Preprocessing Steps

- Look through data dictionary to understand features of the dataset
- Come up with initial guesses of which features are most/not important
- Subset dataset to only include customers that did not just join (Tosin) [SATURDAY]
- Go through dataset to make sure there are no missing features
    - If there are missing values in important features, come up with imputation techniques: mode for factor/categorical column, median for numerical, kNN for any (Abdullah)
-
- STANDARDIZATION OR NORMALIZATION: Scale the dataset (Abdullah) [SUNDAY]

Creating/Training Model

-
- Initial Run: Training and Testing Model with Logistic Regression (Tosin), SVM (Abdullah), Random Forests (Tosin), kNN (Abdullah) using all of the features [TUESDAY]
- Use initial run scores to figure which models to start experimenting with
    - Figure out the most important features using PCA and other techniques (Abdullah)
    - Parameter tuning (Abdullah) [WEDNESDAY]
    - Tbd…

Result Analysis

- Exploring Churn Probabilities instead a binary classification
- Come up with customer retention strategies based on results
- Visualization, Cluster Analysis

We want to get an estimate of the Churn rate of the telecom business. This does not include the customers that just joined. We want to come up with a rate of customers who, after experiencing the service, decide that they want to leave or stay. This can be gotten by just looking at the dataset, however, our goal is to try and come up with estimates as close as possible to the actual rates.

We need to make the dataset only include customers that have experienced the business more than a quarter.

<u>Project Run Down</u>

- Subsetting the dataset to include only either ‘stayed’ or ‘churned’
- Analyze missing values:
  No of missing cells/values:
  FALSE TRUE
  248394 1988
  **This does not include character cells that don’t have any value

