Kaggle客戶流失率預測

Group 3

統計三 陳采宗
統計三 林盈盈
統計三 鄭雅云
統計三 沈冠宇

目錄

1 資料簡介

Raw data

raw data 包含 7043 個觀測值以及 20 個變數

Input format

Input	Format
customerID	string
gender	binary
SeniorCitizen	binary
Partner	binary
Dependents	binary
tenure	integer
PhoneService	binary
MultipleLines	multinomial
InternetService	multinomial
OnlineSecurity	multinomial

Input	Format
OnlineBackup	multinomial
DeviceProtection	multinomial
TechSupport	multinomial
StreamingTV	multinomial
StreamingMovies	multinomial
Contract	multinomial
PaperlessBilling	binary
PaymentMethod	multinomial
MonthlyCharges	numeric
TotalCharges	numeric

Data-preprocessing

➢ NA值處理：使用mice套件填補

➢ 原資料的SeniorCitizen欄位值為1、0，將其改為Yes、No

2 EDA

Churn
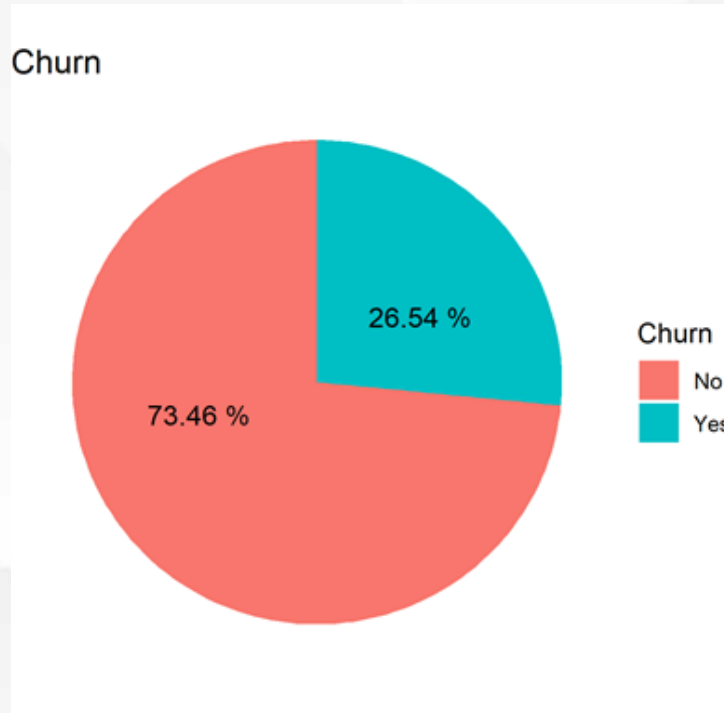
Gender & Churn-gender
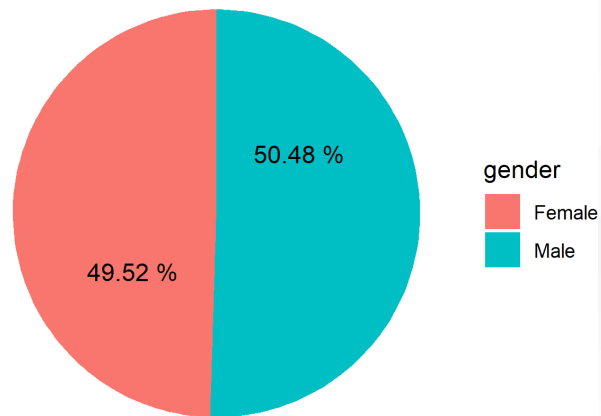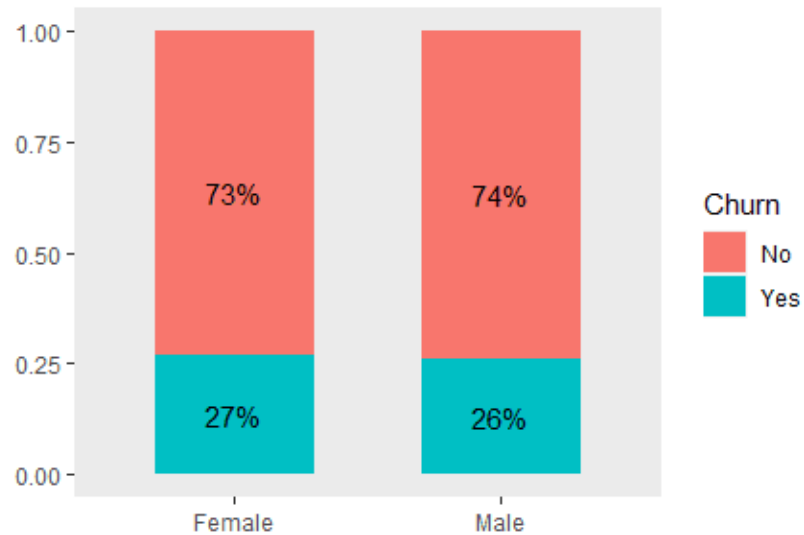
SeniorCitizen & Churn-SeniorCitizen
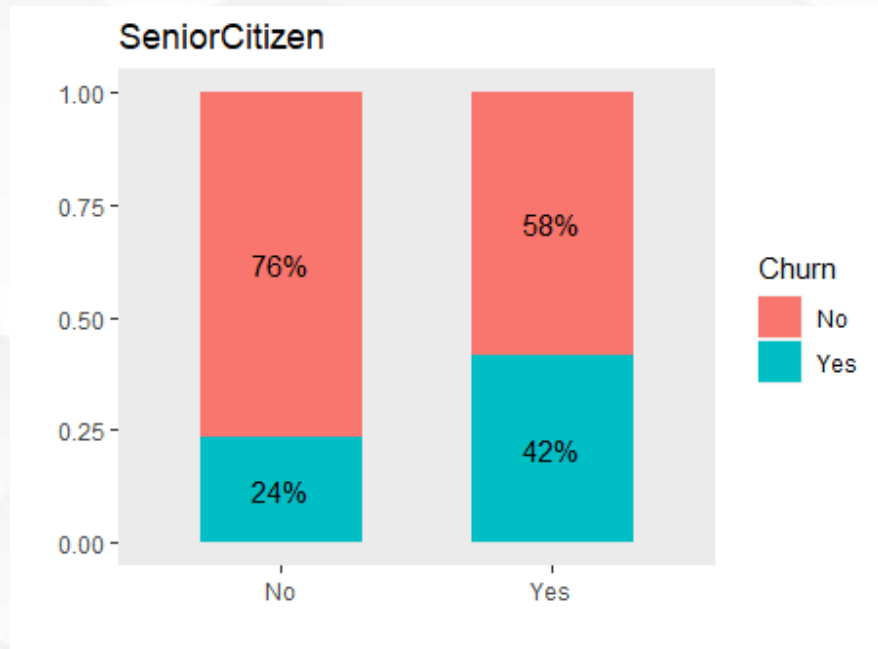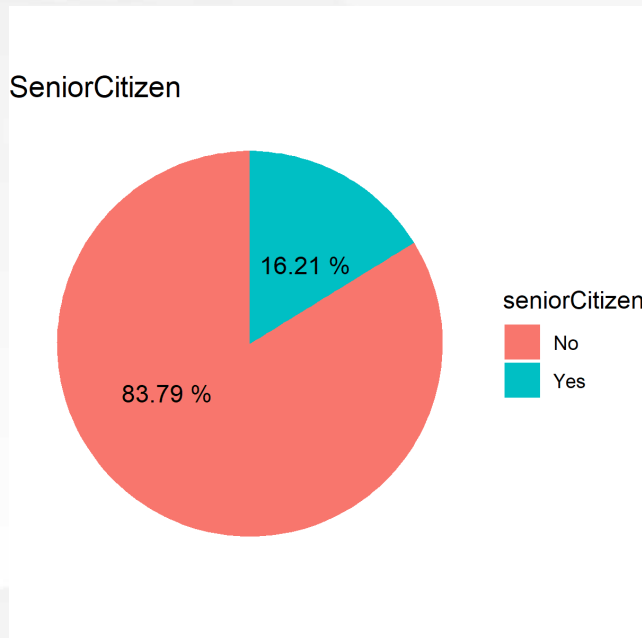
Partner & Churn-Partner
