TELECOM CHURN ANALYSIS

ARE CUSTOMERS WITH HIGHER CHARGES MORE PRONE TO CHURN?

- RITIK PRAKASH NAYAK

AN EXPLORATORY ANALYSIS

WHAT ARE WE INVESTIGATING?

1. A BRIEF INTRODUCTION TO THE DATA AND THE PROBLEM STATEMENT
2. ANALYZING THE DAY, EVENING, NIGHT, AND INTERNATIONAL CHARGES ON EACH CUSTOMERS SEPARATELY AND TOGETHER
3. REPORTING THE PARAMETERS
   a. EFFECT SIZE
   b. HYPOTHESIS TESTING
   c. RELATIVE RISK
4. THE EFFECT OF OTHER VARIABLES
   a. DIGGING DEEP INTO THE INTERNATIONAL CHARGES
   b. THE EFFECT OF VOICEMAILS
   c. A FEW MORE THINGS ABOUT VOICEMAILS
5. AT WHAT TIME ARE THE CUSTOMERS CALLING MORE?
6. CONCLUSION

"Exploratory Data Analysis can never be the whole story, but nothing else can serve as the foundation stone -- as the first step"

- John W. Tukey

A BRIEF INTRODUCTION TO THE DATA AND THE PROBLEM STATEMENT

1. Orange S.A. which was formerly known as France Telecom S.A. is a French multinational telecommunications corporation.
2. The dataset at hand has the records of the customers' activity characterized in as many as 21 variables 3333 records.
3. One of the variables pertains to the churn label specifying whether a customer "cancelled the subscription".
4. The dataset is in 'comma separated values' (.csv) format.
5. No additional information/documentation is provided for the dataset.
6. The next slide contains a snap of the top 5 records of the dataset.

```
df.head()
```

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	Total night charge	Total intl minutes	Total intl calls	Total intl charge	Customer service calls	Churn
0	KS	128	415	No	Yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False
1	OH	107	415	No	Yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False
2	NJ	137	415	No	No	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False
3	OH	84	408	Yes	No	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False
4	OK	75	415	Yes	No	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False

How many variables pertain to the data on charges?

1. 4 variables in the dataset have the information on 4 distinct kind of charges
2. The 4 variables are as follows;
    a. 'Total day charge' - charges on facilities used before 6 pm.
    b. 'Total eve charge' - charges on the facilities used after 6 pm before the night (cannot say until what time).
    c. 'Total night charge' - charges on the facilities employed in the night.
    d. 'Total intl charge' - charges on the total international facilities used (Sans accompanying information, the author assumes that the total international charges have not been included in the former three charges).

Summary of the variables pertaining to charges

STATISTIC	Total day charge	Total eve charge	Total night charge	Total intl charge	Total charge
count	3333	3333	3333	3333	3333
mean	30.562307	17.083540	9.039325	2.764581	59.449754
standard deviation	9.259435	4.310668	2.275873	0.753773	10.502261
minimum	0.000000	0.000000	1.040000	0.000000	22.930000
1st quartile	24.430000	14.160000	7.520000	2.300000	52.380000
median	30.500000	17.120000	9.050000	2.780000	59.470000
3rd quartile	36.790000	20.000000	10.590000	3.270000	66.480000
maximum	59.640000	30.910000	17.770000	5.400000	96.150000

Classification of data in the ‘Churn’ variable

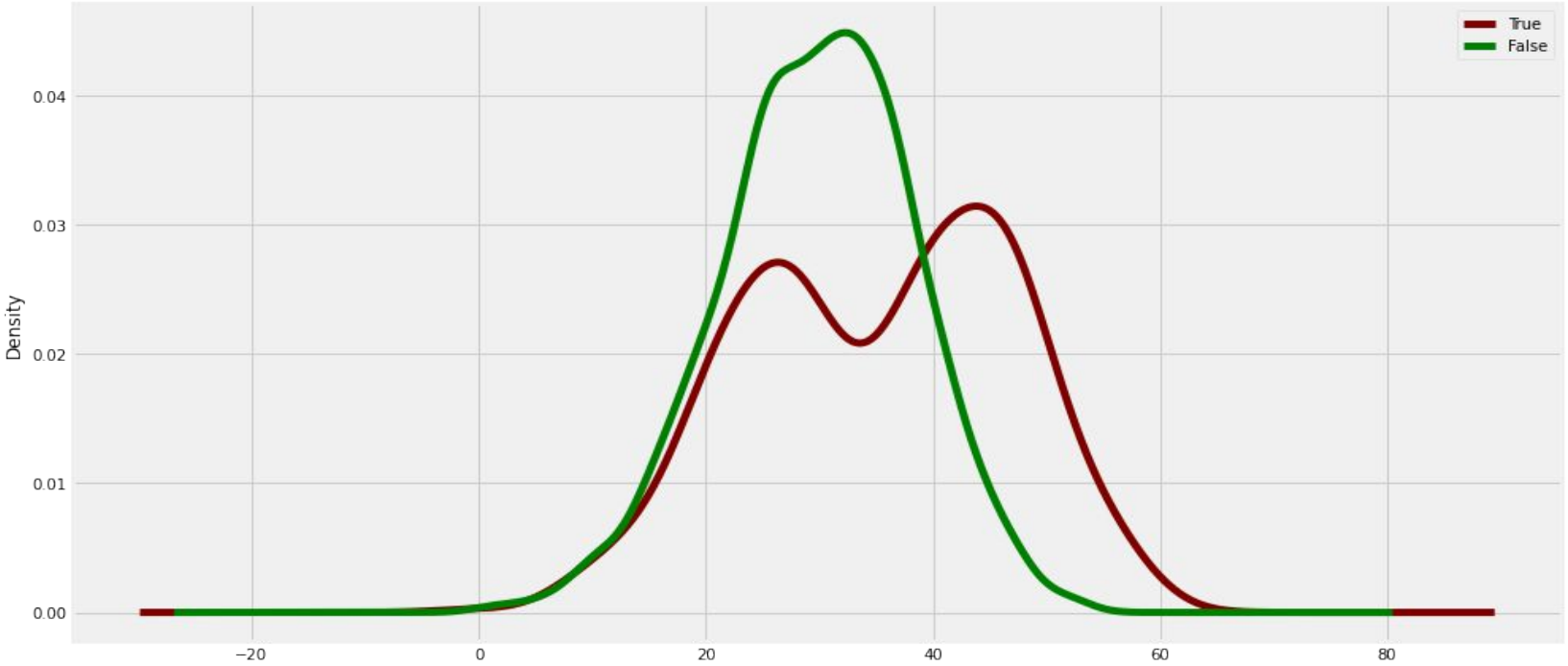
The variable; ‘Churn’, is characterized into 2 classes namely, ‘True’ for records corresponding to the customers that have renounced the services of the company and ‘False’ for those who did not. The following snap gives a summary

```
False    2850
True      483
Name: Churn, dtype: int64
------------------------
False    85.508551
True     14.491449
Name: Churn, dtype: float64
```

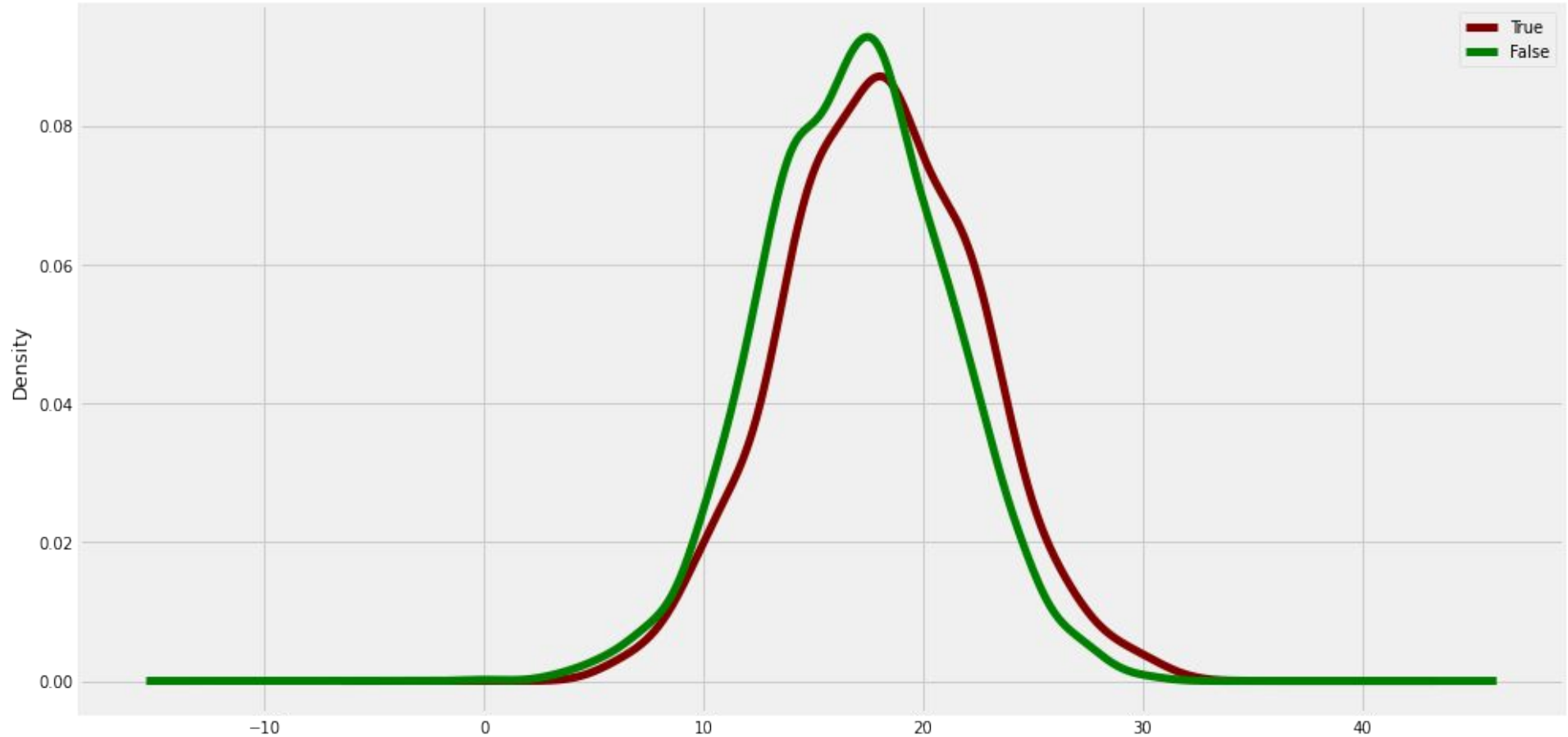
ANALYZING THE DAY, EVENING, NIGHT, AND INTERNATIONAL CHARGES ON EACH CUSTOMERS SEPARATELY AND TOGETHER