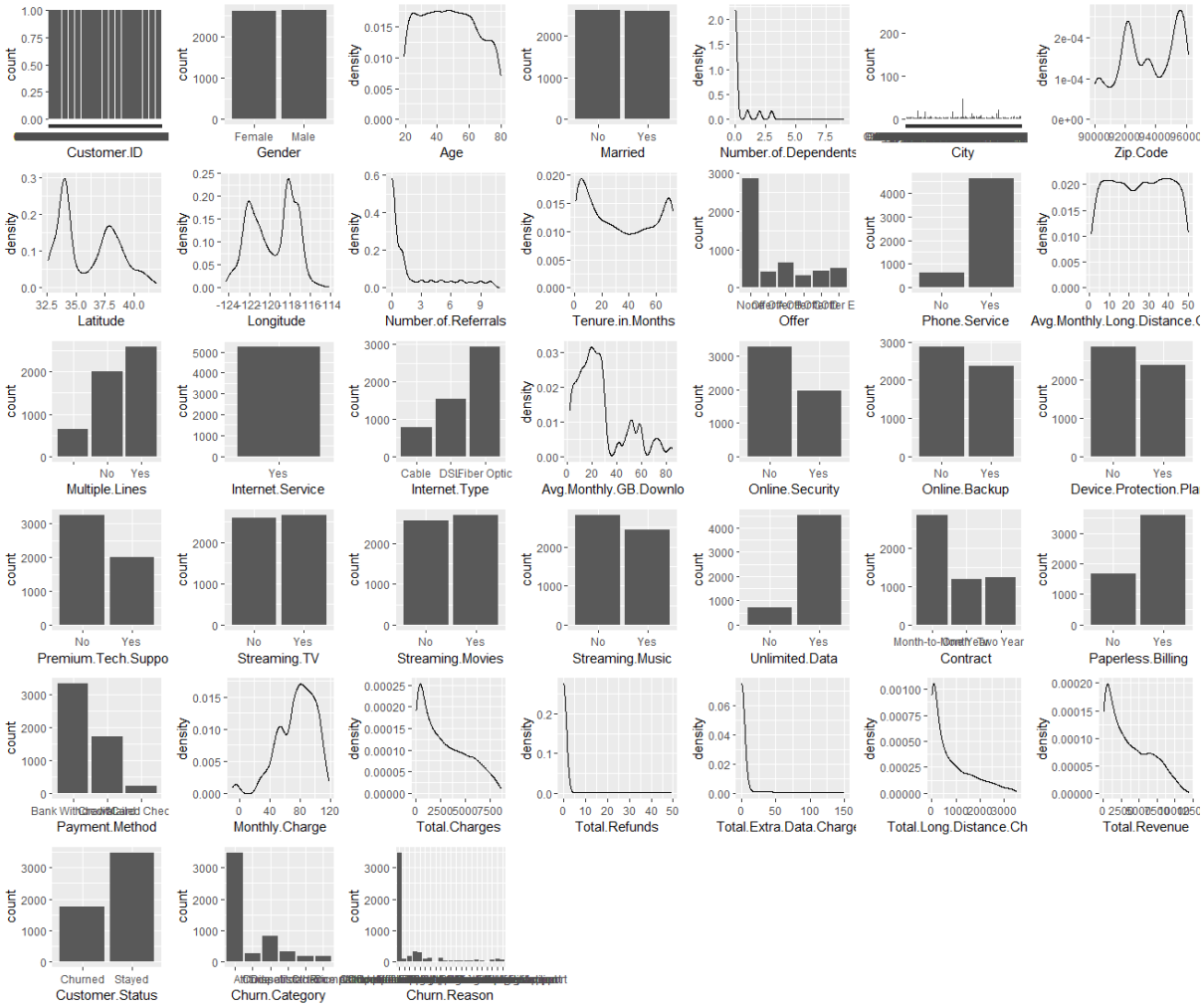
```
> summary(df)
     Customer.ID      Gender          Age        Married    Number.of.Dependents            City         Zip.Code        Latitude       Longitude      Number.of.Referrals
 0002-ORFBO:   1   Female:3277   Min.   :19.00   No :3271   Min.   :0.0000       San Diego    : 278   Min.   :90001   Min.   :32.56   Min.   :-124.3   Min.   : 0.000
 0003-MKNFE:   1   Male  :3312   1st Qu.:33.00   Yes:3318   1st Qu.:0.0000       Los Angeles  : 275   1st Qu.:92103   1st Qu.:33.99   1st Qu.:-121.8   1st Qu.: 0.000
 0004-TLHLJ:   1                Median :46.00              Median :0.0000       San Jose     : 110   Median :93526   Median :36.25   Median :-119.6   Median : 0.000
 0011-IGKFF:   1                Mean   :46.76              Mean   :0.4761       Sacramento   : 102   Mean   :93492   Mean   :36.20   Mean   :-119.8   Mean   : 2.021
 0013-EXCHZ:   1                3rd Qu.:60.00              3rd Qu.:0.0000       San Francisco:  97   3rd Qu.:95333   3rd Qu.:38.17   3rd Qu.:-118.0   3rd Qu.: 3.000
 0013-MHZWF:   1                Max.   :80.00              Max.   :9.0000       Fresno       :  61   Max.   :96150   Max.   :41.96   Max.   :-114.2   Max.   :11.000
 (Other)   :6583                                                                (Other)      :5666
 Tenure.in.Months     Offer      Phone.Service Avg.Monthly.Long.Distance.Charges Multiple.Lines Internet.Service     Internet.Type  Avg.Monthly.GB.Download Online.Security
 Min.   : 1.0     None   :3598   No : 644      Min.   : 1.01                     : 644        No :1344                   :1344   Min.   : 2.00             :1344
 1st Qu.:12.0     Offer A: 520   Yes:5945      1st Qu.:13.14                     No :3019        Yes:5245         Cable     : 774   1st Qu.:13.00          No :3272
 Median :32.0     Offer B: 824                 Median :25.72                     Yes:2926                         DSL       :1537   Median :21.00          Yes:1973
 Mean   :34.5     Offer C: 415                 Mean   :25.50                                                      Fiber Optic:2934   Mean   :26.23
 3rd Qu.:57.0     Offer D: 602                 3rd Qu.:37.69                                                                         3rd Qu.:30.00
 Max.   :72.0     Offer E: 630                 Max.   :49.99                                                                         Max.   :85.00
                                               NA's   :644                                                                           NA's   :1344
 Online.Backup Device.Protection.Plan Premium.Tech.Support Streaming.TV Streaming.Movies Streaming.Music Unlimited.Data          Contract    Paperless.Billing
    :1344         :1344                  :1344               :1344        :1344           :1344           :1344       Month-to-Month:3202   No :2615
 No :2870      No :2855               No :3248            No :2587     No :2562        No :2809        No : 724       One Year      :1526   Yes:3974
 Yes:2375      Yes:2390               Yes:1997            Yes:2658     Yes:2683        Yes:2436        Yes:4521       Two Year      :1861


        Payment.Method Monthly.Charge   Total.Charges     Total.Refunds    Total.Extra.Data.Charges Total.Long.Distance.Charges Total.Revenue      Customer.Status
 Bank Withdrawal:3728   Min.   :-10.00   Min.   :  18.85   Min.   : 0.000   Min.   :  0.00           Min.   :   0.0             Min.   :   21.61   Churned:1869
 Credit Card    :2518   1st Qu.: 35.80   1st Qu.: 544.55   1st Qu.: 0.000   1st Qu.:  0.00           1st Qu.: 106.7             1st Qu.:  835.45   Stayed :4720
 Mailed Check   : 343   Median : 71.05   Median :1563.90   Median : 0.000   Median :  0.00           Median : 472.7             Median : 2376.45
                        Mean   : 65.03   Mean   :2432.04   Mean   : 2.081   Mean   :  7.17           Mean   : 798.1             Mean   : 3235.22
                        3rd Qu.: 90.40   3rd Qu.:4003.00   3rd Qu.: 0.000   3rd Qu.:  0.00           3rd Qu.:1275.1             3rd Qu.: 5106.64
                        Max.   :118.75   Max.   :8684.80   Max.   :49.790   Max.   :150.00           Max.   :3564.7             Max.   :11979.34

        Churn.Category                      Churn.Reason
                :4720                              :4720
 Attitude       : 314   Competitor had better devices: 313
 Competitor     : 841   Competitor made better offer : 311
 Dissatisfaction: 321   Attitude of support person   : 220
 Other          : 182   Don't know                  : 130
 Price          : 211   Competitor offered more data : 117
                        (Other)                     : 778
```