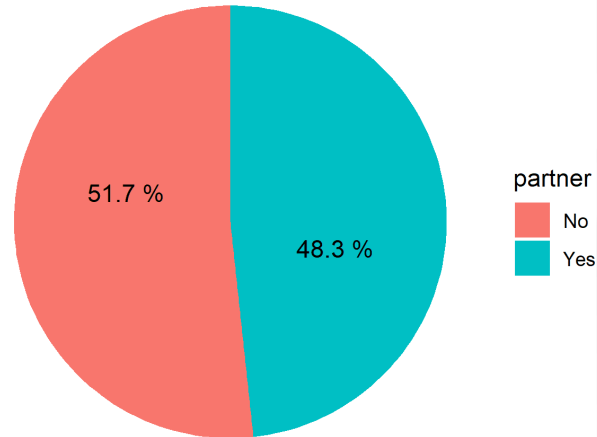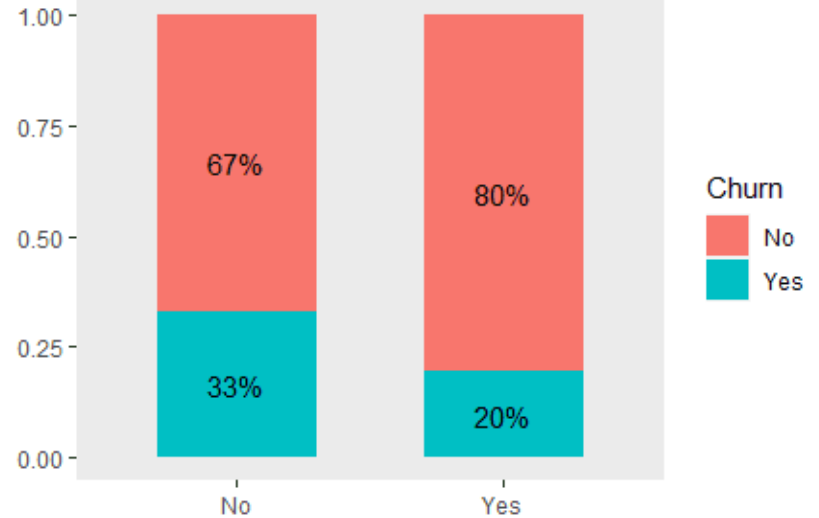
Dependents & Churn-Dependents
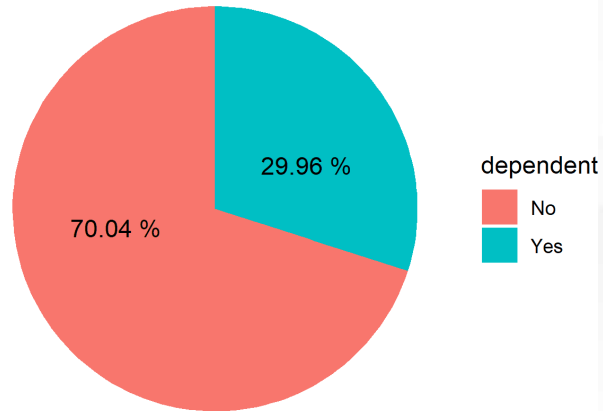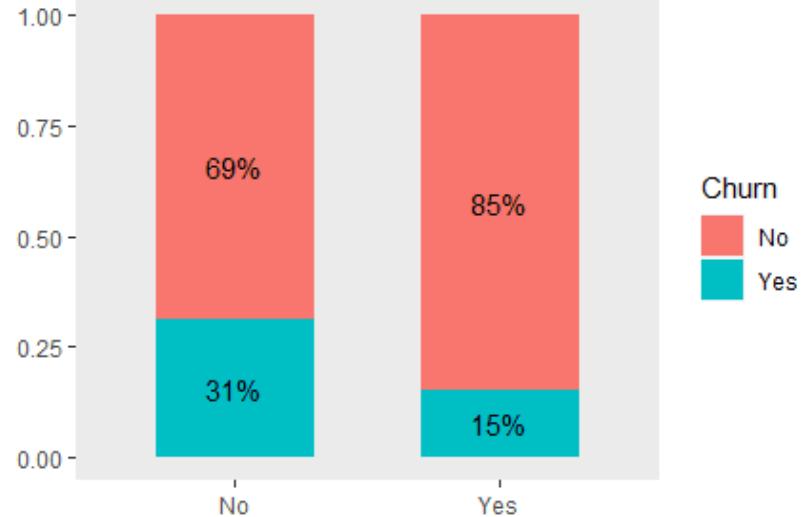
tenure & Churn-tenure
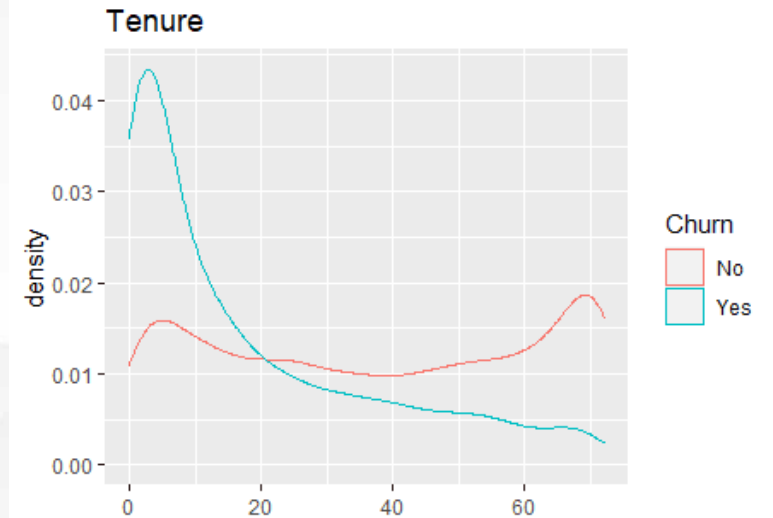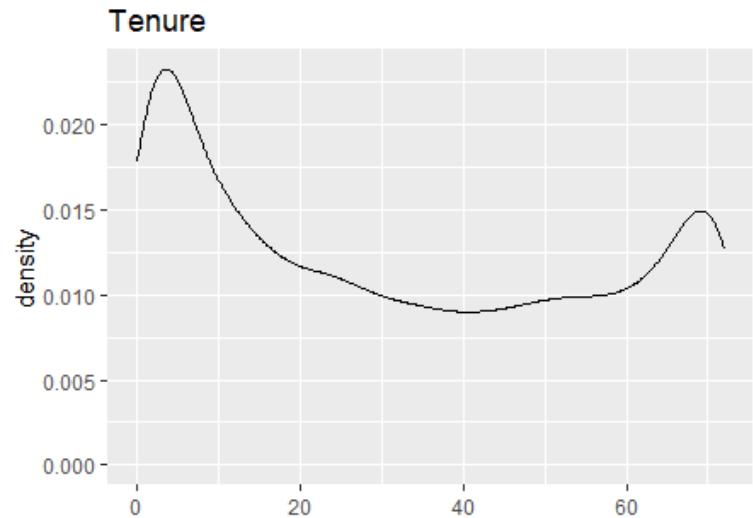
PhoneService & Churn-PhoneService
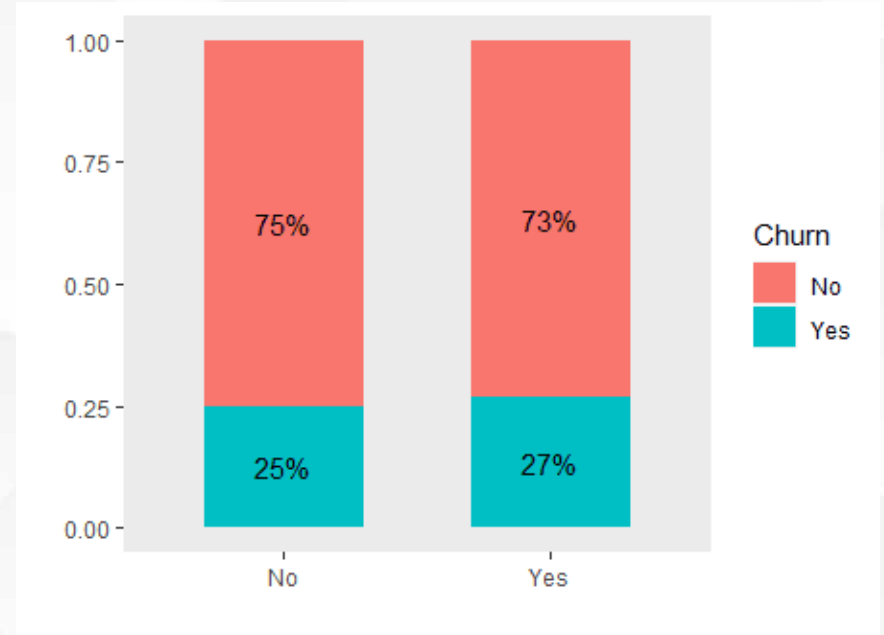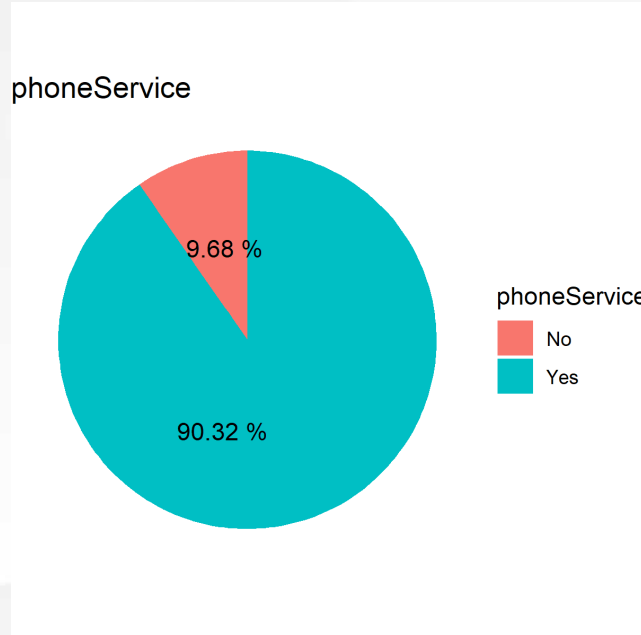
MultipleLines & Churn-MultipleLines
