

고려대학교  
빅데이터 연구회

# KU-BIG

---

경제 경영 데이터 분석

최문규 박인성 최은혁 황예진 김호익 박소현



# 예측 모델링

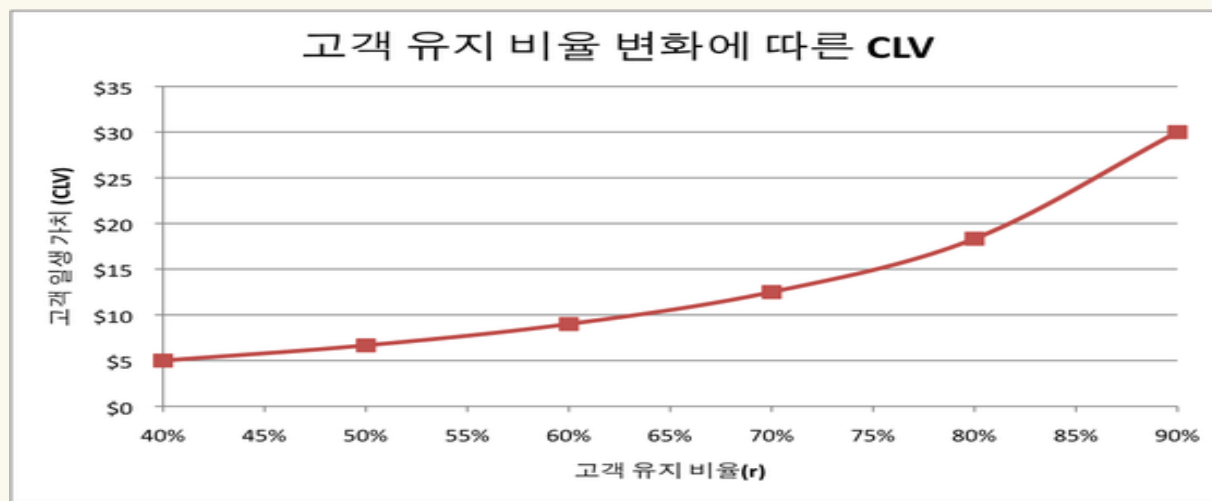
- I 주제 선정
- II 데이터 탐색 및 전처리
- III 모델링
- IV 모델 평가

# I . 주제 선정 : 고객 이탈 여부 (Churn) 예측

## 고객 이탈율 예측이 중요한 이유

- ▶ 고객 이탈율 (Churn Rate)이란?
  - 이탈하는 고객 수 대비 신규 고객 수
  - 고객 이탈율을 정확히 예측하게 되면 고객 생애가치, 마케팅 투자 대비 수익 등을 더욱 더 높일 수 있음
- ▶ 고객 생애가치 (Customer Lifetime Value)

$$CLV = \frac{(M - c)}{1 - r + i} - AC$$



① M : 고객 1인당 평균 매출 ② C : 고객 1인당 평균 비용 ③ R : 고객 유지 비율 ④ i : 이자율 또는 할인율 ⑤ AC : 고객 획득 비용

# I. 주제 선정 : 고객 이탈 여부 (Churn) 예측

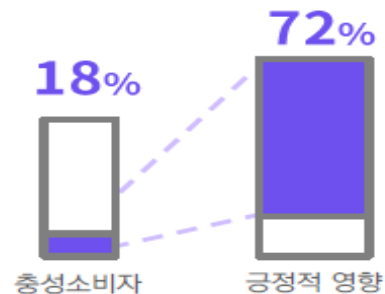
## 고객 이탈을 예측이 중요한 이유

### ▶ 리텐션 마케팅

- 신규 고객이 아닌 기존 고객에게 충성도를 높여서 구매율을 높이는 마케팅 기법
- 구글이나 페이스북 광고 도달률이 점점 낮아짐에 따라 광고비용은 더욱더 증가하게 되고, 그래서 신규 고객 확보가 매우 어려워짐
- **불만족한 고객 이탈로 부정적 서비스 사고 하나를 만회하기 위해서는 12가지 긍정적 사례 필요함**

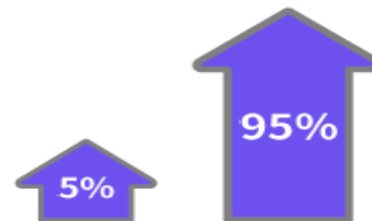
충성 소비자가 타인에게  
긍정적 커뮤니케이션을 하는 비율

\*Harris Interactive (세계 최대 여론조사기관)



고객 충성도를 5%만 올려도  
이익은 95%의 비율로 증가

\*Deloitte (세계 4대 회계법인)



신규 고객 유치 시 기존 고객 대비  
7배 높은 비용 필요

\*Bain & Company




# I. 사례 분석 : 카카오톡 활용 리텐션 마케팅

## 1. 카카오톡 상담서비스 : 접근성 향상

TALK 비즈니스

### 상담톡이란?

플친 관리자 웹/앱이 아니더라도 플러스친구 1:1 채팅 기능을 할 수 있게 하는 API입니다.



#### 상담톡 주요 특징

- 기존 콜센터와 함께 채팅 상담 채널로 운영하면 매우 효과적입니다.
- 콜센터 대기시간이 길 때 상담톡 연결을 유도할 수 있습니다.
- '특상담하기' 버튼을 고객센터나 상품정보 페이지 등 고객이 편히 사용할 곳에 붙일 수 있습니다.
- 챗봇과 연결하여 고객 응대 효율을 높일 수 있습니다.

## 2. 카카오톡 플러스 친구 : 도달률 증가



친구들에게 필요한  
정보와 혜택을 가장 확실하게  
전달할 수 있는 방법

실시간 알림이나 공지, 이벤트, 할인 쿠폰을  
카카오톡 메시지로 발송해보세요.

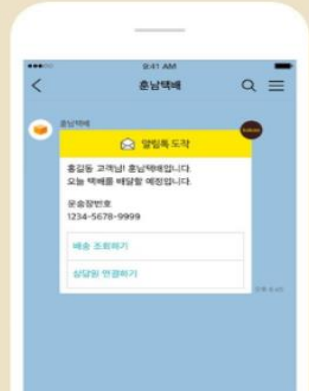
화면을 꽉 채우는 와이드형 메시지로 주목도를 높이고  
정교한 타겟팅으로 친구들의 적극적인 반응을  
이끌어낼 수 있습니다.

## 3. 카카오톡 알림톡 : 긍정적 거래 경험 전달

TALK 비즈니스

### 알림톡이란?

전화번호로 주문, 예약, 결제, 배송 등  
정보성 메시지를 플친으로  
전송할 수 있는 API입니다.



#### 알림톡 주요 특징

- SMS보다 저렴한 가격으로 LMS 만큼 전송 가능합니다.
- 고객에게 귀사의 플친 운영을 알리는 훌륭한 수단이 됩니다.

## II. 변수 설명

### Telecom Customer Churn

▶ 7043개 관측치, 21개 변수

| 변수명           | Type   | 설명                                 |
|---------------|--------|------------------------------------|
| CustomerID    | Factor | 고객 ID: 0002-ORFBO와 같은 형태, 7043개 존재 |
| Gender        | Factor | 성별: Female, Male로 구분               |
| SeniorCitizen | Int    | 고령자 여부: 1이면 고령자, 0이면 아님            |
| Partner       | Factor | 배우자 유무: Yes, No로 구분                |
| Dependents    | Factor | 부양가족 유무: Yes, No로 구분               |
| Tenure        | Int    | 서비스 사용 기간: 단위는 Month               |
| Phoneservice  | Factor | 전화 서비스 사용 유무: Yes, No로 구분          |

## II. 변수 설명

### Telecom Customer Churn

▶ 7043개 관측치, 21개 변수

| 변수명              | Type   | 설명                                                  |
|------------------|--------|-----------------------------------------------------|
| MultipleLines    | Factor | 다중 회선 여부: Yes, No로 구분                               |
| InternetService  | Factor | 인터넷 서비스 사용 유무: dsl, Fiberoptic, No로 구분              |
| OnlineSecurity   | Factor | 인터넷 보안 서비스 사용 유무: No, No internet service, Yes로 구분  |
| OnlineBackup     | Factor | 백업 서비스 사용 유무: No, No internet service, Yes로 구분      |
| DeviceProtection | Factor | 기기 보안 서비스 사용 유무: No, No internet service, Yes로 구분   |
| TechSupport      | Factor | A/S 서비스 사용 유무: No, No internet service, Yes로 구분     |
| StreamingTV      | Factor | 스트리밍 tv 서비스 사용 유무: No, No internet service, Yes로 구분 |

## II. 변수 설명

### Telecom Customer Churn

▶ 7043개 관측치, 21개 변수

| 변수명              | Type    | 설명                                                                    |
|------------------|---------|-----------------------------------------------------------------------|
| StreamingMovies  | Factor  | 영화 스트리밍 서비스 사용 유무: No, No internet service, Yes로 구분                   |
| Contract         | Factor  | 계약 기간: Month-to-month, One year, Two year로 구분                         |
| PaperlessBilling | Factor  | 요금 청구서 방법: Yes, No로 구분                                                |
| PaymentMethod    | Factor  | 지불 방법: Bank transfer, Credit Card, Electronic check, Mailed check로 구분 |
| MonthlyCharges   | Numeric | 월별 지불 금액: 단위는 \$                                                      |
| TotalCharges     | Numeric | 총 지불 금액: 단위는 \$                                                       |
| Churn            | Factor  | 고객 계약 해지 여부: Yes, No로 구분                                              |



## II. 데이터 탐색

### Summary 함수 활용

```
> summary(churn)
 customerID      gender SeniorCitizen  Partner  Dependents    tenure  PhoneService  MultipleLines
0002-ORFBO:      1  Female:3488   Min.   :0.0000   No :3641   No :4933   Min.   : 0.00   No : 682   No :3390
0003-MKNFE:      1   Male :3555   1st Qu.:0.0000   Yes:3402   Yes:2110   1st Qu.: 9.00   Yes:6361   No phone service: 682
0004-TLHLJ:      1                                     Median :0.0000                                     Median :29.00   Yes :2971
0011-IGKFF:      1                                     Mean    :0.1621                                     Mean    :32.37
0013-EXCHZ:      1                                     3rd Qu.:0.0000                                     3rd Qu.:55.00
0013-MHZWF:      1                                     Max.     :1.0000                                     Max.     :72.00
(Other)       :7037
 InternetService  onlineSecurity  onlineBackup  DeviceProtection  TechSupport
DSL              :2421   No              :3498   No              :3088   No              :3095   No              :3473
Fiber optic:3096 No internet service:1526 No internet service:1526 No internet service:1526 No internet service:1526
No              :1526   Yes              :2019   Yes              :2429   Yes              :2422   Yes              :2044

 StreamingTV      StreamingMovies  Contract      PaperlessBilling  PaymentMethod
No                :2810   No              :2785   Month-to-month:3875   No :2872   Bank transfer (automatic):1544
No internet service:1526 No internet service:1526 One year       :1473   Yes:4171   Credit card (automatic) :1522
Yes              :2707   Yes              :2732   Two year       :1695                                     Electronic check :2365
                                                         Mailed check    :1612

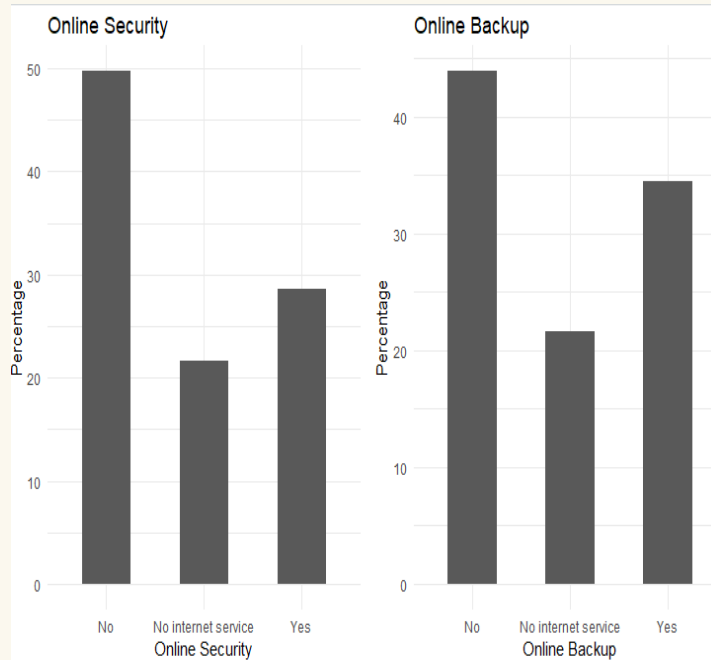
MonthlyCharges  TotalCharges  Churn
Min.   : 18.25  Min.   : 18.8  No :5174
1st Qu.: 35.50  1st Qu.: 401.4 Yes:1869
Median : 70.35  Median :1397.5
Mean   : 64.76  Mean   :2283.3
3rd Qu.: 89.85  3rd Qu.:3794.7
Max.   :118.75  Max.   :8684.8
NA's    :11
```

- TotalCharges에 있는 11개 NA값 -> 전체 데이터 크기 대비 0.1562% 비중이라 Complete.cases 함수로 결측치 제거
- 0과 1만 가능한 변수에서 다른 값 나온 이상치 존재 여부 확인
- Seniorcitizen Factor형 변환 필요 / No internet service도 No로 통합 여부 판단 필요

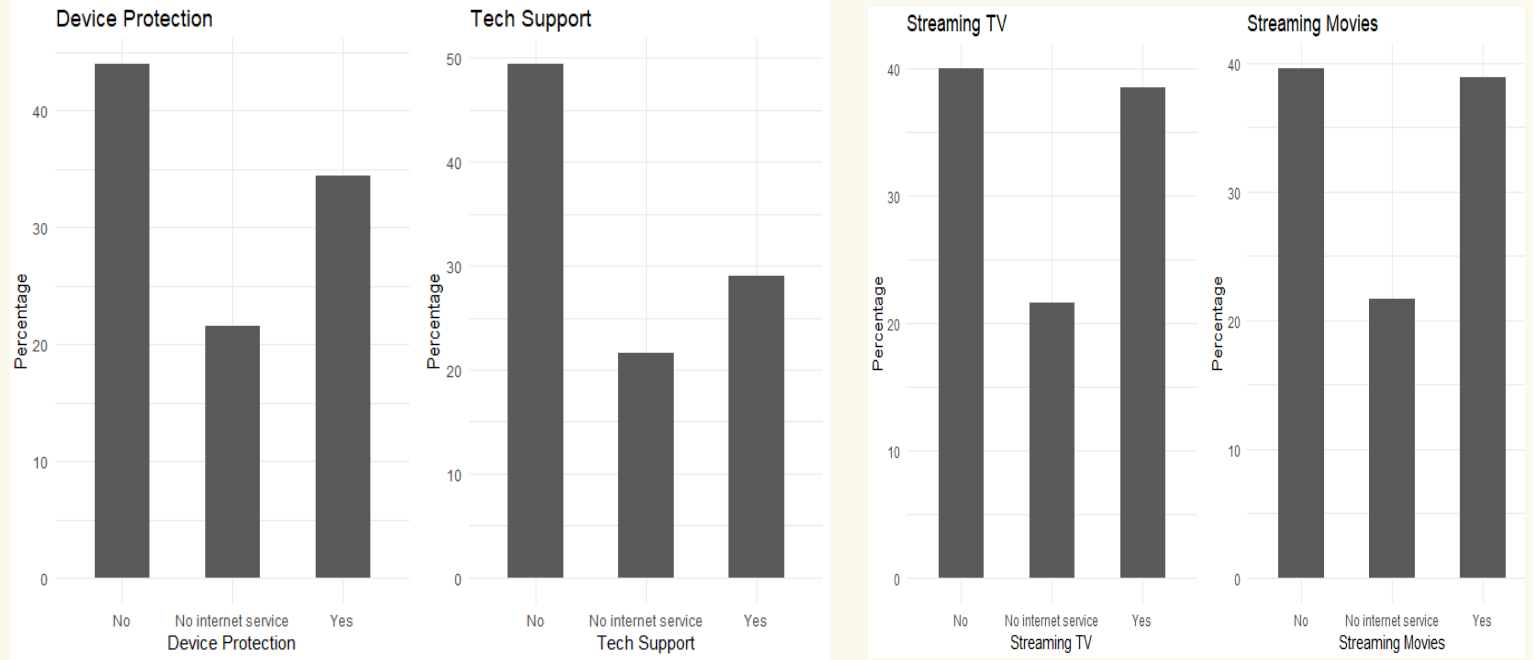
## II. 데이터 전처리

### 범주형 변수 전처리

#### ▶ Online Security, Backup



#### ▶ Device Protection, Tech Support ▶ Streaming TV, Movies

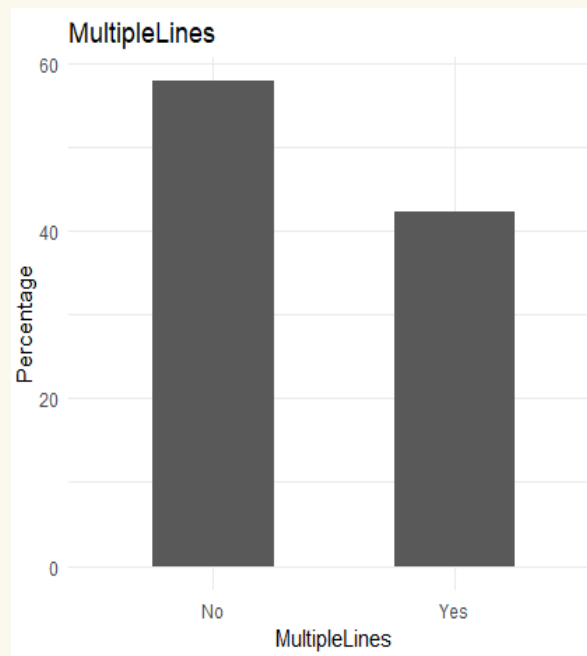
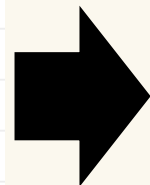
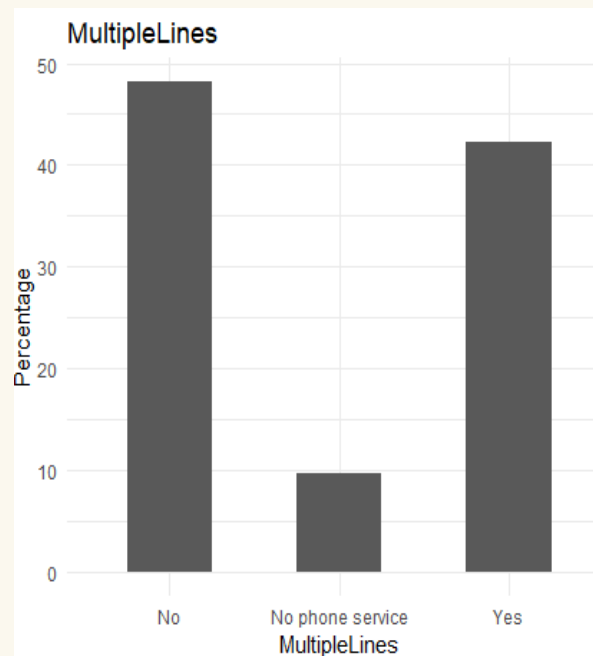


- 'No Internet service' 값의 비중이 낮은 편이 아니기 때문에 굳이 No로 통합할 필요 없어보임
- 모델링 결과에 따라 통합 여부를 다시 결정하는 것으로 판단

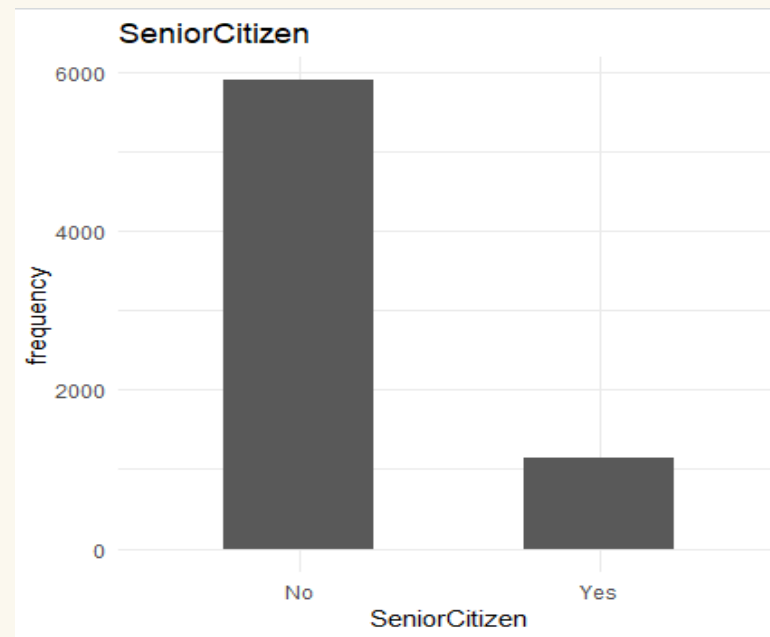
## II. 데이터 전처리

### 범주형 변수 전처리

#### ▶ MultipleLines



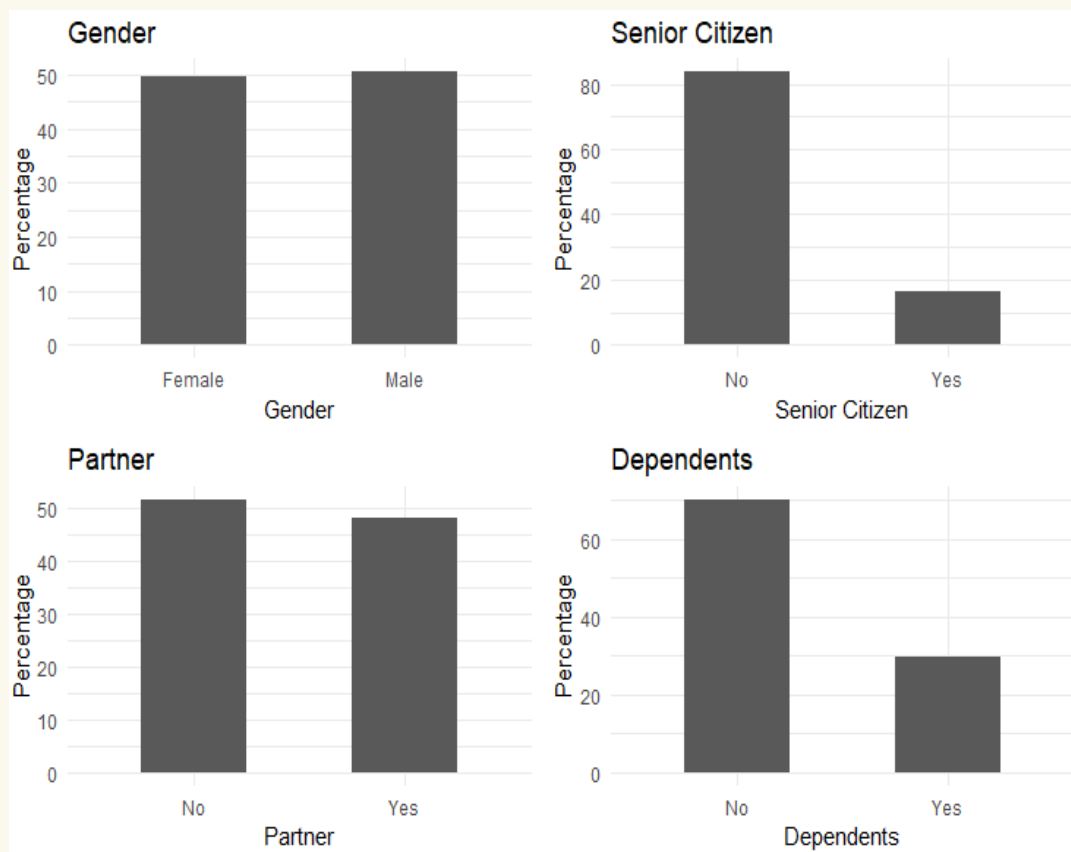
#### ▶ SeniorCitizen Factor형 변환



- 'No Internet service' 값의 비중이 10%도 되지 않기 때문에 No로 통합해도 무관하다고 판단
- SeniorCitizen의 0과 1 값을 No와 Yes의 Factor형으로 변환

