DMBI Mini Project

Problem Statement

From the given churn data, we need to find out the reasons for a customer to churn out of the subscription of the company and give the company business ideas to help them sustain the exising customers

Link for the selected Dataset: https://raw.githubusercontent.com/rushankshah/DMBI-Mini-Project/main/Churn1.csv

In [1]:

```
data <- read.csv('churn1.csv')
```

In [2]:

```
head(data)
```

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	Inter
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	
5575-GNVDE	Male	0	No	No	34	Yes	No	
3668-QPYBK	Male	0	No	No	2	Yes	No	
7795-CFOCW	Male	0	No	No	45	No	No phone service	
9237-HQITU	Female	0	No	No	2	Yes	No	
9305-CDSKC	Female	0	No	No	8	Yes	Yes	

Let's check for the summary of data and preprocess it if at all required

In [3]:

```
summary(data)
```

```
     customerID       gender     SeniorCitizen     Partner    Dependents
0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933
0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110
0004-TLHLJ:   1                 Median :0.0000
0011-IGKFF:   1                 Mean   :0.1621
0013-EXCHZ:   1                 3rd Qu.:0.0000
0013-MHZWF:   1                 Max.   :1.0000
(Other)   :7037
    tenure      PhoneService          MultipleLines     InternetService
Min.   : 0.00   No : 682     No              :3390   DSL        :2421
1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
Median :29.00                Yes             :2971   No         :1526
Mean   :32.37
3rd Qu.:55.00
```

```
 Max.   :72.00                                                            

             OnlineSecurity              OnlineBackup 
 No                 :3498   No                 :3088  
 No internet service:1526   No internet service:1526  
 Yes                :2019   Yes                :2429  




            DeviceProtection              TechSupport  
 No                 :3095    No                 :3473  
 No internet service:1526    No internet service:1526  
 Yes                :2422    Yes                :2044  




              StreamingTV              StreamingMovies           Contract   
 No                 :2810   No                 :2785   Month-to-month:3875  
 No internet service:1526   No internet service:1526   One year      :1473  
 Yes                :2707   Yes                :2732   Two year      :1695  




 PaperlessBilling                   PaymentMethod  MonthlyCharges  
 No :2872         Bank transfer (automatic):1544   Min.   : 18.25  
 Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50  
                  Electronic check         :2365   Median : 70.35  
                  Mailed check             :1612   Mean   : 64.76  
                                                   3rd Qu.: 89.85  
                                                   Max.   :118.75  

  TotalCharges    Churn     
 Min.   :  18.8   No :5174  
 1st Qu.: 401.4   Yes:1869  
 Median :1397.5             
 Mean   :2283.3             
 3rd Qu.:3794.7             
 Max.   :8684.8             
 NA's   :11                 
```

In [4]:

```
install.packages('DataExplorer', dependencies = TRUE)
```

```
package 'DataExplorer' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
	C:\Users\Rushank Shah\AppData\Local\Temp\RtmpyktN2Q\downloaded_packages
```