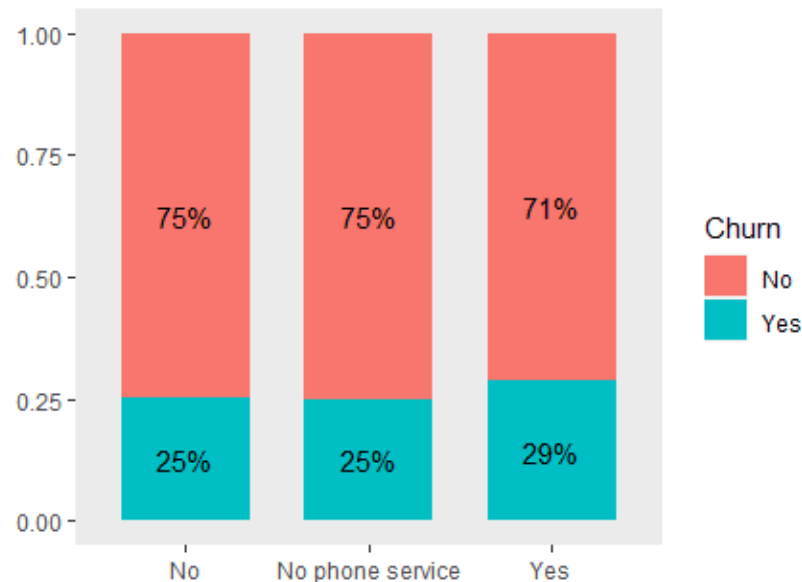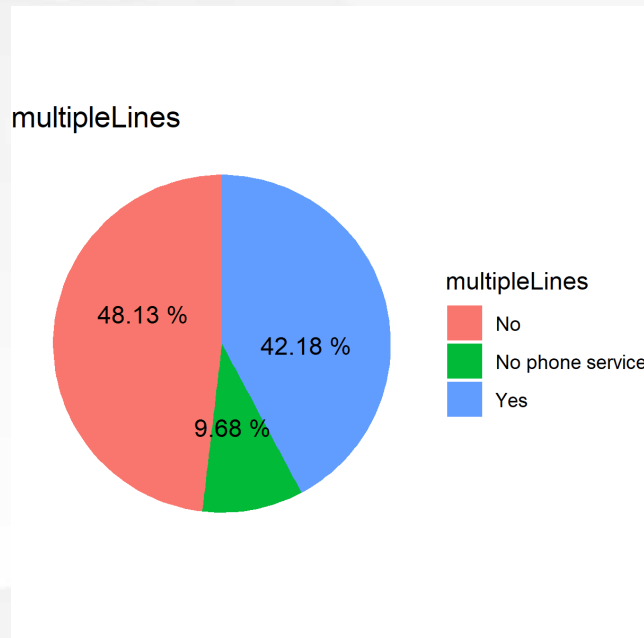
InternetService & Churn-InternetService
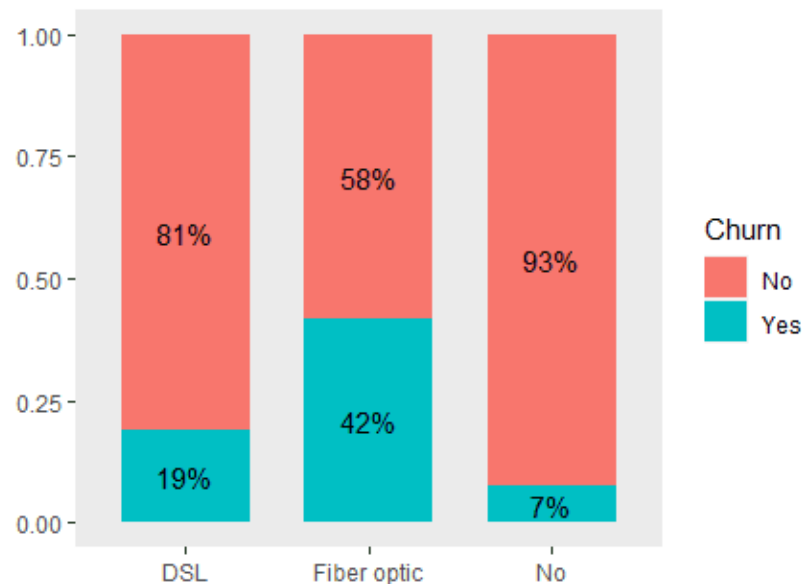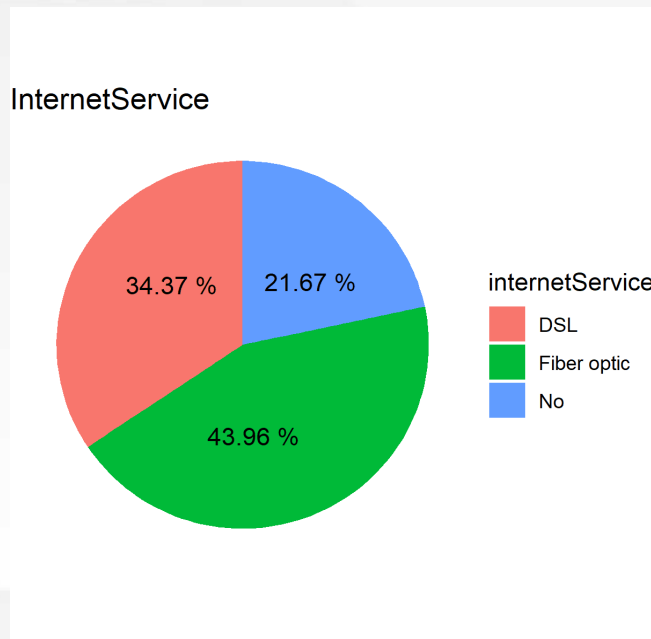
OnlineSecurity & Churn-OnlineSecurity
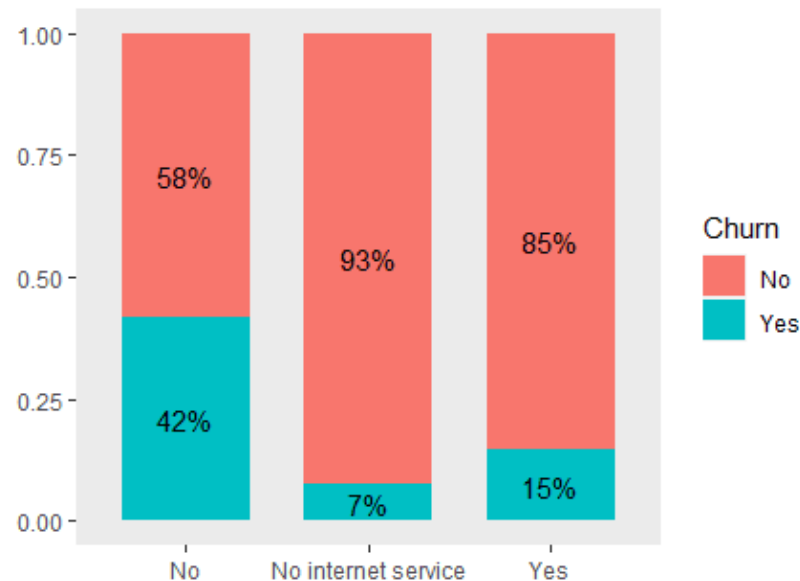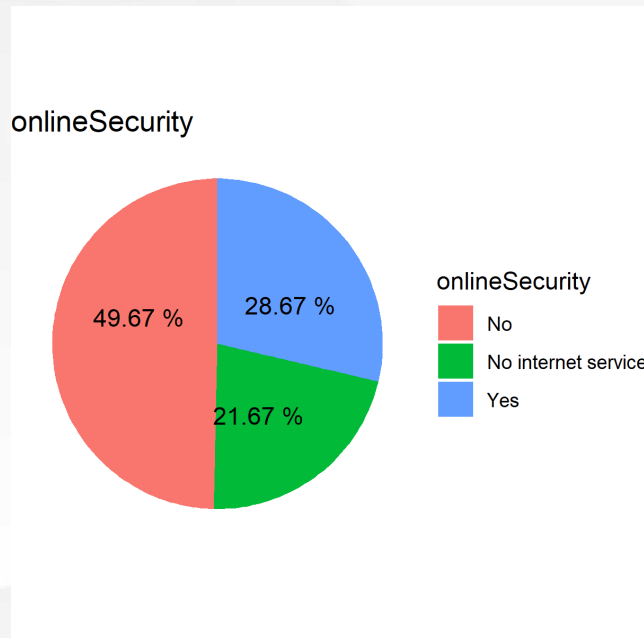
OnlineBackup & Churn-OnlineBackup
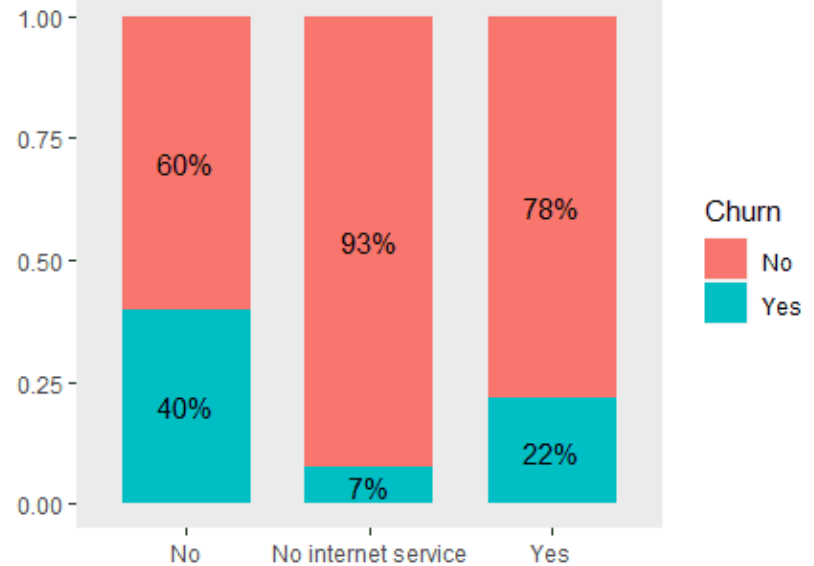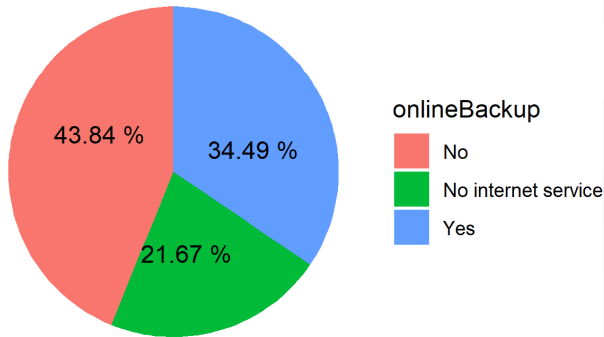
DeviceProtection & Churn-DeviceProtection
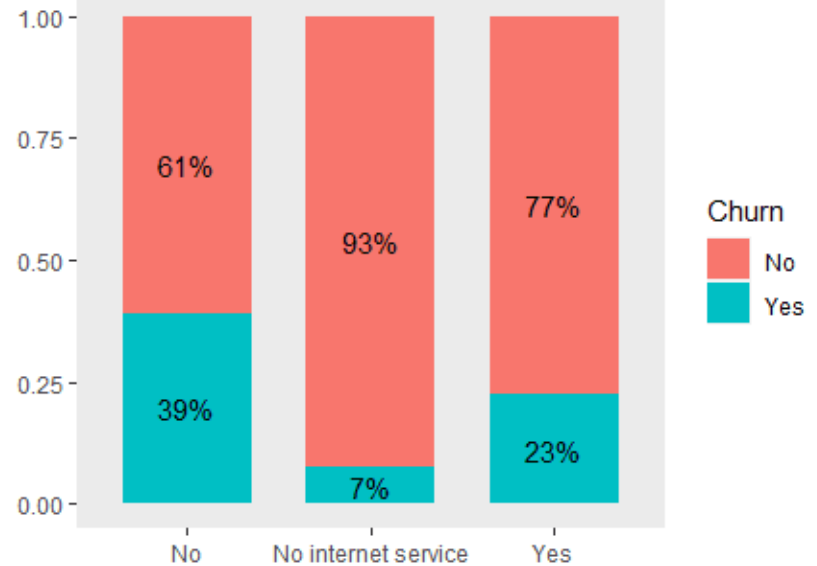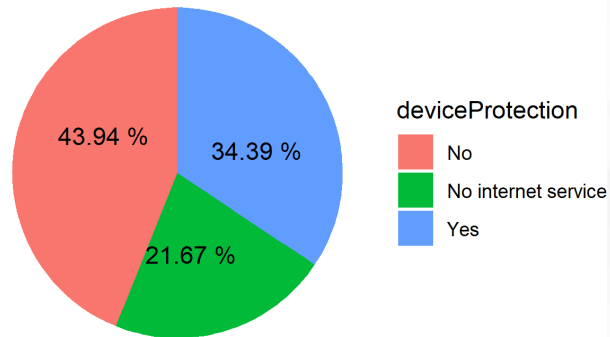