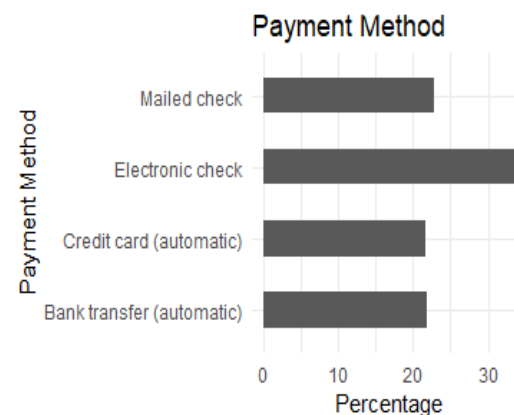
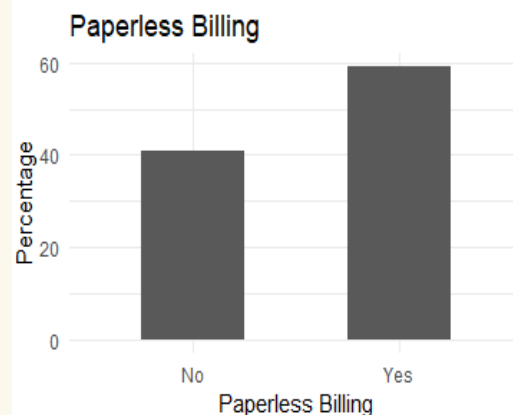
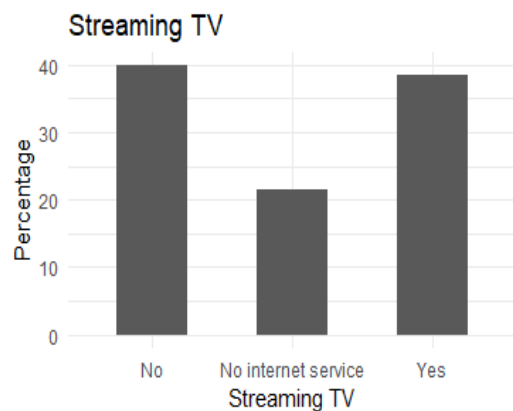
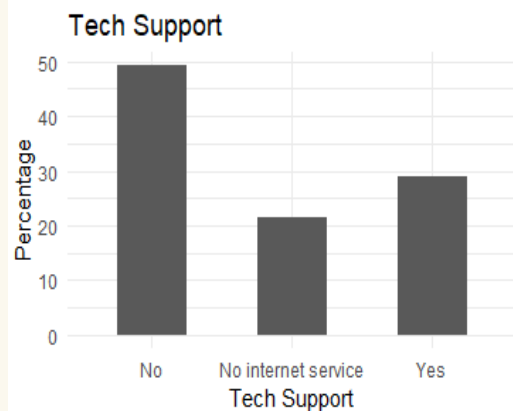
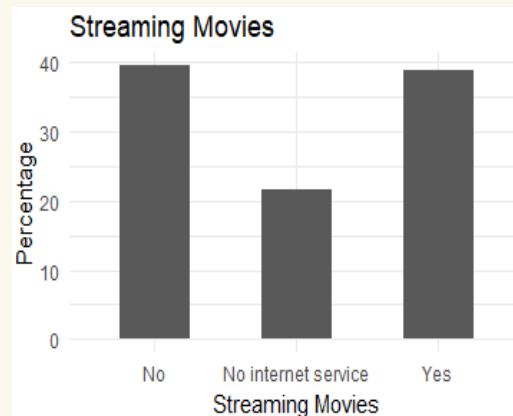
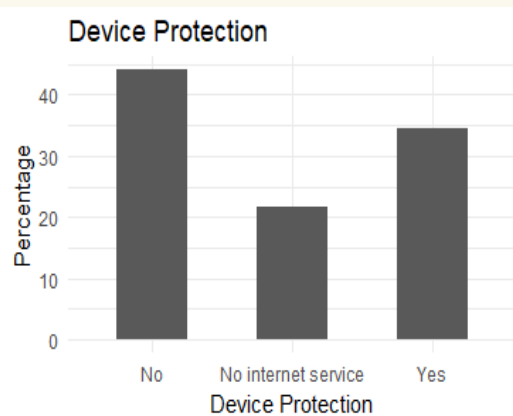
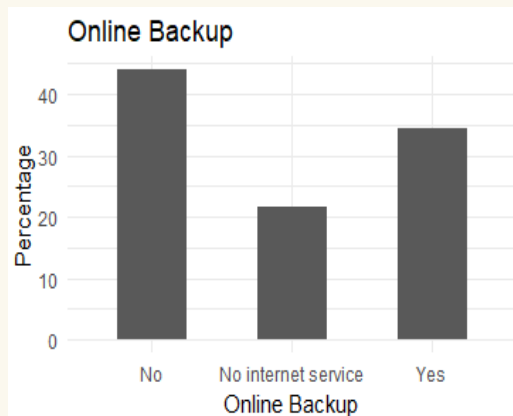
## II. 데이터 전처리

### 범주형 변수 barplot



## II. 데이터 전처리

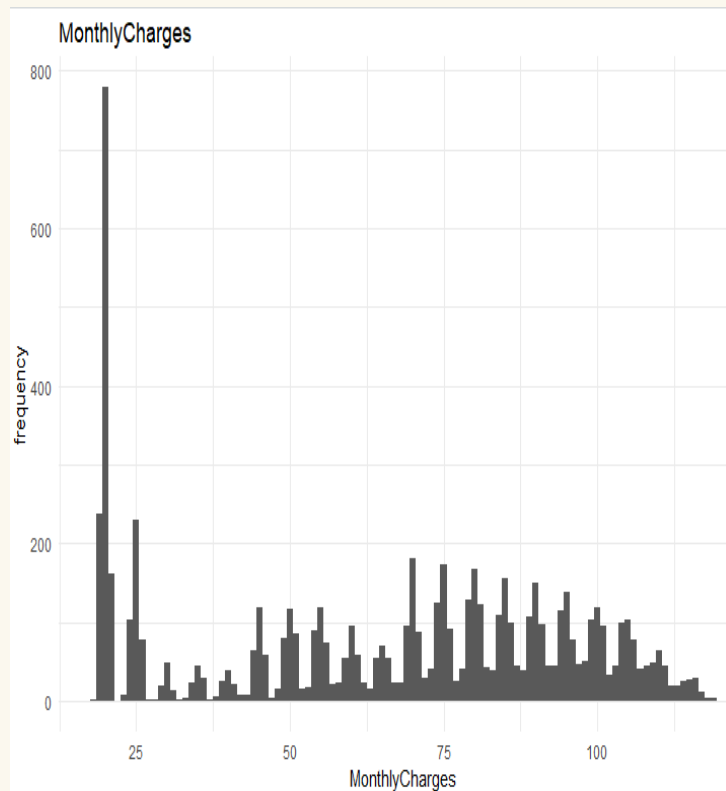
### 범주형 변수 barplot



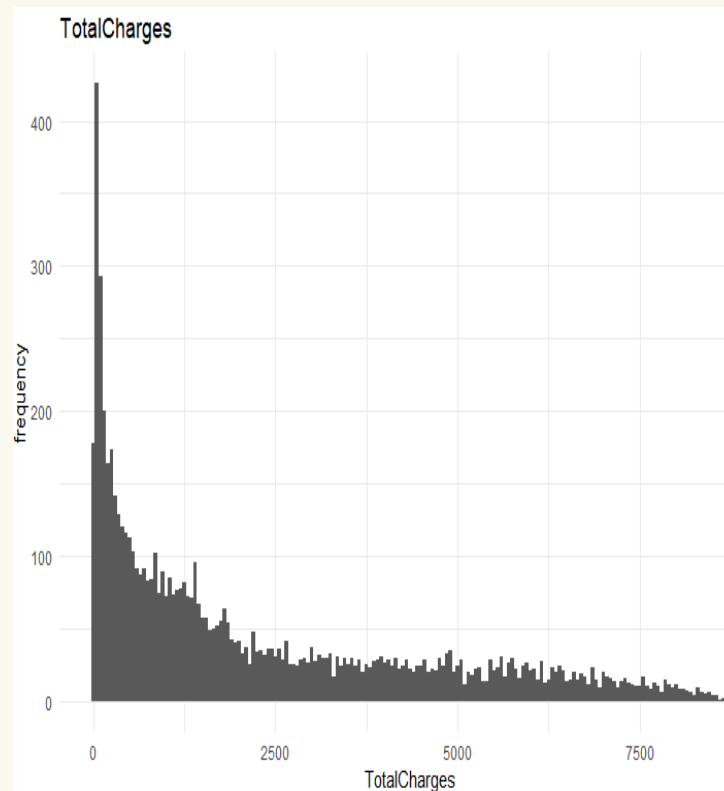
## II. 데이터 전처리

### 연속형 변수 전처리

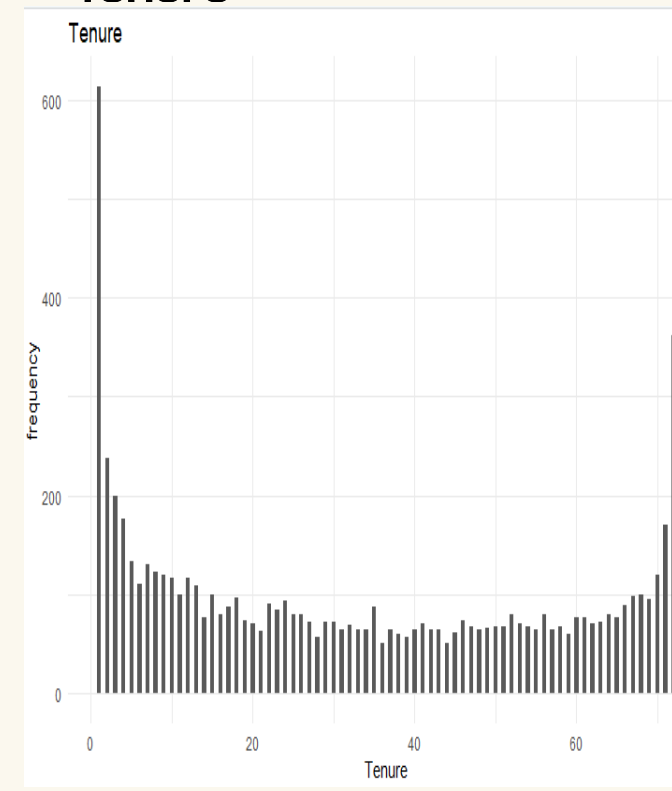
#### ▶ Monthly Charges



#### ▶ Total Charges



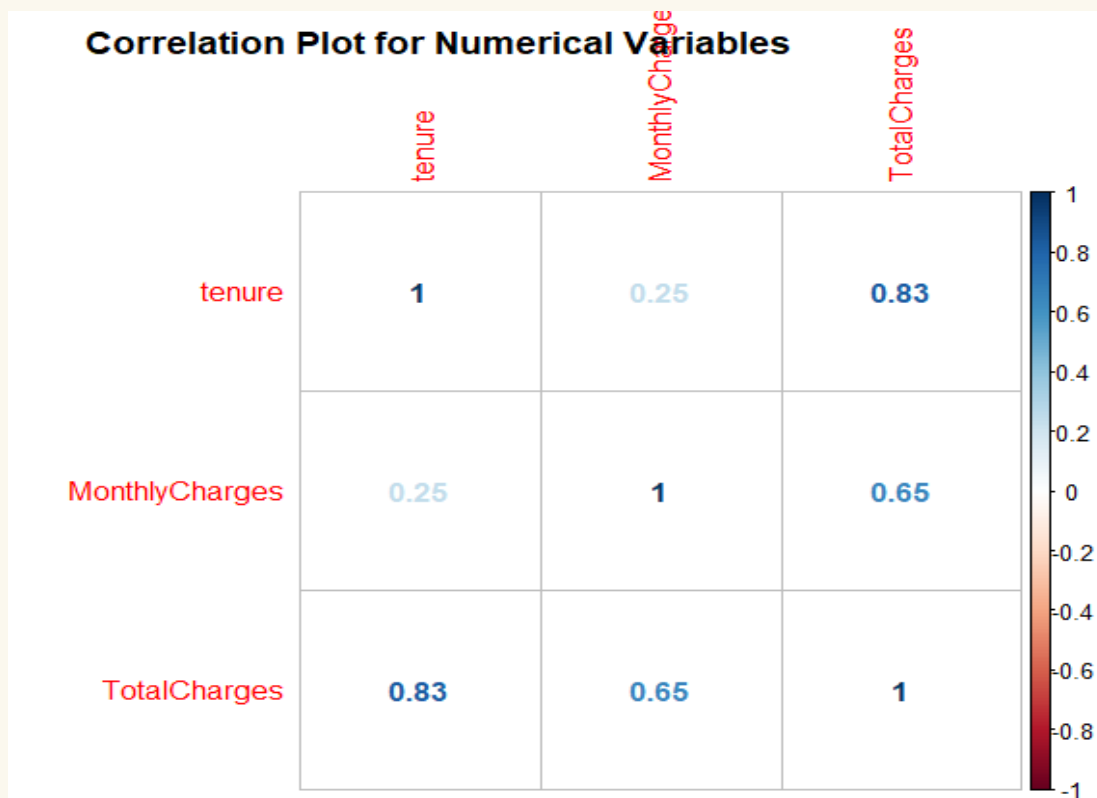
#### ▶ Tenure



## II. 데이터 전처리

### 연속형 변수 전처리

#### ▶ 양적변수 Correlation Plot

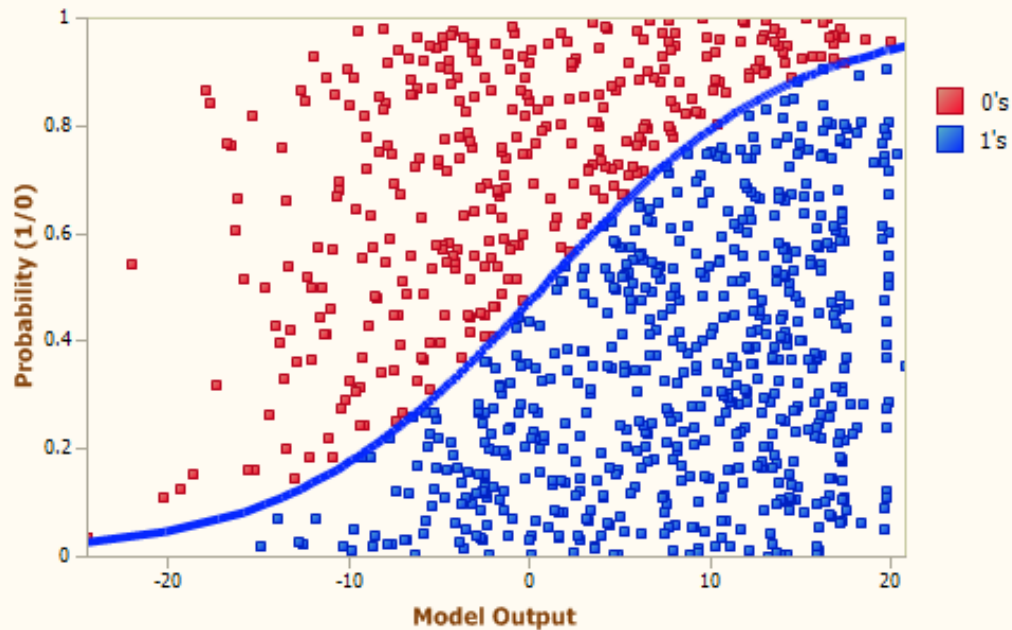


- tenure와 MonthlyCharges 간에 약한 상관관계 존재
- TotalCharges의 경우 tenure, MonthlyCharges 모두 강한 상관관계 존재

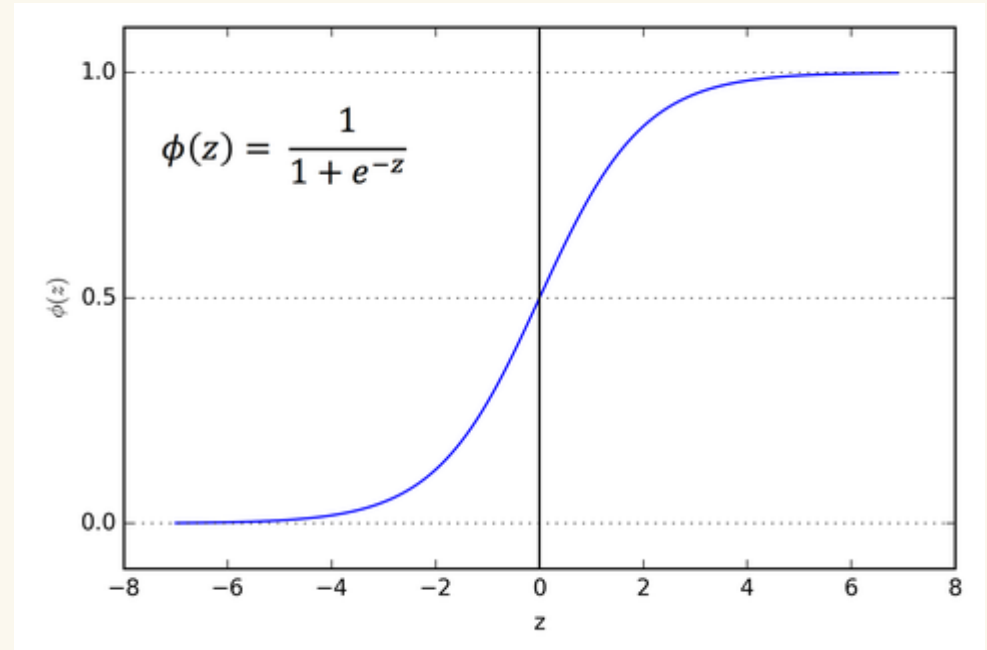
# III. 모델링 – 로지스틱 회귀

## 모델링 정의

### ▶ 로지스틱 회귀



### ▶ 시그모이드 함수



- 종속변수가 연속적이지 않고 이진형인 경우 활용하는 로지스틱 회귀
- 시그모이드 함수를 활용하여 이 값이 0.5를 경계값으로 0 또는 1로 예측



# III. 모델링 – 로지스틱 회귀

## 모델링 결과 : Stepwise 활용

### ▶ 변수 선별 및 예측률

Step: AIC=4102.11  
Churn ~ Dependents + tenure + PhoneService + MultipleLines +  
InternetService + OnlineBackup + DeviceProtection + StreamingTV +  
StreamingMovies + Contract + PaperlessBilling + PaymentMethod +  
MonthlyCharges + TotalCharges

|                    | Df | Deviance | AIC    |
|--------------------|----|----------|--------|
| <none>             |    | 4064.1   | 4102.1 |
| - OnlineBackup     | 1  | 4066.2   | 4102.2 |
| - Dependents       | 1  | 4066.6   | 4102.6 |
| - DeviceProtection | 1  | 4070.4   | 4106.4 |
| - PhoneService     | 1  | 4070.9   | 4106.9 |
| - PaperlessBilling | 1  | 4077.1   | 4113.1 |
| - PaymentMethod    | 3  | 4087.2   | 4119.2 |
| - TotalCharges     | 1  | 4088.7   | 4124.7 |
| - StreamingMovies  | 1  | 4091.3   | 4127.3 |
| - MonthlyCharges   | 1  | 4094.6   | 4130.6 |
| - MultipleLines    | 1  | 4095.2   | 4131.2 |
| - StreamingTV      | 1  | 4095.8   | 4131.8 |
| - InternetService  | 1  | 4122.6   | 4158.6 |
| - Contract         | 2  | 4129.6   | 4163.6 |
| - tenure           | 1  | 4156.6   | 4192.6 |

```
> print(paste('Logistic Regression Accuracy',1-misClassificError))
[1] "Logistic Regression Accuracy 0.817362428842505"
> print("Confusion Matrix for Logistic Regression"); table(testing$churn, fitted.results > 0.5)
[1] "Confusion Matrix for Logistic Regression"
```

```
FALSE TRUE
0 1408 140
1 245 315
```

```
> print(paste('Logistic Regression Accuracy',1-misClassificError2))
[1] "Logistic Regression Accuracy 0.811195445920304"
> print("Confusion Matrix for Logistic Regression"); table(testing$churn, fitted.results2 > 0.5)
[1] "Confusion Matrix for Logistic Regression"
```

```
FALSE TRUE
0 1397 151
1 247 313
```

- Stepwise로 변수 6개 제거 (Gender, Partner, Dependents, Online Security, Online Backup, Tech Support)
- 예측률 측면에서는 모든 변수를 고려한 것과 간단한 모델의 차이가 적음

## III. 모델링 - 로지스틱 회귀

### 변수 중요도 확인

```
> imp_1[order(imp_1$overall,decreasing = T),]
  overall names
5  8.09934511 tenure
17 6.78861033 ContractTwo year
16 6.49692795 ContractOne year
23 4.04902202 TotalCharges
18 3.66117423 PaperlessBillingYes
7  2.73513697 MultipleLinesYes
20 2.71374273 PaymentMethodElectronic check
9  2.43427033 InternetServiceNo
14 2.27064598 StreamingTVYes
8  2.24720480 InternetServiceFiber optic
15 2.03688753 StreamingMoviesYes
2  1.67675566 SeniorCitizenYes
22 1.55814968 MonthlyCharges
12 0.85129715 DeviceProtectionYes
4  0.79480695 DependentsYes
6  0.67843064 PhoneServiceYes
1  0.61905513 genderMale
13 0.29700855 TechSupportYes
11 0.23126399 onlineBackupYes
3  0.16148547 PartnerYes
19 0.14665215 PaymentMethodCredit card (automatic)
10 0.13338445 onlineSecurityYes
21 0.01295932 PaymentMethodMailed check
```