In [5]:

```
# install.packages('scales')
library(DataExplorer)
install.packages('scales')
create_report(data)
```

```
Warning message:
"package 'DataExplorer' was built under R version 3.6.3"
package 'scales' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
	C:\Users\Rushank Shah\AppData\Local\Temp\RtmpyktN2Q\downloaded_packages


processing file: report.rmd
  |..                                                                    |   2%
   inline R code fragments

  |...                                                                   |   5%
```

```
label: global_options (with options)
List of 1
 $ include: logi FALSE

  |.....                                                                 |   7%
  ordinary text without R code

  |.......                                                               |  10%
label: introduce
  |........                                                              |  12%
  ordinary text without R code

  |..........                                                            |  14%
label: plot_intro
  |............                                                          |  17%
  ordinary text without R code

  |.............                                                         |  19%
label: data_structure
  |...............                                                       |  21%
  ordinary text without R code

  |.................                                                     |  24%
label: missing_profile
  |..................                                                    |  26%
  ordinary text without R code

  |....................                                                  |  29%
label: univariate_distribution_header
  |......................                                                |  31%
  ordinary text without R code

  |.......................                                               |  33%
label: plot_histogram
  |.........................                                             |  36%
  ordinary text without R code

  |...........................                                           |  38%
label: plot_density
  |............................                                          |  40%
  ordinary text without R code

  |..............................                                        |  43%
label: plot_frequency_bar
  |................................                                      |  45%
  ordinary text without R code

  |.................................                                     |  48%
label: plot_response_bar
  |...................................                                   |  50%
  ordinary text without R code

  |.....................................                                 |  52%
label: plot_with_bar
  |......................................                                |  55%
  ordinary text without R code

  |........................................                              |  57%
label: plot_normal_qq
  |..........................................                            |  60%
  ordinary text without R code

  |...........................................                           |  62%
label: plot_response_qq
  |.............................................                         |  64%
  ordinary text without R code

  |...............................................                       |  67%
label: plot_by_qq
```

```
  |................................................                      |  69%
  ordinary text without R code

  |..................................................                    |  71%
label: correlation_analysis
  |....................................................                  |  74%
  ordinary text without R code

  |.....................................................                 |  76%
label: principal_component_analysis
  |.......................................................               |  79%
  ordinary text without R code

  |.........................................................             |  81%
label: bivariate_distribution_header
  |..........................................................            |  83%
  ordinary text without R code

  |............................................................          |  86%
label: plot_response_boxplot
  |..............................................................        |  88%
  ordinary text without R code

  |...............................................................       |  90%
label: plot_by_boxplot
  |.................................................................     |  93%
  ordinary text without R code

  |...................................................................   |  95%
label: plot_response_scatterplot
  |....................................................................  |  98%
  ordinary text without R code

  |......................................................................| 100%
label: plot_by_scatterplot

output file: E:/R Programming/DMBI Mini Project/report.knit.md

"C:/Users/RUSHAN~1/ANACON~1/envs/DMBI_S~1/Scripts/pandoc" +RTS -K512m -RTS "E:/R Pro
gramming/DMBI Mini Project/report.utf8.md" --to html4 --from markdown+autolink_bare_
uris+tex_math_single_backslash --output pandoc33b868023df0.html --lua-filter "C:\Use
rs\RUSHAN~1\ANACON~1\envs\DMBI_S~1\Lib\R\library\RMARKD~1\RMARKD~1\lua\PAGEBR~1.LUA"
--lua-filter "C:\Users\RUSHAN~1\ANACON~1\envs\DMBI_S~1\Lib\R\library\RMARKD~1\RMARKD
~1\lua\LATEX-~1.LUA" --self-contained --variable bs3=TRUE --standalone --section-div
s --table-of-contents --toc-depth 6 --template "C:\Users\RUSHAN~1\ANACON~1\envs\DMBI
_S~1\Lib\R\library\RMARKD~1\rmd\h\DEFAUL~1.HTM" --no-highlight --variable highlightj
s=1 --variable theme=yeti --include-in-header "C:\Users\RUSHAN~1\AppData\Local\Temp
\RtmpyktN2Q\rmarkdown-str33b8610da3e.html" --mathjax --variable "mathjax-url:http
s://mathjax.rstudio.com/latest/MathJax.js?config=TeX-AMS-MML_HTMLorMML"
Output created: report.html
```

https://github.com/rushankshah/DMBI-Mini-Project/blob/main/Profile%20Report.pdf

Pre-Process the data

In [6]:

```
data <- na.omit(data)
```

In [7]:

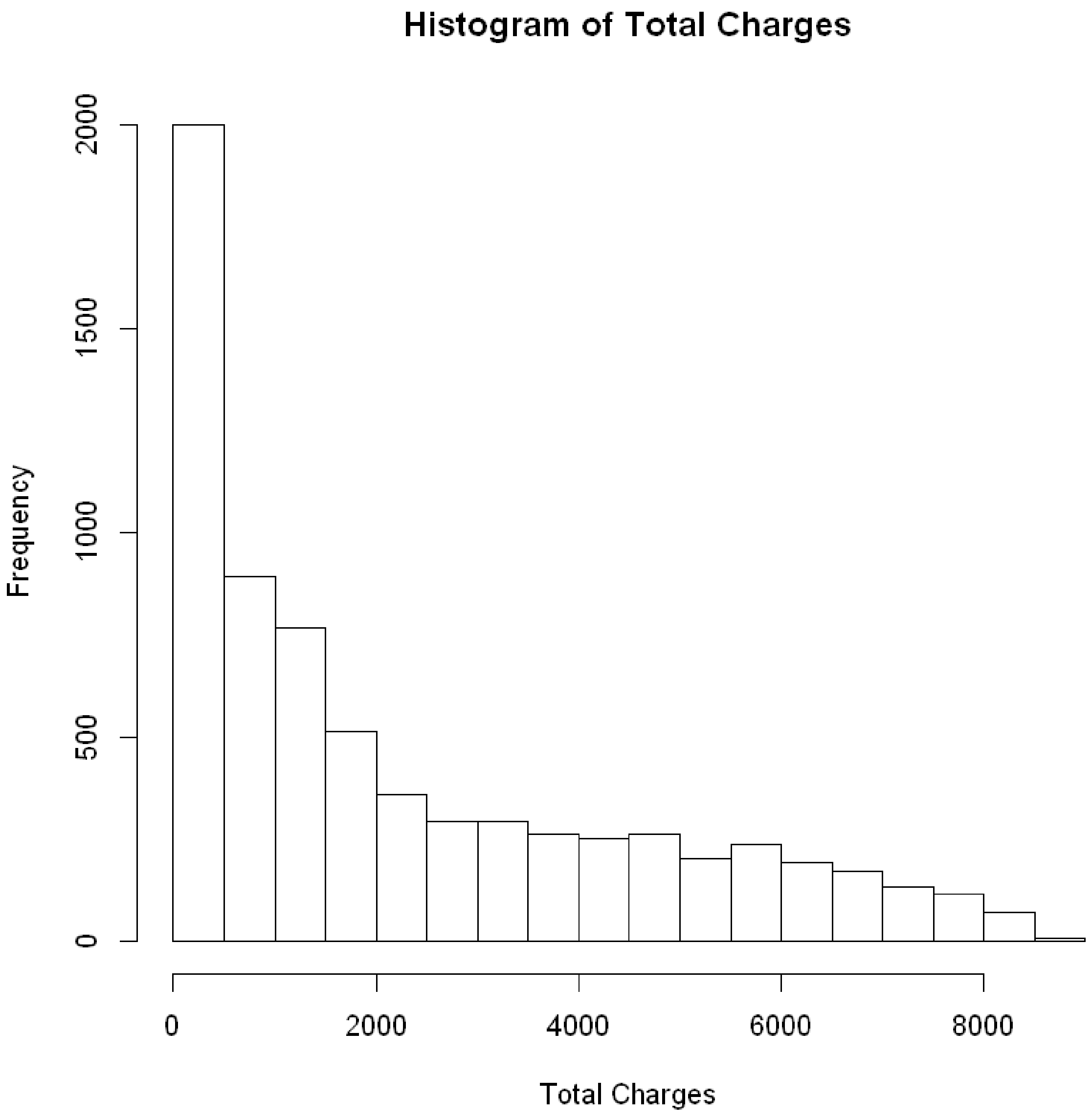
```
head(data)
```

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	Inter
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	Inter
5575-GNVDE	Male	0	No	No	34	Yes	No	
3668-QPYBK	Male	0	No	No	2	Yes	No	
7795-CFOCW	Male	0	No	No	45	No	No phone service	
9237-HQITU	Female	0	No	No	2	Yes	No	
9305-CDSKC	Female	0	No	No	8	Yes	Yes	