***The variables ‘Phone Service’ and ‘Internet Service’ tells us a lot about potential NA values in other variables. For instance, if you examine the summary of the dataframe above, you will notice that the number of lines with no Phone service is 644, which is the same as the number of NA values in Avg.Monthly.Long.Distance.Charges and Multiple.Lines. The number of lines with no Internet Service is also the same number of NA values in 10 other variables.

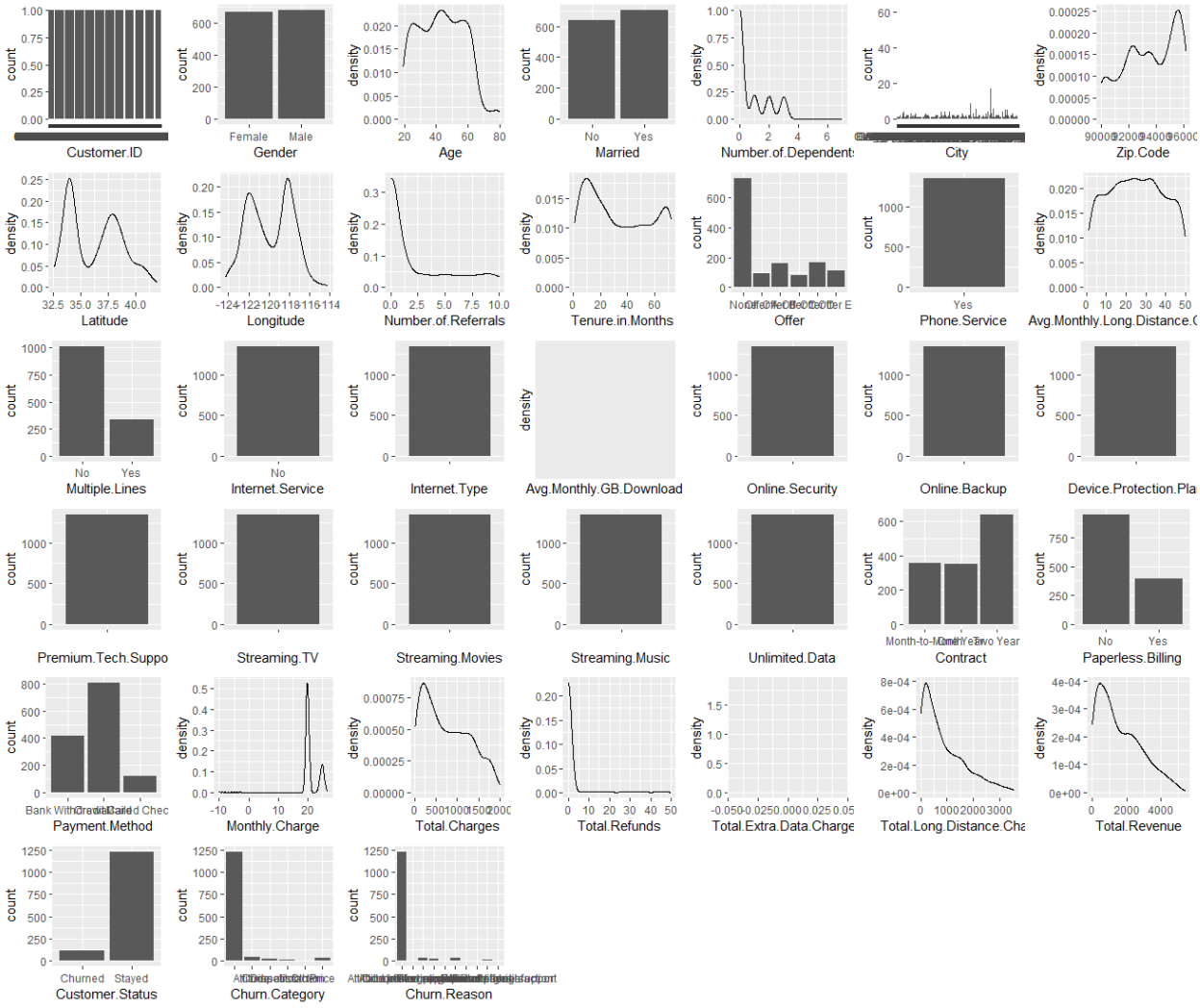
Also looks like no phone service actually guarantees that the line will definitely have internet service.

