Max Tjen

3/4

Data Preparation Assignment

Preparing the Data

I would begin by importing the packages and features necessary in order to prepare the given data for modeling and then I would read the CSV file containing the data and look at its shape (how many rows and columns), which would then tell how many samples and features the data contains. After that, I would check to see if there are any samples with null values, and if so, which features have an abundance of them as if certain features had a lot of null values, I would see what the best options are to supplement them. Following that, I would check the data types of each feature and if anything was a type other than a float or integer and could be made binary, I would use the replace function to give a numerical representation. By doing so, the data can be used by various functions and address each feature, as if the feature was for instance a string, it wouldn't be able to compare each feature to other ones that are represented by numbers. Once I've replaced the sample values that aren't numbers, I would get descriptive statistics of the data to look at values like the mean, count, and standard deviation of each feature. Similarly, I would check the balance of the target feature because if it's very unbalanced, the model may incorrectly learn some things and create assumptions during its training. After that, I would look at each feature's correlation with other features, as well as use a heatmap which would effectively do the same thing except that it would be more clear to the viewer. Next, I would look at a variety of plots to see the distribution of each feature and determine if there are any outlier samples that may need to be removed from the data by also looking at skewness. Lastly, I would rescale and then standardize the data for modeling, as normalizing can't be done without treating null values.

Data Exploration

⇒ see figures of results at end of document

⇒ imports

```
import pandas as pd
import numpy as np
import seaborn as sea
from matplotlib import pyplot
from pandas import read_csv
from numpy import unique
from numpy import set_printoptions
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
```

⇒ read the file and see the data's shape and head

```
fileName = "/Users/mtjen/Desktop/Churn_Train.csv"
customerDF = read_csv(fileName)
shape = customerDF.shape
head = customerDF.head
print(shape)
print(head)
```

⇒ check for null values and each features data type

```
isNullValues = customerDF.isnull().sum()
types = customerDF.dtypes
print(isNullValues)
print(types)
```

⇒ replace churn values with binary values

```
customerDF = customerDF.replace(['no', 'yes'], [0, 1])
```

⇒ simple descriptive statistics

```
description = customerDF.describe()
print(description)
```

⇒ get the balance of the target feature

```
churnBalance = customerDF.groupby('churn').size()
print(churnBalance)
```

⇒ get the correlations between each feature

```
correlation = customerDF.corr(method = 'pearson')
print(correlation)
```

⇒ create a heatmap of correlations

```
matrix = np.triu(customerDF.corr())
sea.heatmap(customerDF.corr(), annot = True, fmt = '0.1g', mask = matrix)
```

⇒ use plots to find feature distribution

```
customerDF.hist(layout = (6,3), figsize = (10,10))
customerDF.plot(kind = 'box', layout = (6,3), subplots = True, sharex = False, sharey = False,
                figsize = (10,10))
pyplot.show()
```

⇒ get skew values

```
skew = customerDF.skew()
print(skew)
```

⇒ scale and normalize the data

```
customer = customerDF.drop(['area_code', 'state'], 1)   # remove to scale the numerical features
customer = customer.replace(['no', 'yes'], [0, 1])
array = customer.values
X = array[:,0:17]
Y = array[:,17]
scaler = MinMaxScaler(feature_range=(0, 1))
rescaledX = scaler.fit_transform(X)
set_printoptions(precision=3)
print(rescaledX[0:5,:])
print()
scaler = StandardScaler().fit(X)
rescaledX = scaler.transform(X)
set_printoptions(precision=3)
print(rescaledX[0:5,:])
```

Overview of Data

By doing basic exploratory data analysis, it was found that there are a number of samples with some missing data, although relative to the entire data set, it is a small percentage. With that though, the leading feature of missing values is the sample's account length, which is one of the most important factors to know. The 'churn rate' of the samples is also unbalanced and leans towards not churning. However, the percentage of people is somewhat alarming as about 14.5% of the samples have switched to a competition's services. A couple of features within the data also have high correlation, telling how some of the features studied were not that necessary because of the overlap with others. Even though some of them are highly correlated though, most of the features may still play a large role as they allow the company to see if the customers that they are losing are high value ones that they should be trying to retain. While looking at histograms of the distribution of the samples by each feature, most of the features are relatively evenly distributed which is good for studying the data. Combining the histograms with the descriptive statistics shows a small problem though, as a couple of the features have a high standard deviation and the data is not distributed very evenly. As such, looking at the box plots, a number of the features like total international calls, total number of day minutes, and number of voicemail messages have a lot of outliers. With those outliers, the data is skewed and the model may not be able to be trained as accurately as possible, particularly when figuring out which features are the most important to weigh. To further the idea of outliers and how the data may be skewed, a couple of features were heavily skewed and needed to be fixed. There were only four features that were heavily skewed, but there were also five other features that were also skewed to a lesser degree but still need to be fixed.