Comparative Study 1: Comparing the PDFs for Both the Classes

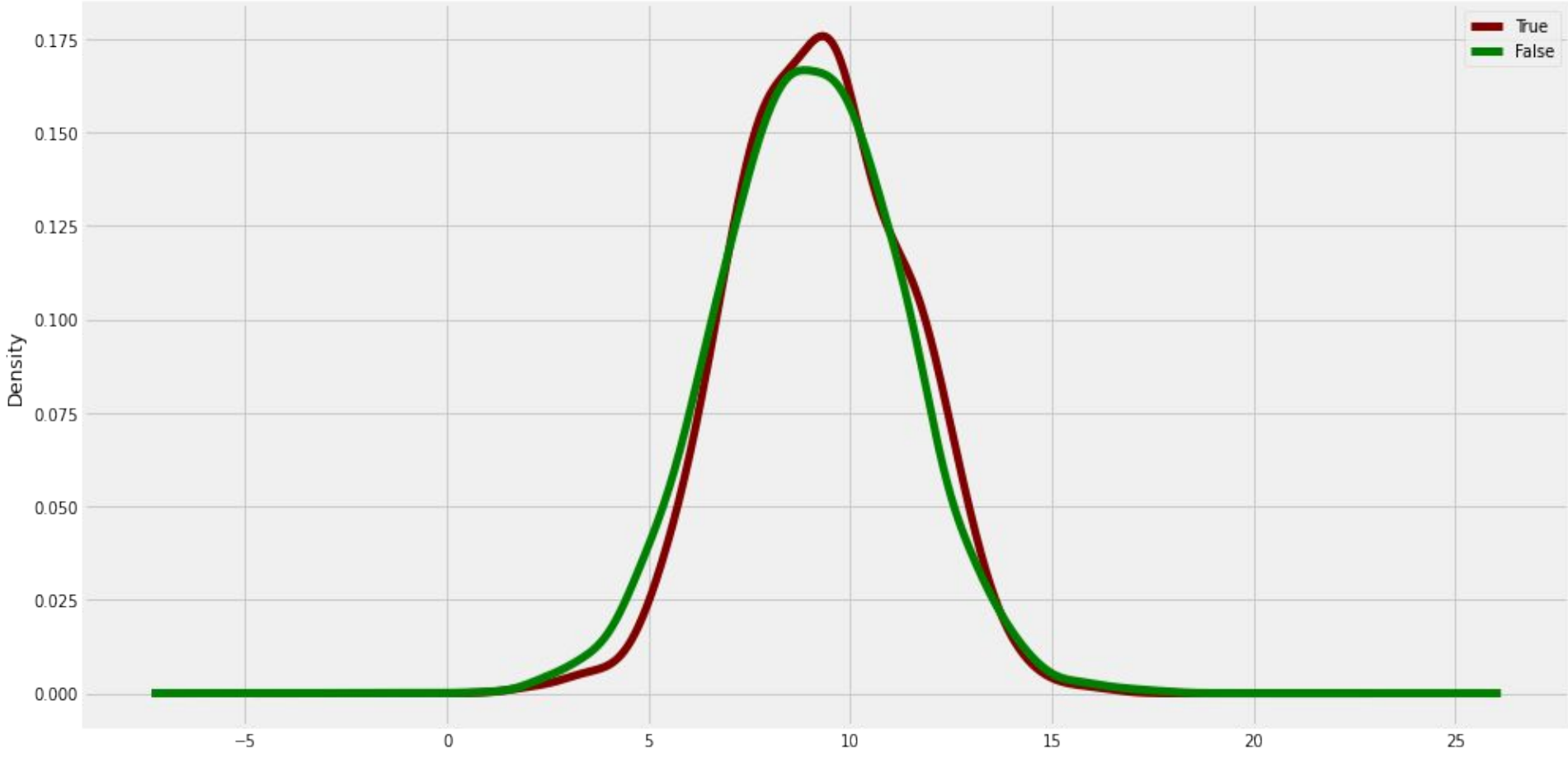
1. Total day charge



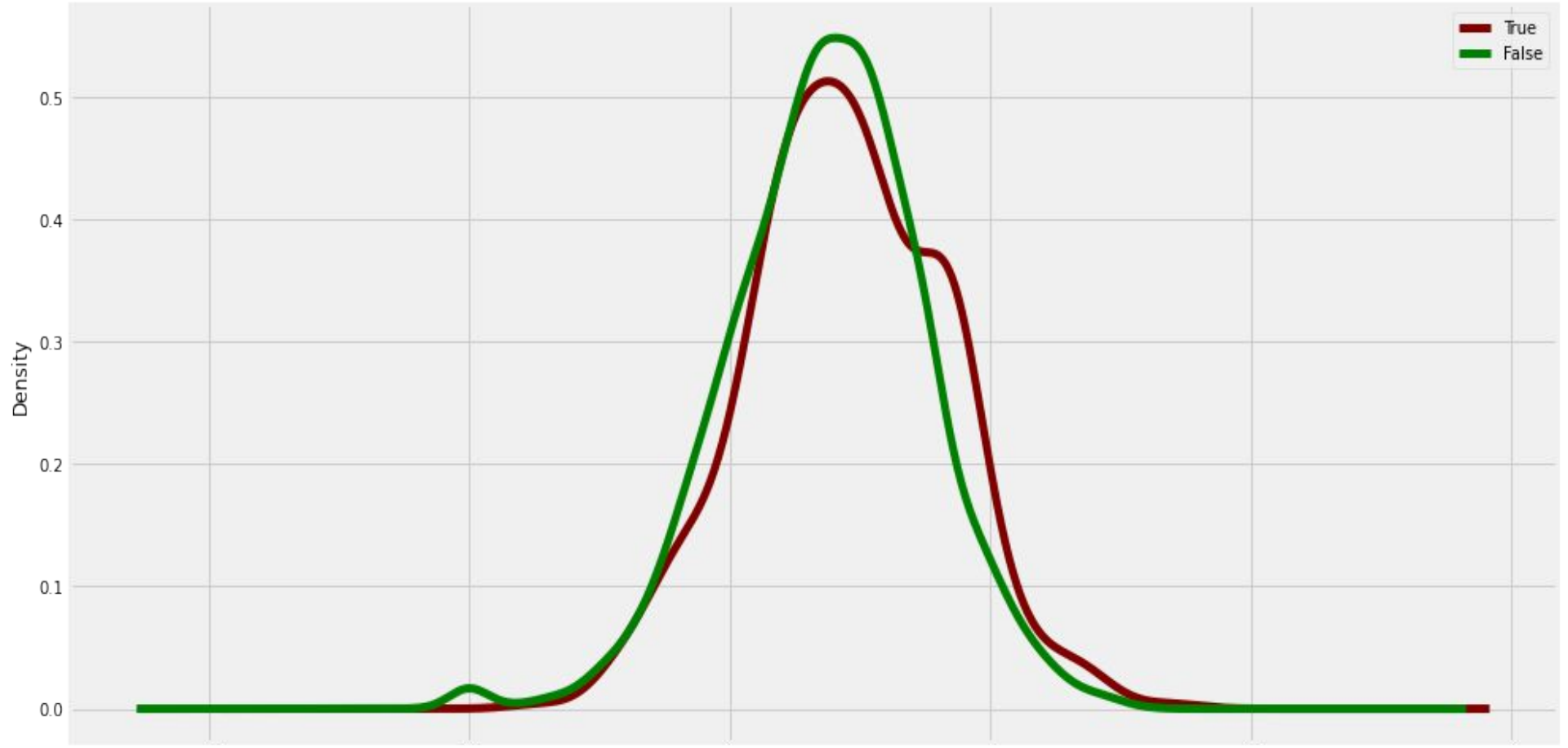
2. Total eve charge



3. Total night charge

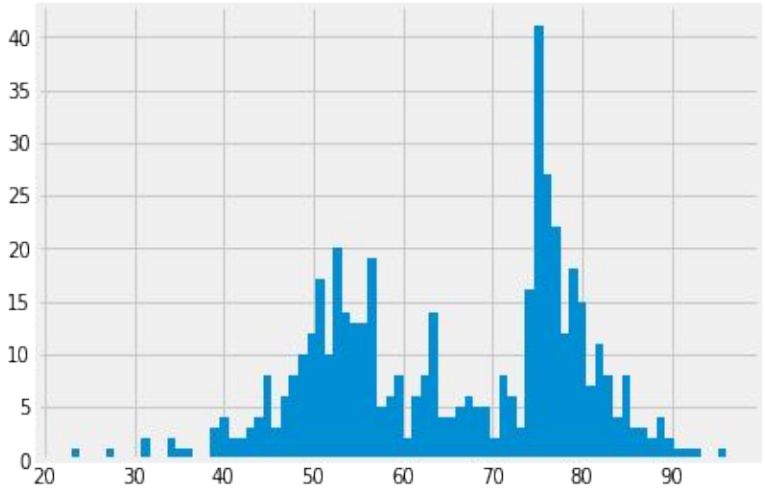


4. Total intl charge

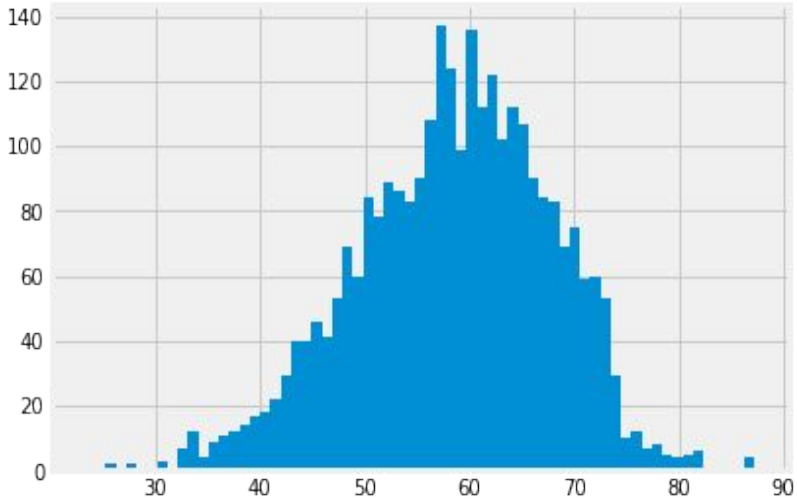


Histogram of the records in total charges

Of the customers that churned