TechSupport & Churn-TechSupport
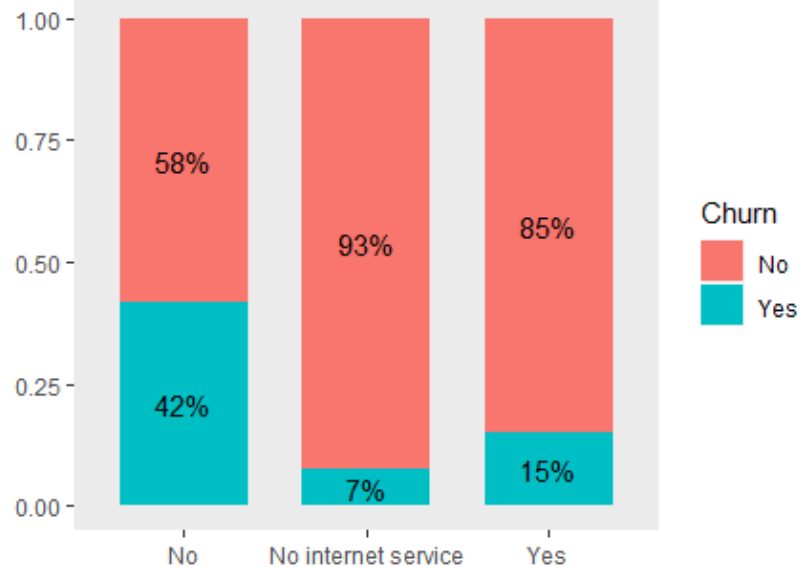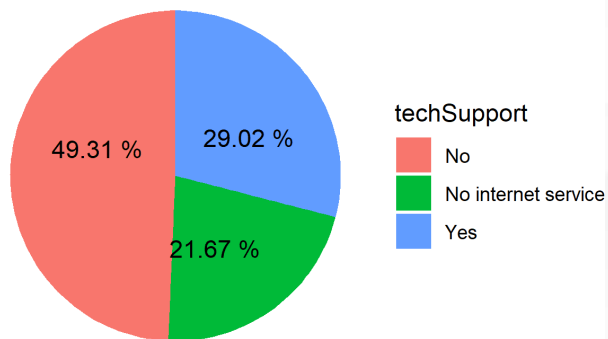
StreamingTV & Churn-StreamingTV
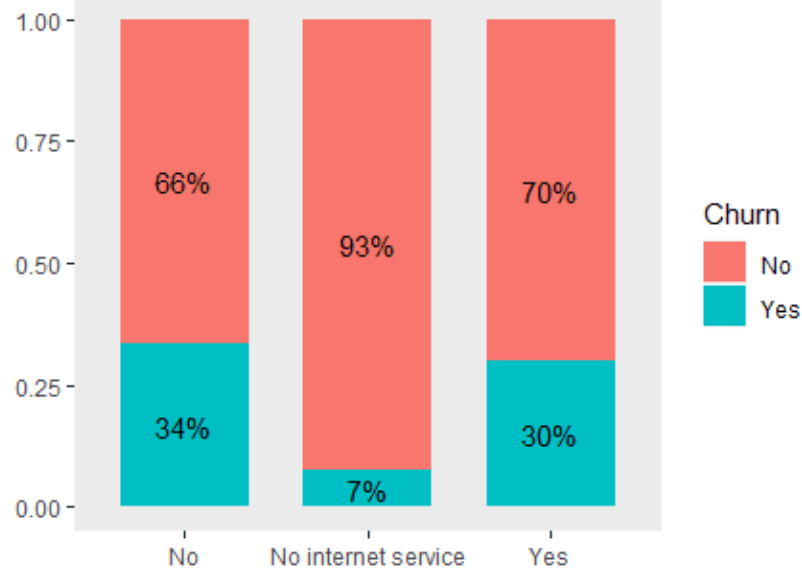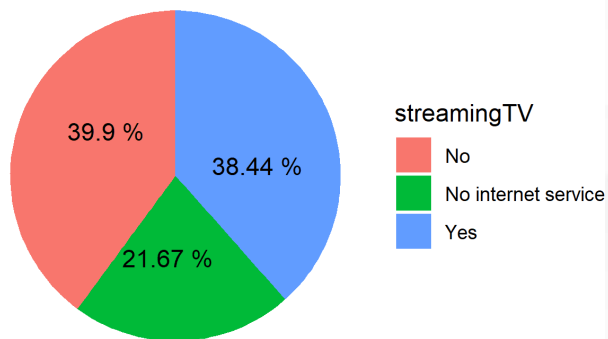
StreamingMovies & Churn-StreamingMovies
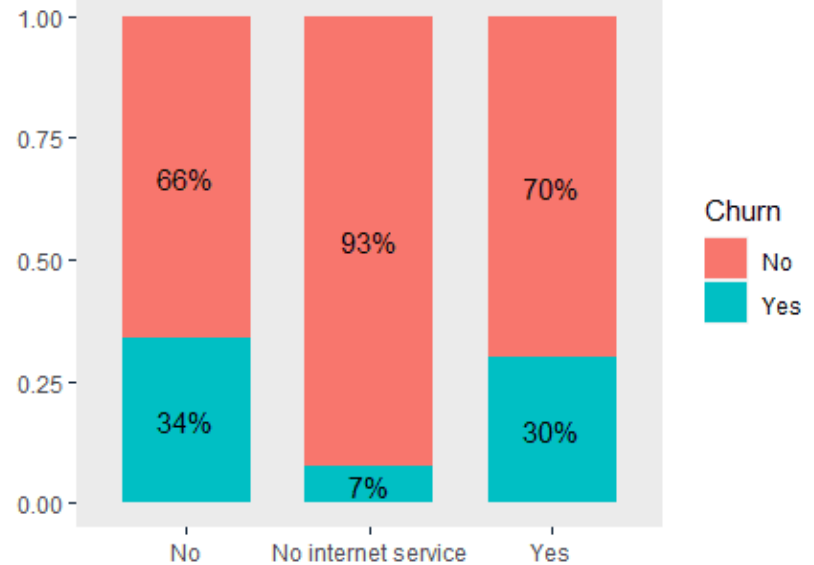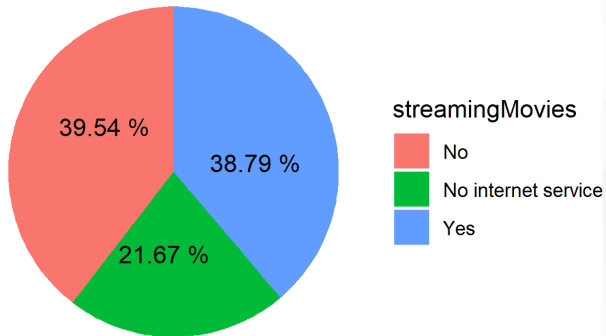
Contract & Churn-Contract
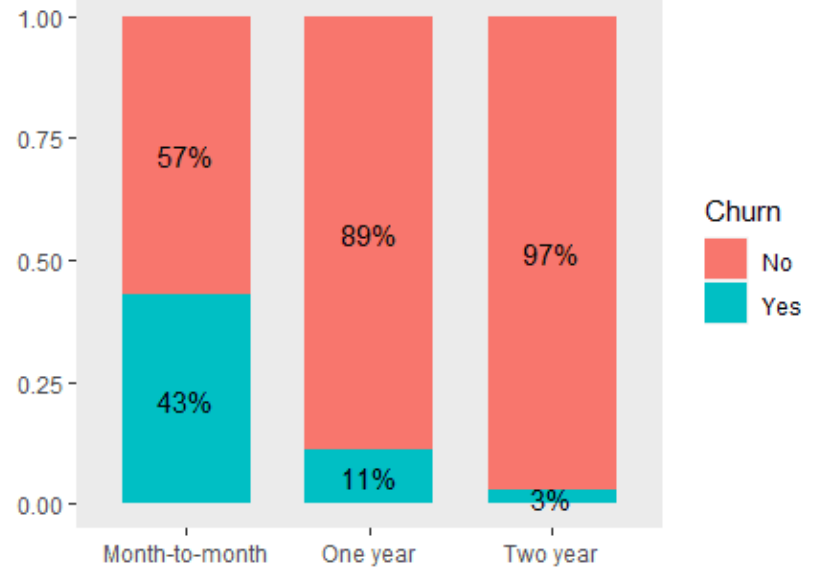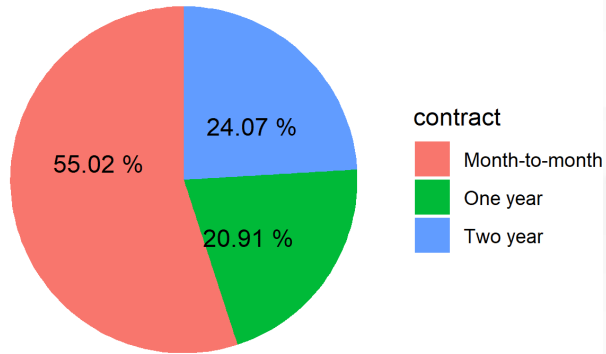