Recommendations and Potential Corrective Action

To fully prepare the data set for modeling, there are a couple of corrective actions that I would take to hopefully make the model more accurate. The first thing that I would do is to make sure that every feature's values are numerical by assigning a numerical value for instances like state and area code and by giving binary values for yes or no questions. By doing this, each feature would be able to be compared with each other and graphics like map overlays would convey more useful and effective information. The next thing that I would do is figure out the best way to handle samples with null values within the data. Some of the features may be able to have extrapolated values used, but others with numerous null values may have to just be removed because of the randomness of other samples' values. Moreover, null values within features that are skewed would likely not be able to be accurately extrapolated based on other samples' values. After that, I would remove a couple of the features from the data because of its high correlation and overlap with others. There are four features in particular that have very high correlation, and as such can be removed to improve efficiency within the model. I would also scale and normalize the data, as even though most of the features are evenly distributed, a number of them are skewed to one side or another. Furthermore, the scale and range of some of the features are much greater than others, so it would be useful to have all the data on the same scale. Lastly, I would remove heavy outliers from the data set so that the data is more evenly distributed. Without heavy outliers that greatly skew the data, the model would be able to more accurately and precisely predict whether or not a customer may leave for another provider.

Data Figures

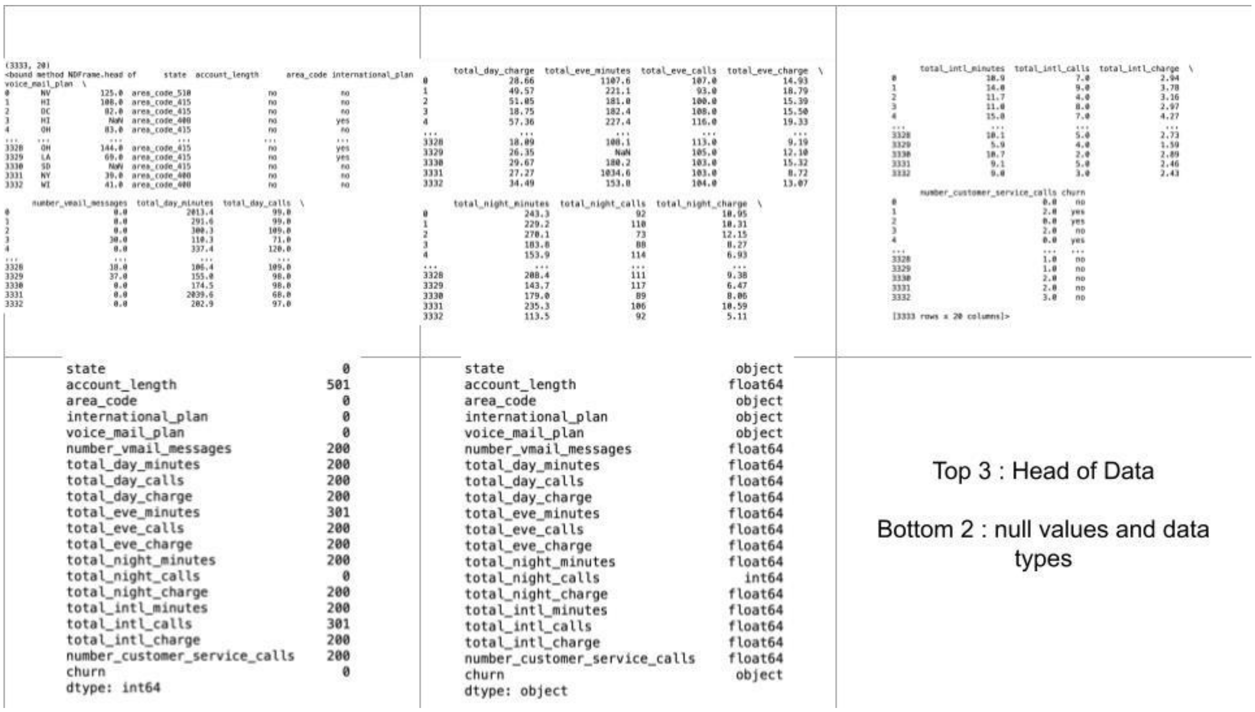
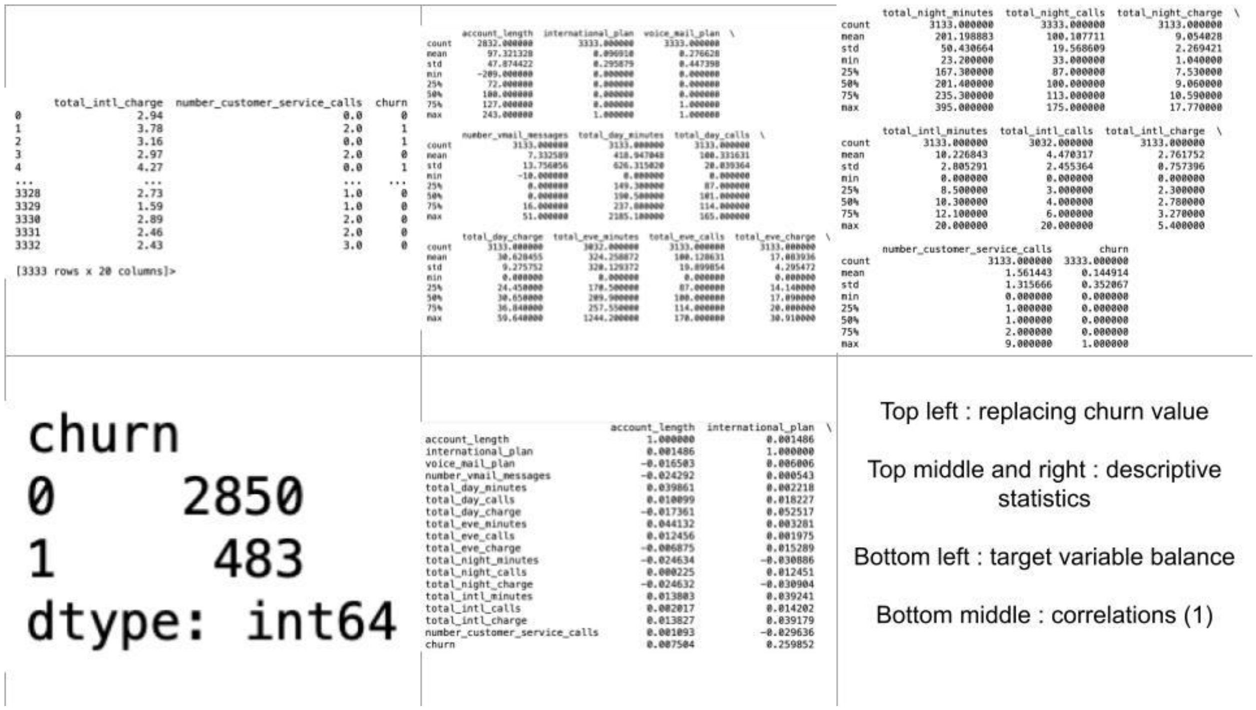


Figure 1



	total_intl_charge	number_customer_service_calls	churn
0	2.94	0.0	0
1	3.78	2.0	1
2	3.16	0.0	1
3	2.97	2.0	0
4	4.27	0.0	1
...	...	...	...
3328	2.73	1.0	0
3329	1.59	1.0	0
3330	2.89	2.0	0
3331	2.46	2.0	0
3332	2.43	3.0	0

[3333 rows x 20 columns]>

	account_length	international_plan	voice_mail_plan
count	2832.000000	3333.000000	3333.000000
mean	97.321328	0.096910	0.276628
std	47.874422	0.295879	0.447398
min	-209.000000	0.000000	0.000000
25%	72.000000	0.000000	0.000000
50%	100.000000	0.000000	0.000000
75%	127.000000	0.000000	1.000000
max	243.000000	1.000000	1.000000

	number_vmail_messages	total_day_minutes	total_day_calls
count	3133.000000	3133.000000	3133.000000
mean	7.332589	418.947048	100.331631
std	13.756056	626.315020	20.039364
min	-10.000000	0.000000	0.000000
25%	0.000000	149.300000	87.000000
50%	0.000000	190.500000	101.000000
75%	16.000000	237.000000	114.000000
max	51.000000	2185.100000	165.000000