POSSIBLE SOLUTION: We create different models for lines with internet service and lines without internet service. Disadvantage of that is, the dataset of lines without internet service is imbalanced.

- Visualizing distribution of each feature:
  Lines with internet service:



Lines without internet service:



- Getting rid of (predicted) unimportant features:
- For dataset with Internet Service Lines: Customer ID, Longitude, Latitude, Zipcode, Internet Service, Churn Category, Churn Reason
- For dataset without Internet Service Lines: Customer ID, Longitude, Latitude, Zipcode, Internet Service, Phone.Service, Internet.Type, Avg.Monthly.GB.Download, Online.Security, Online.Backup, Device.Protection.Plan, Premium.Tech.Support, Streaming.TV, Streaming.Movies, Streaming.Music, Unlimited.Data, Total Extra Data Charges, Churn Category, Churn Reason
- How much imputation would we need to do for both datasets, after we have now removed some features?
- IMPUTATION: In the ISdf dataset, there are two features still with missing values: Average Monthly Long Distance Charges and Multiple Lines (Yes or No). These two features have the same number of NA's (644) which also coincides with the number of Lines that dont have a phone service (As mentioned previously). We will impute the NA values in these two features with 0 and 'No' respectively, because it makes sense for a

line without phone service to have no amount charged for long distance calling, and also for such an account to not have multiple lines.

- There does not seem to be any missing values in the dataset for accounts with no Internet Service
- STANDARDIZATION: Standardizing both datasets before splitting into train and test. Alternatively, we could use K-Fold Cross Validation, because of the dilemma of handling missing values (how should missing values be handled for the test dataset?)
- Both datasets were split into training and testing datasets with a split ratio of 75:25. We also maintained the event rate so that the same percentage of Churn in training dataset is also the same in the testing dataset
- CREATING MODEL: We first train both datasets on a Random Forests model to see how it performs with prediction:
  We need to adjust 'city' feature as it contains too many categories for the rf model to handle. If the count for a particular city is less than 30, its value changes to 'Other.' But first, we delete the feature to see how the model performs without that information.
- RESULTS:
- For ISdf:
  after training **random forest model**, we get the following confusion matrix:

	correctResponse_A	
response_A	Churned	Stayed
Churned	293	61
Stayed	146	811

```
Confusion Matrix and Statistics

          Reference
Prediction Churned Stayed
   Churned     293     61
   Stayed      146    811

               Accuracy : 0.8421
                 95% CI : (0.8212, 0.8614)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6276

 Mcnemar's Test P-Value : 5.27e-09

            Sensitivity : 0.6674
            Specificity : 0.9300
         Pos Pred Value : 0.8277
         Neg Pred Value : 0.8474
              Precision : 0.8277
                 Recall : 0.6674
                     F1 : 0.7390
             Prevalence : 0.3349
         Detection Rate : 0.2235
   Detection Prevalence : 0.2700
      Balanced Accuracy : 0.7987

       'Positive' Class : Churned
```