Start the Data Visualization

In [8]:
```
hist(data$TotalCharges, main = 'Histogram of Total Charges', xlab = 'Total Charges')
```



Adding Data visualization libraries

In [10]:

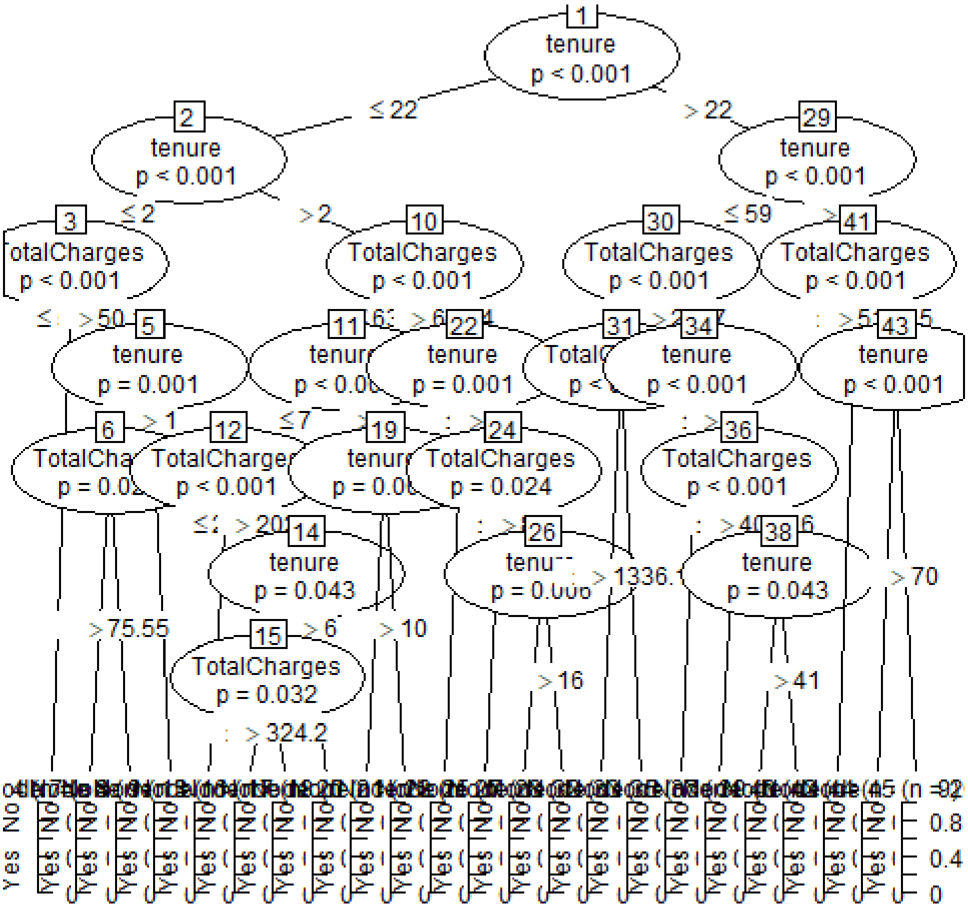
```
install.packages("party")
library(party)
```

```
package 'party' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Rushank Shah\AppData\Local\Temp\RtmpyktN2Q\downloaded_packages
Warning message:
"package 'party' was built under R version 3.6.3"Loading required package: grid
Loading required package: mvtnorm
Warning message:
"package 'mvtnorm' was built under R version 3.6.3"Loading required package: modelto
ols
Warning message:
"package 'modeltools' was built under R version 3.6.3"Loading required package: stat
s4
Loading required package: strucchange
Warning message:
"package 'strucchange' was built under R version 3.6.3"Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

Loading required package: sandwich
Warning message:
"package 'sandwich' was built under R version 3.6.3"
```