PaperlessBilling & Churn-PaperlessBilling
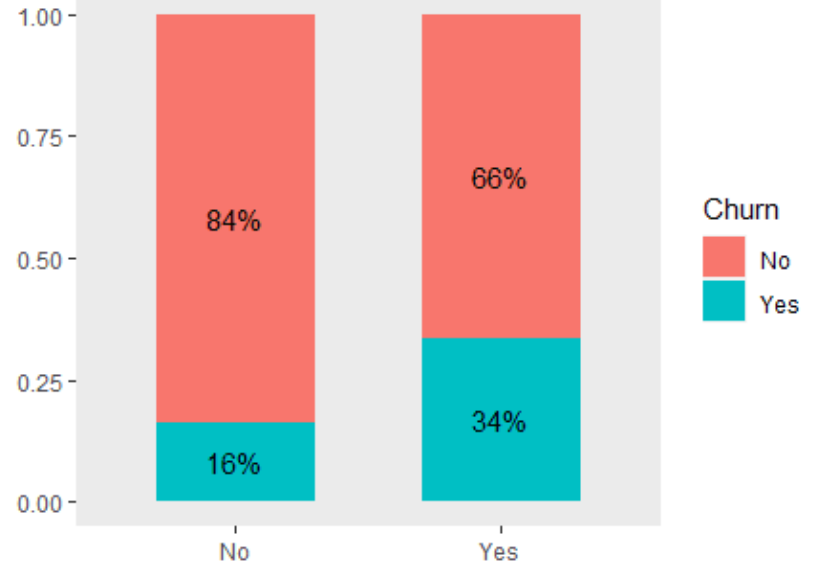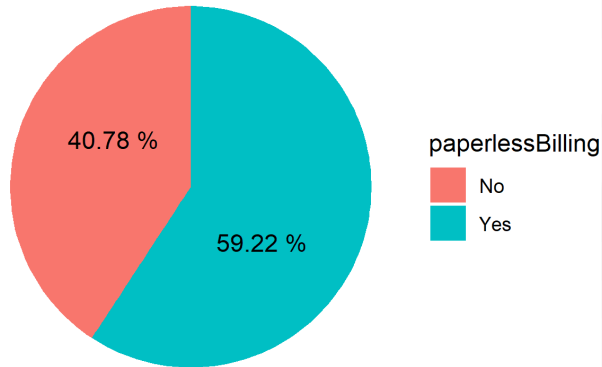
PaymentMethod & Churn-PaymentMethod
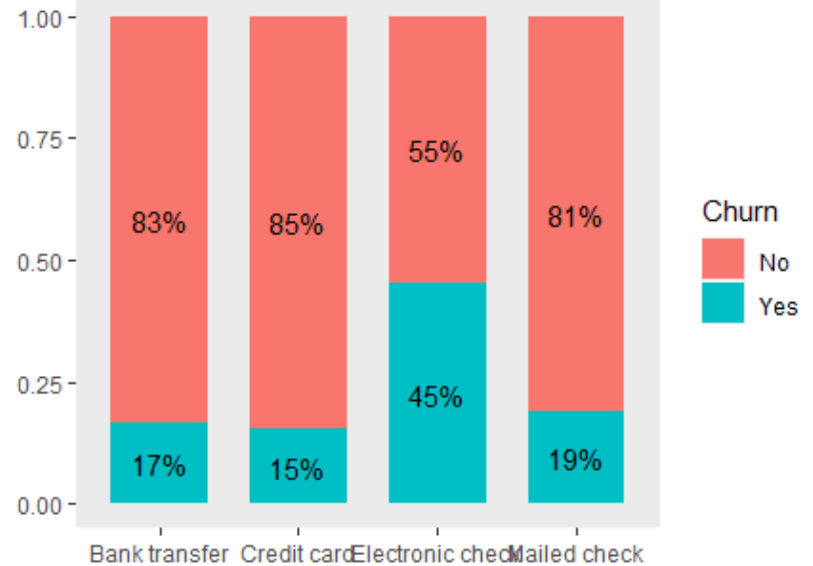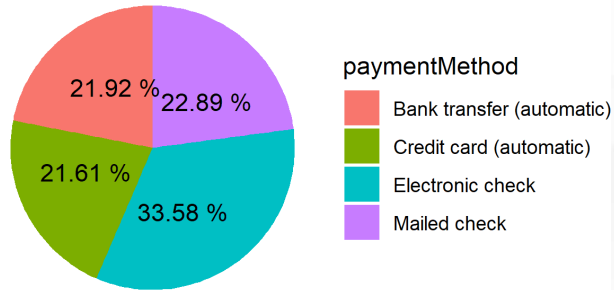
MonthlyCharges & Churn-MonthlyCharges
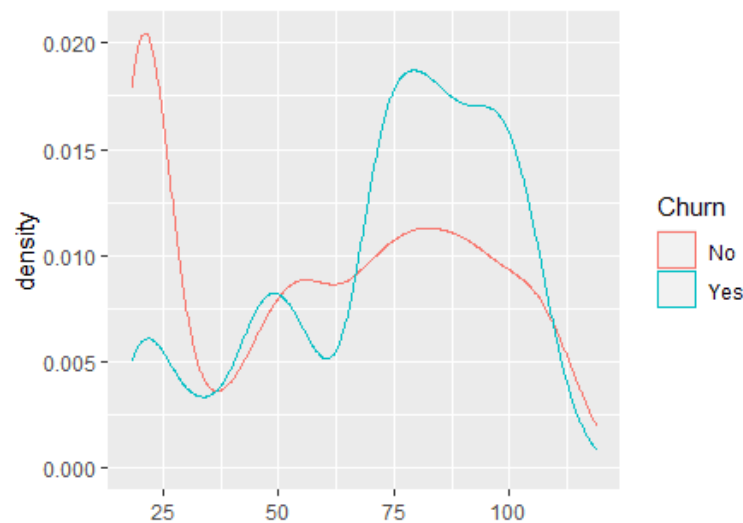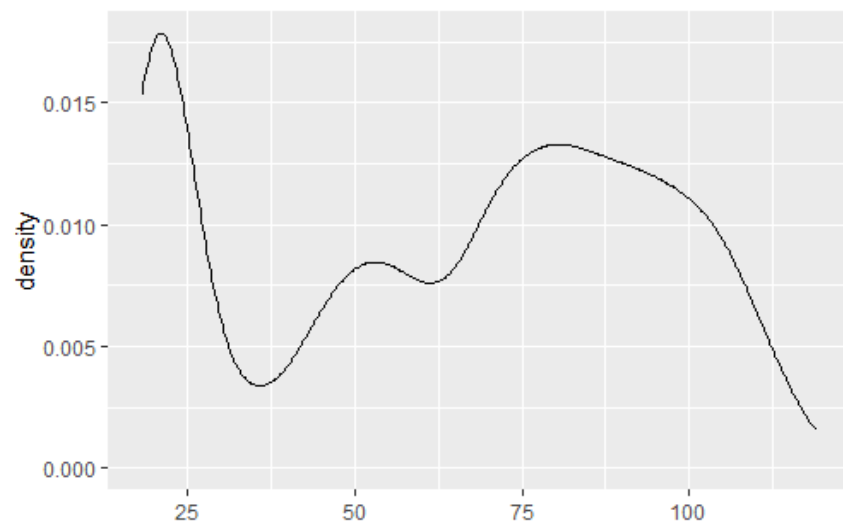
TotalCharges & Churn-TotalCharges
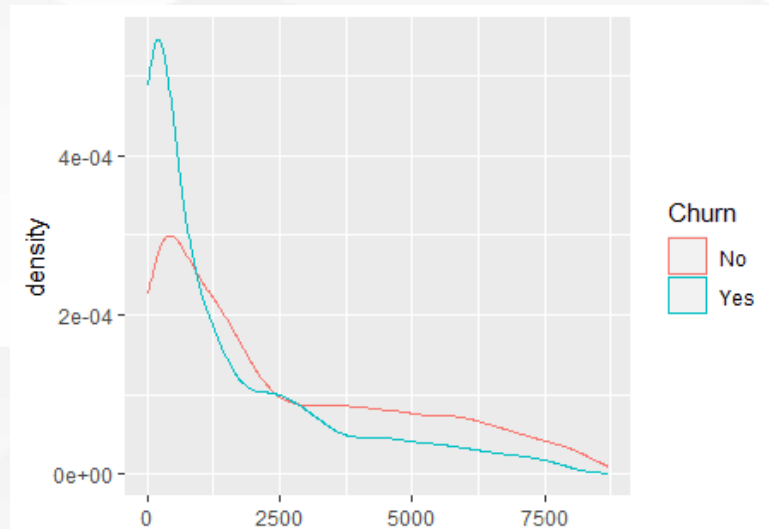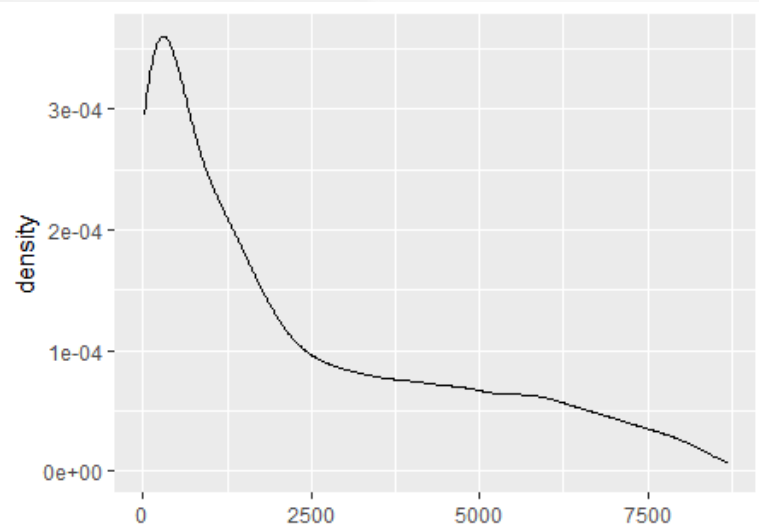
3 Model