	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge
count	3133.000000	3032.000000	3133.000000	3133.000000
mean	30.620455	324.250872	100.120631	17.003936
std	9.275752	320.129372	19.899054	4.295472
min	0.000000	0.000000	0.000000	0.000000
25%	24.450000	170.500000	87.000000	14.140000
50%	30.650000	209.900000	100.000000	17.090000
75%	36.840000	257.550000	114.000000	20.000000
max	59.640000	1244.200000	170.000000	30.910000

	total_night_minutes	total_night_calls	total_night_charge
count	3133.000000	3333.000000	3133.000000
mean	201.198883	100.107711	9.054028
std	50.430664	19.568609	2.269421
min	23.200000	33.000000	1.040000
25%	167.300000	87.000000	7.530000
50%	201.400000	100.000000	9.060000
75%	235.300000	113.000000	10.590000
max	395.000000	175.000000	17.770000

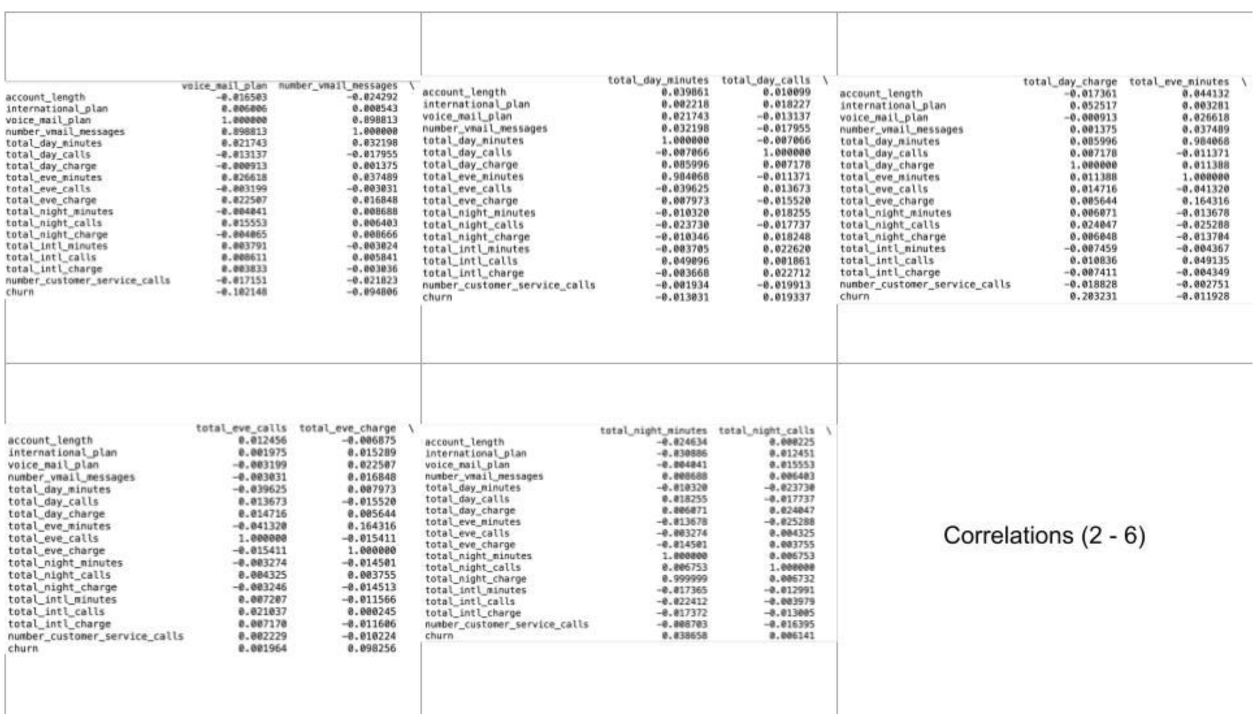
	total_intl_minutes	total_intl_calls	total_intl_charge
count	3133.000000	3032.000000	3133.000000
mean	10.226843	4.470317	2.761752
std	2.805291	2.455364	0.757396
min	0.000000	0.000000	0.000000
25%	8.500000	3.000000	2.300000
50%	10.300000	4.000000	2.780000
75%	12.100000	6.000000	3.270000
max	20.000000	20.000000	5.400000