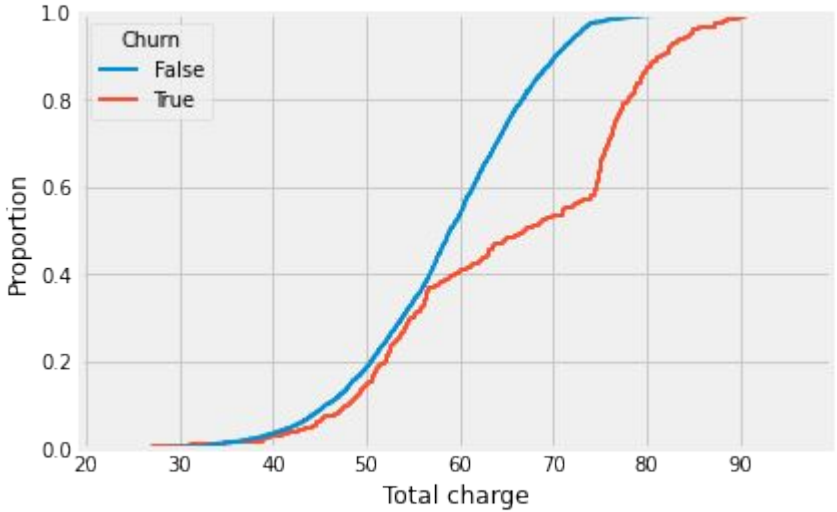


Of the customers that did not churn

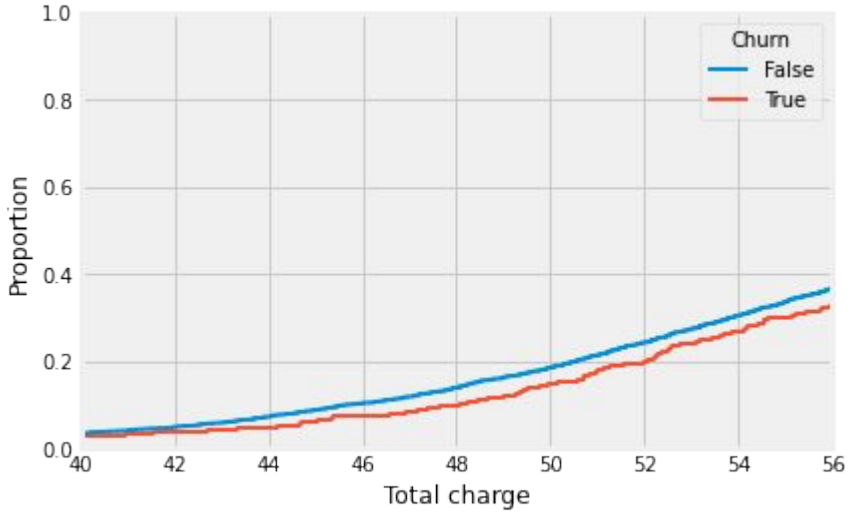


Hued CDF of the total charges

Original CDF

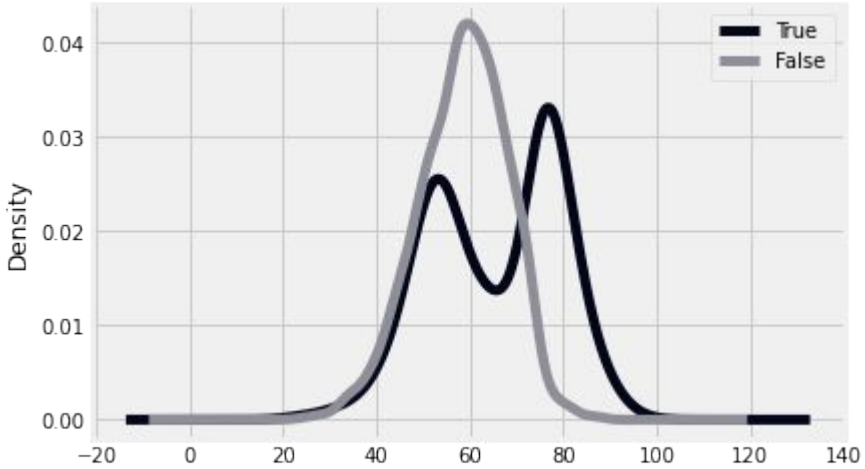


Zoomed CDF

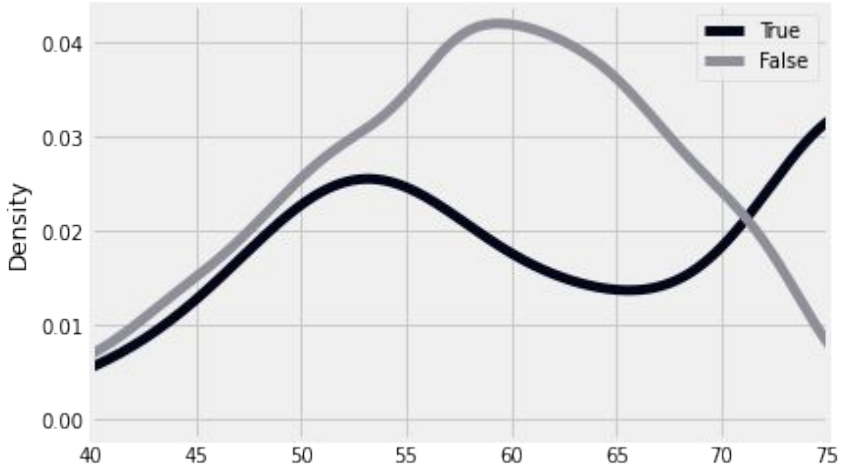


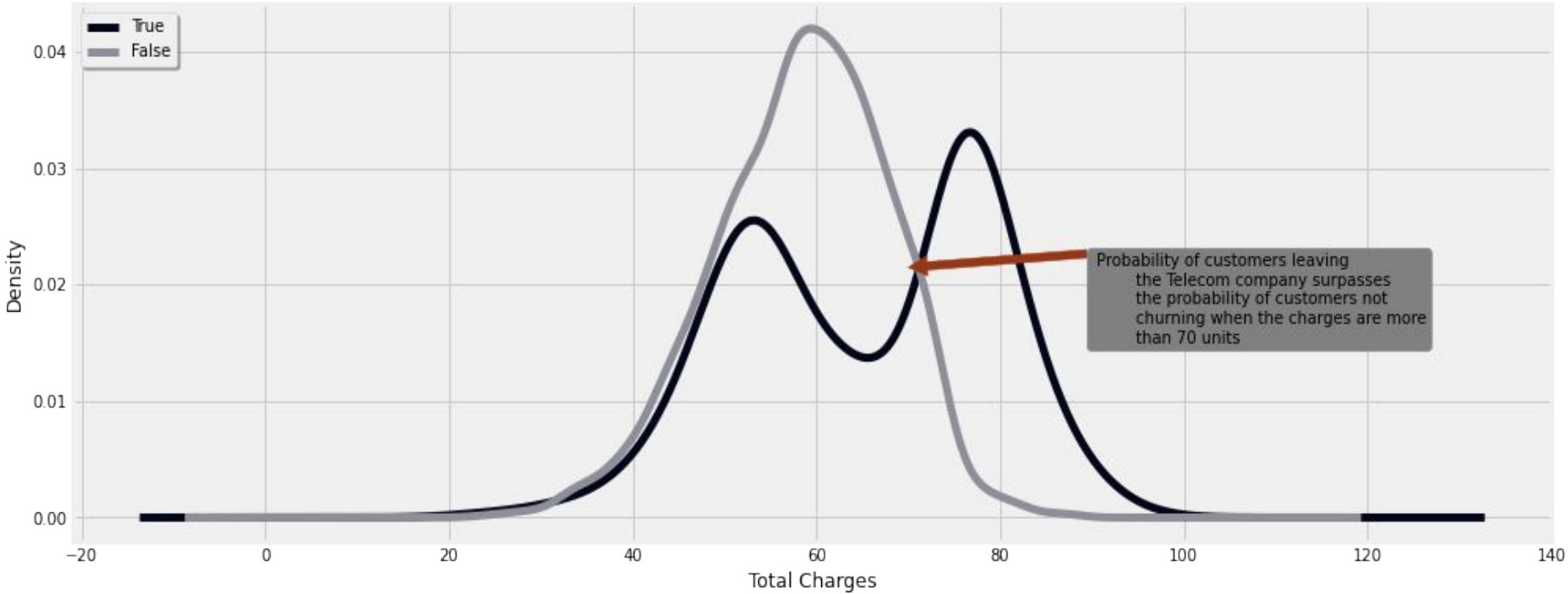
Comparative Study 2: Comparing the PDFs for Both the Classes of the ‘Total charge’ variable