- varImp() 함수로 판단

- 사용 및 계약기간과 관련된 변수가  
상위 중요변수로 나타남

# III. 모델링 - 로지스틱 회귀

## Odds Ratio

```
> exp(cbind(OR=coef(LogModel), confint(LogModel)))
waiting for profiling to be done...
```

|                                      | OR         | 2.5 %      | 97.5 %     |
|--------------------------------------|------------|------------|------------|
| (Intercept)                          | 7.04061640 | 1.06520828 | 46.7018624 |
| genderMale                           | 0.97703390 | 0.83913574 | 1.1376101  |
| SeniorCitizenYes                     | 1.13450462 | 0.93020156 | 1.3832125  |
| PartnerYes                           | 0.95062578 | 0.79156800 | 1.1418658  |
| DependentsYes                        | 0.89875309 | 0.72911046 | 1.1066739  |
| tenure                               | 0.93511872 | 0.92088529 | 0.9490127  |
| PhoneServiceYes                      | 2.01010733 | 0.44599055 | 9.0789751  |
| MultipleLinesYes                     | 1.79978576 | 1.19055470 | 2.7240547  |
| InternetServiceFiber optic           | 9.98502643 | 1.56826021 | 63.9644361 |
| InternetServiceNo                    | 0.07494898 | 0.01149157 | 0.4870943  |
| OnlineSecurityNo internet service    | NA         | NA         | NA         |
| OnlineSecurityYes                    | 0.95121491 | 0.62789724 | 1.4407291  |
| OnlineBackupNo internet service      | NA         | NA         | NA         |
| OnlineBackupYes                      | 1.13708162 | 0.75644234 | 1.7099424  |
| DeviceProtectionNo internet service  | NA         | NA         | NA         |
| DeviceProtectionYes                  | 1.29126906 | 0.85584882 | 1.9492843  |
| TechSupportNo internet service       | NA         | NA         | NA         |
| TechSupportYes                       | 0.99184372 | 0.65440988 | 1.5026081  |
| StreamingTVNo internet service       | NA         | NA         | NA         |
| StreamingTVYes                       | 2.33167084 | 1.09186438 | 4.9916138  |
| StreamingMoviesNo internet service   | NA         | NA         | NA         |
| StreamingMoviesYes                   | 2.17670801 | 1.02084572 | 4.6514145  |
| ContractOne year                     | 0.48581785 | 0.37585394 | 0.6246959  |
| ContractTwo year                     | 0.24127608 | 0.15624972 | 0.3628856  |
| PaperlessBillingYes                  | 1.37452805 | 1.15354767 | 1.6388870  |
| PaymentMethodCredit card (automatic) | 0.98014914 | 0.75111672 | 1.2787734  |
| PaymentMethodElectronic check        | 1.38458439 | 1.10957972 | 1.7304112  |
| PaymentMethodMailed check            | 0.86075499 | 0.65478791 | 1.1320471  |
| MonthlyCharges                       | 0.93584195 | 0.86929611 | 1.0073180  |
| TotalCharges                         | 1.00041674 | 1.00024901 | 1.0005889  |

- 로지스틱 회귀분석에서는 회귀계수보다는  
오즈비로 영향력 크기를 해석

- ①  $\text{Exp}(B) > 1$  : (+) 효과
- ②  $\text{Exp}(B) = 1$  : 효과 x
- ③  $\text{Exp}(B) < 1$  : (-) 효과

# III. 모델링 - 로지스틱 회귀

## 문제점 : 계수 NA값

```
> print(summary(fit.logit))

Call:
glm(formula = Churn ~ Dependents + tenure + PhoneService + MultipleLines +
  InternetService + OnlineBackup + DeviceProtection + StreamingTV +
  StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
  MonthlyCharges + TotalCharges, family = binomial, data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8349  -0.6716  -0.2762   0.7463   3.4548

Coefficients: (4 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.108e+00  4.202e-01   5.016 5.27e-07 ***
DependentsYes  -1.500e-01  9.504e-02  -1.578 0.114540
tenure         -6.712e-02  7.649e-03  -8.774 < 2e-16 ***
PhoneServiceYes 8.116e-01  3.116e-01   2.605 0.009193 **
MultipleLinesYes 6.239e-01  1.126e-01   5.543 2.98e-08 ***
InternetServiceFiber optic 2.470e+00  3.284e-01   7.523 5.34e-14 ***
InternetServiceNo -2.746e+00  4.057e-01  -6.769 1.30e-11 ***
OnlineBackupNo internet service NA      NA      NA      NA
OnlineBackupYes 1.621e-01  1.132e-01   1.432 0.152163
DeviceProtectionNo internet service NA      NA      NA      NA
DeviceProtectionYes 2.889e-01  1.156e-01   2.500 0.012429 *
StreamingTVNo internet service NA      NA      NA      NA
StreamingTVYes 9.107e-01  1.634e-01   5.572 2.52e-08 ***
StreamingMoviesNo internet service NA      NA      NA      NA
StreamingMoviesYes 8.430e-01  1.631e-01   5.167 2.38e-07 ***
ContractOne year -7.309e-01  1.291e-01  -5.662 1.50e-08 ***
ContractTwo year -1.436e+00  2.131e-01  -6.740 1.58e-11 ***
PaperlessBillingYes 3.218e-01  8.944e-02   3.597 0.000321 ***
PaymentMethodCredit card (automatic) -1.718e-02  1.355e-01  -0.127 0.899074
PaymentMethodElectronic check 3.317e-01  1.131e-01   2.931 0.003374 **
PaymentMethodMailed check -1.507e-01  1.392e-01  -1.082 0.279080
MonthlyCharges -7.254e-02  1.325e-02  -5.476 4.35e-08 ***
TotalCharges 4.160e-04  8.662e-05   4.803 1.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- But... 모델링 결과 계수값이 NA값이 나온 결과 발생
- dummy variable 처리 여부와 관련해서 해결해야 할 문제
- NA값 나온 변수들 분포 확인 필요

# III. 모델링 - 로지스틱 회귀

문제점 : 계수 NA값

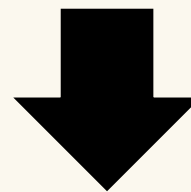
```
> print(summary(LogModel))

Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training,
     x = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8464  -0.6810  -0.2839   0.7454   3.3208

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.308e-02  2.710e-01  -0.196  0.844683
genderMale     1.925e-02  7.747e-02   0.249  0.803726
SeniorCitizenYes 2.795e-01  1.001e-01   2.793  0.005222 **
PartnerYes     2.845e-03  9.387e-02   0.030  0.975821
DependentsYes  -1.725e-01  1.087e-01  -1.587  0.112484
tenure         -5.913e-02  7.425e-03  -7.964  1.67e-15 ***
PhoneServiceYes -7.817e-01  1.760e-01  -4.442  8.90e-06 ***
MultipleLinesYes 2.218e-01  9.851e-02   2.251  0.024374 *
InternetServiceFiber optic 6.819e-01  1.665e-01   4.095  4.21e-05 ***
InternetServiceNo -4.300e-01  2.271e-01  -1.894  0.058276 .
OnlineSecurityNo internet service NA      NA      NA      NA
OnlineSecurityYes -4.422e-01  1.033e-01  -4.280  1.87e-05 ***
ContractOne year -8.053e-01  1.317e-01  -6.113  9.75e-10 ***
ContractTwo year -1.396e+00  2.048e-01  -6.818  9.25e-12 ***
PaperlessBillingYes 3.074e-01  8.865e-02   3.468  0.000525 ***
PaymentMethodCredit card (automatic) -2.380e-01  1.360e-01  -1.750  0.080095 .
PaymentMethodElectronic check 1.815e-01  1.118e-01   1.623  0.104496
PaymentMethodMailed check -1.796e-01  1.365e-01  -1.316  0.188222
MonthlyCharges  7.735e-03  5.086e-03   1.521  0.128343
TotalCharges     2.710e-04  8.436e-05   3.213  0.001314 ***
```

1. 비슷한 분포 가진 변수 중 1개  
(Online Security)만 남기고  
모델링 진행



**실패!!**



# III. 모델링 - 로지스틱 회귀

문제점 : 계수 NA값

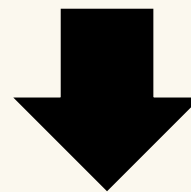
```
> print(summary(LogModel))

Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training,
     x = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9232  -0.6823  -0.2882   0.7373   3.3311

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.186e-01  2.981e-01   1.740  0.081932 .
genderMale      2.019e-02  7.745e-02   0.261  0.794342
SeniorCitizenYes 2.960e-01  9.996e-02   2.961  0.003063 **
PartnerYes      3.111e-03  9.384e-02   0.033  0.973556
DependentsYes   -1.817e-01  1.086e-01  -1.673  0.094257 .
tenure          -6.007e-02  7.431e-03  -8.083  6.31e-16 ***
PhoneServiceYes -3.537e-01  1.975e-01  -1.792  0.073212 .
MultipleLinesYes 3.281e-01  1.003e-01   3.272  0.001066 **
InternetServiceFiber optic 1.284e+00  1.905e-01   6.741  1.57e-11 ***
InternetServiceNo -9.634e-01  2.605e-01  -3.698  0.000217 ***
StreamingTVNo internet service NA      NA      NA      NA
StreamingTVYes   5.229e-01  1.260e-01   4.148  3.35e-05 ***
ContractOne year -8.085e-01  1.316e-01  -6.143  8.12e-10 ***
ContractTwo year -1.400e+00  2.049e-01  -6.835  8.20e-12 ***
PaperlessBillingYes 3.143e-01  8.857e-02   3.548  0.000387 ***
PaymentMethodCredit card (automatic) -2.323e-01  1.360e-01  -1.709  0.087494 .
PaymentMethodElectronic check  1.815e-01  1.118e-01   1.623  0.104548
PaymentMethodMailed check    -1.850e-01  1.363e-01  -1.357  0.174901
Monthlycharges    -1.518e-02  6.680e-03  -2.273  0.023043 *
Totalcharges      2.844e-04  8.448e-05   3.366  0.000763 ***
```

2. 비슷한 분포를 가진 변수들 중 1개  
(Streaming TV)만 남기고  
모델링 진행



**실패!!**

### III. 모델링 – 로지스틱 회귀

문제점 : 계수 NA값

```
> nrow(churn[churn$InternetService != "No", ])  
[1] 5512  
> nrow(churn[churn$OnlineSecurity != "No internet service", ])  
[1] 5512
```

- 범주형 변수들 간에 비슷한 분포를 가진 것이 문제가 아닌 것인가??
- Internet Service의 수준 Yes = Online Security의 수준 Yes + 수준 No
- 범주형 변수의 수준 = 다른 범주형 변수의 수준의 **선형 결합**으로 표현됨  
=> **다중공선성 의심**
- InternetService 변수 제외하고 OnlineSecurity 변수는 포함해서 진행해보자!

# III. 모델링 - 로지스틱 회귀

## 문제점 : 계수 NA값

```
> print(summary(LogModel))
```

Call:

```
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training,
     x = TRUE)
```

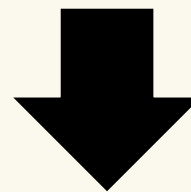
Deviance Residuals:

| Min     | 1Q      | Median  | 3Q     | Max    |
|---------|---------|---------|--------|--------|
| -1.8999 | -0.6869 | -0.2810 | 0.7403 | 3.3505 |

Coefficients:

|                                      | Estimate   | Std. Error | z value | Pr(> z ) |     |
|--------------------------------------|------------|------------|---------|----------|-----|
| (Intercept)                          | -0.5765036 | 0.2399144  | -2.403  | 0.016263 | *   |
| genderMale                           | 0.0223812  | 0.0772532  | 0.290   | 0.772036 |     |
| SeniorCitizenYes                     | 0.3105530  | 0.0996249  | 3.117   | 0.001826 | **  |
| PartnerYes                           | 0.0048145  | 0.0936383  | 0.051   | 0.958994 |     |
| DependentsYes                        | -0.1893453 | 0.1083849  | -1.747  | 0.080642 | .   |
| tenure                               | -0.0583639 | 0.0074219  | -7.864  | 3.73e-15 | *** |
| PhoneServiceYes                      | -0.8670819 | 0.1739357  | -4.985  | 6.19e-07 | *** |
| MultipleLinesYes                     | 0.1738785  | 0.0972758  | 1.787   | 0.073860 | .   |
| OnlineSecurityNo internet service    | -0.0957134 | 0.2118394  | -0.452  | 0.651398 |     |
| OnlineSecurityYes                    | -0.5523101 | 0.0997089  | -5.539  | 3.04e-08 | *** |
| ContractOne year                     | -0.8914457 | 0.1300496  | -6.855  | 7.15e-12 | *** |
| ContractTwo year                     | -1.5114204 | 0.2029523  | -7.447  | 9.54e-14 | *** |
| PaperlessBillingYes                  | 0.3100709  | 0.0883080  | 3.511   | 0.000446 | *** |
| PaymentMethodCredit card (automatic) | -0.2451562 | 0.1357871  | -1.805  | 0.071005 | .   |
| PaymentMethodElectronic check        | 0.1900016  | 0.1116198  | 1.702   | 0.088714 | .   |
| PaymentMethodMailed check            | -0.1927317 | 0.1358612  | -1.419  | 0.156018 |     |
| MonthlyCharges                       | 0.0227325  | 0.0035614  | 6.383   | 1.74e-10 | *** |
| TotalCharges                         | 0.0002402  | 0.0000842  | 2.852   | 0.004345 | **  |

3. Internet Service 변수 제거 후  
NA가 나왔던 Online Security를  
포함하고 모델링 진행



**성공!!**



# III. 모델링 - 로지스틱 회귀

문제점 : 계수 NA값

```
> print(summary(LogModel))
```

Call:

```
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training,
     x = TRUE)
```

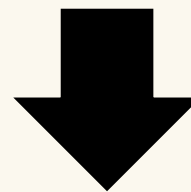
Deviance Residuals:

| Min     | 1Q      | Median  | 3Q     | Max    |
|---------|---------|---------|--------|--------|
| -1.9037 | -0.6867 | -0.2807 | 0.7382 | 3.3509 |

Coefficients: (1 not defined because of singularities)

|                                      | Estimate   | Std. Error | z value | Pr(> z ) |     |
|--------------------------------------|------------|------------|---------|----------|-----|
| (Intercept)                          | -5.685e-01 | 2.429e-01  | -2.341  | 0.019242 | *   |
| genderMale                           | 2.248e-02  | 7.725e-02  | 0.291   | 0.771033 |     |
| SeniorCitizenYes                     | 3.116e-01  | 9.975e-02  | 3.124   | 0.001785 | **  |
| PartnerYes                           | 4.735e-03  | 9.364e-02  | 0.051   | 0.959676 |     |
| DependentsYes                        | -1.898e-01 | 1.084e-01  | -1.751  | 0.079985 | .   |
| tenure                               | -5.835e-02 | 7.423e-03  | -7.862  | 3.79e-15 | *** |
| PhoneServiceYes                      | -8.526e-01 | 1.866e-01  | -4.570  | 4.88e-06 | *** |
| MultipleLinesYes                     | 1.762e-01  | 9.789e-02  | 1.800   | 0.071819 | .   |
| OnlineSecurityNo internet service    | -1.091e-01 | 2.209e-01  | -0.494  | 0.621278 |     |
| OnlineSecurityYes                    | -5.507e-01 | 1.000e-01  | -5.507  | 3.66e-08 | *** |
| StreamingTVNo internet service       | NA         | NA         | NA      | NA       |     |
| StreamingTVYes                       | 2.271e-02  | 1.063e-01  | 0.214   | 0.830867 |     |
| ContractOne year                     | -8.935e-01 | 1.304e-01  | -6.851  | 7.35e-12 | *** |
| ContractTwo year                     | -1.513e+00 | 2.032e-01  | -7.449  | 9.41e-14 | *** |
| PaperlessBillingYes                  | 3.098e-01  | 8.832e-02  | 3.507   | 0.000453 | *** |
| PaymentMethodCredit card (automatic) | -2.452e-01 | 1.358e-01  | -1.806  | 0.070968 | .   |
| PaymentMethodElectronic check        | 1.895e-01  | 1.116e-01  | 1.697   | 0.089638 | .   |
| PaymentMethodMailed check            | -1.930e-01 | 1.359e-01  | -1.421  | 0.155399 |     |
| MonthlyCharges                       | 2.232e-02  | 4.050e-03  | 5.512   | 3.55e-08 | *** |
| TotalCharges                         | 2.398e-04  | 8.423e-05  | 2.848   | 0.004405 | **  |

4. Internet Service 변수 제거 후  
NA가 나왔던 Online Security,  
Streaming TV 포함해 모델링 진행




**실패!!**

**다중공선성 해결 필요**

### III. 모델링 - 로지스틱 회귀

#### 다중공선성의 문제점

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- 행렬  $X$  에서 Column 간의 선형 결합을 통해 어떤 Column이 설명!  
(Ex.  $X_1 = X_2 + X_3$ )
  - 행렬  $X'X$  가 **full rank** 아님!
  - $X'X$  의 **역행렬**이 존재하지 않아 **B 계수**를 추정 불가능(OLS)
- 

# III. 모델링 - 로지스틱 회귀

## 범주형 변수 다중공선성 해결법

### ▶ 연속형 변환, 랜덤 에러 더하기

```
> print(summary(LogModel_err))

Call:
glm(formula = Churn ~ ., data = training2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8974  -0.7177  -0.0068   0.7169   4.2998

Coefficients: (6 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.030e-01  8.600e-01   0.585  0.558648
genderMale     -2.990e-03  3.055e-02  -0.098  0.922027
SeniorCitizen  9.628e-02  4.385e-02   2.196  0.028172 *
PartnerYes     2.851e-02  3.697e-02   0.771  0.440768
DependentsYes  -2.452e-02  3.899e-02  -0.629  0.529504
tenure         -5.660e-03  1.706e-03  -3.318  0.000915 ***
PhoneServiceYes 4.520e-02  3.039e-01   0.149  0.881769
MultipleLinesYes 4.410e-02  8.378e-02   0.526  0.598649
ContractOne year -5.674e-02  4.790e-02  -1.185  0.236256
ContractTwo year -3.734e-04  5.862e-02  -0.006  0.994918
PaperlessBillingYes 6.132e-02  3.397e-02   1.805  0.071137 .
PaymentMethodCredit card (automatic) 1.754e-02  4.650e-02   0.377  0.706050
PaymentMethodElectronic check 9.725e-02  4.513e-02   2.155  0.031223 *
PaymentMethodMailed check 9.242e-02  5.012e-02   1.844  0.065269 .
MonthlyCharges -2.672e-03  1.490e-02  -0.179  0.857711
TotalCharges    -1.572e-06  2.195e-05  -0.072  0.942897
ins_dsl         1.240e-01  4.851e-01   0.256  0.798172
ins_fib         2.224e-01  6.704e-01   0.332  0.740087
os_no           4.916e-02  4.245e-02   1.158  0.246876
os_nis          NA         NA         NA         NA
ob_no           1.431e-02  4.193e-02   0.341  0.732899
ob_nis          NA         NA         NA         NA
dp_no           -2.982e-02  4.205e-02  -0.709  0.478324
dp_nis          NA         NA         NA         NA
ts_no           6.106e-02  4.273e-02   1.429  0.153022
ts_nis          NA         NA         NA         NA
st_no           -4.089e-02  7.689e-02  -0.532  0.594897
st_nis          NA         NA         NA         NA
sm_no           -4.900e-02  7.699e-02  -0.637  0.524466
sm_nis          NA         NA         NA         NA
```

- 다중 공선성 의심되는 7개 변수(NA값이 나왔던 6개 변수와 InternetService 변수)를 더미변수로 변환 후  $N(0,1)$ 을 따르는 랜덤 에러를 더해줌.
- 랜덤 에러를 더해준 변수들을  $(-1, 1)$ 로 스케일링해줌.
- Glm함수로 적합해 보았으나 계수에 NA값이 나타남.

### III. 모델링 – 그룹 라쏘

#### Group LASSO – logistic regression

-범주들 간에 다중공선성이 존재하는 경우, 다중공선성이 존재하는 범주들을 묶어서  
“**그룹 라쏘**”를 시행하면, 공선성 문제를 해결할 수 있다.(Lukas et al, 2007)

#### 그룹 라쏘 모델

$$S_{\lambda}(\beta) = -l(\beta) + \lambda \sum_{g=1}^G s(df_g) \|\beta_g\|_2$$

where

$$l(\beta) = \sum_{i=1}^n y_i \eta_{\beta}(\mathbf{x}_i) - \log[1 + \exp\{\eta_{\beta}(\mathbf{x}_i)\}]$$

### III. 모델링 – 그룹 라쏘

Group LASSO – logistic regression

“Block Co-ordinate Gradient Descent” 를 사용하여 계산함.

$$M_{\lambda}^{(t)}(\mathbf{d}) = -\{l(\hat{\beta}^{(t)}) + \mathbf{d}^T \nabla l(\hat{\beta}^{(t)}) + \frac{1}{2} \mathbf{d}^T H^{(t)} \mathbf{d}\} + \lambda \sum_{g=1}^G s(df_g) \|\hat{\beta}_g^{(t)} + \mathbf{d}_g\|_2$$

$$\approx S_{\lambda}(\hat{\beta}^{(t)} + \mathbf{d}),$$

그룹라쏘에서 손실 함수 부분에 테일러 근사시킨 식 넣어주고,

d값에 베타 초기값 넣고, 적합해서 베타 업데이트해 서 다시 넣고, 수렴할 때 까지 반복...

### III. 모델링 - 그룹 라쏘

#### Group LASSO - logistic regression

If  $\|\nabla l(\hat{\beta}^{(t)})_g - \tilde{h}_g^{(t)} \hat{\beta}_g^{(t)}\|_2 \leq \lambda s(df_g)$ , the minimizer of equation (2.3) is

$$\mathbf{d}_g^{(t)} = -\hat{\beta}_g^{(t)}.$$

Otherwise

$$\mathbf{d}_g^{(t)} = -\frac{1}{h_g^{(t)}} \left\{ \nabla l(\hat{\beta}^{(t)})_g - \lambda s(df_g) \frac{\nabla l(\hat{\beta}^{(t)})_g - h_g^{(t)} \hat{\beta}_g^{(t)}}{\|\nabla l(\hat{\beta}^{(t)})_g - h_g^{(t)} \hat{\beta}_g^{(t)}\|_2} \right\}.$$

BCGD 알고리즘: Tseng and Yun, (2007)

A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization

로지스틱 모델에 적용: Lukas et al, (2008)

The group lasso for Logistic regression

# III. 모델링 – 그룹 라쏘

사례 분석 : Group lasso를 통한 중학생 삶의 만족도에 영향을 미치는 변수 탐색

별점 회귀모형 분석 결과 회귀계수가 0이 아닌 15개 설명변수와 척도

| 순 | 변수명     | 변수 설명                                                     | 척도                                    | 계수     |
|---|---------|-----------------------------------------------------------|---------------------------------------|--------|
| 1 | PSY2A01 | 자아인식: 자아존중감-나는나에게 만족한다                                    | Likert: 1(매우 그렇다)<br>~4(전혀 그렇지 않다)    | -0.886 |
| 2 | PSY1E03 | 정서문제: 우울-걱정이 많다                                           | Likert: 1(매우 그렇다)<br>~4(전혀 그렇지 않다)    | 0.821  |
| 3 | PSY2A10 | 자아인식: 자아존중감-나는 나에 대해 긍정적인 태도를 가지고 있다                      | Likert: 1(매우 그렇다)<br>~4(전혀 그렇지 않다)    | -0.781 |
| 4 | INT1D   | 전체성적 만족도                                                  | Likert: 1(매우 만족한다)<br>~4(전혀 만족하지 않는다) | -0.471 |
| 5 | FAM2F01 | 양육방식: 합리적설명-부모님(보호자)의 결정을 무조건 따르게 하기 보다는 왜 그래야 하는지 설명해주신다 | Likert: 1(매우 그런 편이다)<br>~4(전혀 그렇지 않다) | -0.463 |
| 6 | COM1A05 | 지역사회인식: 나는 우리 동네 사람들과 지내는 것이 좋다                           | Likert: 1(매우 그런 편이다)<br>~4(전혀 그렇지 않다) | -0.463 |
| 7 | GENDER  | 성별                                                        | 0(남자), 1(여자)                          | -0.409 |
| 8 | PSY1C06 | 정서문제: 신체증상-자주 피곤해한다                                       | Likert: 1(매우 그런 편이다)<br>~4(전혀 그렇지 않다) | 0.336  |

## ▶ 모델링 결과

- 삶의 만족도를 0,1로 구분(높음, 낮음 수준)하여  
설문 조사를 결과에 적용

- 결과:

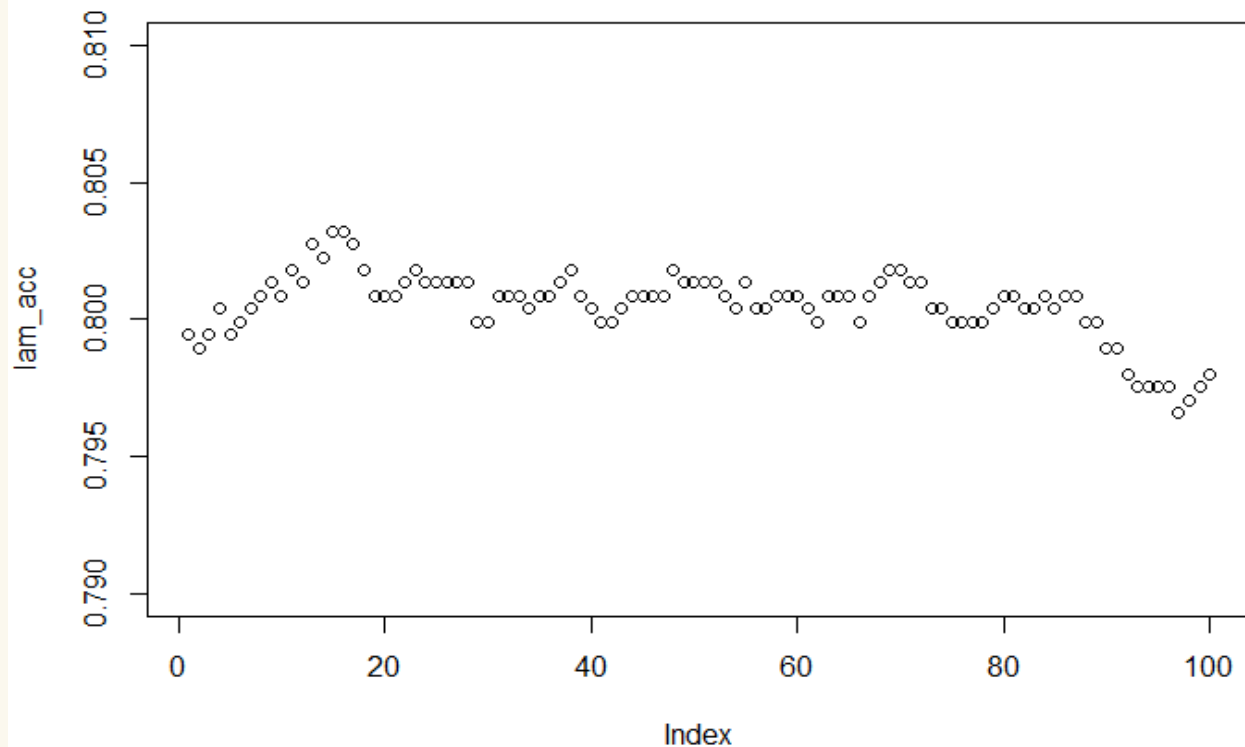
338개의 설명변수를 15개로 감소 (변수 중요도에 따라)

정확도는 78%

## III. 모델링 - 그룹 라쏘

### Churn 데이터 적용

#### ▶ 벌점 모수( $\lambda$ ) 결정



- 람다가 1~100일 때 Group LASSO 적합 시 모델의 정확성을 알아봄.
- 람다가 15, 16일 때 Group LASSO 모델의 정확성이 가장 높았다.



# III. 모델링 - 그룹 라쏘

## Churn 데이터 적용