原始Model

```
> str(d)
'data.frame':    7043 obs. of  20 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..:
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)"
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1
```

➢ 使用mice套件填補NA
➢ 全變數選用

```
> print(accuracy)
         [,1]            [,2]        [,3]          [,4]
 [1,]  "set"          "training"  "validation"  "test"
 [2,]  "logistic1"    "0.79"      "0.79"        "0.81"
 [3,]  "logistic2"    "0.79"      "0.78"        "0.79"
 [4,]  "logistic3"    "0.79"      "0.81"        "0.79"
 [5,]  "rpart4"       "0.79"      "0.77"        "0.79"
 [6,]  "rpart5"       "0.79"      "0.8"         "0.77"
 [7,]  "rpart6"       "0.8"       "0.78"        "0.81"
 [8,]  "randomforest7" "0.98"     "0.8"         "0.79"
 [9,]  "randomforest8" "0.98"     "0.8"         "0.8"
[10,]  "randomforest9" "0.98"     "0.82"        "0.8"
[11,]  "ave."         "0.85"      "0.79"        "0.79"
```

Model 1

➤ 使用EDA挑選變數

```
> str(d)
'data.frame':	7043 obs. of  13 variables:
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 2
 $ tenure          : int  13 66 11 63 26 62 61 16 55 4 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 3
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 3 1
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 1 2
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)"
 $ MonthlyCharges  : num  48.8 25.1 25.2 39.4 69 ...
 $ TotalCharges    : num  633 1698 321 2395 1816 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1
```

```
> print(accuracy)
       [,1]            [,2]        [,3]         [,4]
 [1,] "set"          "training"  "validation" "test"
 [2,] "logistic1"    "0.78"      "0.82"       "0.82"
 [3,] "logistic2"    "0.79"      "0.77"       "0.81"
 [4,] "logistic3"    "0.79"      "0.8"        "0.77"
 [5,] "rpart4"       "0.8"       "0.79"       "0.8"
 [6,] "rpart5"       "0.8"       "0.75"       "0.8"
 [7,] "rpart6"       "0.8"       "0.77"       "0.75"
 [8,] "randomforest7" "0.96"     "0.81"       "0.78"
 [9,] "randomforest8" "0.96"     "0.78"       "0.81"
[10,] "randomforest9" "0.95"     "0.82"       "0.79"
[11,] "ave."         "0.85"      "0.79"       "0.79"
```