Original PDF



Zoomed PDF




True
False
0.04
0.03
0.02
0.01
0.00
Density
−20
0
20
40
60
80
100
120
140
Total Charges
Probability of customers leaving the Telecom company surpasses the probability of customers not churning when the charges are more than 70 units

Binning

Bin name	1	2	3	4	5
Data	<= 40	<= 50	<= 60	<= 70	> 70

The data depicts the difference between the probability of finding the values of the respective bins in the True category and in that of the False category.

The bar graphs demonstrate the result more robustly. As one goes eastward, one could the valley grows shallower. However, suddenly for charges more than 70 units, a steep hill is encountered in the favor of the 'True' class. That is, a customer being charged more than 70 units is 36% more likely to churn.

REPORTING THE PARAMETERS

a. Effect Size

An “effect size is a summary statistic intended to describe the size of an effect”.

It is a robust way to “compare the difference between groups to the variability within groups”

- Allen B. Downey in Think Stats

Formula for effect size (Cohen’s d):

$$d = (\bar{x}_1 - \bar{x}_2) / s$$

Where; xbar1, and xbar2 are the sample means of the groups and s is the “pooled standard deviation”.

```
[ ]  data1 = df[df['Churn'] == True]['Total charge']
     data2 = df[df['Churn'] == False]['Total charge']
```

```
def cohen_d(group1, group2):
  diff = group1.mean() - group2.mean()
  var1 = group1.var()
  var2 = group2.var()
  n1, n2 = len(group1), len(group2)
  pooled_var = (n1 * var1 + n2 * var2) / (n1 + n2)
  d = diff / np.sqrt(pooled_var)
  return d

cohen_d(data1, data2)
```

```
0.6758829380675058
```

By Cohen’s own standards, an effect size of 0.6 suggests a medium effect

b. Hypothesis Testing

> “Hypothesis testing involves collecting data from a sample and evaluating the sample. Then the statistician makes a decision as to whether or not there is sufficient evidence… to reject the null hypothesis”

- Illowsky and Dean in Introductory Statistics

```
data1 = df[df['Churn'] == True]['Total charge']
data2 = df[df['Churn'] == False]['Total charge']

ht = HypothesisTest((data1, data2))
pvalue = ht.PValue()
print(pvalue)
```

```
0.13933333333333334
```

Null Hypothesis: “The customers that are being charged more do not churn more”

Alternative Hypothesis: “The customers that are being charged more, churn more”

A p-value of 0.13 suggests that we cannot reject the null hypothesis. Perhaps we need more data to support our claim

c. Relative Risk

Taking cue from the comparative study, we could say - as we have said earlier- that customers being charged less than 70 are more likely to stay as compared to those that are charged more than that. Considering 70 units as a threshold, we could make two groups. Probability of the customers churning when charges <= 70, and probability of the same event when charges > 70.

Relative Risk is a ratio of two probabilities. Say for instance, the ratio of both groups turns out to be 1.22. This means that the customers being charged less than 70 are 22% more likely to churn. The following is the result from our sample

```
Customers being charged less than 70 units are about 68% MORE LIKELY TO STAY
Customers being charged more than 70 units are about 80% LESS LIKELY TO STAY
```

THE EFFECT OF OTHER VARIABLES

a. Digging deep into International Charges

The following are the observations:

1. The per minute charge on international calls is around 0.26 units.
2. The per minute charge for the day, evening, and night charges however are around 0.17, 0.08, and 0.04 units respectively.
3. No difference in the per minute charge was noted between the customers availing and not availing for the international plan.
4. The no. of calls made by the customers in each group (international plan and no international plan) is evenly distributed between 1-20 calls suggesting that the customers in the second group have intentionally used the international services.
5. Apparently, the company does not have plans for the no. of calls. The charges seem completely influenced by the no. of minutes.

6. What is the churning percentage for both the groups?

```
False    88.504983
True     11.495017
Name: Churn, dtype: float64
--------------------------------
False    57.585139
True     42.414861
Name: Churn, dtype: float64
--------------------------------
--------------------------------
False    2664
True      346
Name: Churn, dtype: int64
--------------------------------
False    186
True     137
Name: Churn, dtype: int64
```

In the image, the topmost table refers to the customers that did not opt for the international plan while the one below is for those that did

7. What is the percentage of the customers with international plans in both the classes (Churn and not Churn / True and False)?

```
No     71.635611
Yes    28.364389
Name: International plan, dtype: float64
------------------------------
No     93.473684
Yes     6.526316
Name: International plan, dtype: float64
------------------------------
------------------------------
No     346
Yes    137
Name: International plan, dtype: int64
------------------------------
No     2664
Yes     186
Name: International plan, dtype: int64
```

Top: For the customers that are churning

Below: For the customers that are not churning

8. What difference will it make if we remove the international charges from the total charges?

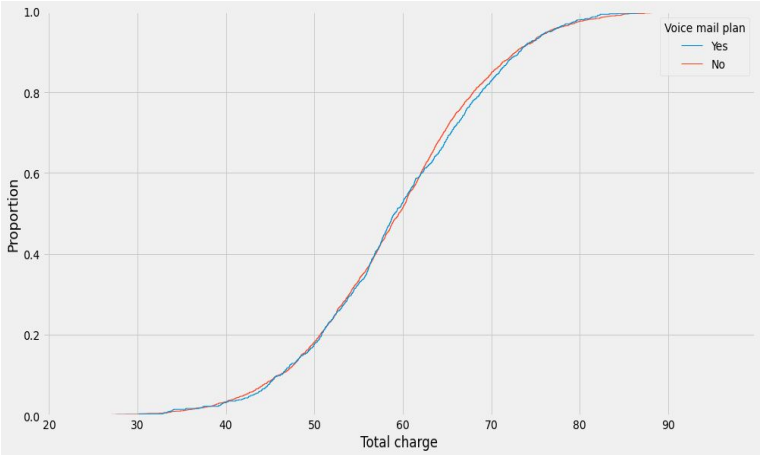
a. If we remove the intl charges from the total charges, the mean of the total charge comes down to 0.09 from 0.10.

b. The corresponding figures for day, evening, and night charges respectively are 0.04, 0.07, and 0.08.

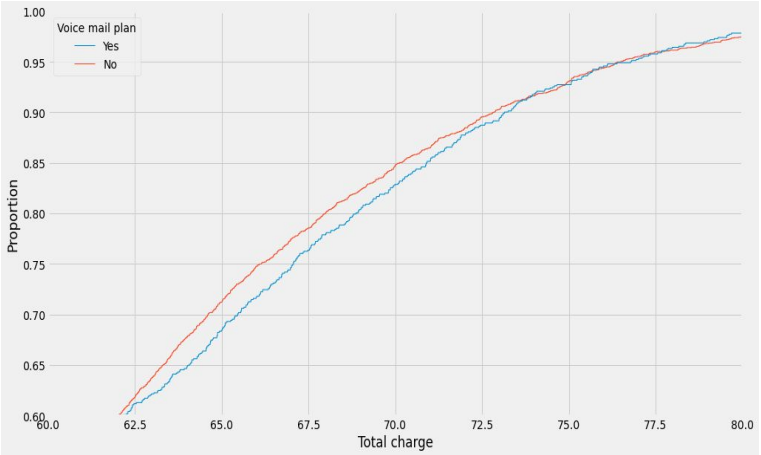
c. This is in part because of the less no. of intl calls made as compared to its other counterparts.