For NISdf:

```
Confusion Matrix and Statistics

          Reference
Prediction Churned Stayed
   Churned      20      0
   Stayed        8    307

               Accuracy : 0.9761
                 95% CI : (0.9535, 0.9896)
    No Information Rate : 0.9164
    P-Value [Acc > NIR] : 4.66e-06

                  Kappa : 0.8209

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 0.71429
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.97460
              Precision : 1.00000
                 Recall : 0.71429
                     F1 : 0.83333
             Prevalence : 0.08358
         Detection Rate : 0.05970
   Detection Prevalence : 0.05970
      Balanced Accuracy : 0.85714

       'Positive' Class : Churned
```

- **Logistic Regression Model Results**:
- For ISdf:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 325 113
         1 114 759

               Accuracy : 0.8268
                 95% CI : (0.8053, 0.847)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6111

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7403
            Specificity : 0.8704
         Pos Pred Value : 0.7420
         Neg Pred Value : 0.8694
              Precision : 0.7420
                 Recall : 0.7403
                     F1 : 0.7412
             Prevalence : 0.3349
         Detection Rate : 0.2479
   Detection Prevalence : 0.3341
      Balanced Accuracy : 0.8054

       'Positive' Class : 0
```

For NISdf:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  15   7
         1  13 300

               Accuracy : 0.9403
                 95% CI : (0.9093, 0.9632)
    No Information Rate : 0.9164
    P-Value [Acc > NIR] : 0.06431

                  Kappa : 0.5682

 Mcnemar's Test P-Value : 0.26355

            Sensitivity : 0.53571
            Specificity : 0.97720
         Pos Pred Value : 0.68182
         Neg Pred Value : 0.95847
              Precision : 0.68182
                 Recall : 0.53571
                     F1 : 0.60000
             Prevalence : 0.08358
         Detection Rate : 0.04478
   Detection Prevalence : 0.06567
      Balanced Accuracy : 0.75646

       'Positive' Class : 0
```

- Using kNN (k=70):
  ISdf=

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 304 151
         1 135 721

               Accuracy : 0.7818
                 95% CI : (0.7585, 0.8039)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5147

 Mcnemar's Test P-Value : 0.3751

            Sensitivity : 0.6925
            Specificity : 0.8268
         Pos Pred Value : 0.6681
         Neg Pred Value : 0.8423
              Precision : 0.6681
                 Recall : 0.6925
                     F1 : 0.6801
             Prevalence : 0.3349
         Detection Rate : 0.2319
   Detection Prevalence : 0.3471
      Balanced Accuracy : 0.7597

       'Positive' Class : 0
```

k=30

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 316 148
         1 123 724

               Accuracy : 0.7933
                 95% CI : (0.7703, 0.8149)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5424

 Mcnemar's Test P-Value : 0.1449

            Sensitivity : 0.7198
            Specificity : 0.8303
         Pos Pred Value : 0.6810
         Neg Pred Value : 0.8548
              Precision : 0.6810
                 Recall : 0.7198
                     F1 : 0.6999
             Prevalence : 0.3349
         Detection Rate : 0.2410
   Detection Prevalence : 0.3539
      Balanced Accuracy : 0.7750

       'Positive' Class : 0

Confusion Matrix and Statistics
```

k=10

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 308 145
         1 131 727

               Accuracy : 0.7895
                 95% CI : (0.7664, 0.8113)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5311

 Mcnemar's Test P-Value : 0.4339

            Sensitivity : 0.7016
            Specificity : 0.8337
         Pos Pred Value : 0.6799
         Neg Pred Value : 0.8473
              Precision : 0.6799
                 Recall : 0.7016
                     F1 : 0.6906
             Prevalence : 0.3349
         Detection Rate : 0.2349
   Detection Prevalence : 0.3455
      Balanced Accuracy : 0.7677

       'Positive' Class : 0
```

k=100

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 305 145
         1 134 727

               Accuracy : 0.7872
                 95% CI : (0.764, 0.8091)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5252

 Mcnemar's Test P-Value : 0.5494

            Sensitivity : 0.6948
            Specificity : 0.8337
         Pos Pred Value : 0.6778
         Neg Pred Value : 0.8444
              Precision : 0.6778
                 Recall : 0.6948
                     F1 : 0.6862
             Prevalence : 0.3349
         Detection Rate : 0.2326
   Detection Prevalence : 0.3432
      Balanced Accuracy : 0.7642

       'Positive' Class : 0
```

- Do summary and analysis of model results
- After initial Testing of models, analysis shows that it will be most effective to utilize a random forest model for prediction.
- Further testing using rf
- Correlation Matrix:

Total Extra Data Charges and Unlimited Data, DSL and Fiber Optic Internet Type, Total Revenue and Tenure in Months, Total Charges and Tenure in Months, are pairs that look to be correlated. We will delete Total Revenue, Total Extra Data Charges, and DSL Internet type.