	number_customer_service_calls	churn
count	3133.000000	3333.000000
mean	1.561443	0.144914
std	1.315666	0.352067
min	0.000000	0.000000
25%	1.000000	0.000000
50%	1.000000	0.000000
75%	2.000000	0.000000
max	9.000000	1.000000

```
churn
0    2850
1     483
dtype: int64
```

	account_length	international_plan
account_length	1.000000	0.001486
international_plan	0.001486	1.000000
voice_mail_plan	-0.016503	0.006006
number_vmail_messages	-0.024292	0.000543
total_day_minutes	0.039861	0.002218
total_day_calls	0.010099	0.018227
total_day_charge	-0.017361	0.052517
total_eve_minutes	0.044132	0.003281
total_eve_calls	0.012456	0.001975
total_eve_charge	-0.006875	0.015289
total_night_minutes	-0.024634	-0.030886
total_night_calls	0.000225	0.012451
total_night_charge	-0.024632	-0.030904
total_intl_minutes	0.013803	0.039241
total_intl_calls	0.002017	0.014202
total_intl_charge	0.013827	0.039179
number_customer_service_calls	0.001093	-0.029636
churn	0.007504	0.259852

Figure 2



	voice_mail_plan	number_vmail_messages
account_length	-0.016503	-0.024292
international_plan	0.006006	0.000543
voice_mail_plan	1.000000	0.898813
number_vmail_messages	0.898813	1.000000
total_day_minutes	0.021743	0.032198
total_day_calls	-0.013137	-0.017955
total_day_charge	-0.000913	0.001375
total_eve_minutes	0.026618	0.037489
total_eve_calls	-0.003199	-0.003031
total_eve_charge	0.022507	0.016848
total_night_minutes	-0.004041	0.008688
total_night_calls	0.015553	0.006403
total_night_charge	-0.004065	0.008666
total_intl_minutes	0.003791	-0.003024
total_intl_calls	0.000611	0.005841
total_intl_charge	0.003833	-0.003036
number_customer_service_calls	-0.017151	-0.021823
churn	-0.102148	-0.094806

	total_day_minutes	total_day_calls
account_length	0.039861	0.010099
international_plan	0.002218	0.018227
voice_mail_plan	0.021743	-0.013137
number_vmail_messages	0.032198	-0.017955
total_day_minutes	1.000000	-0.007066
total_day_calls	-0.007066	1.000000
total_day_charge	0.085996	0.007178
total_eve_minutes	0.984068	-0.011371
total_eve_calls	-0.039625	0.013673
total_eve_charge	0.007973	-0.015520
total_night_minutes	-0.010320	0.018255
total_night_calls	-0.023730	-0.017737
total_night_charge	-0.010346	0.018248
total_intl_minutes	-0.003705	0.022620
total_intl_calls	0.049096	0.001861
total_intl_charge	-0.003668	0.022712
number_customer_service_calls	-0.001934	-0.019913
churn	-0.013031	0.019337

	total_day_charge	total_eve_minutes
account_length	-0.017361	0.044132
international_plan	0.052517	0.003281
voice_mail_plan	-0.000913	0.026618
number_vmail_messages	0.001375	0.037489
total_day_minutes	0.085996	0.984068
total_day_calls	0.007178	-0.011371
total_day_charge	1.000000	0.011388
total_eve_minutes	0.011388	1.000000
total_eve_calls	0.014716	-0.041320
total_eve_charge	0.005644	0.164316
total_night_minutes	0.006071	-0.013678
total_night_calls	0.024047	-0.025288
total_night_charge	0.006048	-0.013704
total_intl_minutes	-0.007459	-0.004367
total_intl_calls	0.010836	0.049135
total_intl_charge	-0.007411	-0.004349
number_customer_service_calls	-0.018828	-0.002751
churn	0.203231	-0.011928

	total_eve_calls	total_eve_charge
account_length	0.012456	-0.006875
international_plan	0.001975	0.015289
voice_mail_plan	-0.003199	0.022507
number_vmail_messages	-0.003031	0.016848
total_day_minutes	-0.039625	0.007973
total_day_calls	0.013673	-0.015520
total_day_charge	0.014716	0.005644
total_eve_minutes	-0.041320	0.164316
total_eve_calls	1.000000	-0.015411
total_eve_charge	-0.015411	1.000000
total_night_minutes	-0.003274	-0.014501
total_night_calls	0.004325	0.003755
total_night_charge	-0.003246	-0.014513
total_intl_minutes	0.007207	-0.011566
total_intl_calls	0.021037	0.000245
total_intl_charge	0.007170	-0.011606
number_customer_service_calls	0.002229	-0.010224
churn	0.001964	0.098256