```
> round(grpls$coefficients, 4)
              15
(Intercept)  -0.1950
genderMale    -0.0327
SeniorCitizenYes  0.2909
PartnerYes    -0.0244
DependentsYes  -0.0755
tenure        -0.0343
PhoneServiceYes -0.3619
MultipleLinesYes  0.2962
InternetServiceFiber optic  0.7633
InternetServiceNo -0.1042
OnlineSecurityNo internet service -0.3880
OnlineSecurityYes -0.3814
OnlineBackupNo internet service -0.1575
OnlineBackupYes -0.1405
DeviceProtectionNo internet service 0.0000
DeviceProtectionYes 0.0000
TechSupportNo internet service -0.2284
TechSupportYes -0.2223
StreamingTVNo internet service -0.0347
StreamingTVYes 0.1818
StreamingMoviesNo internet service -0.0370
StreamingMoviesYes 0.2136
ContractOne year -0.6610
ContractTwo year -1.1503
PaperlessBillingYes 0.3687
PaymentMethodCredit card (automatic) -0.0959
PaymentMethodElectronic check 0.3301
PaymentMethodMailed check -0.0252
MonthlyCharges 0.0000
TotalCharges 0.0000
```

```
> print(paste('Group LASSO Accuracy',1-misClasificError1))
[1] "Group LASSO Accuracy 0.803224276908487"
> print("Confusion Matrix for Group LASSO Logistic Regression"); table(testing1$churn, fitted.results1)
[1] "Confusion Matrix for Group LASSO Logistic Regression"
      fitted.results1
      0      1
0 1408  168
1   247  286
```

```
> print(paste('Logistic Regression Accuracy',1-misClasificError2))
[1] "Logistic Regression Accuracy 0.811195445920304"
> print("Confusion Matrix for Logistic Regression"); table(testing$churn, fitted.results2 > 0.5)
[1] "Confusion Matrix for Logistic Regression"
      FALSE TRUE
0  1397  151
1   247  313
```

1. 계수 NA값 문제 해결

2. 모델 정확도 측면에서도 로지스틱 회귀와 비슷한 성능 보임

# III. 모델링 – 로지스틱 회귀 vs 그룹 라쏘

## 계수 NA값 문제점 해결

Step: AIC=4078.93  
Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +  
InternetService + OnlineSecurity + TechSupport + StreamingTV +  
StreamingMovies + Contract + PaperlessBilling + PaymentMethod +  
MonthlyCharges + TotalCharges

|                    | Df | Deviance | AIC    |
|--------------------|----|----------|--------|
| <none>             |    | 4040.9   | 4078.9 |
| - Dependents       | 1  | 4043.6   | 4079.6 |
| - OnlineSecurity   | 1  | 4044.7   | 4080.7 |
| - TechSupport      | 1  | 4045.4   | 4081.4 |
| - SeniorCitizen    | 1  | 4047.6   | 4083.6 |
| - PaymentMethod    | 3  | 4054.0   | 4086.0 |
| - PaperlessBilling | 1  | 4051.3   | 4087.3 |
| - TotalCharges     | 1  | 4053.7   | 4089.7 |
| - MultipleLines    | 1  | 4055.0   | 4091.0 |
| - MonthlyCharges   | 1  | 4061.2   | 4097.2 |
| - StreamingMovies  | 1  | 4064.5   | 4100.5 |
| - StreamingTV      | 1  | 4065.7   | 4101.7 |
| - Contract         | 2  | 4100.5   | 4134.5 |
| - InternetService  | 2  | 4105.8   | 4139.8 |
| - tenure           | 1  | 4117.2   | 4153.2 |

‘No Internet service’ 수준을 ‘No’ 수준에 통합 후 Stepwise

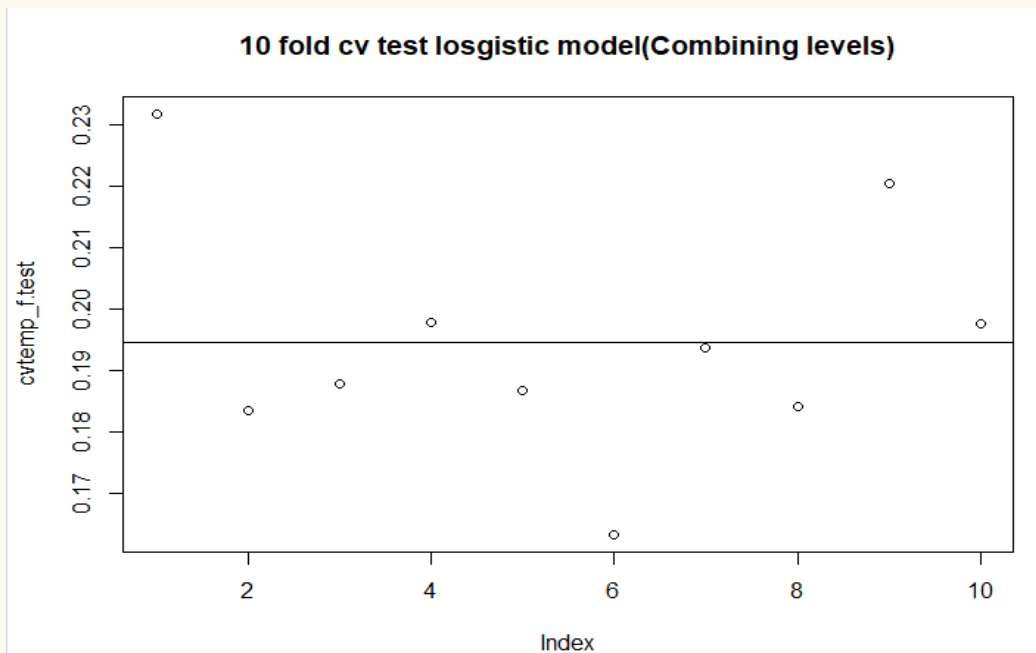
```
> print(paste('Logistic Regression(Combining levels) Accuracy',1-misClasificError_f))
[1] "Logistic Regression(Combining levels) Accuracy 0.787950664136622"
> print("Confusion Matrix for Logistic Regression(Combining levels)"); table(testing_f$churn, fitted.results_f )
[1] "Confusion Matrix for Logistic Regression(Combining levels)"
      fitted.results_f
      No  Yes
No  1381  167
Yes   280  280
> print(paste('Group LASSO Accuracy',1-misClasificError1))
[1] "Group LASSO Accuracy 0.803224276908487"
> print("Confusion Matrix for Group LASSO Logistic Regression"); table(testing1$churn, fitted.results1 )
[1] "Confusion Matrix for Group LASSO Logistic Regression"
      fitted.results1
      0    1
0  1408  168
1   247  286
```

- 변수 5개 제거 (Gender, Partner, Phone Service, Online Backup, Device Protection)
- 둘다 NA 계수값 문제 해결했으나, Group LASSO 적용한 모델의 정확성이 더 높게 나옴.

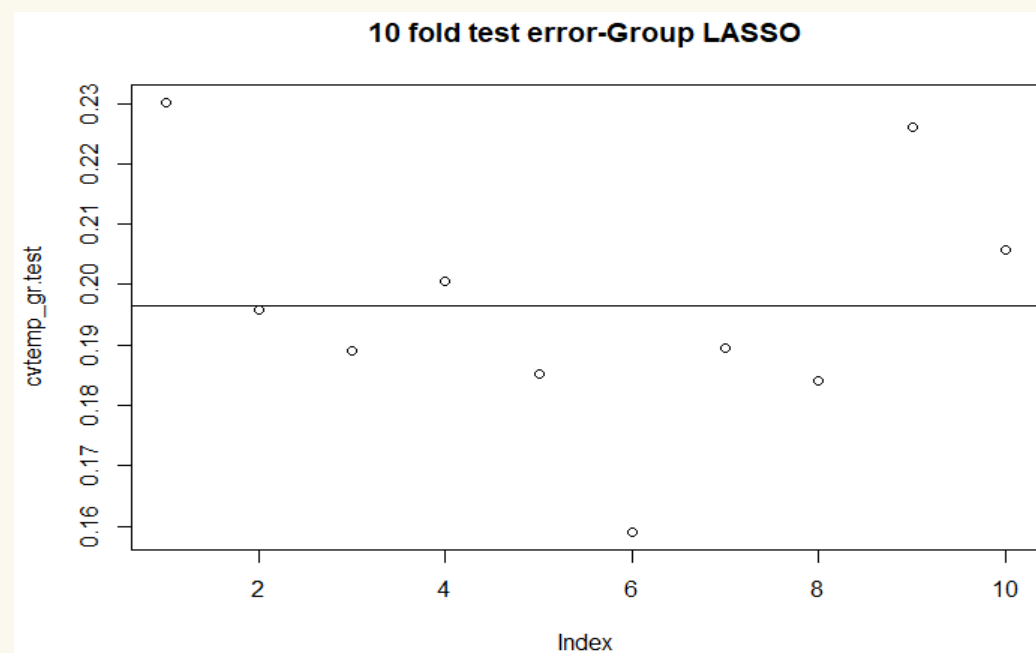
# III. 모델링 – 로지스틱 회귀 vs 그룹 라쏘

## 10-fold validation

### ▶ 로지스틱 회귀



### ▶ Group LASSO

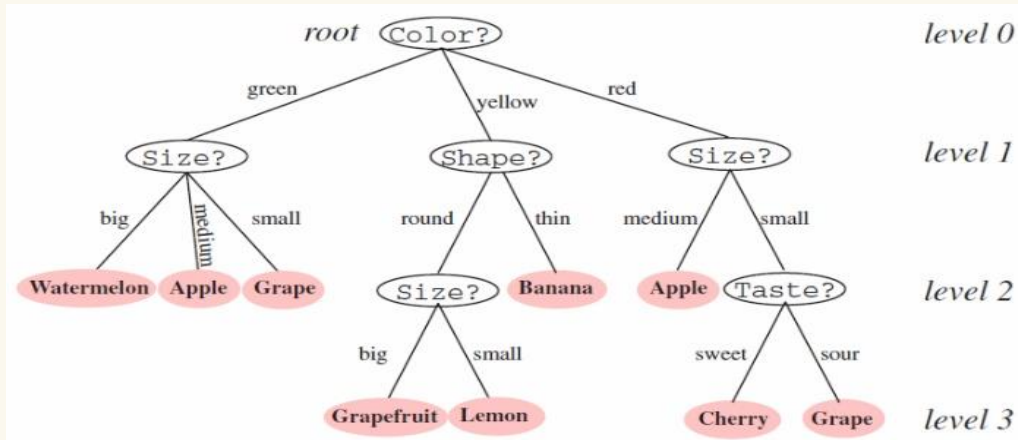


1. 로지스틱 회귀 : mean.error값이 0.1945, varianc는 0.00037
2. Group LASSO : mean.error값이 0.1965, varianc는 0.00043

# III. 모델링 – 의사결정 나무(Decision Tree)

## 모델링 정의

### ▶ 의사결정 나무(Decision Tree)란?



- 기계학습 중 하나로 특정 항목에 대한 의사 결정 규칙을 나무 형태로 분류해 나가는 분석기법
- 장점 : ① 구조가 단순하여 결과해석이 쉬움  
② 선형성, 정규성, 등분산성 가정이 불필요
- 단점 : ① 기준값의 경계선 근방 자료 값에 대해 오차가 클 수 있음  
② 새로운 자료에 대한 예측 불안정하고 선형성 미흡함

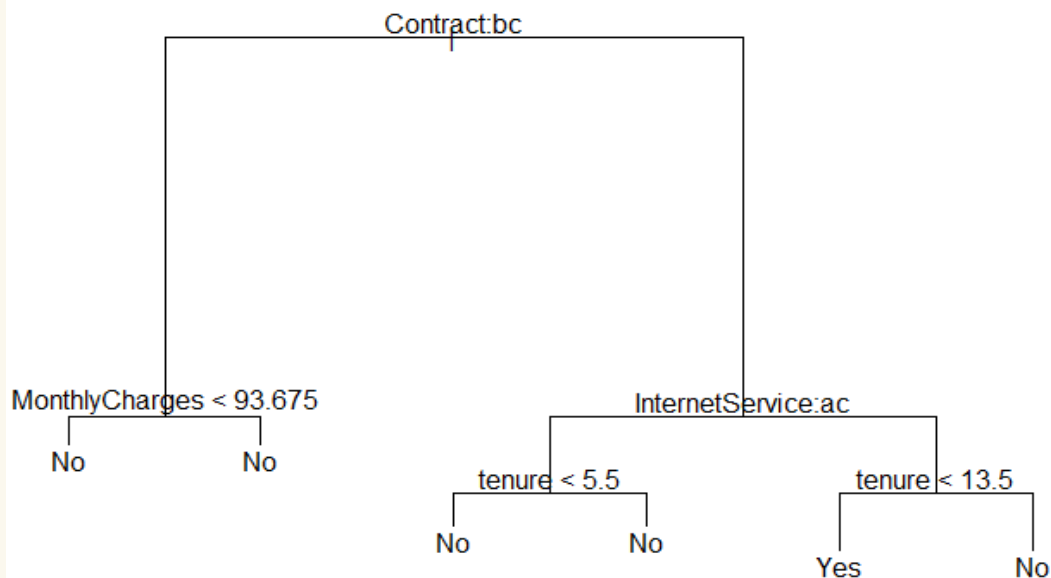
### ▶ R 패키지 비교

1. tree 패키지(binary recursive partitioning 방법)와 rpart 패키지(CART 방법)은 엔트로피, 지니계수 기준으로 가지치기할 변수 결정하기 때문에 연산 속도는 빠르지만, **과적합 위험이 있어 가지치기 과정이 따로 필요함**
2. Party 패키지(Unbiased recursive partitioning based on permutation tests 방법)은 p-test 기반 중요도 기준으로 가지치기할 변수 정해주기 때문에 **별도의 가지치기 과정 불필요**

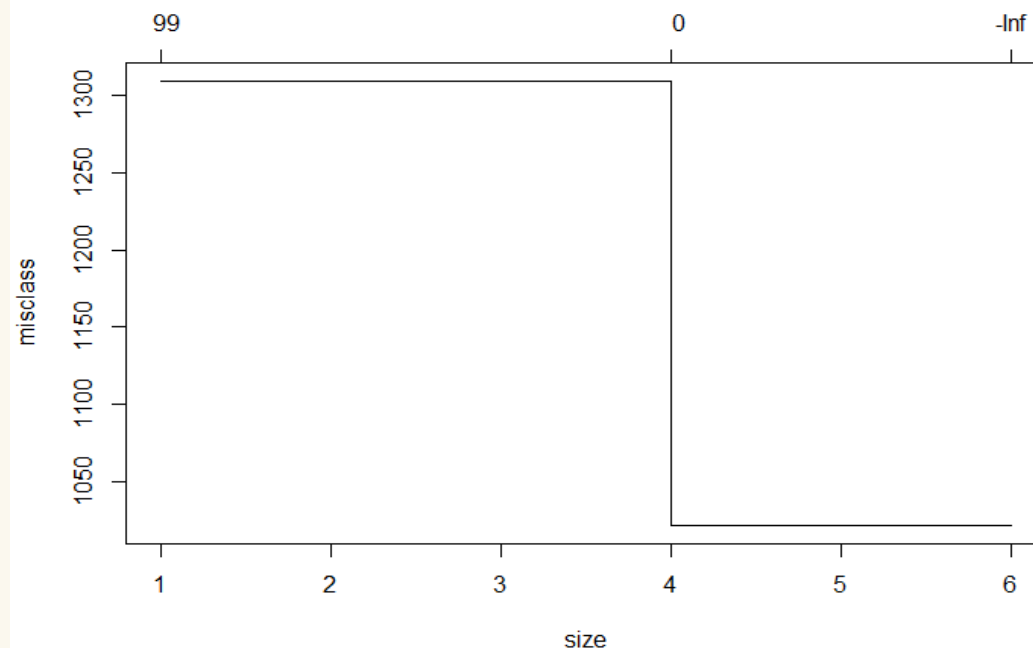
# III. 모델링 - 의사결정 나무(Decision Tree)

## tree 패키지

### ▶ 의사 결정 나무(training data 이용)



### ▶ tree size 결정

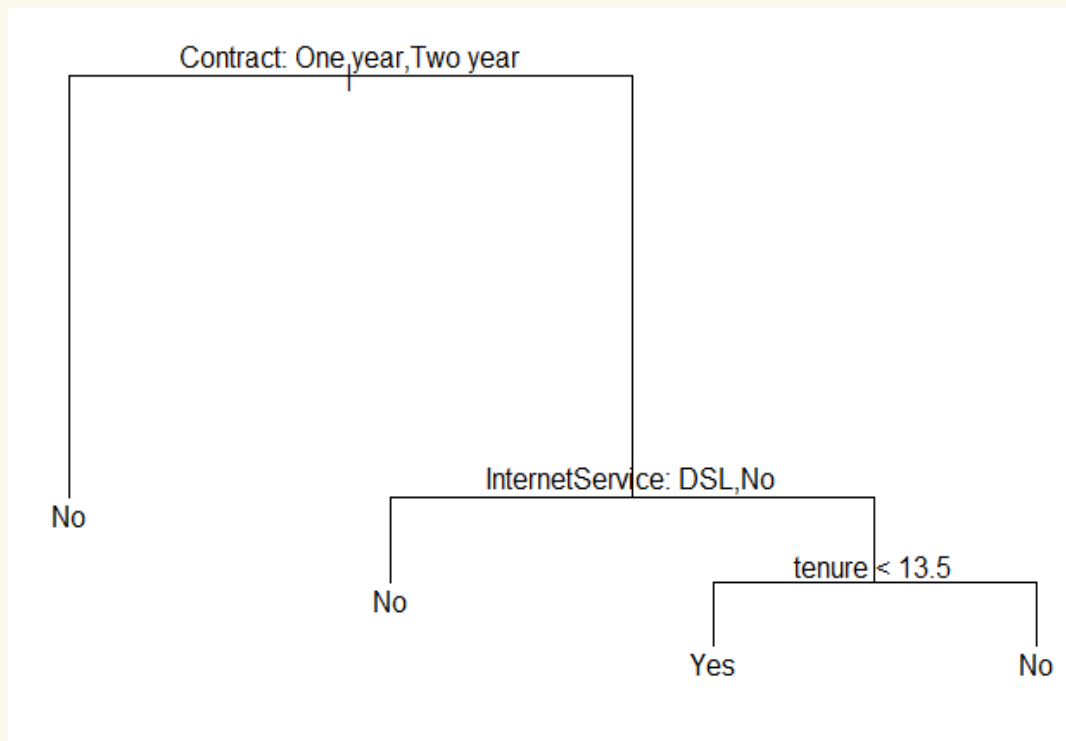


- tree 함수 이용해서 의사결정 나무 만든 결과, MonthlyCharges나 tenure<5.5는 같은 결과값 출력 -> 의미 x
- Pruning(가지치기) 과정을 위해 cross-validation 과정 통해 **misclass가 낮아지는 4개 tree size 결정**

# III . 모델링 – 의사결정 나무(Decision Tree)

## tree 패키지

### ▶ 모델링 결과