b. The Effect of Voicemails

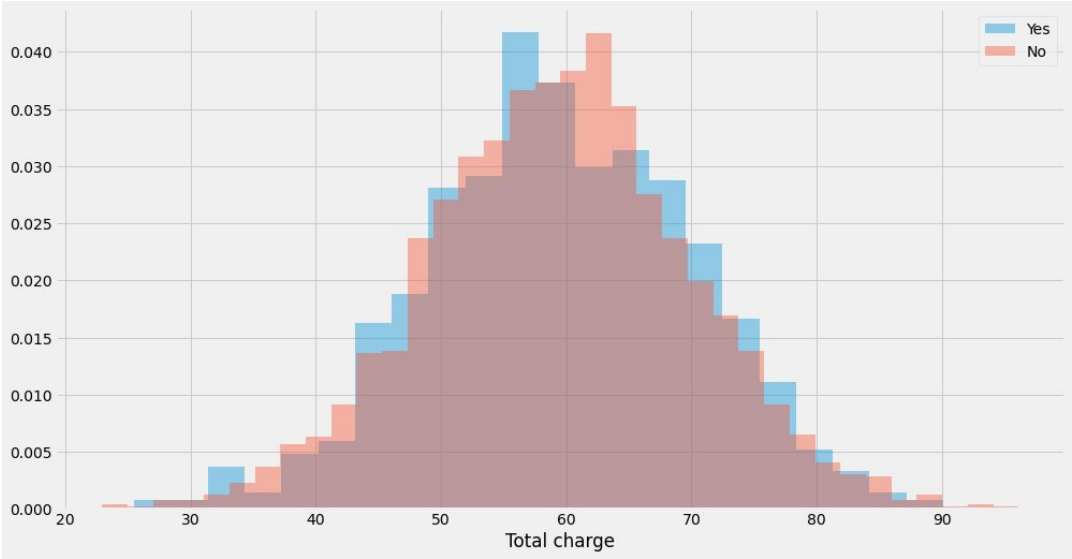
Original CDF



Zoomed CDF

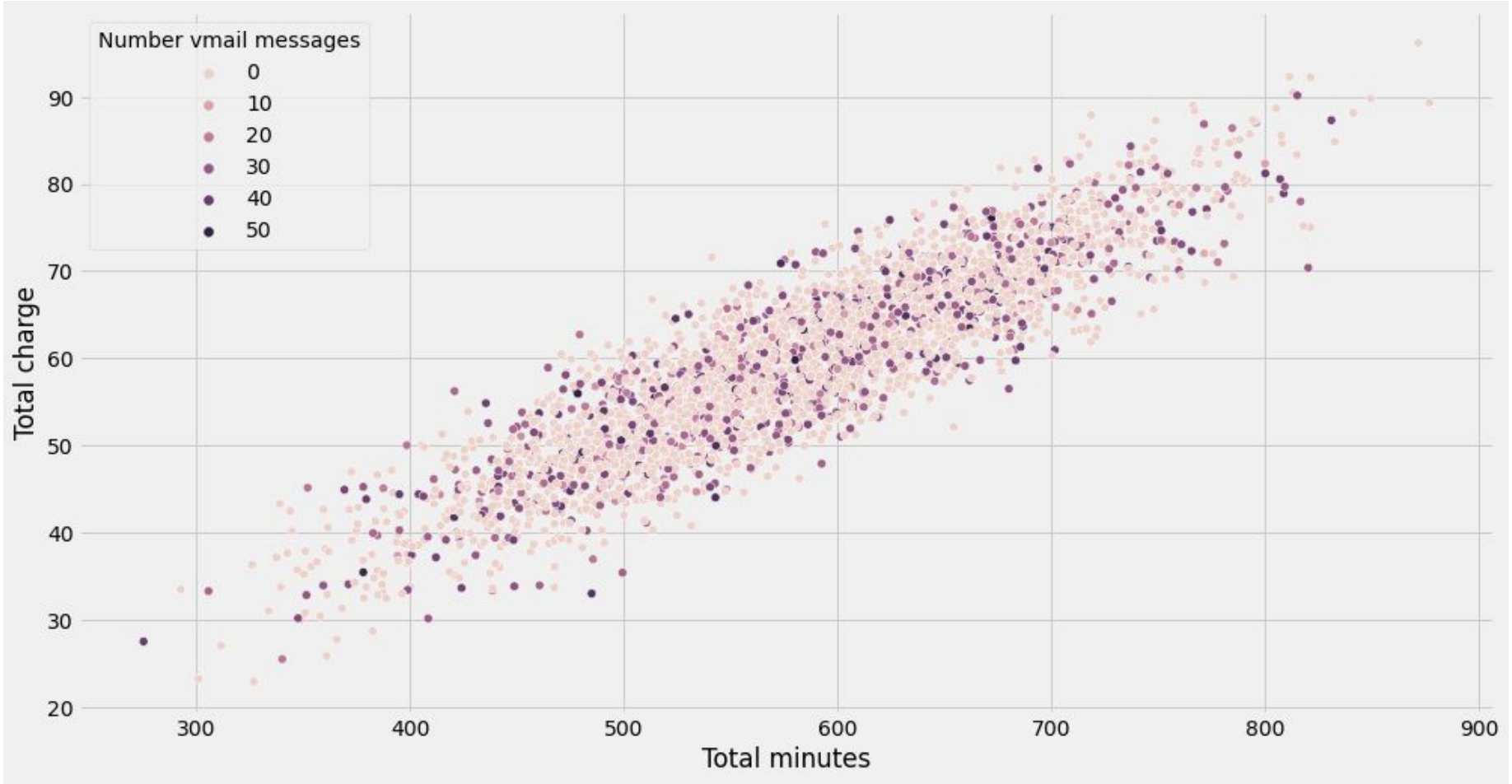


1. From the CDF, I suspect that the charges for voice mails are unusually less perhaps for which the charges of it weigh more when the overall charges are more.
2. Let's further dwell on this claim and find out whether it is true or not. For that, I'll do a comparative analysis between the customers that have opted for the voice mail services and those that didn't. Given below is the result



3. As opposed to my assumption, the trend for the higher charges is not as much biased towards the 'Yes' class as for me to conclude in the favor of my hypothesis. But I could not reject that entirely also. Because, be little, there's still an inconsistency towards the higher values of charges that seem to fall in favor of 'Yes'. Of course, it might be because of the other variables that are at play.

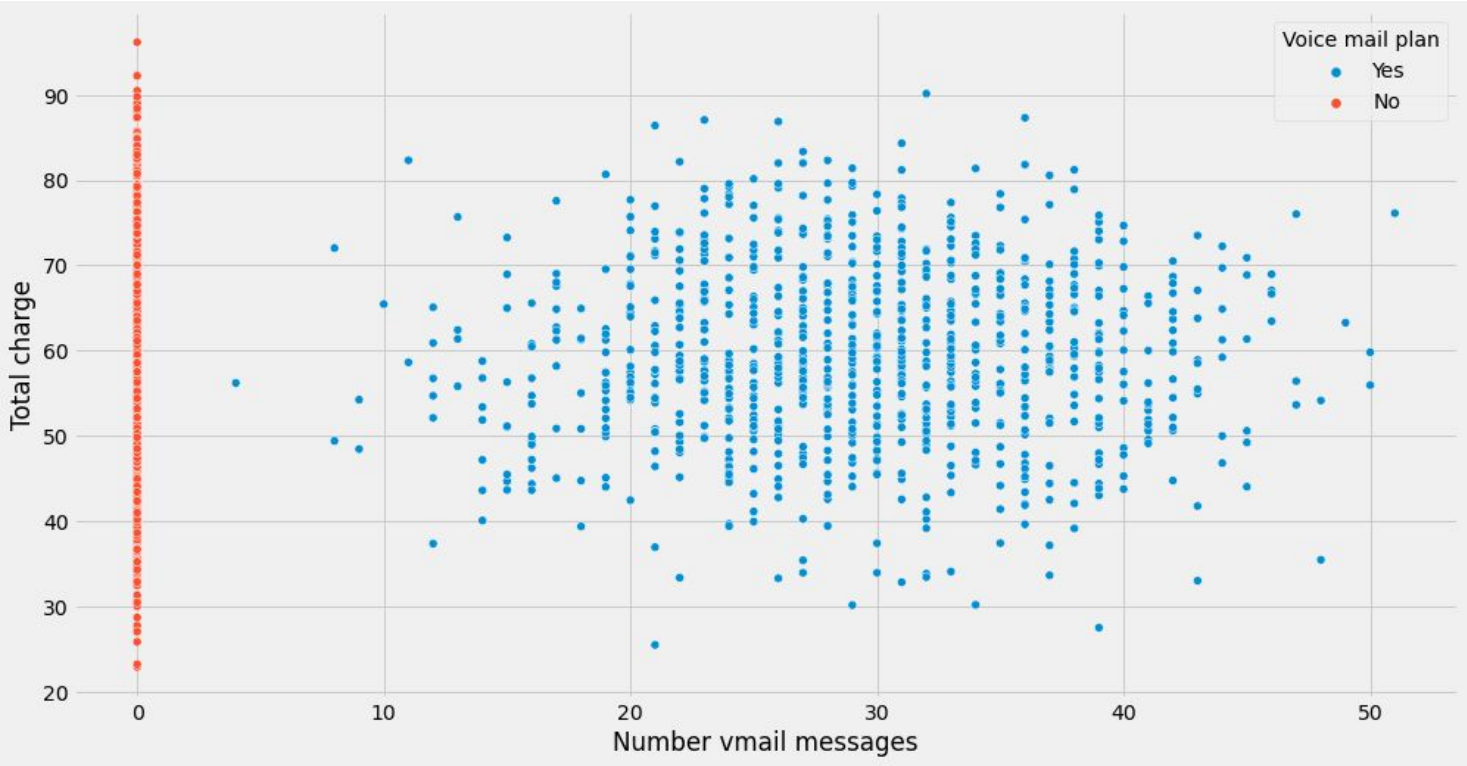
4. Whether the customers making more voice mails are the ones with higher charges or not? The result is in the next slide.