PCA Results

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     2.5544 1.82945 1.38223 1.34359 1.31521 1.22570 1.20489 1.07469 1.06223 1.04497  1.0375 1.00734 0.99733  0.9655 0.95222 0.92013 0.89688
Proportion of Variance 0.1812 0.09297 0.05307 0.05015 0.04805 0.04173 0.04033 0.03208 0.03134 0.03033  0.0299 0.02819 0.02763  0.0259 0.02519 0.02352 0.02234
Cumulative Proportion  0.1812 0.27422 0.32729 0.37743 0.42548 0.46722 0.50754 0.53962 0.57097 0.60130  0.6312 0.65939 0.68702  0.7129 0.73810 0.76162 0.78396
                          PC18    PC19    PC20    PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28    PC29    PC30    PC31    PC32    PC33    PC34
Standard deviation     0.88243 0.86504 0.85836 0.84793 0.83633 0.8073 0.77332 0.70705 0.65191 0.62989 0.58451 0.54772 0.51109  0.4570 0.39658 0.35549 0.31226
Proportion of Variance 0.02163 0.02079 0.02047 0.01997 0.01943 0.0181 0.01661 0.01389 0.01181 0.01102 0.00949 0.00833 0.00726  0.0058 0.00437 0.00351 0.00271
Cumulative Proportion  0.80559 0.82638 0.84684 0.86681 0.88624 0.9043 0.92096 0.93484 0.94665 0.95767 0.96716 0.97549 0.98275  0.9886 0.99292 0.99643 0.99914
                          PC35      PC36
Standard deviation     0.17601 3.266e-15
Proportion of Variance 0.00086 0.000e+00
Cumulative Proportion  1.00000 1.000e+00
```

Since the first 26 variables in the data set seem to do a good job in explaining the total variance, We will remove the last 10 variables which are:

```
Online.Security.Yes
Online.Backup.Yes
Device.Protection.Plan.Yes
Premium.Tech.Support.Yes
Streaming.TV.Yes
Streaming.Movies.Yes
Streaming.Music.Yes
Unlimited.Data.Yes
Contract.One.Year
Contract.Two.Year
```

Updated Random Forest Results:

```
Confusion Matrix and Statistics

          Reference
Prediction Churned Stayed
   Churned     268     85
   Stayed      171    787

               Accuracy : 0.8047
                 95% CI : (0.7822, 0.8259)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5392

 Mcnemar's Test P-Value : 1.081e-07

            Sensitivity : 0.6105
            Specificity : 0.9025
         Pos Pred Value : 0.7592
         Neg Pred Value : 0.8215
              Precision : 0.7592
                 Recall : 0.6105
                     F1 : 0.6768
             Prevalence : 0.3349
         Detection Rate : 0.2044
   Detection Prevalence : 0.2693
      Balanced Accuracy : 0.7565

       'Positive' Class : Churned
```

Since we get lower scores for subsetted training data, we include some previously deleted features to see if model makes any improvement:

```
Confusion Matrix and Statistics

          Reference
Prediction Churned Stayed
   Churned     272     79
   Stayed      167    793

               Accuracy : 0.8124
                 95% CI : (0.7901, 0.8332)
    No Information Rate : 0.6651
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5567

 Mcnemar's Test P-Value : 2.908e-08

            Sensitivity : 0.6196
            Specificity : 0.9094
         Pos Pred Value : 0.7749
         Neg Pred Value : 0.8260
              Precision : 0.7749
                 Recall : 0.6196
                     F1 : 0.6886
             Prevalence : 0.3349
         Detection Rate : 0.2075
   Detection Prevalence : 0.2677
      Balanced Accuracy : 0.7645

       'Positive' Class : Churned
```

It looks like including more features from previously deleted variables makes the model more accurate. We can conclude that the initial selection of features is best for prediction.