```
> confusionMatrix(tree.pred.churn, testing$churn)
Confusion Matrix and Statistics
```

|            | Reference |     |
|------------|-----------|-----|
| Prediction | No        | Yes |
| No         | 1460      | 382 |
| Yes        | 88        | 178 |

```

Accuracy : 0.777
95% CI : (0.7587, 0.7947)
No Information Rate : 0.7343
P-Value [Acc > NIR] : 3.535e-06
```

```

Kappa : 0.3135
McNemar's Test P-Value : < 2.2e-16
```

```

Sensitivity : 0.9432
Specificity : 0.3179
Pos Pred value : 0.7926
Neg Pred value : 0.6692
Prevalence : 0.7343
Detection Rate : 0.6926
Detection Prevalence : 0.8738
Balanced Accuracy : 0.6305
```

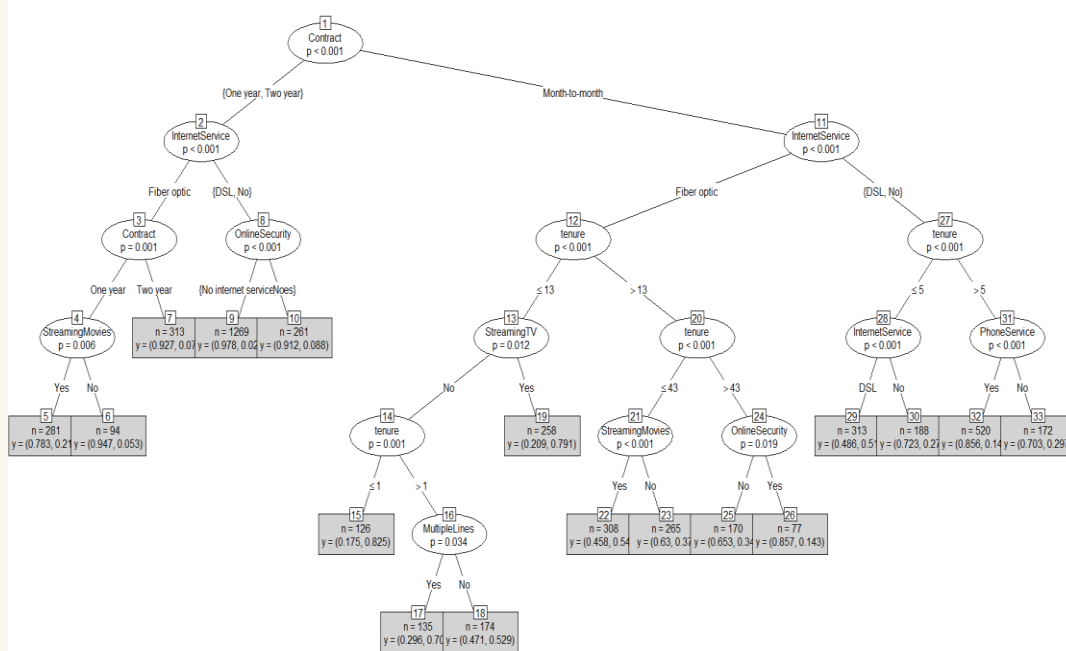
```
'Positive' class : No
```

- 가지치기 이후 모델로 예측 결과, confusion matrix 기반 예측 정확도 77.7% 보임

# III. 모델링 – 의사결정 나무(Decision Tree)

## party 패키지

### ▶ 모델링 결과



```
> confusionMatrix(tree.pred.churn, testing$churn)
Confusion Matrix and Statistics
```

|            | Reference |     |
|------------|-----------|-----|
| Prediction | No        | Yes |
| No         | 1320      | 252 |
| Yes        | 228       | 308 |

Accuracy : 0.7723  
 95% CI : (0.7538, 0.79)  
 No Information Rate : 0.7343  
 P-Value [Acc > NIR] : 3.439e-05

Kappa : 0.4083  
 McNemar's Test P-Value : 0.2938

Sensitivity : 0.8527  
 Specificity : 0.5500  
 Pos Pred value : 0.8397  
 Neg Pred value : 0.5746  
 Prevalence : 0.7343  
 Detection Rate : 0.6262  
 Detection Prevalence : 0.7457  
 Balanced Accuracy : 0.7014

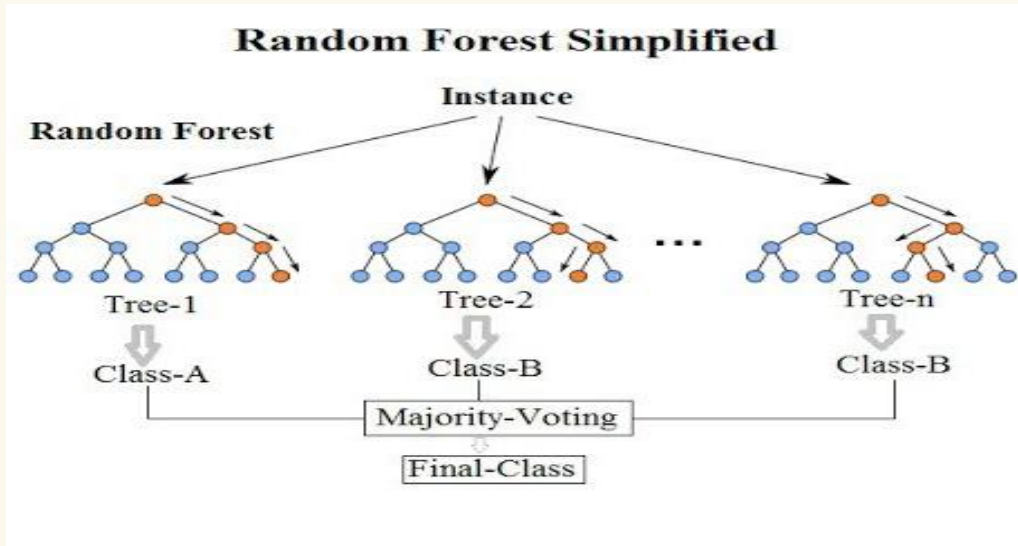
'Positive' Class : No

- party 패키지의 ctree 함수 기반 의사결정 나무 모델 예측률은 77.2%를 보임
- 근소하게나마 가지치기 과정을 수행해주는 tree 패키지가 party 패키지보다 높은 성능을 보임

# III. 모델링 – 랜덤 포레스트(Random Forest)

## 모델링 정의

### ▶ 랜덤 포레스트(Random Forest)란?



- 다수의 의사결정 나무를 결합하여 하나의 모형을 생성하는 방법
- 장점 : ① 다양성 극대화하여 예측력이 높음  
② 다수 tree의 예측 결과를 종합하여 안정성도 우수
- 단점 : ① 다수 tree 이용해 의사결정 내려 기존 의사결정 나무가 갖는 장점인 설명력이 떨어짐

### ▶ R 패키지

- randomForest 패키지 활용
- 보통 regression의 경우 변수 개수/3, classificatio의 경우 sqrt(변수 개수)로 사용



# III. 모델링 – 랜덤 포레스트(Random Forest)

## 트리 개수 결정

### ▶ 모든 변수 고려

```
> confusionMatrix(pred_rf, testing$churn)
Confusion Matrix and Statistics
```

|            | Reference |     |
|------------|-----------|-----|
| Prediction | No        | Yes |
| No         | 1394      | 280 |
| Yes        | 154       | 280 |

```

      Accuracy : 0.7941
    95% CI : (0.7762, 0.8112)
  No Information Rate : 0.7343
    P-Value [Acc > NIR] : 1.058e-10

      Kappa : 0.4315
  Mcnemar's Test P-Value : 1.971e-09

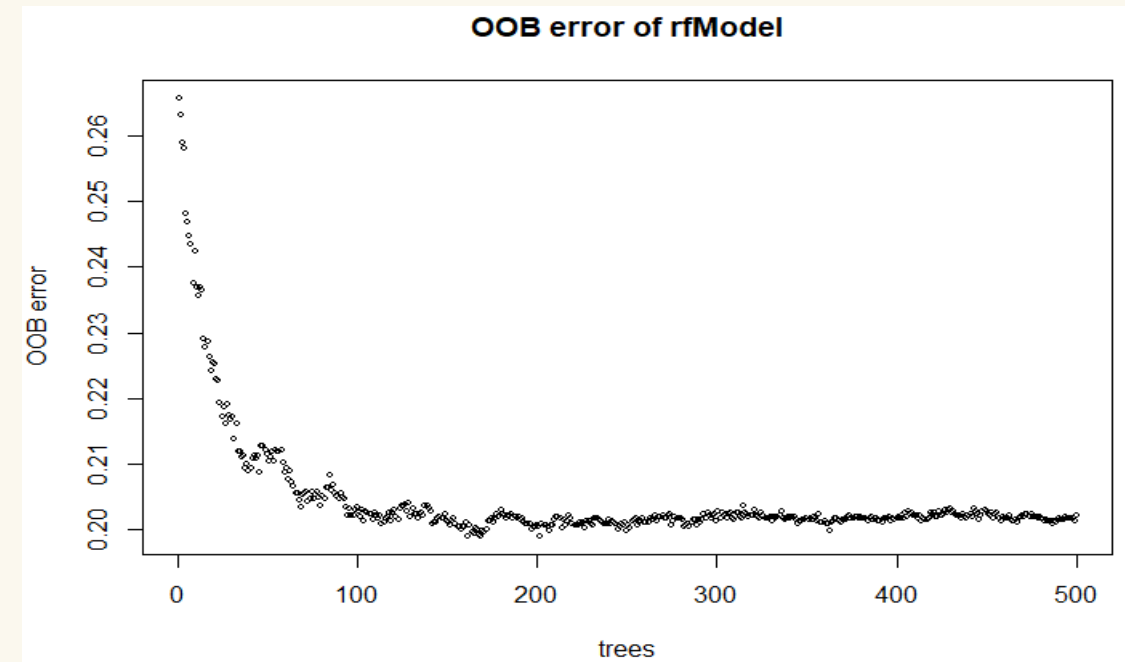
    sensitivity : 0.9005
   specificity : 0.5000
  Pos Pred Value : 0.8327
  Neg Pred Value : 0.6452
    Prevalence : 0.7343
  Detection Rate : 0.6613
Detection Prevalence : 0.7941
Balanced Accuracy : 0.7003

'Positive' Class : No

```

- 정확성은 0.7941로 Decision Tree보다 높게 나옴
- 민감도는 0.9, 특이도는 0.5

### ▶ OOB Error

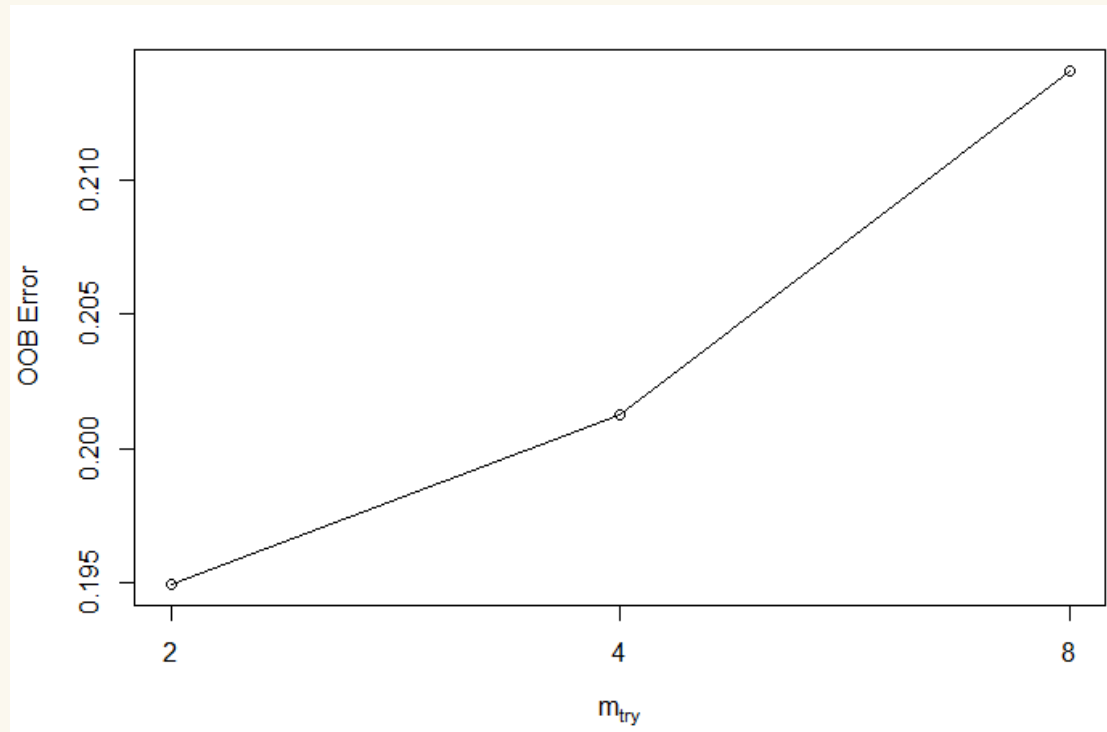


- OOB Error 트리 개수가 200개부터 안정화됨
- 적정 트리 개수 200개로 선정

# III. 모델링 – 랜덤 포레스트(Random Forest)

## P값 구하기

```
> t <- tuneRF(training[, -20], training[, 20], stepFactor = 0.5,
+             plot = TRUE, ntreeTry = 200, trace = TRUE, improve = 0.05)
mtry = 4 OOB error = 20.13%
Searching left ...
mtry = 8 OOB error = 21.41%
-0.06357215 0.05
Searching right ...
mtry = 2 OOB error = 19.5%
0.03128153 0.05
```



- Random Forest 각각의 tree마다 몇 개의 feature를 사용할 것인지를 정하는 과정.
- 분류 트리이므로 Default  $m_{try} = \sqrt{19} \approx 4$ 개일 때부터 확인하여 최적 개수로는 2개를 선정

# III. 모델링 – 랜덤 포레스트(Random Forest)

## 모델링 결과

### ▶ 트리 개수=200, p=2 적용

```
> confusionMatrix(pred_rf_new, testing$churn)
Confusion Matrix and Statistics

              Reference
Prediction    No  Yes
   No      1392  289
   Yes     156  271

              Accuracy : 0.7889
              95% CI   : (0.7708, 0.8061)
   No Information Rate : 0.7343
   P-Value [Acc > NIR] : 3.727e-09

              Kappa    : 0.4146
   Mcnemar's Test P-Value : 3.914e-10

              Sensitivity : 0.8992
              Specificity : 0.4839
   Pos Pred Value   : 0.8281
   Neg Pred Value   : 0.6347
   Prevalence       : 0.7343
   Detection Rate   : 0.6603
   Detection Prevalence : 0.7974
   Balanced Accuracy : 0.6916

   'Positive' Class : No
```

```
> print(rfModel_new)
```

Call:

```
randomForest(formula = Churn ~ ., data = training, ntree = 200, mtry = 2, importance = TRUE, proximity = TRUE)
Type of random forest: classification
Number of trees: 200
No. of variables tried at each split: 2
```

OOB estimate of error rate: 19.82%

Confusion matrix:

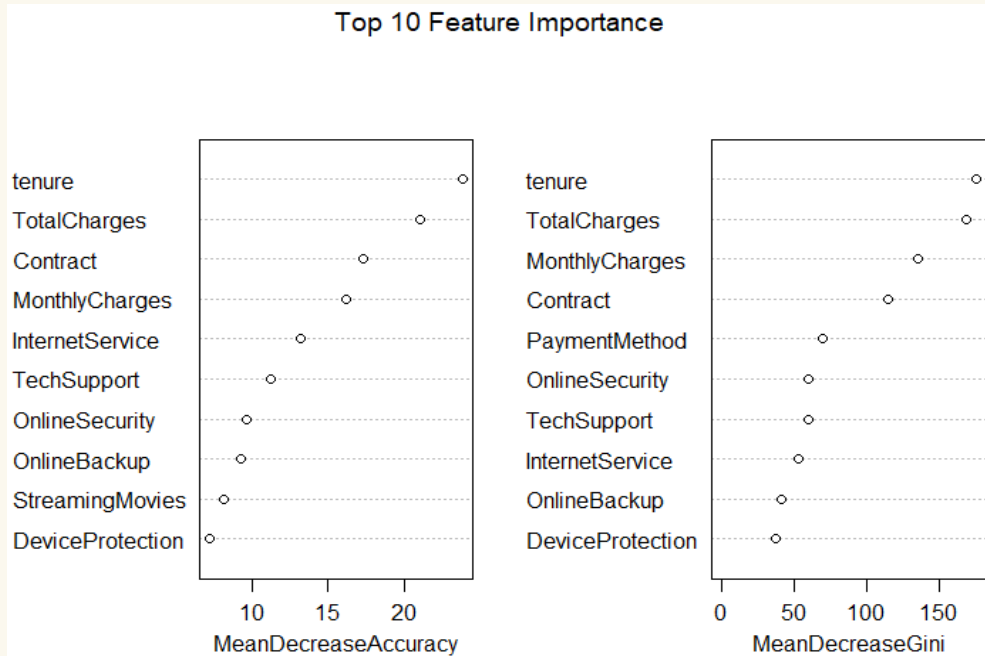
```
      No Yes class.error
No 3270 345 0.09543568
Yes 631 678 0.48204736
```

- 정확성은 0.7889로 모든 변수를 포함했을 때의 모델 정확성과 0.052밖에 차이나지 않음
- OOB Error는 0.2045에서 0.1982로 0.063만큼 감소

# III. 모델링 – 랜덤 포레스트(Random Forest)

## 변수 중요도

### ▶ 랜덤 포레스트



### ▶ 로지스틱 회귀

