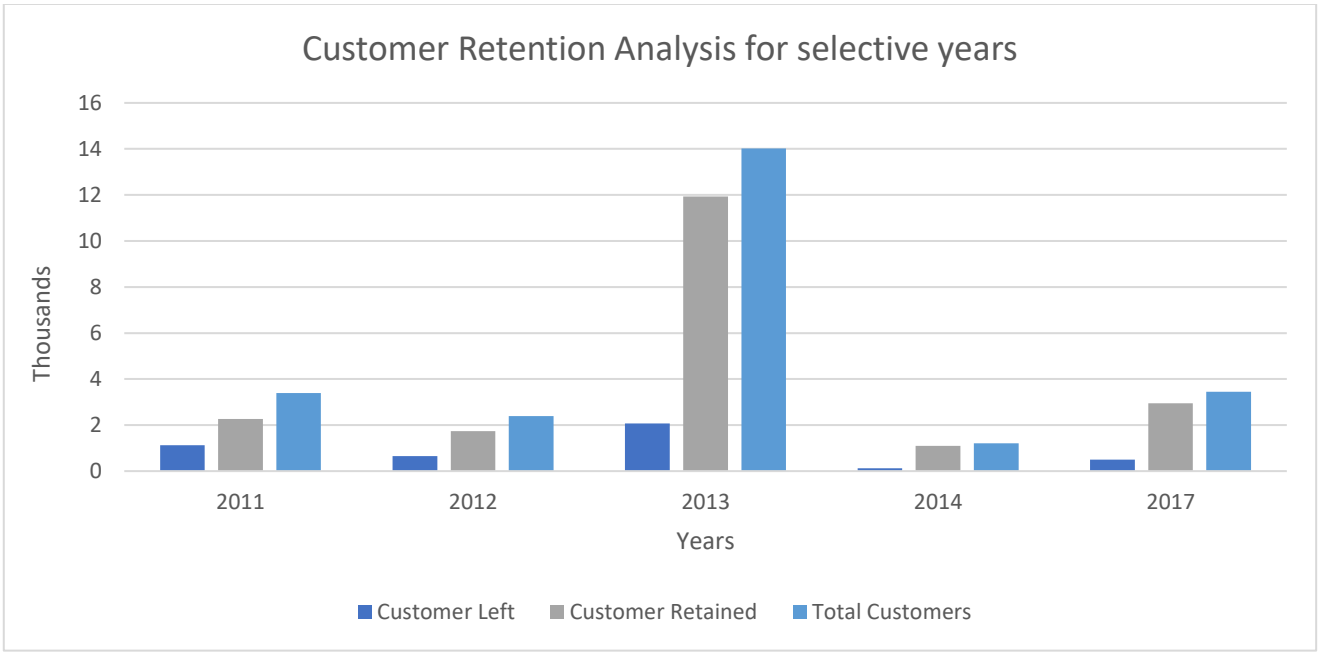


Figure No. (13)

Regression Test between Email Sent and Customer retained

Based on correlation metric, we could see a strong correlation between variables ‘esent’ and ‘retained’. We can evaluate this relation using regression method of following values from the table:

To conduct the test, sum of email sent in each is compared against number of customers retained in those same year.

	X	Y
Year	**esent**	**Retained**
2008	464	14
2009	4214	122
2010	23382	589
2011	86186	2273
2012	74610	1737
2013	427803	11937
2014	17952	1093
2015	22156	564
2016	20885	486
2017	105587	2941
2018	3848	284

Table No. (8)

Following graph shows the graph of predicted retention vs actual retention graph bases on number of emails sent. And we can see the correlation is extraordinarily strong between these variables.