Number vmail messages
0
10
20
30
40
50
Total charge
Total minutes
20
30
40
50
60
70
80
90
300
400
500
600
700
800
900

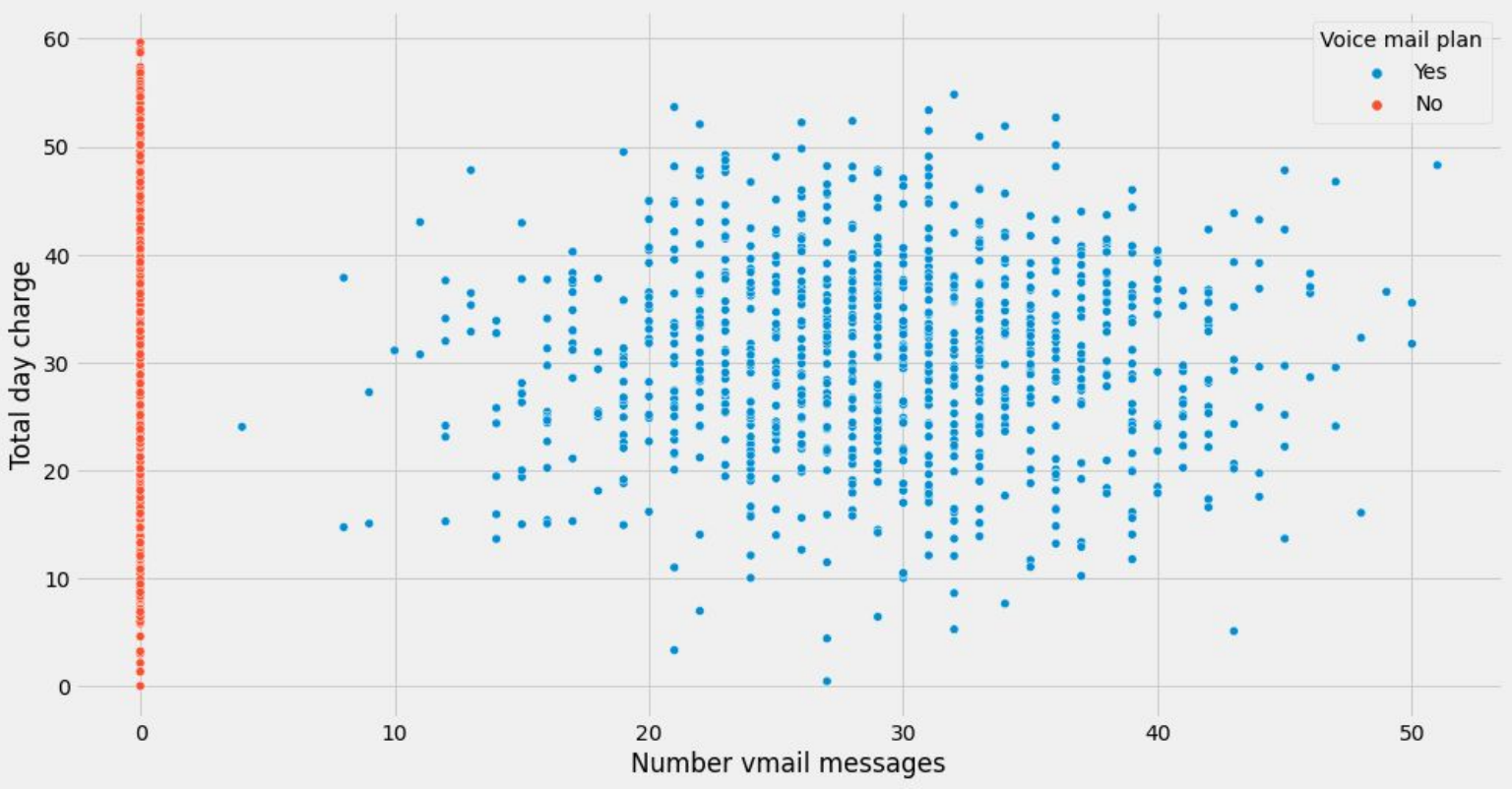
Surprisingly, the even distribution of points suggest that the total charges are not at all influenced by the no. of vmails. This reassures my belief in my earlier hypothesis - the company levies paltry (not even modest) charges on the vmails apparently.