```
> imp_l[order(imp_l$overall,decreasing = T),]
      overall names
5  8.09934511 tenure
17 6.78861033 ContractTwo year
16 6.49692795 Contractone year
23 4.04902202 Totalcharges
18 3.66117423 PaperlessBillingYes
7  2.73513697 MultipleLinesYes
20 2.71374273 PaymentMethodElectronic check
9  2.43427033 InternetServiceNo
14 2.27064598 StreamingTVYes
8  2.24720480 InternetServiceFiber optic
15 2.03688753 StreamingMoviesYes
2  1.67675566 SeniorCitizenYes
22 1.55814968 MonthlyCharges
12 0.85129715 DeviceProtectionYes
4  0.79480695 DependentsYes
6  0.67843064 PhoneserviceYes
1  0.61905513 genderMale
13 0.29700855 TechSupportYes
11 0.23126399 onlineBackupYes
3  0.16148547 PartnerYes
19 0.14665215 PaymentMethodCredit card (automatic)
10 0.13338445 onlinesecurityYes
21 0.01295932 PaymentMethodMailed check
```

- 분류 트리이므로 지니계수 기준으로 변수 중요도 파악
- tenure, TotalCharges, MontlyCharges, Contract가 특히 중요한 변수 (로지스틱 회귀 변수 중요도 결과와 유사)

## IV. 모델 평가

### Classification 평가 척도

#### ▶ Confusion Matrix : 암판정 예시로 이해하기

• TP / TN / FP / FN

| 1000 | 정상판정                  | 암판정                 |
|------|-----------------------|---------------------|
| 정상환자 | 988 <small>TN</small> | 2 <small>FP</small> |
| 암환자  | 1 <small>FN</small>   | 9 <small>TP</small> |

#### ▶ Accuracy로 해석할 경우

$$Acc = \frac{TP + TN}{All}$$

① TP : 올바른 예측값 대해서 실제값도 올바른 경우

② TN : 거절해야 할 예측값 대해서 실제값도 거절되어야 하는 경우

③ FP : 올바른 예측값 대해서 실제값 거절한 경우

④ FN : 거절해야 할 예측값 대해서 실제값 올바른 경우

① 정상 정확도 :  $988/990=99.8\%$  ② 암환자 정확도 :  $9/10=90\%$

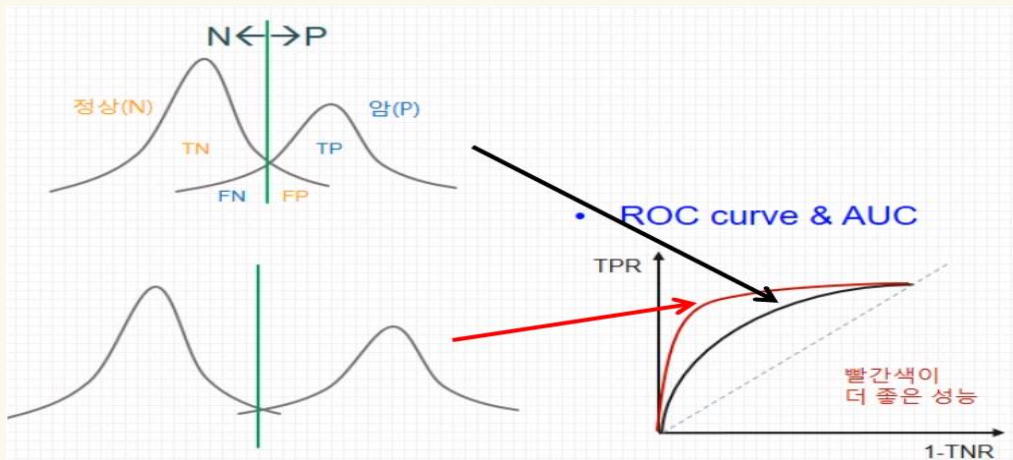
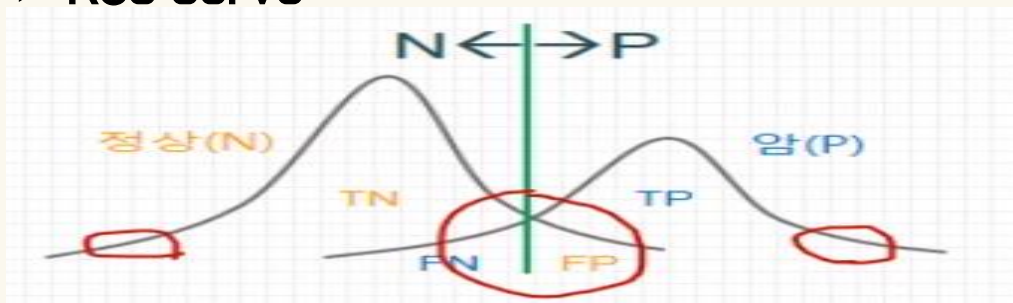
- 전체 정확도 : 99.7%로 암환자 정확도가 정확하게 반영 x

- 이렇듯 클래스별 분포가 불균형할 때 정확도만 고려하는 건 부적절

## IV. 모델 평가

### Classification 평가 척도

#### ▶ ROC Curve

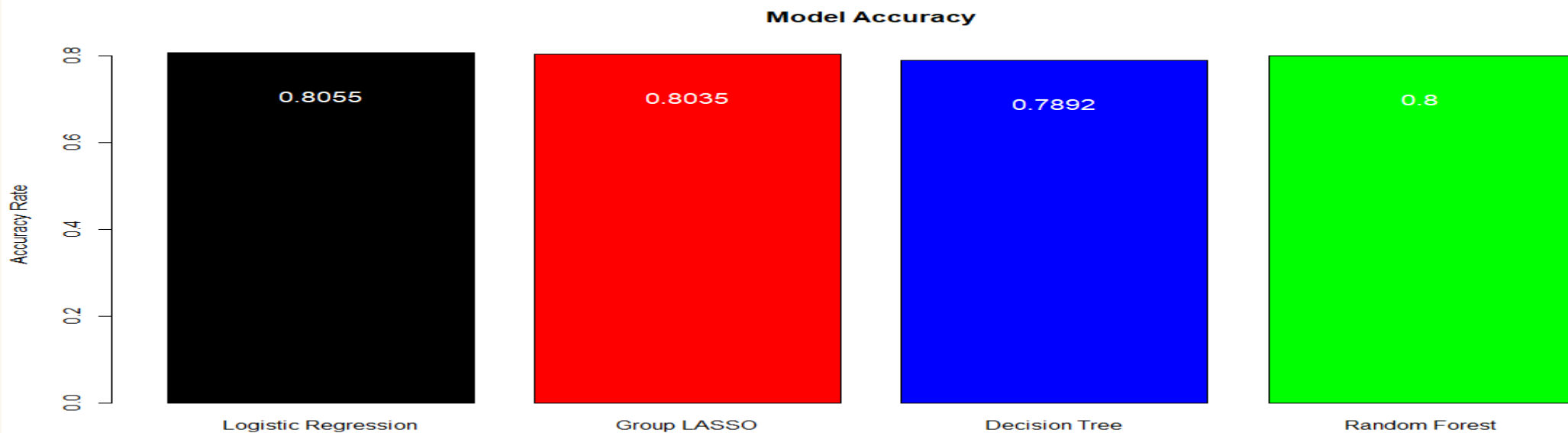


- 앞의 암환자 예시처럼 불균형 분포에서 양 끝쪽은 확실히 정상이나 암환자로 구분이 가능하나 가운데 쪽은 불분명한 부분이 생김
- 최선의 판단(초록색 선을 긋는 행위)하려면 필연적으로 error 포함
- 판단선(초록색 선)을 내릴 때 왼쪽이나 오른쪽으로 움직일수록 어느 정도로 안 좋은 결과 초래하는지 보여주는 것이 ROC Curve
- 옆의 위쪽 분포처럼 많이 겹치는 부분 많을수록 직선에 가까워짐
- ROC Curve가 상단에 붙을수록(직선에 멀어질수록) 성능 좋은 것

## IV. 모델 평가

### 모델링 결과 비교

#### ▶ 예측률

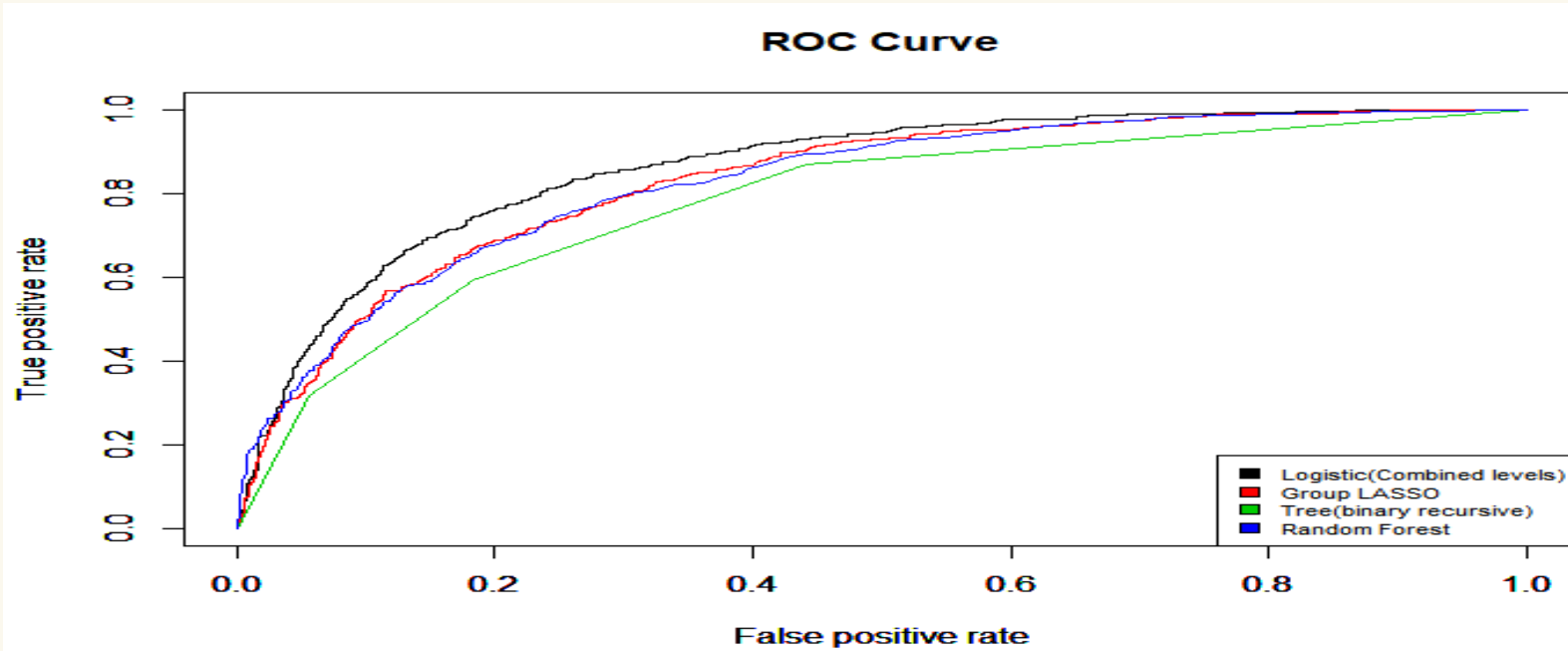


- 10-fold로 돌린 예측률 결과, 로지스틱 회귀 모델이 성능이 가장 뛰어남
- 그러나 대부분 모델 예측률에 큰 차이가 없으며 클래스 분포도 불균형한 부분이 있었으므로 ROC Curve 추가적으로 고려

## IV. 모델 평가

### 모델링 결과 비교

#### ▶ ROC Curve



```
> auc_fit.logit  
[1] 0.86087  
> auc_grpls  
[1] 0.8286665  
> auc_tree  
[1] 0.7769841  
> auc_rf  
[1] 0.8261686
```

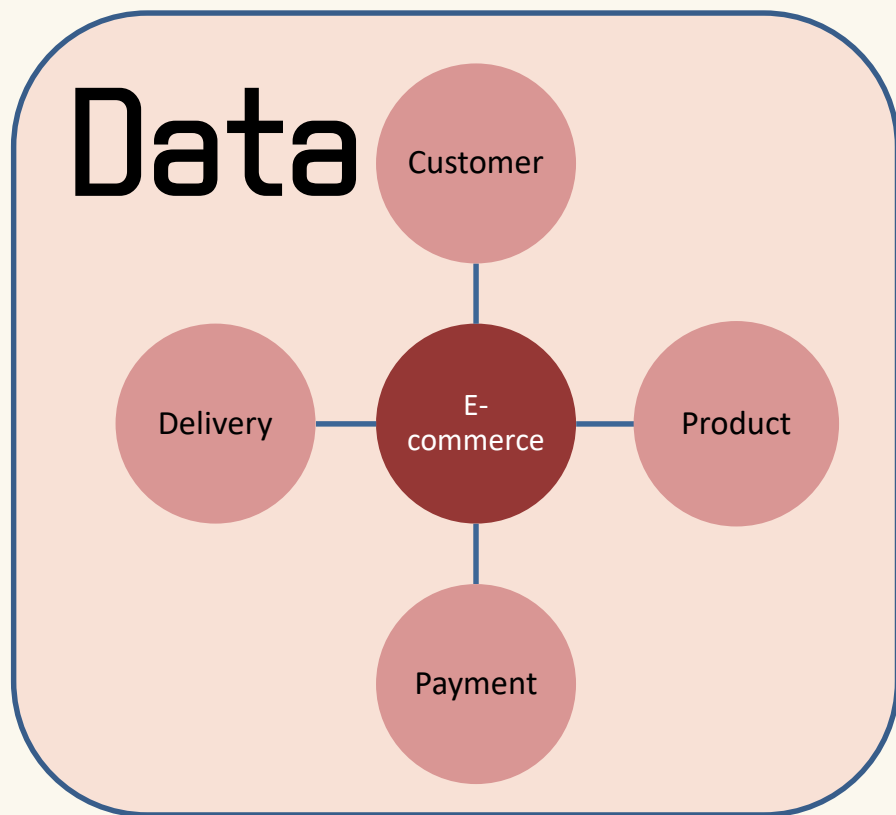
- ROC Curve 결과 역시 로지스틱 회귀 모델이 성능이 가장 좋음을 확인할 수 있음
- 예측률과 ROC Curve 함께 고려한 결과, 4가지 모델 중 고객 이탈률 예측 측면에서 로지스틱 회귀 모델이 가장 뛰어남



# 군집 분석

- I 주제 선정
- II 데이터 탐색 및 전처리
- III 군집 분석의 실패사례와 EDA
- IV 선호도 지수

## I. 주제 선정



Analysis



**Conclusion**

기업이 **마케팅**에  
효율적으로 활용할 수 있는 방안은?

## II. 데이터 탐색 및 전처리

### Brazilian E-Commerce Public Dataset by Olist

- 판매자들을 위한 브라질의 온라인 E-commerce 사이트 (우리나라의 위메프 같은 사이트)
- 2016년~2018년 사이에 들어온 100K개의 주문 내역



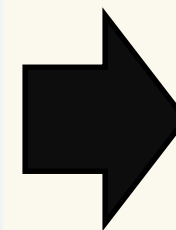
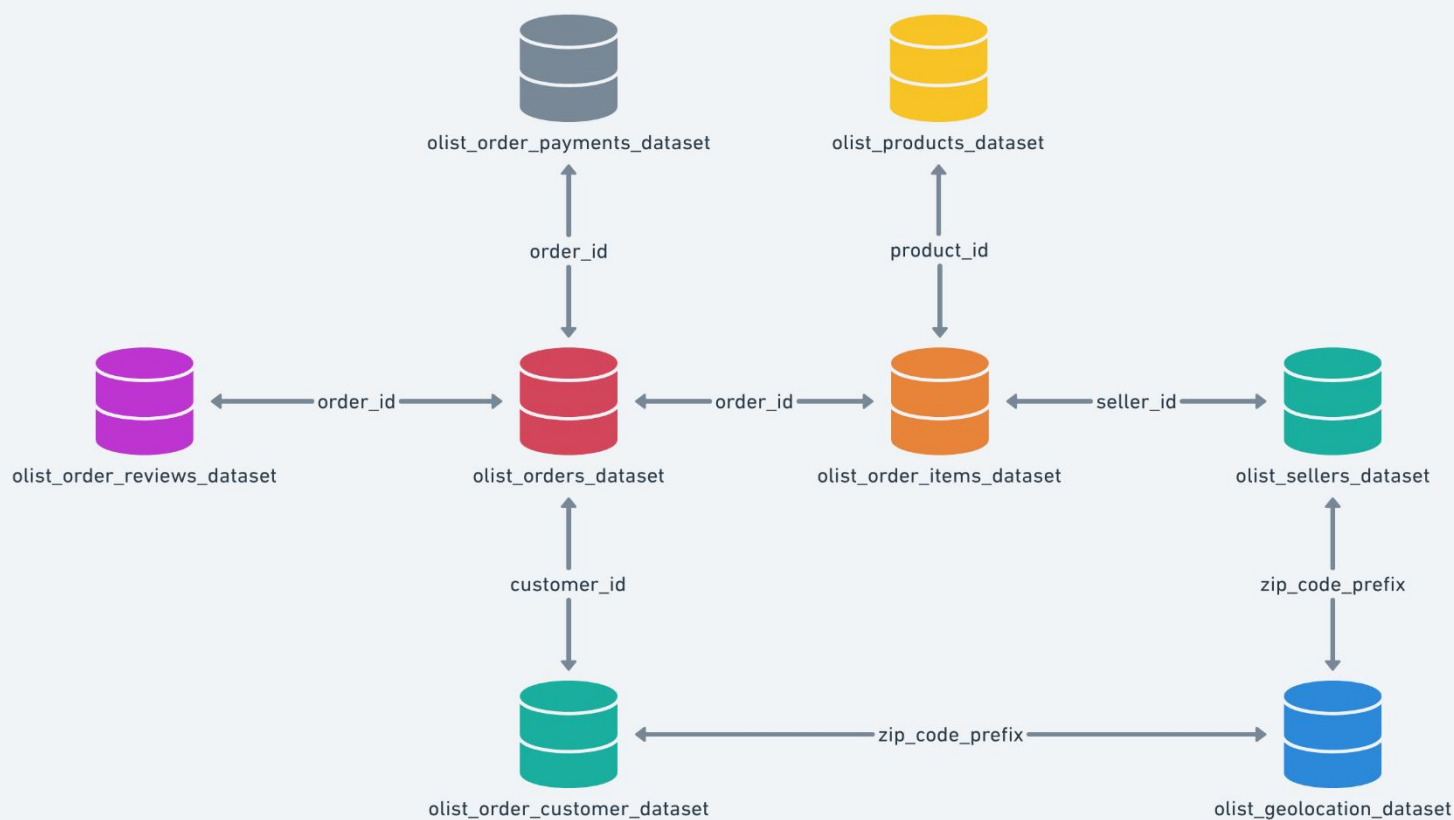
## II. 데이터 탐색 및 전처리

### 데이터셋 구조

- Orders : 구매자, 주문번호, 세부주문번호, 배송방법, 구매시각, 배송시작시각, 배송완료시각,...
- Order payments : 결제 정보 - 주문번호, 결제 수단, 할부 개월, 결제수단 가짓수(쿠폰+상품권+신용카드)
- Order items : 주문상품 정보 - 주문번호, 상품번호, 구매자, 판매자, 상품가격, 배송료, 배송기한
- Order products : 상품번호, 상품 설명길이, 설명사진개수, 높이, 무게, ...
- Order reviews : 상품후기 정보 - 평점, 후기 제목, 후기 내용, 후기 업로드 시각, 판매자 답변 시각
- Customers : 구매자 정보- 구매자, 거주 행정구역, 도시, 우편번호
- Sellers : 판매자 정보- 판매자, 거주 행정구역, 도시
- Geolocation : 우편번호, 위도, 경도, 행정구역, 도시

## II. 데이터 탐색 및 전처리

### 데이터셋 구조



공통변수 기준으로 병합

## II. 데이터 탐색 및 전처리

### 데이터셋 구조

```
data.frame': 116581 obs. of 41 variables:
 $ order_id      : Factor w/ 97255 levels "00010242fe8c5a6d1ba2dd792cb16214",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ x            : int  31349 101822 42207 76544 102609 45867 51885 41468 79082 84845 ...
 $ seller_id     : Factor w/ 3033 levels "0015a82c2db000af6aaaf3ae2ecb0532",...: 839 2625 1099 1883 2644 1201 1343 1072 1
 $ product_category_name : Factor w/ 71 levels "agro_industria_e_comercio",...: 27 63 55 62 41 71 69 41 12 50 ...
 $ product_id    : Factor w/ 32328 levels "00066f42aeeb9f3007548bb9d3f33c38",...: 8469 29041 25185 15046 21684 30263 1785
 } ...
 $ customer_id   : Factor w/ 97398 levels "00012a2ce6f8dcda20d059ce98491703",...: 3786 3787 3788 3789 3789 3790 3791 3792
 $ customer_unique_id : Factor w/ 94087 levels "0000366f3b9a7992bf8c76cfd3221e2",...: 49702 86538 20709 64669 37121 49244 366
 '87 ...
 $ customer_zip_code_prefix : int  28013 15775 35661 12952 13226 38017 16700 11702 11075 6636 ...
 $ customer_city   : Factor w/ 4095 levels "abadia dos dourados",...: 738 3270 2650 329 4004 3928 1513 2985 3397 1914 ...
 $ customer_state  : Factor w/ 27 levels "AC","AL","AM",...: 19 26 11 26 26 11 26 26 26 26 ...
 $ order_status    : Factor w/ 7 levels "approved","canceled",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ order_purchase_timestamp : Factor w/ 96720 levels "2016-09-04 21:15:19",...: 24117 7255 46762 92524 1289 9309 40576 84780 61724 8
 $ order_approved_at : Factor w/ 88964 levels "", "2016-10-04 09:43:32",...: 23562 7042 44640 85151 1260 9187 38888 78692 5846
 $ order_delivered_carrier_date : Factor w/ 79802 levels "", "2016-10-08 10:34:01",...: 22233 7019 42787 77428 1692 8321 36766 72959 5859
 $ order_delivered_customer_date : Factor w/ 94337 levels "", "2016-10-11 13:46:32",...: 22800 7100 43929 89287 1906 8179 37348 81740 5856
 $ order_estimated_delivery_date : Factor w/ 449 levels "2016-10-20 00:00:00",...: 194 97 278 410 58 113 257 392 313 390 ...
 $ order_item_id   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ shipping_limit_date : Factor w/ 92052 levels "2016-09-19 00:15:34",...: 23618 7028 44884 88518 1600 9168 38968 80625 59124 8
 $ price           : num  58.9 239.9 199 13 199.9 ...
 $ freight_value   : num  13.3 19.9 17.9 12.8 18.1 ...
 $ payment_sequential : int  1 1 1 1 1 1 1 1 1 1 ...
 $ payment_type     : Factor w/ 4 levels "boleto","credit_card",...: 2 2 2 2 2 1 2 2 2 2 ...
 $ payment_installments : int  2 3 5 2 3 1 1 10 3 1 ...
 $ payment_value    : num  72.2 259.8 216.9 25.8 218 ...
 $ product_name_lenght : int  58 56 59 42 59 36 52 39 59 52 ...
 $ product_description_lenght : int  598 239 695 480 409 558 815 1310 493 1192 ...
```

## II. 데이터 탐색 및 전처리

### 데이터셋 구조

- NA 존재하지 않음 - 리뷰 데이터에서 리뷰 남기지 않은 경우 제외
- 문자열로 취급되는 날짜/시각 데이터에 전처리 필요
- 상품 크기, 가격 등의 데이터 순서형 변수 변환 고려 필요
- 한 주문에 여러 개의 상품이 포함된 경우의 처리 방식 고려 필요
- 브라질어 리뷰를 분석에 어떻게 반영할지 고려 필요

## II. 데이터 탐색 및 전처리

### 변수 탐색

상품, 판매자, 구매자, 배송, 금액, 후기 등의 여러가지 성질을 알려주는 변수들이 있음.

- 상품: **카테고리**, 상품명 또는 상품 설명의 길이, 상품의 무게, 길이, 높이, 넓이, 상품 사진의 개수
- 판매자/ 구매자: 주소(우편번호, **도시, 주**)
- 배송: **배송상태**, **소비자 구매시간**, 구매 확인시간, 운송장 도착시간, **구매자 수령시간**, **예정 배달시간**,

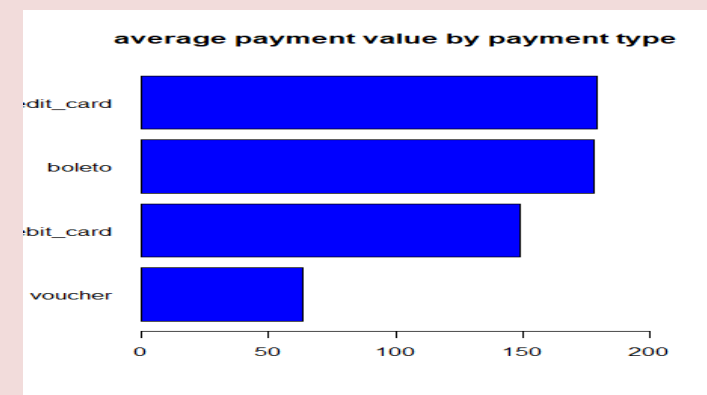
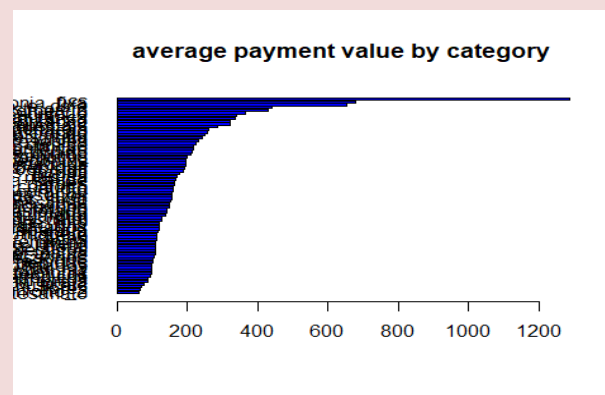
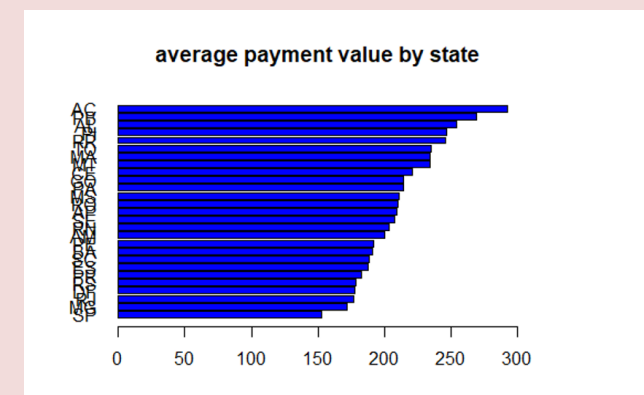
판매자 발송 완료 제한시간

- 금액: **상품 하나당 금액**, 배송료, **결제 방법**, 할부 기간, **총 결제금액**
- 후기: **접수**, 남긴 시간, 답글 시간

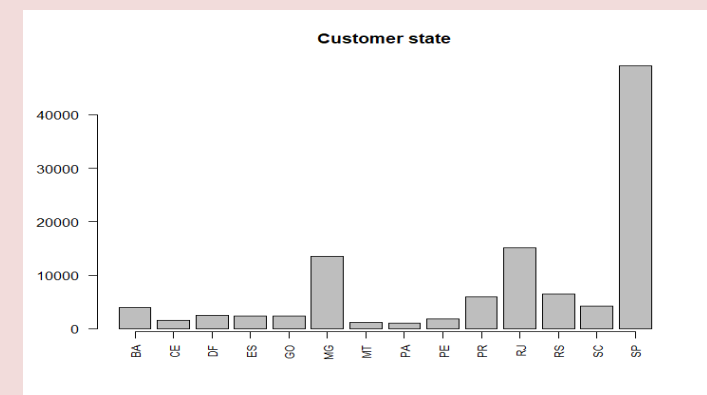
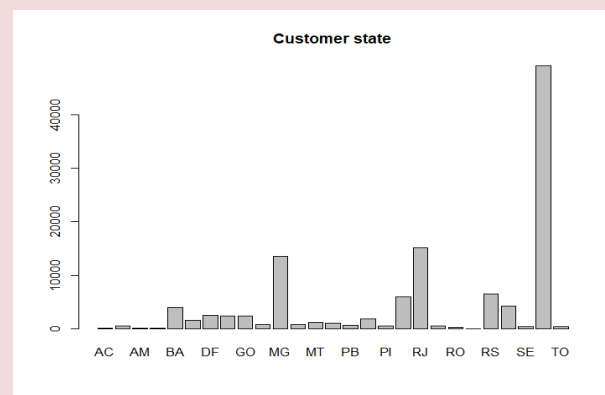


## II. 데이터 탐색 및 전처리

### 데이터 탐색

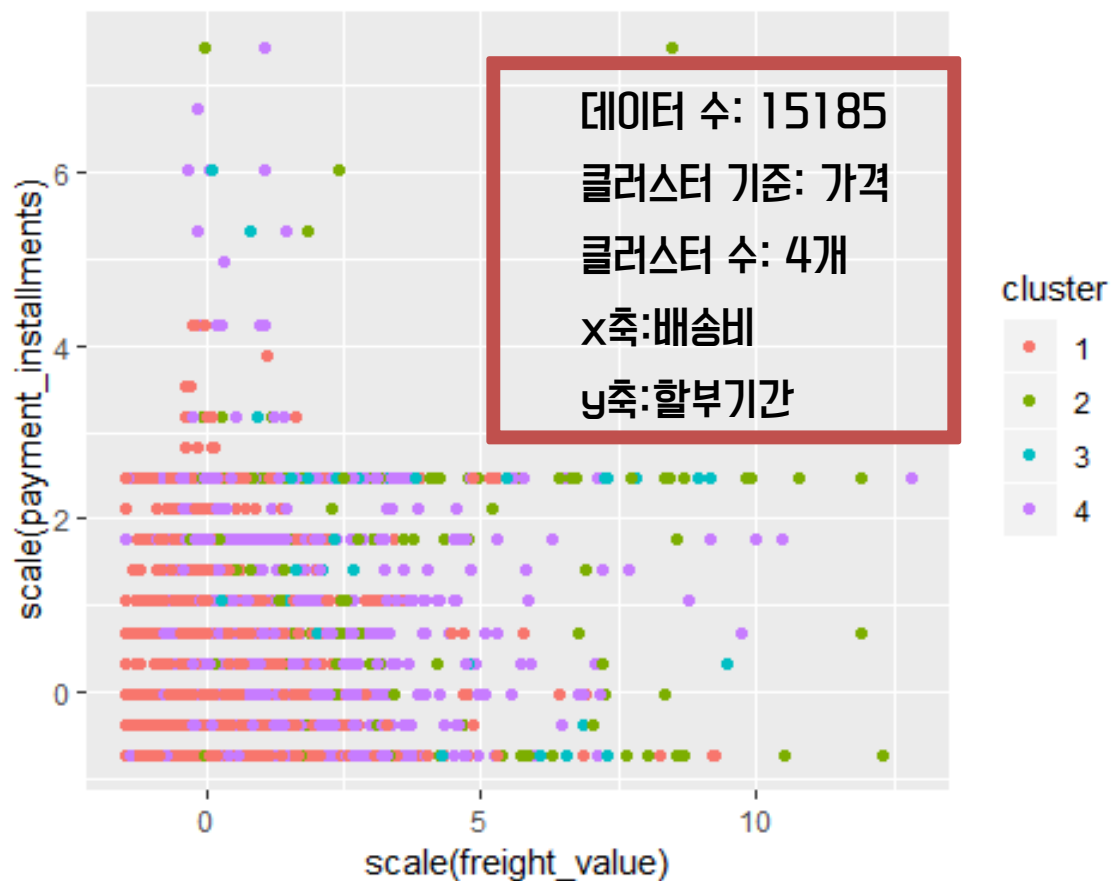


- 지역별 평균 지불 금액
- 상품 카테고리별 가격 평균
- 지불 방식에 따른 평균 지불 금액
- 고객들의 지역 빈도표
- 주문 횟수 1000회 이상 지역들의 빈도표



# III. 군집분석의 실패사례와 EDA

## 군집분석 1: RJ state