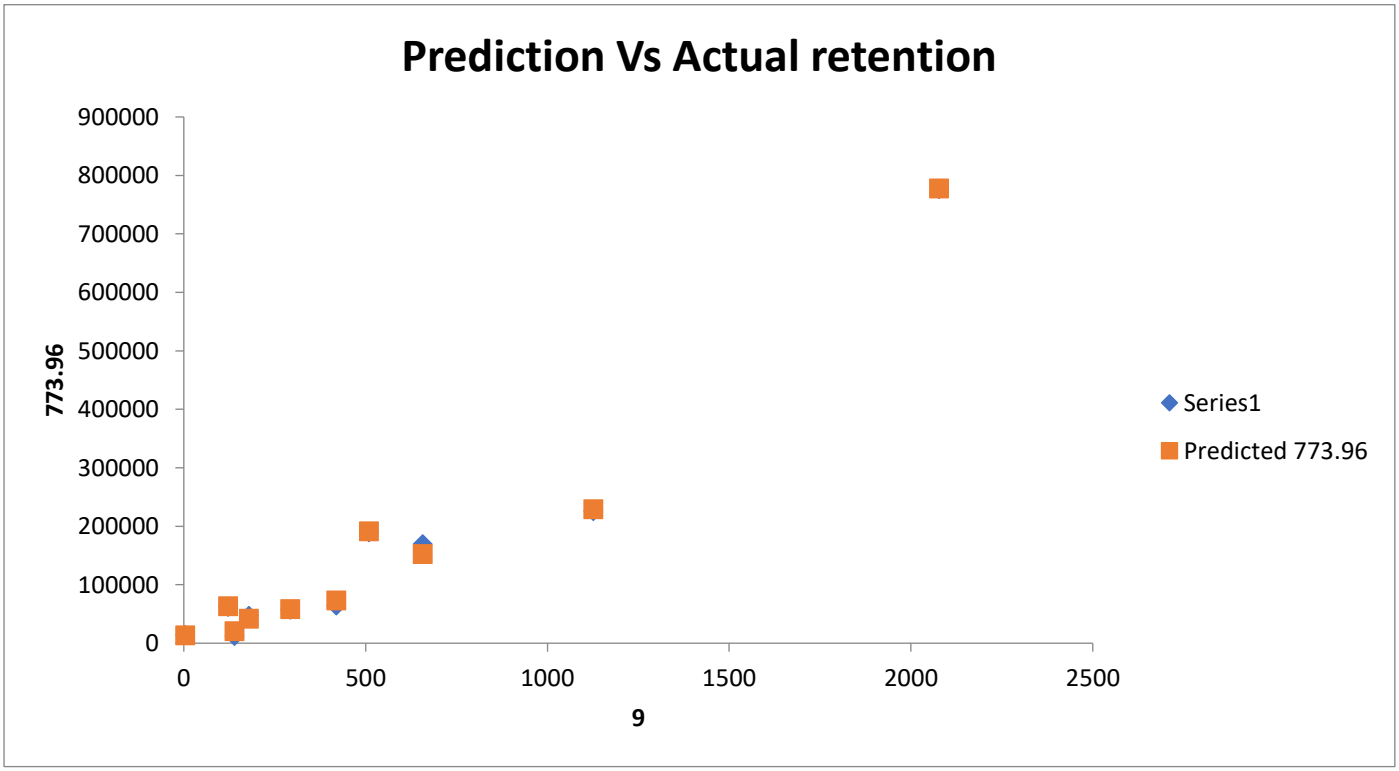


Figure No. (14)

Conclusion

From this study, it is clear that retention strategies of a business play huge role in creating a loyal and large customer base. The tea retail store has taken advantage of email communication to retain majority of their customers. But substantial portion of customers have not opted for paperless communication, refill, and doorstep services (*Figure No. (8*)). The proportion of customers retained, or left is uneven for different cities. Customers prefer to get delivery of the goods during weekdays as compared to weekends. An internal factor inside the business has left to drop in average number of orders served after year 2013.

Recommendations

- The business should consider doing a thorough analysis of email sent for different years, especially for email sent during year 2012-2013, to understand the retention strategy used in this year and how it could be used by revising it for coming years. Also investigate email marketing done after year 2013, which might reveal facts about drop in average orders.
- Due to ambiguity in the context of email, the study is unable to provide with accurate diagnostic analysis of the data. It is recommended to store a more descriptive email data consisting of highlights of marketing strategy used, word count in the email and number of responses received for marketing emails.
- Research should be conducted to understand and develop an email marketing strategy based on locations of the customer, since managing content of emails based on geographical location of the customers could be advantageous in retaining existing customers as well as attracting new ones from respective regions.
- Data analysis found presence of huge number of outliers; these need to be investigated for respective variables to understand their significance in the various operations in the business.