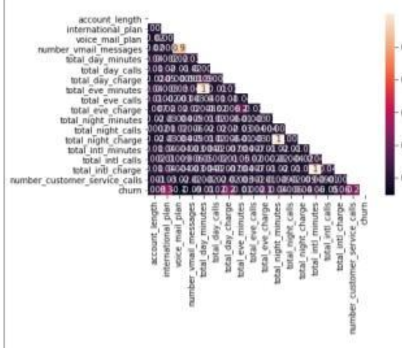
	total_night_minutes	total_night_calls
account_length	-0.024634	0.000225
international_plan	-0.030886	0.012451
voice_mail_plan	-0.004041	0.015553
number_vmail_messages	0.008688	0.006403
total_day_minutes	-0.010320	-0.023730
total_day_calls	0.018255	-0.017737
total_day_charge	0.006071	0.024047
total_eve_minutes	-0.013678	-0.025288
total_eve_calls	-0.003274	0.004325
total_eve_charge	-0.014501	0.003755
total_night_minutes	1.000000	0.006753
total_night_calls	0.006753	1.000000
total_night_charge	0.999999	0.006732
total_intl_minutes	-0.017365	-0.012991
total_intl_calls	-0.022412	-0.003979
total_intl_charge	-0.017372	-0.013005
number_customer_service_calls	-0.008703	-0.016395
churn	0.038658	0.006141

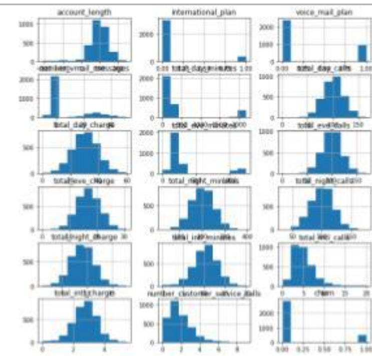
Figure 3

```
                              total_night_charge  total_intl_minutes  \
account_length                         -0.024632            0.013803
international_plan                     -0.030904            0.039241
voice_mail_plan                        -0.004065            0.003791
number_vmail_messages                   0.008666           -0.003024
total_day_minutes                      -0.010346           -0.003705
total_day_calls                         0.018248            0.022620
total_day_charge                        0.006048           -0.007459
total_eve_minutes                      -0.013704           -0.004367
total_eve_calls                        -0.003246            0.007207
total_eve_charge                       -0.014513           -0.011566
total_night_minutes                     0.999999           -0.017365
total_night_calls                       0.006732           -0.012991
total_night_charge                      1.000000           -0.017374
total_intl_minutes                     -0.017374            1.000000
total_intl_calls                       -0.022382            0.039101
total_intl_charge                      -0.017381            0.999993
number_customer_service_calls          -0.008688           -0.009266
churn                                   0.038656            0.064780
```

```
                              total_intl_calls  total_intl_charge  \
account_length                        0.002017           0.013827
international_plan                    0.014202           0.039179
voice_mail_plan                       0.008611           0.003833
number_vmail_messages                 0.005841          -0.003036
total_day_minutes                     0.049096          -0.003668
total_day_calls                       0.001861           0.022712
total_day_charge                      0.010836          -0.007411
total_eve_minutes                     0.049135          -0.004349
total_eve_calls                       0.021037           0.007170
total_eve_charge                      0.000245          -0.011606
total_night_minutes                  -0.022412          -0.017372
total_night_calls                    -0.003979          -0.013005
total_night_charge                   -0.022382          -0.017381
total_intl_minutes                    0.039101           0.999993
total_intl_calls                      1.000000           0.039144
total_intl_charge                     0.039144           1.000000
number_customer_service_calls        -0.012842          -0.009293
churn                                -0.050340           0.064792
```

```
                              number_customer_service_calls     churn
account_length                                     0.001093  0.007504
international_plan                                -0.029636  0.259852
voice_mail_plan                                   -0.017151 -0.102148
number_vmail_messages                             -0.021823 -0.094806
total_day_minutes                                 -0.001934 -0.013031
total_day_calls                                   -0.019913  0.019337
total_day_charge                                  -0.018828  0.203231
total_eve_minutes                                 -0.002751 -0.011928
total_eve_calls                                    0.002229  0.001964
total_eve_charge                                  -0.010224  0.098256
total_night_minutes                               -0.008703  0.038658
total_night_calls                                 -0.016395  0.006141
total_night_charge                                -0.008688  0.038656
total_intl_minutes                                -0.009266  0.064780
total_intl_calls                                  -0.012842 -0.050340
total_intl_charge                                 -0.009293  0.064792
number_customer_service_calls                      1.000000  0.205828
churn                                              0.205828  1.000000
```