```
table(rj$cluster, rj$payment_type)
```

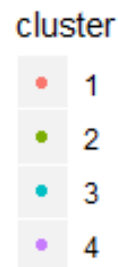
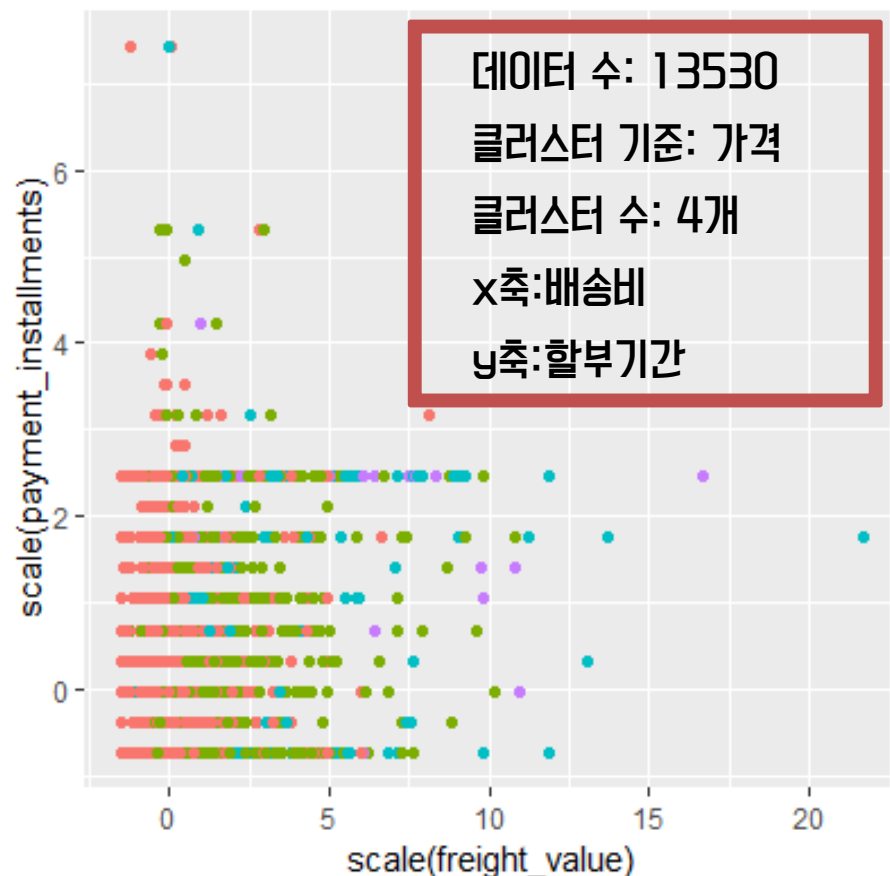
```
##      ↓
##      boleto credit_card debit_card voucher ↓
## 1  1940      8383      152      790 ↓
## 2    53       340         5       27 ↓
## 3     9        92         3        2 ↓
## 4   506      2693        46     144 ↓
```

Qplot 결과:

- 랜덤한 플롯이 나옴.
- 실제 지불수단과 클러스터 간의 table을 만들어보면 큰 관련 X

# III. 군집분석의 실패사례와 EDA

## 군집분석 2: MG state



```
table(mg$cluster,mg$payment_type)
```

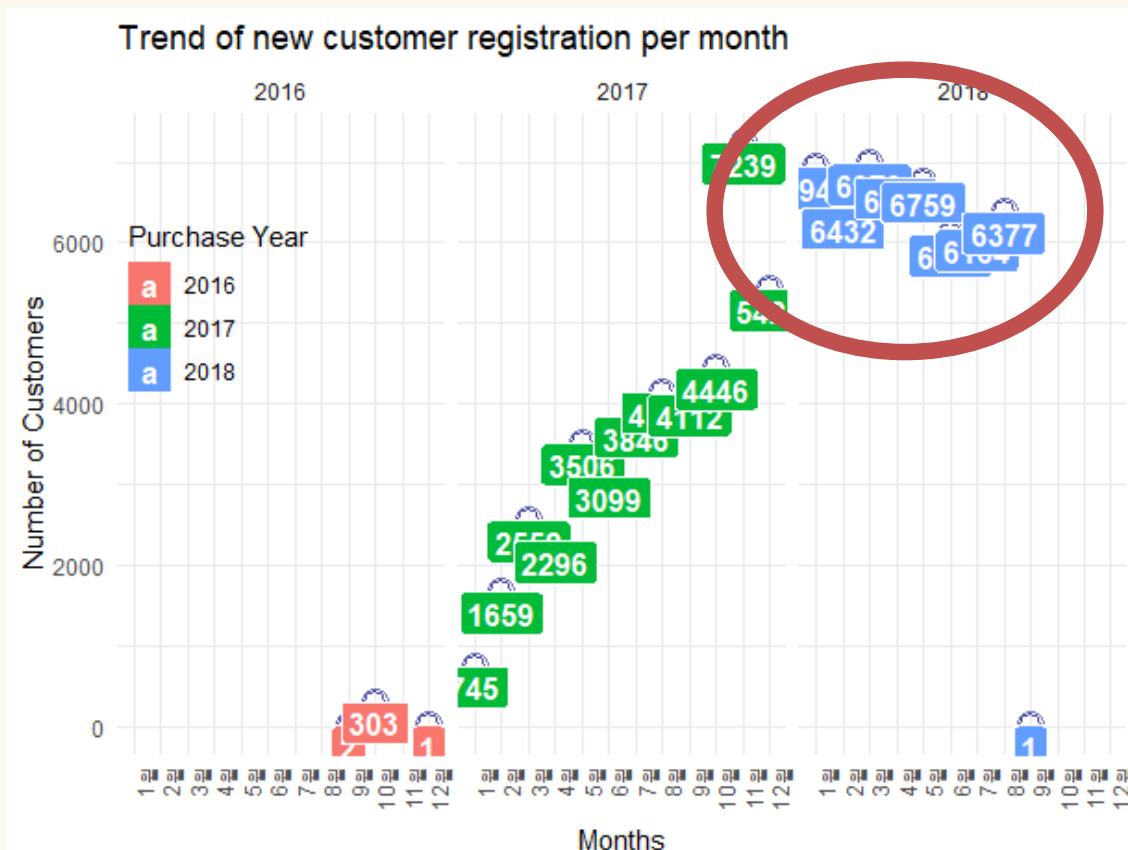
| ## |   | boleto | credit_card | debit_card | voucher |
|----|---|--------|-------------|------------|---------|
| ## | 1 | 2129   | 7475        | 119        | 518     |
| ## | 2 | 492    | 2239        | 30         | 105     |
| ## | 3 | 58     | 298         | 2          | 4       |
| ## | 4 | 8      | 51          | 1          | 1       |

Qplot 결과:

- 랜덤한 플롯
- 실제 지불수단과 클러스터 간의 table을 만들어보면 큰 관련X

### III. 군집분석의 실패사례와 EDA

#### EDA 1-1: Trend of new customer registration per month

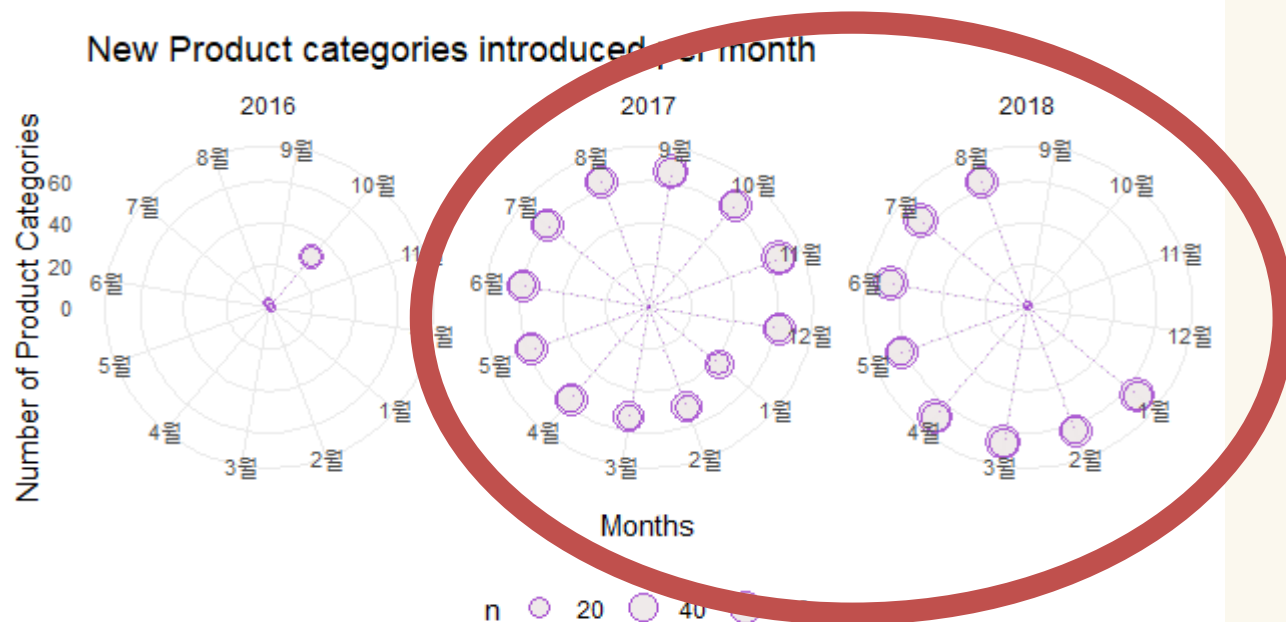


Customer의 월별 증가량 분석 결과:

- 2017년에 급속도로 늘어나기 시작
- 2018년도에는 매월 6000명 이상의 새로운 소비자가 등록 (조사 달 제외)

### III. 군집분석의 실패사례와 EDA

#### EDA 1-2: New Product categories introduced per month

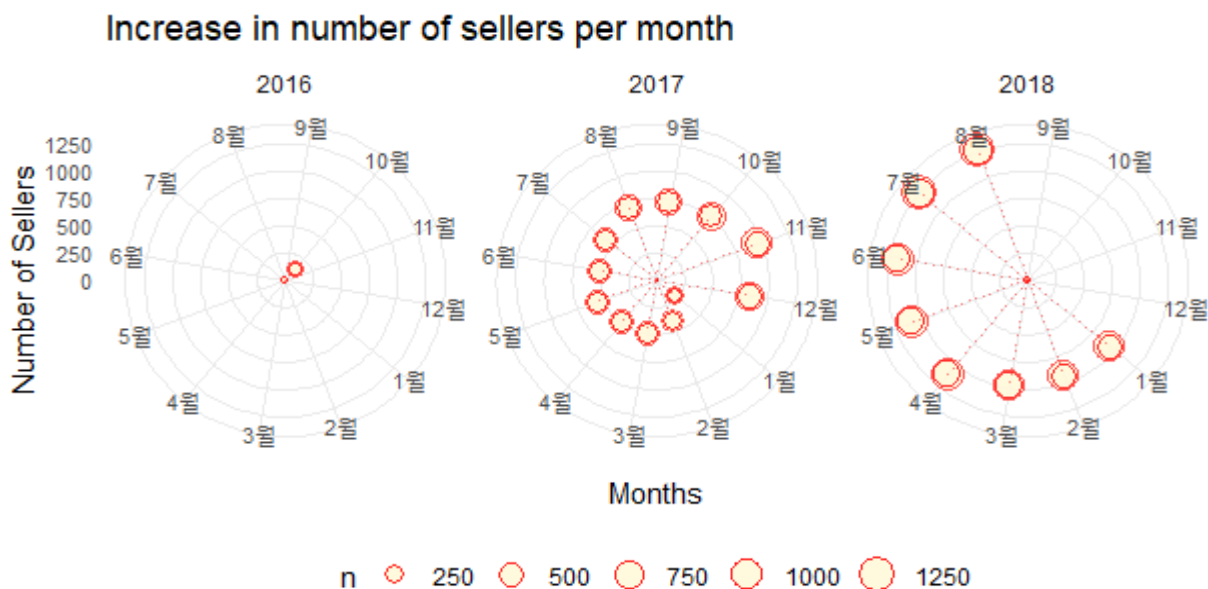


상품 카테고리 개수 증가량 분석 결과:

- 2017년부터 점진적으로 증가
- 새로운 소비자의 증가량 추이와 같은 모습을 보임.

### III. 군집분석의 실패사례와 EDA

EDA 1-3: Increase in number of sellers per month

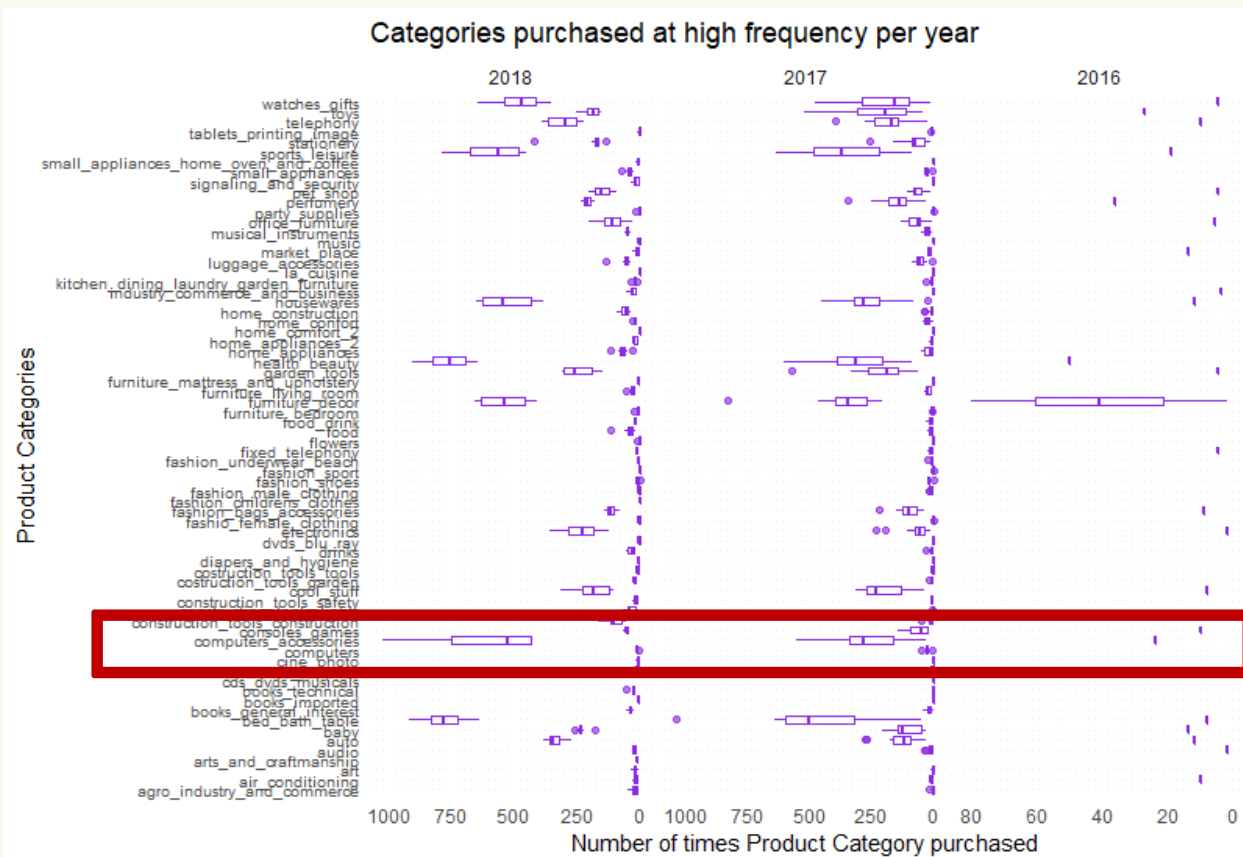


Seller id 증가량 분석 결과:  
-2017년부터 점진적으로 증가  
-처음보다 약 3000명이 넘는  
판매자가 등록

즉, 소비자, 판매자, 상품 카테고리 모두  
2017년에 점진적이고 급속도로 증가하여  
가장 최근까지 증가하는 추세를 보임을 확인할  
수 있다.

### III. 군집분석의 실패사례와 EDA

#### EDA 2: Categories purchased at high frequency per year



-연도별로 가장 자주 구매한  
카테고리를 확인 가능!

-카테고리의 수&전체적인 구매량  
2016년에 비해 많이 증가  
Ex) computer accessories  
눈에 띄는 증가

### III. 군집분석의 실패사례와 EDA

#### EDA 3: ONE order for one Product Category/Month/Year



-2017년에 많은 카테고리가 생기기 시작하면서  
각 년마다 월별로 한번만 구매하는 카테고리가  
일시적으로 증가

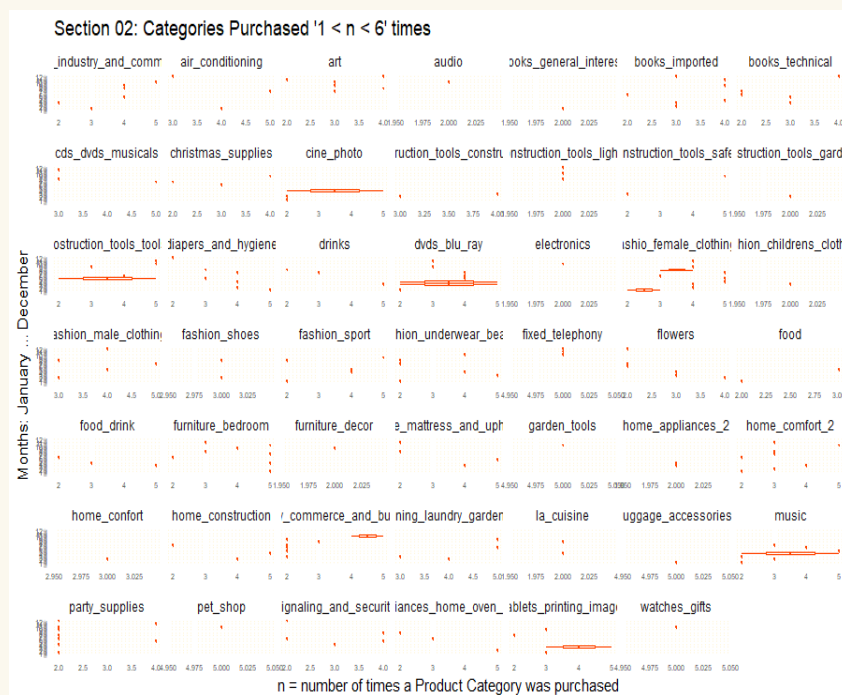
- 2018년에는 구매자 또한 많아졌으므로  
적은 구매 빈도를 가지는 카테고리가 줄어들었음



# III. 군집분석의 실패사례와 EDA

## EDA 4: Categories Purchased 'x < n < y' times

- 1 < n < 6 times



- 5 < n < 21 times



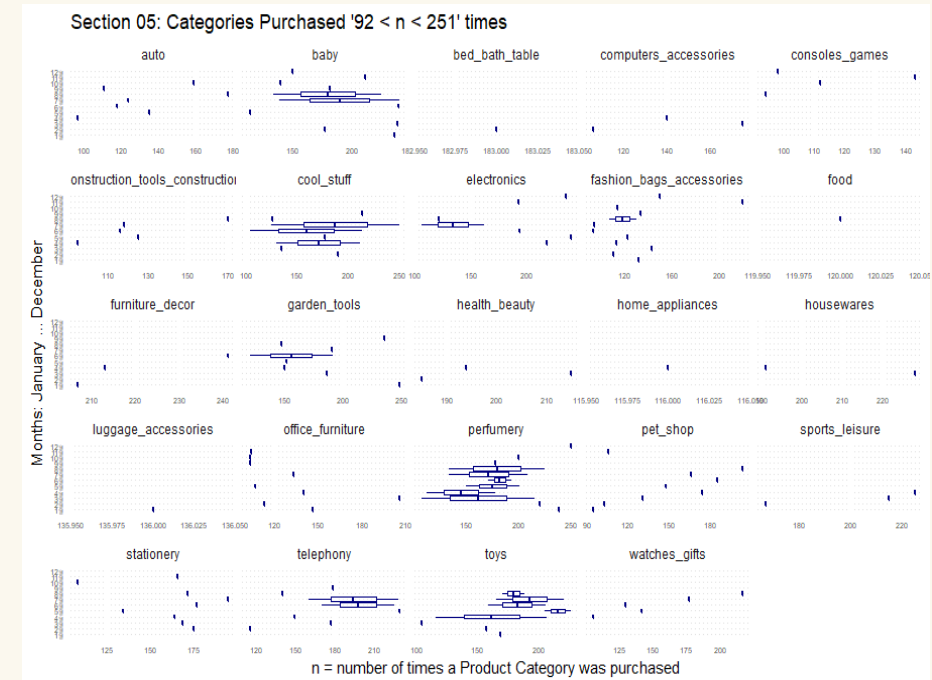
# III. 군집분석의 실패사례와 EDA

## EDA 4: Categories Purchased 'x < n < y' times

-20 < n < 93 times

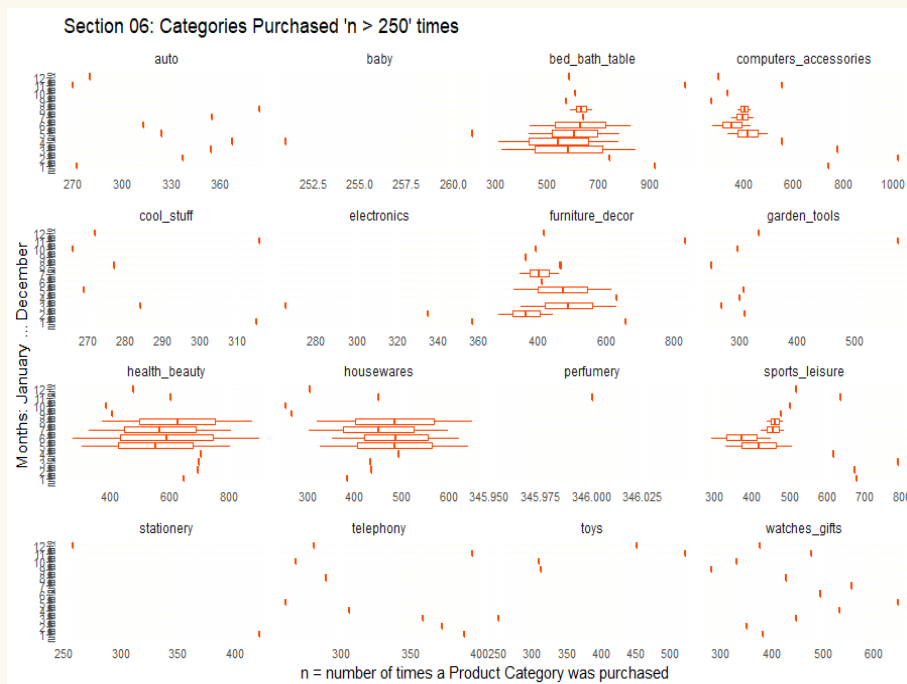


-92 < n < 251 times



# III. 군집분석의 실패사례와 EDA

## EDA 4: Categories Purchased 'x < n < y' times



- 1 < n < 6: Construction Tools, Cine Photo, DVDs Blue Ray, Fashion Female Clothing, Music
- 5 < n < 21: Air Conditioning, Construction tools Garden, Fixed Telephony, Home Appliances 2, Market Place
- 20 < n < 93: Books general interest, Console games, Home appliances, Luggage accessories, Musical instruments, Small appliances
- 92 < n < 251: Baby, Cool stuff, Garden tools, Perfumery, Toys
- 250 < n: Bed bath table, Computer accessories, Furniture décor, Health beauty, Housewares, Sports leisure

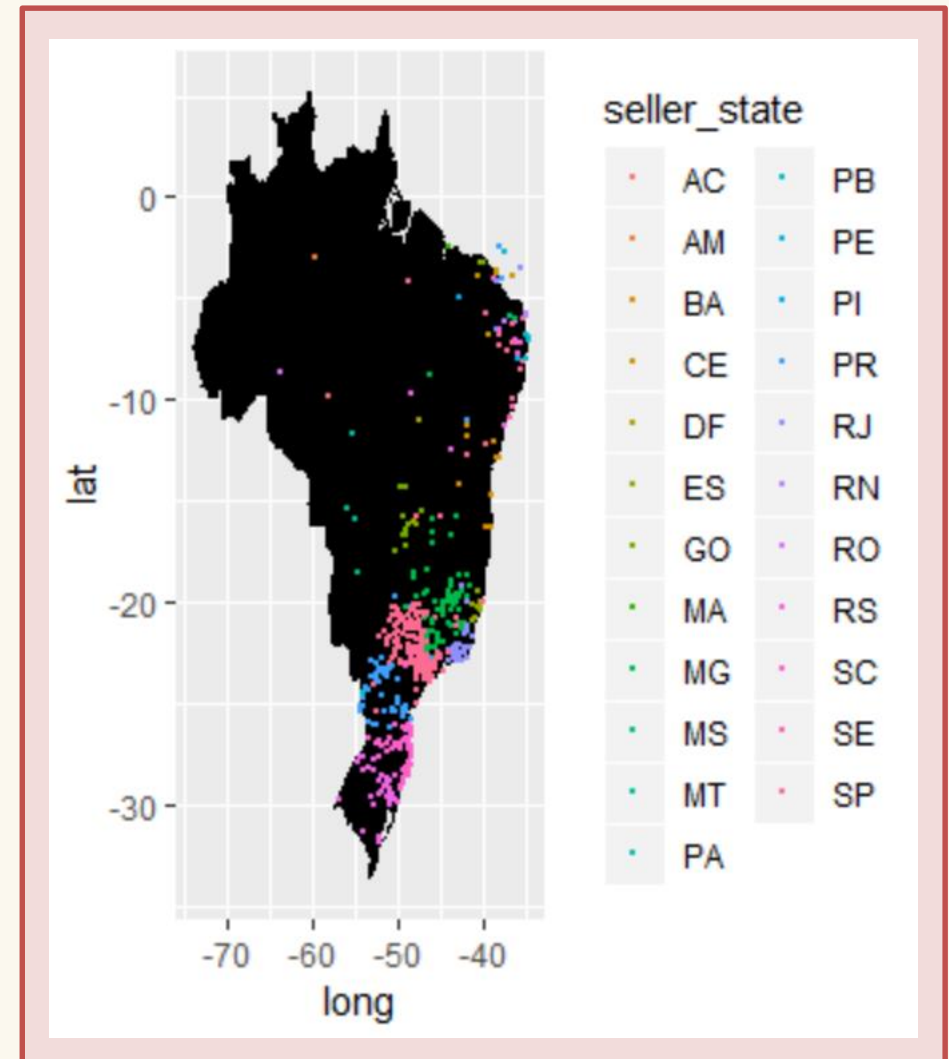
### III. 군집분석의 실패사례와 EDA

#### EDA 5: Map with Seller states