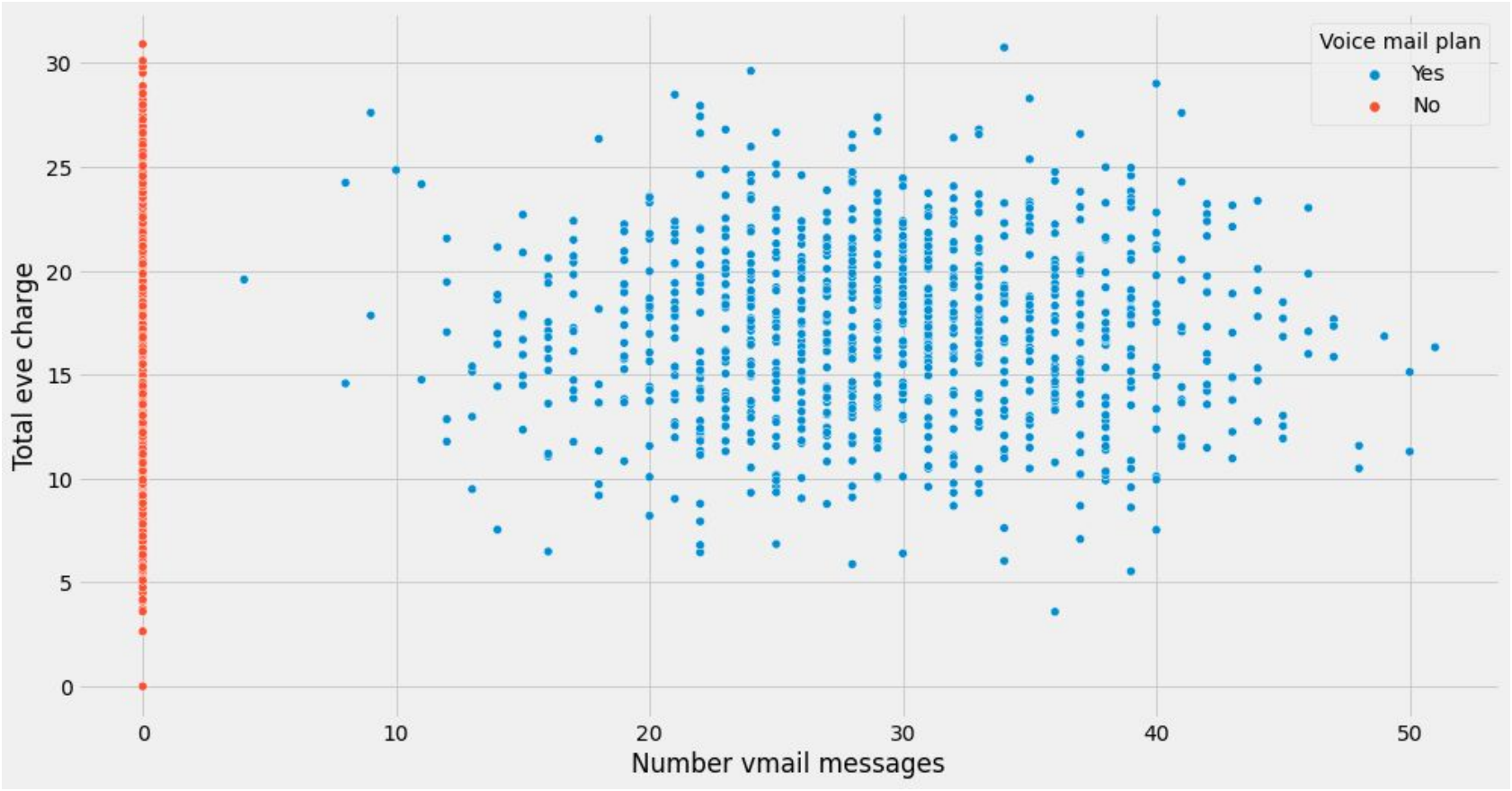
c. A Few More Things About Voicemails

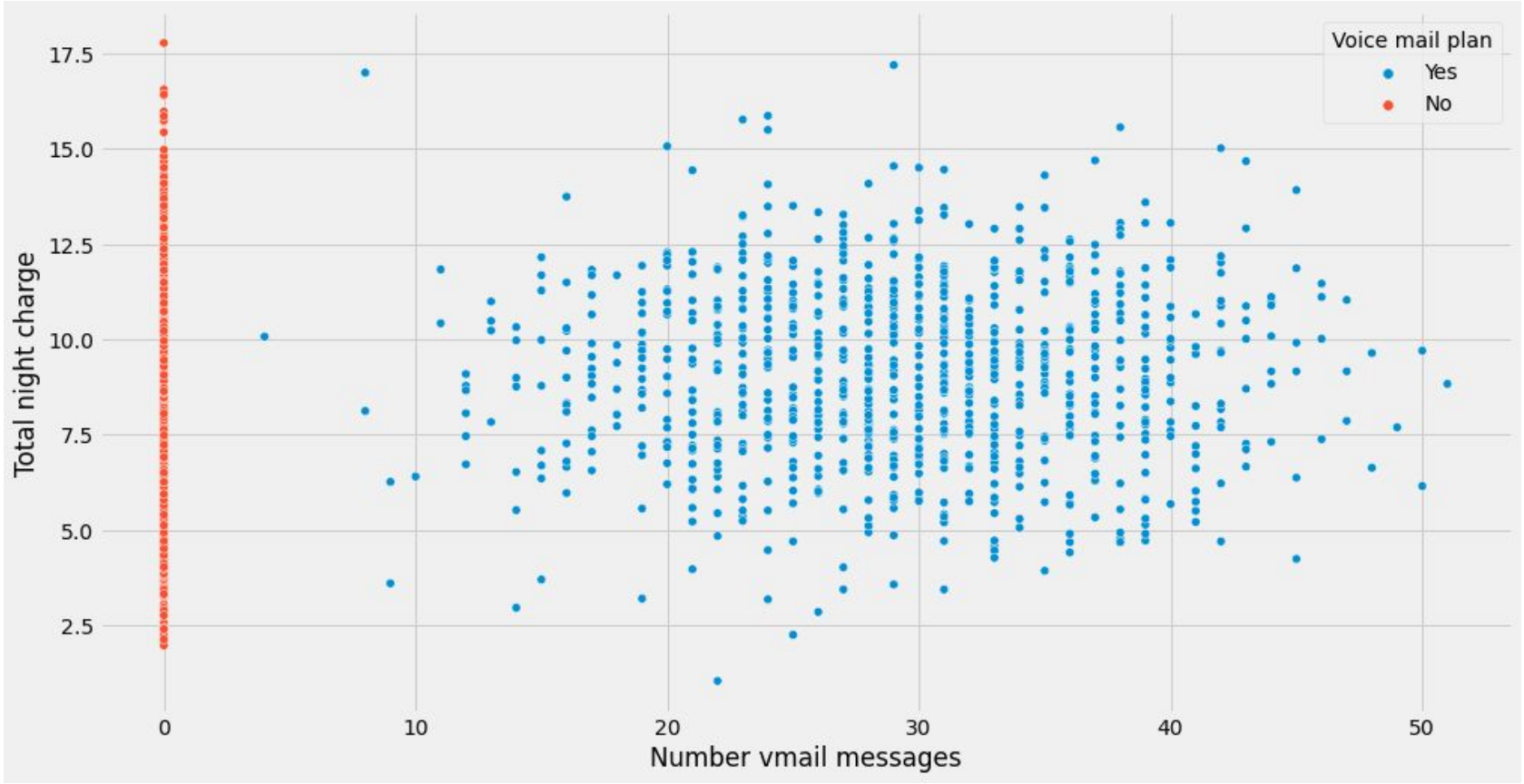
1. Correlation with total charge

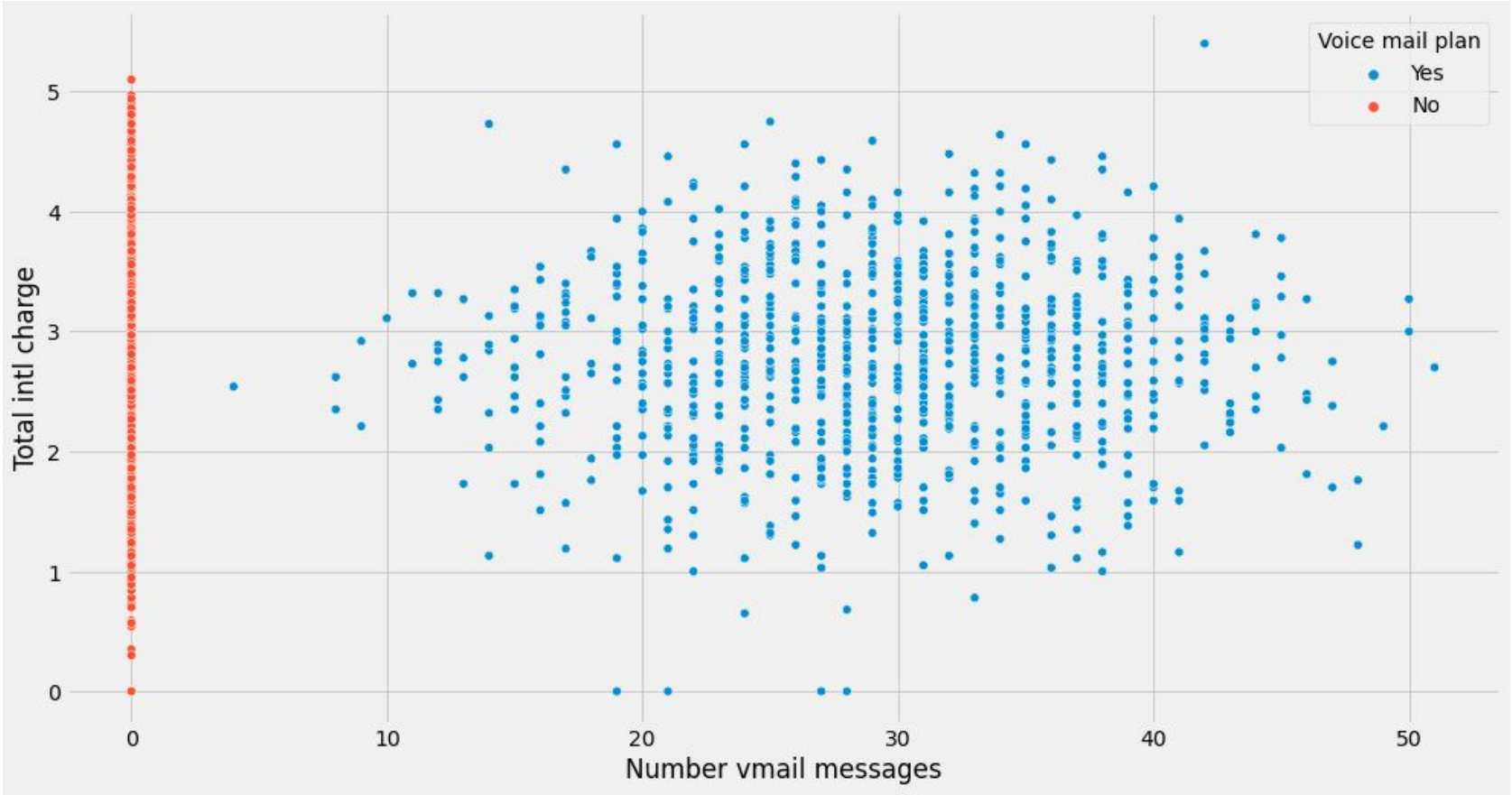


2. Correlation with total day, evening, night, and intl charge

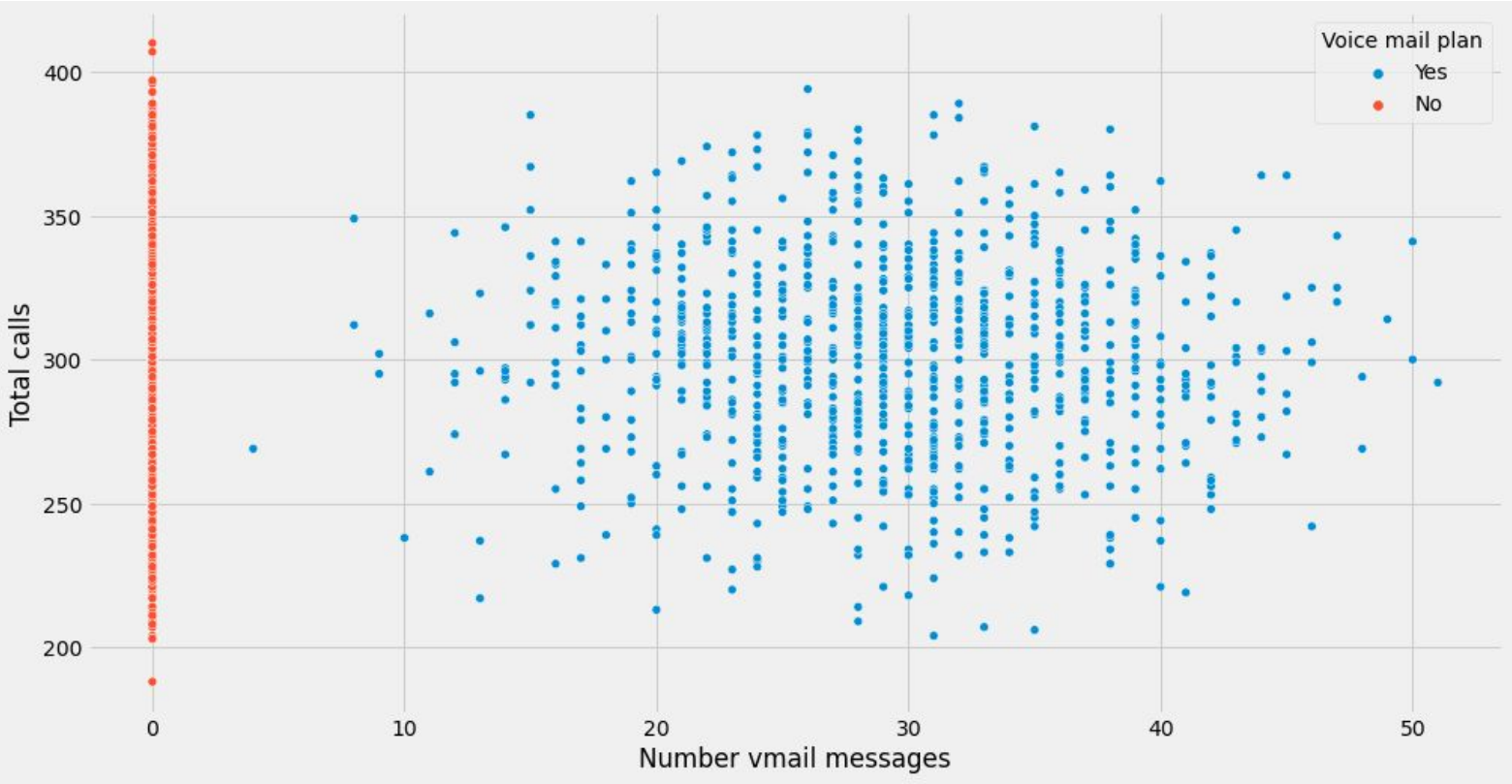



Voice mail plan
Yes
No
Total eve charge
Number vmail messages
0
5
10
15
20
25
30
0
10
20
30
40
50