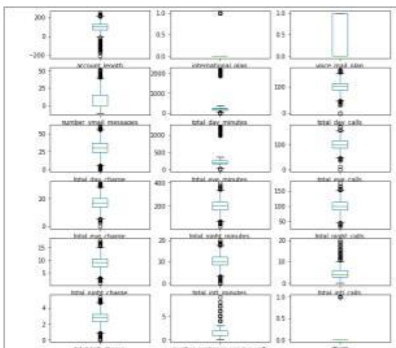


Top row : correlations (7 - 9)

Bottom left : correlation heat map matrix

Bottom middle : histogram of feature value distribution

Figure 4



```
account_length                  -1.168594
international_plan               2.726332
voice_mail_plan                  0.999140
number_vmail_messages            1.285841
total_day_minutes                2.203279
total_day_calls                 -0.125103
total_day_charge                -0.027607
total_eve_minutes                2.089711
total_eve_calls                 -0.062598
total_eve_charge                 0.002546
total_night_minutes              0.006069
total_night_calls                0.032500
total_night_charge               0.006017
total_intl_minutes              -0.249428
total_intl_calls                 1.323692
total_intl_charge               -0.249691
number_customer_service_calls    1.068860
churn                            2.018356
dtype: float64
```

```
[[0.739 0.    0.    0.164 0.921 0.6   0.481 0.89  0.629 0.483 0.592 0.415
  0.592 0.545 0.35  0.544 0.   ]
 [0.701 0.    0.    0.164 0.133 0.6   0.831 0.178 0.547 0.608 0.554 0.542
  0.554 0.7   0.45  0.7   0.222]
 [0.644 0.    0.    0.164 0.137 0.661 0.856 0.145 0.588 0.498 0.664 0.282
  0.664 0.585 0.2   0.585 0.   ]
 [  nan 0.    1.    0.656 0.05  0.43  0.314 0.147 0.635 0.501 0.432 0.387
  0.432 0.55  0.4   0.55  0.222]
 [0.646 0.    0.    0.164 0.154 0.727 0.962 0.183 0.682 0.625 0.352 0.57
  0.352 0.79  0.35  0.791 0.   ]]
```

```
[[ 0.578 -0.328 -0.618 -0.533  2.546 -0.066 -0.212  2.447  0.345 -0.502
   0.835 -0.414  0.836  0.24   1.03   0.235 -1.187]
 [ 0.223 -0.328 -0.618 -0.533 -0.203 -0.066  2.042 -0.322 -0.358  0.397
   0.555  0.506  0.554  1.345  1.845  1.345  0.333]
 [-0.32  -0.328 -0.618 -0.533 -0.189  0.433  2.202 -0.448 -0.006 -0.394
   1.366 -1.385  1.364  0.525 -0.192  0.526 -1.187]
 [   nan -0.328  1.617  1.648 -0.493 -1.464 -1.281 -0.443  0.396 -0.369
  -0.345 -0.619 -0.346  0.276  1.438  0.275  0.333]
 [-0.299 -0.328 -0.618 -0.533 -0.13   0.982  2.882 -0.303  0.798  0.523
  -0.938  0.71  -0.936  1.987  1.03   1.992 -1.187]]
```

Top left : box plots of feature values

Top middle : skew values of features

Top right : scaling the data

Bottom left : normalizing the data

Figure 5