```
library(ggplot2)
Brazil<-map_data("world")%>%filter(region=="Brazil")

ggplot() +
  geom_polygon(data = Brazil, aes(x=long, y = lat, group = group), fill="black")+
  geom_point(data= complete2, aes(x=selllng,y=selllat,color=seller_state),size=0.2)
ggsave("geo3.png", plot = last_plot())
```

- Seller**들은 주로 브라질의 **남부 및 남동부** 지역에 밀집
- 지리적 특성 잘 반영
- (**남동부 지역**- 농수산업, 광업, 공업 고루 발달, 광물 자원 풍부/  
**남부 지역**-농수산업, 공업 발전)



# III. 군집분석의 실패사례와 EDA

## EDA 5: Map with Customer states

```
library(ggplot2)
Brazil<-map_data("world")%>%filter(region=="Brazil")

ggplot() +
  geom_polygon(data = Brazil, aes(x=long, y = lat, group = group), fill="black")+
  geom_point(data= complete2,aes(x=custlng,y=custlat,color=customer_state),size=0.2)
ggsave("geo4.png", plot = last_plot())
```

-**Customer**들은 주로 브라질의 **동부와 남부** 지역에 밀집

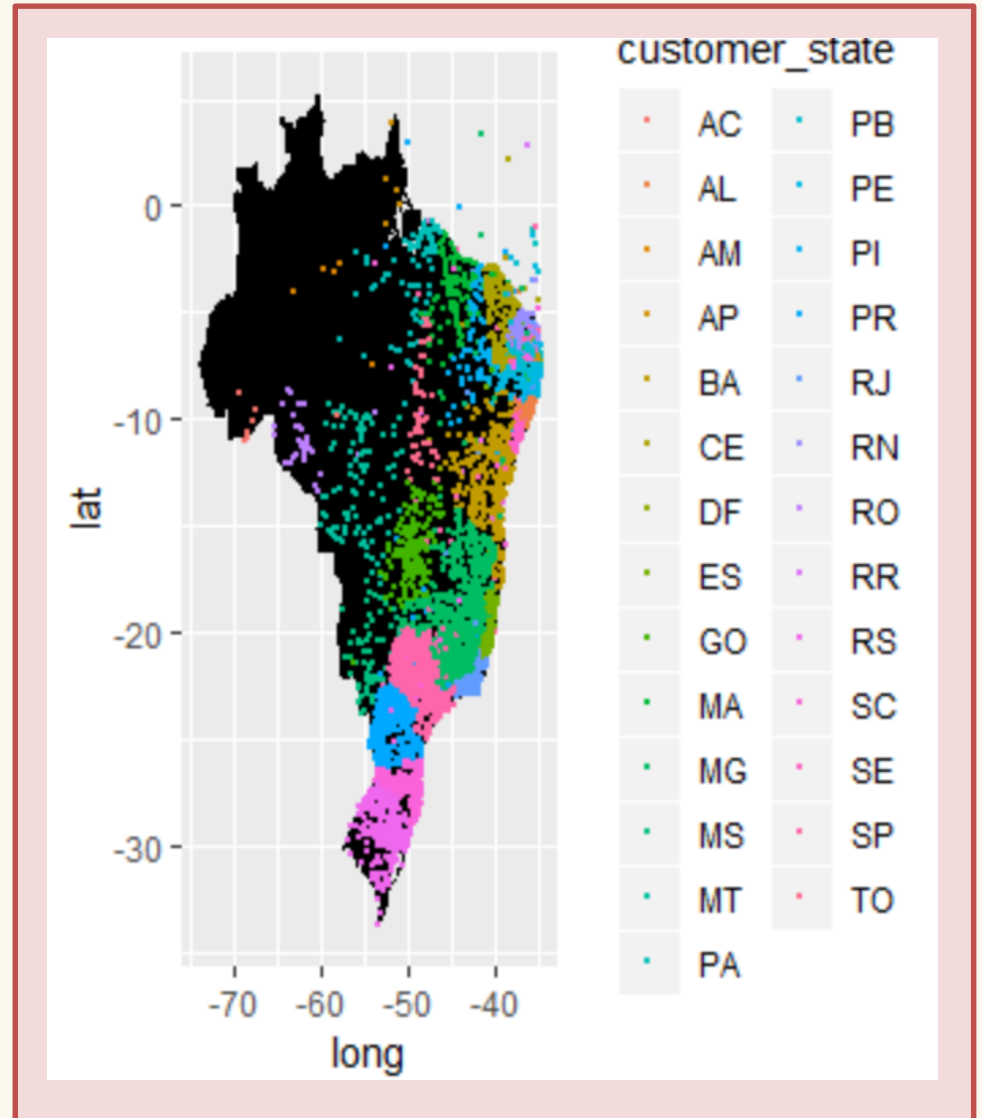
-지리적 특성 잘 반영

(**북부** 지역-열대 우림 지역으로 대부분 미개발 상태이며

사람들이 많이 거주 하고 있지 않음.

**동부** 지역-북동부에는 총 인구의 29%가 거주하며, 비옥한 토양  
지대와 내륙 지방

남부-**상파울루**는 브라질 최대 소비도시, 공업중심지)



## IV. 선호도 지수

### 1. 의도와 계획

- 71개의 **product category**를 12개의 대그룹으로 분류
- 각 소비자의 **그룹별 소비 패턴** 분석: score 만들기
- **총 가격, 빈도, 최근 구매날짜**를 바탕으로 그룹별 선호 점수를 산출

```
olist<-read.csv("olist_with_review.csv",header = TRUE)
rj<-subset(olist,customer_state=="RJ")
rj$big_category<-0
rj$big_category<-as.character(rj$big_category)
rj$product_category_name_english<-as.character(rj$product_category_name_english)
a<-c("fashion_shoes", "fashion_bags_accessories", "watches_gifts","luggage_accessories")
rj$big_category[which(rj$product_category_name_english %in% a)]<-"fashion_accessories"
```

## IV. 선호도 지수

### 1. 의도와 계획

-각 고객의 카테고리별 구매 가격 합 구하기

```
###가격 sum 구하는것
rj.fd<-subset(rj,big_category=="food")
dim(rj.fd) #970
length(unique(rj.fd$customer_id)) #933

for(i in 1:nrow(rj.fd)){
  b<-which(rj.fd$customer_id==rj.fd$customer_id[i])
  sum<-0
  for(j in 1:length(b)){
    sum<-sum+rj.fd$price[b[j]]
  }
  rj.fd$price[i]<-sum
}
```

-각 고객의 카테고리별 구매 빈도 구하기

```
###빈도(비율) 구하기

rj.fd$freq<-0
rj.fd$proportion<-0

for(i in 1:nrow(rj.fd)){
  b<-which(rj$customer_id==rj.fd$customer_id[i])
  total_freq<-length(b)
  c<-which(rj.fd$customer_id==rj.fd$customer_id[i])
  category_freq<-length(c)
  rj.fd$freq[i]<-category_freq
  rj.fd$proportion[i]<-category_freq/total_freq
}
```

## IV. 선호도 지수

### 1. 의도와 계획

- 각 고객의 카테고리별 **가장 최근 구매 날짜** 구하기

### 가장 최근 구매 날짜 구하기

```
rj.fd$timediff<-0
rj.fd$order_purchase_timestamp<-as.numeric(rj.fd$order_purchase_timestamp)
lastorder<-as.numeric(rj$order_purchase_timestamp[which.max(rj$order_purchase_timestamp)])

for(i in 1:nrow(rj.fd)){
  b<-which(rj.fd$customer_id==rj.fd$customer_id[i])
  time<-vector()
  for(j in 1:length(b)){
    time<-c(time,rj.fd$order_purchase_timestamp[b[j]])
  }
  rj.fd$order_purchase_timestamp[i]<-time[which.max(time)]
  rj.fd$timediff[i]<-lastorder-rj.fd$order_purchase_timestamp[i]
}
```



## IV. 선호도 지수

### 2. NA처리와 4분위 점수 부여

- NA 처리 후 총 금액, 빈도, 최근 구매날짜 분포의 4분위마다 각각 1점부터 4점까지 부여

```
##NA##
for(i in 1:nrow(rj.fd)){
  b<-which(rj.fd$customer_id==rj.fd$customer_id[i])
  if(length(b)>=2){
    for(m in 2:length(b)){
      rj.fd$price[b[m]]<-NA
      rj.fd$freq[b[m]]<-NA
      rj.fd$proportion[b[m]]<-NA
      rj.fd$timediff[b[m]]<-NA
    }
  }
}
```

```
###price score###
rj.fd$price_score<-0
q<-as.numeric(summary(rj.fd$price))

for(i in 1:nrow(rj.fd)){
  if(rj.fd$price[i]<=q[2]) {rj.fd$price_score[i]<-1}
}
for(i in 1:nrow(rj.fd)){
  if((q[2]<rj.fd$price[i]) & (rj.fd$price[i]<=q[3])) {rj.fd$price_score[i]<-2}
}
for(i in 1:nrow(rj.fd)){
  if((q[3]<rj.fd$price[i]) & (rj.fd$price[i]<=q[5])) {rj.fd$price_score[i]<-3}
}
for(i in 1:nrow(rj.fd)){
  if(q[5]<rj.fd$price[i]) {rj.fd$price_score[i]<-4}
}
```

- 구매 금액에 따라 스코어 부여하기

## IV. 선호도 지수

### 2. NA처리와 4분위 점수 부여

-구매 빈도에 따라 스코어 부여하기

```
###proportion score###
rj.fd$proportion_score<-0

q<-as.numeric(summary(rj.fd$proportion))

for(i in 1:nrow(rj.fd)){
  if(rj.fd$proportion[i]<=q[2]) {rj.fd$proportion_score[i]<-1}
}
for(i in 1:nrow(rj.fd)){
  if((q[2]<rj.fd$proportion[i]) & (rj.fd$proportion[i]<=q[3])) {rj.fd$proportion_score[i]<-2}
}
for(i in 1:nrow(rj.fd)){
  if((q[3]<rj.fd$proportion[i]) & (rj.fd$proportion[i]<=q[5])) {rj.fd$proportion_score[i]<-3}
}
for(i in 1:nrow(rj.fd)){
  if(q[5]<rj.fd$proportion[i]) {rj.fd$proportion_score[i]<-4}
}
```

## IV. 선호도 지수

### 2. NA처리와 4분위 점수 부여

-최근 구매 날짜에 따라 스코어 부여하기

```
###time diff score###
rj.fd<-rj.fd[complete.cases(rj.fd),]
rj.fd$time_score<-0

q<-as.numeric(summary(rj.fd$timediff))

for(i in 1:nrow(rj.fd)){
  if(rj.fd$timediff[i]<=q[2]) {rj.fd$time_score[i]<-4}
}
for(i in 1:nrow(rj.fd)){
  if((q[2]<rj.fd$timediff[i]) & (rj.fd$timediff[i]<=q[3])) {rj.fd$time_score[i]<-3}
}
for(i in 1:nrow(rj.fd)){
  if((q[3]<rj.fd$timediff[i]) & (rj.fd$timediff[i]<=q[5])) {rj.fd$time_score[i]<-2}
}
for(i in 1:nrow(rj.fd)){
  if(q[5]<rj.fd$timediff[i]) {rj.fd$time_score[i]<-1}
}
.
```

## IV. 선호도 지수

### 3. 각 score에 가중치 부여 후 최종 score 생성

- 총 금액 0.5, 빈도 0.1, 최근 구매날짜 0.4 로 가중치 부여

```
###최종합수###
preference_score<-function(name){
  b1<-which(rj.in$customer_id==name)
  if(length(b1)>=1){
    in.score<-0.5*rj.in$price_score[b1]+0.1*rj.in$proportion_score[b1]+0.4*rj.in$time_score[b1]
  } else {
    in.score<-0
  }
  .
  .
  .
  vec<-c(in.score,cu.score,ct.score,sp.score,td.score,ba.score,bt.score,ra.score,hl.score,ha.score,fr.score,fc.score)
  names(vec)<-c("in.score","cu.score","ct.score","sp.score","fd.score","ba.score","bt.score",
               "fa.score","hl.score","ha.score","fr.score","fc.score")
  return(vec)
}
```

12개 카테고리 각각에 대한 점수 산출

## IV. 선호도 지수

## 4. 소비자 개별 score 산출 결과와 해석

[illegible]

## IV. 선호도 지수

### 4. 소비자 개별 score 산출 결과와 해석

#### 1) 결과 해석

- 소비자 개인별로 한번도 구매 내역이 없으면 0점, 한번이라도 구매한 경험があれば 1점에서 4점 사이의 점수를 부여함.
- 이 소비자가 각 그룹별 선호도가 어느 정도인지 수치화해서 보여줄 수 있음.

#### 2) 특징

- 소비자 각각 한 카테고리에서만 점수가 산출됨.
- 원인: EDA에서 보여졌듯이 소비자, 판매자, 카테고리 모두 급증한지 1년 정도 밖에 되지 않았기 때문에 다양한 카테고리에서 많은 소비를 한 소비자를 찾기 힘들기 때문.
- 그러나 선호도 점수를 사용하여 그룹 마케팅을 할 때 고객 타겟팅하기 쉽다는 장점이 있음.

## IV. 선호도 지수

마케팅 활용 방안

각 고객의  
관심 카테고리  
파악

관심 카테고리별  
고객  
segmentation

고객 맞춤 마케팅  
ex.)  
관심 카테고리 관련  
상품 추천/  
광고 배너 배치

|          | pcsp1 |
|----------|-------|
| in.score | 0     |
| cu.score | 0     |
| ct.score | 0     |
| sp.score | 2.5   |
| fd.score | 0     |
| ba.score | 0     |
| bt.score | 0     |
| fa.score | 0     |
| hl.score | 0     |
| ha.score | 0     |
| fr.score | 0     |
| fc.score | 0     |

Ex. 스포츠 점수가  
가장 높은 고객



A stylized illustration of a person from the chest up, wearing a grey suit jacket, a white shirt, and a dark tie. The person's face is partially visible at the top, showing a red nose and a smiling mouth. A large, white speech bubble with a thick black outline is positioned on the left side of the image, pointing towards the person's mouth. The background is a solid light beige color.

Do you have any  
**question?**

Thank you  
for your attention

.