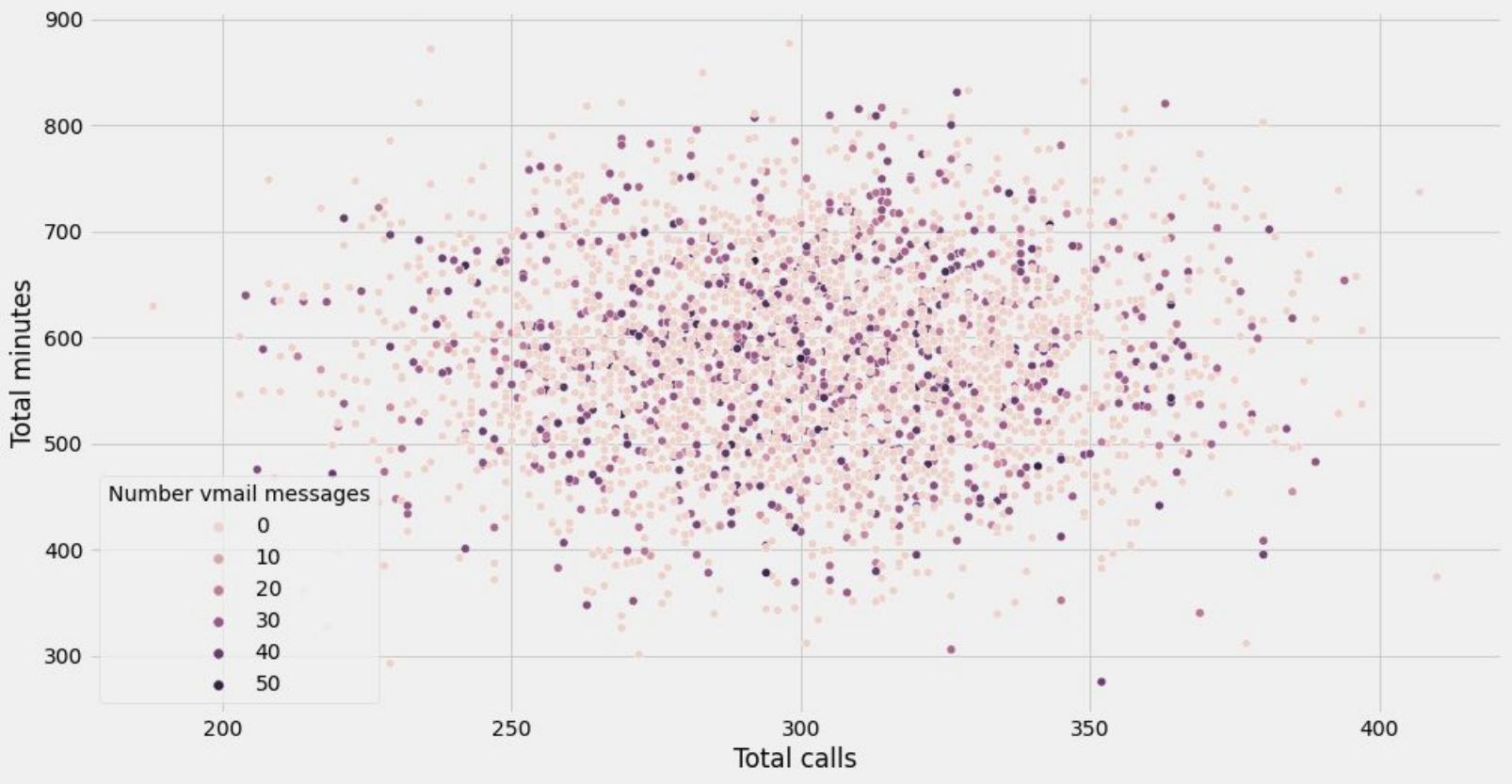

Voice mail plan
Yes
No
Total night charge
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
10
20
30
40
50
Number vmail messages


Voice mail plan
Yes
No
Total intl charge
Number vmail messages
0
1
2
3
4
5
0
10
20
30
40
50

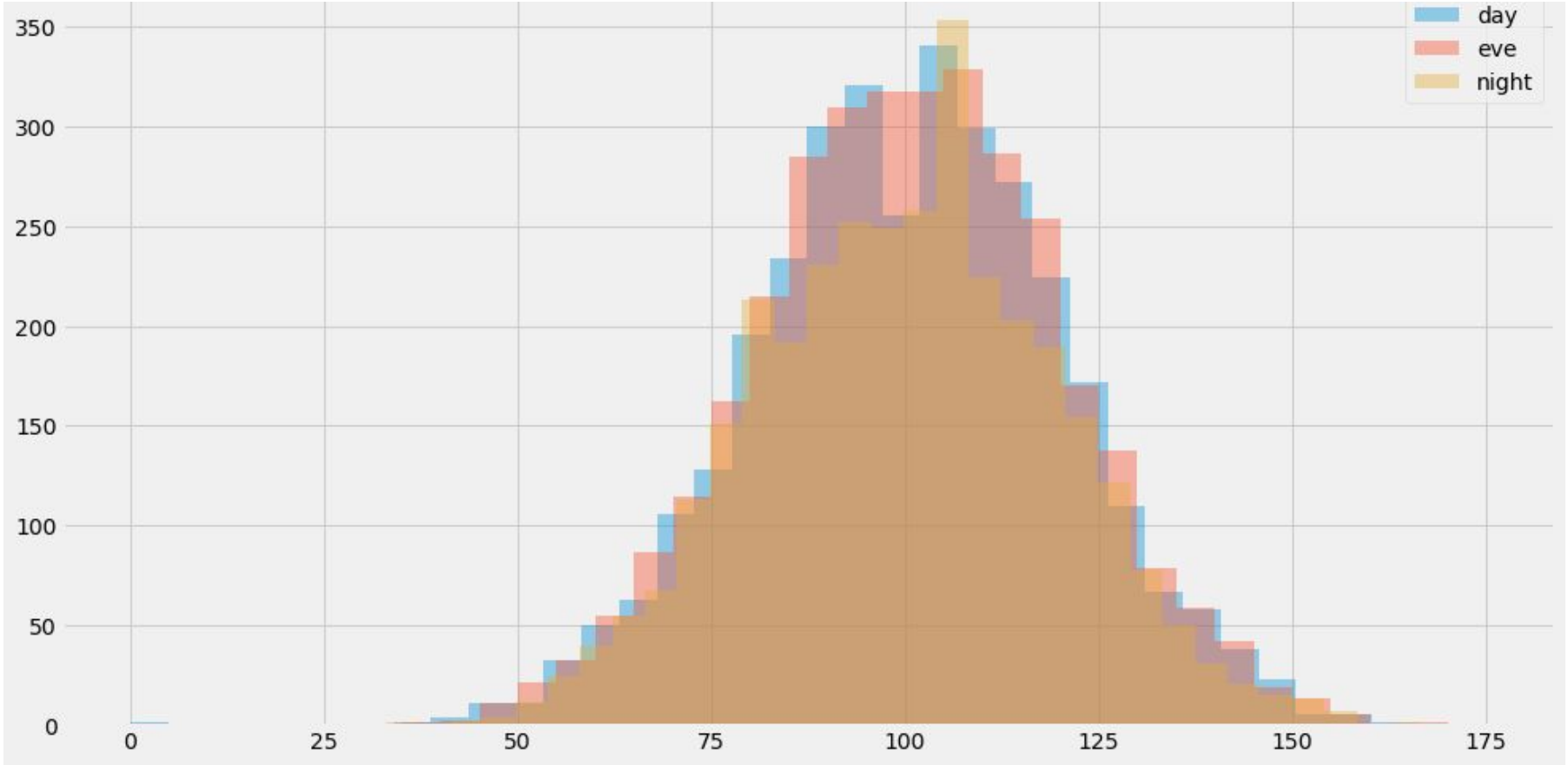
3. Correlation with total no. of phone calls



4. Correlation between total calls and total minutes hued by the no. of voice mails



AT WHAT TIME ARE THE CUSTOMERS CALLING MORE?


day
eve
night
350
300
250
200
150
100
50
0
0
25
50
75
100
125
150
175

The customers are apparently making more calls in the evening as compared to day and night. It should also be noted that least no. of calls are being made in the night. So an increasing order based on the total no. of calls being made would be;

Evening > Day > Night

CONCLUSION

From the above EDA, I conclude that my earlier hypothesis - Customers that are being charged more, tend to churn more - stands unvalidated for the population. It is true that a huge difference was noted in this sample, when a comparative study demonstrated that after a threshold charge of 70 units, the probability for the customers churning surpasses the probability of its counterparts. But, that effect might have transpired by chance. Perhaps we need more data to confirm the hypothesis, but for now there is no reason to believe in that.

I want to give a few suggestions to the company based on my analysis.

1. It seems to me that there's no perks for the customers that are availing for international plans, nor is there a difference between the international charges scheme for the customers opting and not opting for the plan.
2. The charges for the voice calls seems abysmally small. Perhaps the company might lower the international charges in order to adjust the charges in their voice call scheme.
3. The company should consider increasing the evening charges and lower the day charges as the customers seem to be calling more during the evening.

“TO DO [EDA] WELL, WE MUST LIMIT OUR OBJECTIVES”

- JOHN W. TUKEY

THANK YOU!