In [11]:

```
input.data <- data[c(1:4000),]
png(filename = "Decision Tree.png")
output.tree <- ctree(Churn ~ gender + TotalCharges + tenure, data = input.data)
plot(output.tree)
dev.off()
```

png: 2



In [12]:

```
install.packages('randomForest')
library(randomForest)
```

```
package 'randomForest' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Rushank Shah\AppData\Local\Temp\RtmpyktN2Q\downloaded_packages
Warning message:
"package 'randomForest' was built under R version 3.6.3"randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:ggplot2':

    margin
```

In [13]:

```
library(party)
library(randomForest)
output.forest <- randomForest(Churn ~ TotalCharges + gender + Contract + MultipleLin
print(output.forest)
print(importance(output.forest,type = 2))
```

```
Call:
 randomForest(formula = Churn ~ TotalCharges + gender + Contract +      MultipleLine
s + PhoneService, data = data)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 23.44%
Confusion matrix:
      No Yes class.error
No  4907 256  0.04958358
Yes 1392 477  0.74478331
              MeanDecreaseGini
TotalCharges        223.163677
gender                6.446596
Contract            380.203607
MultipleLines        34.376710
PhoneService          4.226486
```

From above we can conclude that TotalCharges and Contract are the only important attributes.

In [14]:

```
output <- glm(formula = Churn ~  PhoneService + MultipleLines + OnlineSecurity + Dev
print(summary(output))
```

```
Call:
glm(formula = Churn ~ PhoneService + MultipleLines + OnlineSecurity + 
    DeviceProtection + TechSupport + StreamingTV + StreamingMovies + 
    Contract + PaperlessBilling, family = binomial, data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4753  -0.7452  -0.2958   0.9401   2.9584  

Coefficients: (5 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         -0.50694    0.11571  -4.381 1.18e-05 ***
PhoneServiceYes                      0.18927    0.11059   1.712 0.086983 .  
MultipleLinesNo phone service             NA         NA      NA       NA    
MultipleLinesYes                     0.08989    0.06949   1.294 0.195812    
OnlineSecurityNo internet service   -1.43362    0.12103 -11.845  < 2e-16 ***
OnlineSecurityYes                   -0.71703    0.07948  -9.021  < 2e-16 ***
DeviceProtectionNo internet service       NA         NA      NA       NA    
DeviceProtectionYes                 -0.19211    0.07369  -2.607 0.009136 ** 
TechSupportNo internet service            NA         NA      NA       NA    
TechSupportYes                      -0.58770    0.08058  -7.294 3.01e-13 ***
StreamingTVNo internet service            NA         NA      NA       NA    
StreamingTVYes                       0.25056    0.07427   3.374 0.000742 ***
StreamingMoviesNo internet service        NA         NA      NA       NA    
StreamingMoviesYes                   0.26179    0.07409   3.533 0.000411 ***
ContractOne year                    -1.48349    0.09612 -15.433  < 2e-16 ***
ContractTwo year                    -2.61219    0.15985 -16.341  < 2e-16 ***
PaperlessBillingYes                  0.39307    0.07034   5.588 2.30e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8143.4  on 7031  degrees of freedom
Residual deviance: 6276.7  on 7020  degrees of freedom
AIC: 6300.7

Number of Fisher Scoring iterations: 6
```