References

Customer retention retail. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/uttamp/store-data

Tea production rises 6.5 per cent in 2013: Indian tea association. (2014, February 2). *The Economic Times*. https://economictimes.indiatimes.com/news/economy/agriculture/tea-production-rises-6-5-per-cent-in-2013-indian-tea-association/articleshow/29769270.cms?from=mdr

Chronology of Indian Tea (Source : Tea Board of India). (n.d.). Home. https://www.indiatea.org/chronology_pdf/Chronology-of-IT.pdf

India Brand Equity Foundation. (n.d.). *Business Opportunities in India: Investment Ideas, Industry Research, Reports | IBEF*. https://www.ibef.org/download/The-Rise-and-Rise-of-E-commerce-in-India.pdf

E-Commerce in India. (2011, January 21). In *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/wiki/E-commerce_in_India#cite_note-9

Appendix:

1. Null values from the dataset:

```
df.isnull().sum() # We have 20 null customer IDs, lastorder  firstorder created
```

```
custid            20
retained           0
retained_status    0
created           20
firstorder        20
lastorder         20
esent              0
eopenrate          0
eclickrate         0
avgorder           0
ordfreq            0
paperless          0
refill             0
doorstep           0
favday             0
city               0
Active_days       20
dtype: int64
```

2. Dropping all null values:

```
df = df.dropna(how='any') # To drop all rows having ANY null values in them
```

3. Outlier detection using boxplot() method

```
#Using boxplot() to see distribution of the data and outliers
plt.figure(figsize=(15,10))
plt.subplot(2,3,1)
sns.boxplot(x='esent', data = df,color='red')
plt.subplot(2,3,2)
sns.boxplot(x='eclickrate', data = df,color='green')
plt.subplot(2,3,3)
sns.boxplot(x='eopenrate', data = df,color='blue')
plt.subplot(2,3,4)
sns.boxplot(x='avgorder', data = df,color='brown')
plt.subplot(2,3,5)
sns.boxplot(x='ordfreq', data = df,color='orange')
plt.show()
```

4. Outlier detection and treatment

```
#Outlier Treatment
#Extracting all outliers from all the columns and get their indices
def outliers(data, ft):
    Q1 = data[ft].quantile(0.25)
    Q3 = data[ft].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 3 * IQR
    upper_bound = Q3 + 3 * IQR


    list = data.index[(data[ft] < lower_bound) | (data[ft] >upper_bound)]
    return list

#Create an empty list to return output indices from multiple columns
index_list = []
for  order in ['esent','eopenrate','eclickrate','avgorder', 'ordfreq']:
    index_list.extend(outliers(df, order))

#Define a function called 'remove' which returns a cleaned dataset without any outliers
def remove(data, list):
    list = sorted(set(list))
    data = data.drop(list)
    return data
```

5. Skewness and Kurtosis

```
#Skewness of the variables

skew(df_cleaned['esent']) -0.479649928200444
skew(df_cleaned['eclickrate']) 1.8222323439599148
skew(df_cleaned['eopenrate']) 1.2981722822282233
skew(df_cleaned['avgorder']) 1.3586971417930462
skew(df_cleaned['ordfreq']) 1.9200019100871861

#Kurtosis of the variables
kurtosis(df_cleaned['esent']) -0.9654720469858256
kurtosis(df_cleaned['eclickrate']) 2.8754454616376037
kurtosis(df_cleaned['eopenrate']) 0.6636424062086763
kurtosis(df_cleaned['avgorder']) 1.8984700903313128
kurtosis(df_cleaned['ordfreq']) 2.6977570987578154

#Normality Check

st.shapiro(df_cleaned['esent']) ShapiroResult(statistic=0.9164579510688782, pvalue=0.0)
st.shapiro(df_cleaned['eclickrate']) ShapiroResult(statistic=0.7110521793365479, pvalue=0.0)
st.shapiro(df_cleaned['eopenrate']) ShapiroResult(statistic=0.804276704788208, pvalue=0.0)
st.shapiro(df_cleaned['avgorder']) ShapiroResult(statistic=0.8792672753334045, pvalue=0.0)
st.shapiro(df_cleaned['ordfreq']) ShapiroResult(statistic=0.6160522103309631, pvalue=0.0)
```

7. Google Drive link for python programming file:
https://drive.google.com/drive/folders/1ML_nrKsxDVA7XcFkcjSfkU_FQo3wt1JH?usp=sharing