Model 2

➢ 挑選羅吉斯顯著之變數

```
Coefficients: (2 not defined because of singularities)
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.589e+00  4.389e-01   3.622 0.000293 ***
SeniorCitizenYes                 2.783e-01  8.925e-02   3.118 0.001822 **
DependentsYes                   -1.474e-01  8.543e-02  -1.725 0.084461 .
tenure                          -5.888e-02  6.534e-03  -9.011 < 2e-16 ***
MultipleLinesNo phone service   -3.093e-01  2.131e-01  -1.451 0.146672
MultipleLinesYes                 5.124e-01  9.273e-02   5.526 3.28e-08 ***
InternetServiceFiber optic       2.013e+00  2.135e-01   9.431 < 2e-16 ***
InternetServiceNo               -1.949e+00  2.848e-01  -6.843 7.73e-12 ***
StreamingTVNo internet service        NA        NA      NA       NA
StreamingTVYes                   6.933e-01  1.207e-01   5.742 9.34e-09 ***
StreamingMoviesNo internet service    NA        NA      NA       NA
StreamingMoviesYes               7.323e-01  1.199e-01   6.109 1.01e-09 ***
ContractOne year                -5.889e-01  1.113e-01  -5.291 1.22e-07 ***
ContractTwo year                -1.486e+00  1.880e-01  -7.905 2.68e-15 ***
PaperlessBillingYes              3.047e-01  7.875e-02   3.869 0.000109 ***
PaymentMethodCredit card (automatic) -3.036e-02 1.204e-01 -0.252 0.800890
PaymentMethodElectronic check    3.831e-01  9.974e-02   3.841 0.000123 ***
PaymentMethodMailed check       -3.334e-02  1.208e-01  -0.276 0.782509
MonthlyCharges                  -4.888e-02  8.298e-03  -5.891 3.83e-09 ***
TotalCharges                     3.212e-04  7.445e-05   4.314 1.60e-05 ***
```

```
> print(accuracy)
       [,1]           [,2]         [,3]           [,4]
 [1,] "set"          "training"   "validation"   "test"
 [2,] "logistic1"    "0.8"        "0.81"         "0.78"
 [3,] "logistic2"    "0.79"       "0.81"         "0.81"
 [4,] "logistic3"    "0.79"       "0.78"         "0.81"
 [5,] "rpart4"       "0.8"        "0.81"         "0.78"
 [6,] "rpart5"       "0.79"       "0.76"         "0.8"
 [7,] "rpart6"       "0.79"       "0.79"         "0.76"
 [8,] "randomforest7" "0.96"      "0.82"         "0.77"
 [9,] "randomforest8" "0.96"      "0.8"          "0.82"
[10,] "randomforest9" "0.97"      "0.8"          "0.79"
[11,] "ave."         "0.85"       "0.8"          "0.79"
```

Model 3

➤ 挑選stepwise結果

```
> print(accuracy)
       [,1]               [,2]          [,3]           [,4]
 [1,]  "set"              "training"    "validation"   "test"
 [2,]  "logistic1"        "0.8"         "0.79"         "0.8"
 [3,]  "logistic2"        "0.8"         "0.8"          "0.78"
 [4,]  "logistic3"        "0.79"        "0.8"          "0.8"
 [5,]  "rpart4"           "0.79"        "0.81"         "0.8"
 [6,]  "rpart5"           "0.79"        "0.78"         "0.81"
 [7,]  "rpart6"           "0.79"        "0.78"         "0.78"
 [8,]  "randomforest7"    "0.96"        "0.8"          "0.79"
 [9,]  "randomforest8"    "0.96"        "0.81"         "0.81"
[10,]  "randomforest9"    "0.96"        "0.8"          "0.81"
[11,]  "ave."             "0.85"        "0.8"          "0.8"
```

Data processing

```
> str(d)
'data.frame':   7043 obs. of  20 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 1 1 2 2 2 1 2 1 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 1 2 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 2 1 1 2 ...
 $ tenure          : Factor w/ 2 levels "long","short": 2 2 1 2 2 2 2 2 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 1 1 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 2 2 1 2 1 1 1 1 3
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 1 2 1 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 2 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 1 1 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 2 1 2 ...
 $ Contract        : Factor w/ 2 levels "long","short": 2 2 1 2 2 2 1 2 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 1 1 ...
 $ PaymentMethod   : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 2 1 1 ...
 $ MonthlyCharges  : Factor w/ 4 levels "<36",">66","36~50",..: 2 2 2 2 4 4 2 3 1
 $ TotalCharges    : Factor w/ 3 levels "<800",">2500",..: 1 1 3 3 1 1 3 1 1 3 ..
 $ Churn           : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 1 ...
```

Model

➢ 反覆挑選變數測試後test的accuracy仍無法突破0.8

```
> print(accuracy)
       [,1]               [,2]        [,3]          [,4]
 [1,] "set"             "training" "validation" "test"
 [2,] "logistic1"       "0.79"     "0.79"       "0.79"
 [3,] "logistic2"       "0.79"     "0.79"       "0.79"
 [4,] "logistic3"       "0.79"     "0.79"       "0.79"
 [5,] "rpart4"          "0.79"     "0.82"       "0.79"
 [6,] "rpart5"          "0.79"     "0.79"       "0.82"
 [7,] "rpart6"          "0.79"     "0.79"       "0.79"
 [8,] "randomforest7"   "0.84"     "0.76"       "0.79"
 [9,] "randomforest8"   "0.83"     "0.81"       "0.76"
[10,] "randomforest9"   "0.83"     "0.79"       "0.81"
[11,] "ave."            "0.8"      "0.79"       "0.79"
```

4 Demo

Challenge

➢ 無法突破 null model的accuracy

➢ feature大多都是類別型變數 操作空間不大

➢ EDA觀察看出的趨勢並不等於model幫我們挑出的顯著變數

➢ 使用stepwise挑選變數後依然無顯著提升

5 結論

結論

各方面分析重點：
1.客戶基本資料(客戶本身)：
(1)gender(性別)不是影響客戶流失率的主因
(2)老年、單身、無家屬用戶為流失的重點對象
改進建議：
針對高流失率客群ex:老人、單身、無家屬用戶制定專屬方案

2.使用服務內容(服務品質)：
(1)使用Fiber optic(光纖通訊)的用戶流失率極高
改進建議：
因使用Fiber optic的客戶流失率極高，因此建議該公司的技術部門與業務部門可以共同合作，從使用者端取得反饋，並由技術部門進行服務優化，以提升服務品質，降低客戶流失率

結論

3.合約期間與價格(市場定價)：
(1)Contract為Month-to-month的流失率極高
(2)tenure(使用期數)為20月是重要的分界點，tenure小於20時流失率極高，高於20的流失率就逐漸下降並趨於穩定
(3)MonthlyCharge在70-100之間時，流失率極高
改進建議：
(1)重新思考合約方案，透過增加誘因ex:價格優惠、提高服務品質、優質行銷，吸引客戶可以長期使用服務(簽訂較長的合約1、2年)，減少月費用戶的流失
(2)調降月費為70-100的用戶方案、維持月費但提供額外服務來增加合約cp值，或進行更多市場調查，合理估計使用客群的價格接受區間再行定

結論

整體結論：
Tenure 、Contract及MonthlyCharge 是影響客戶流失率最大的因素

模型使用：
透過我們所建置的模型可以預測該客戶未來將流失的可能性，若被模型判定為流失的客群，可以提前針對性地實施策略(客製化服務)ex:降低MonthlyCharge、更改合約期數、增加額外服務等，使客戶能接受該方案並穩定使用後(提高tenure)，便能大幅降低客戶的流失率

6 參考資料

參考資料

- Telco Customer Churn
https://www.kaggle.com/blastchar/telco-customer-churn
- 我如何分析客戶流失預測？Kaggle比賽思路分享
https://reurl.cc/3N1MgM
- Shiny Data-Tables Demo
https://shiny.rstudio.com/gallery/datatables-demo.html

Thank you!