PhoneService : Yes

MultipleLines : Yes

InternetService : Fiber optic

OnlineSecurity : No

OnlineBackup : No

DeviceProtection : No

TechSupport : No

StreamingTV : No

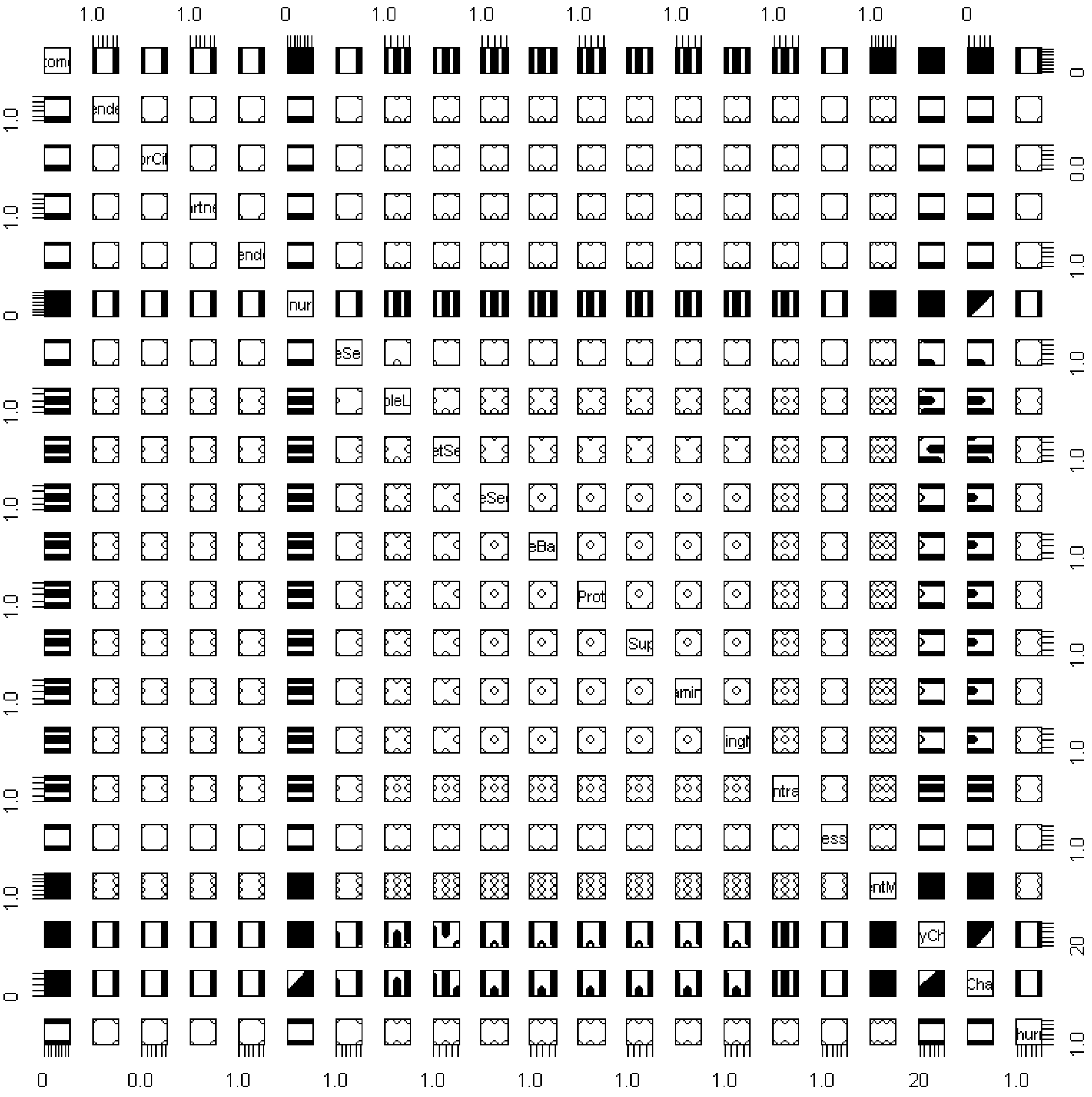
StreamingMovies : No

PaperlessBilling : Yes

These are the services which are important for the company to know about.

In [15]:

```
plot(data)
```



Conclusion

From the above analysis we can conclude that:

- The company should focus on the products which it's customer are loving the most and remove some services which the customers are not using.
- The company should lower it's Total Charges as it is one of the reasons for churning out the user

- The company can provide some services like Online Backup and Online Security at a very nominal cost because these the services it's customers are loving and churned ou customers are those who have not subscribed to these services

In [ ]: