UNIVERSIDAD
COMPLUTENSE
MADRID

Universidad Complutense de Madrid

Master en Big Data y Business Analytics

PRACTICA MINERIA DE DATOS Y MODELIZACION PREDICTIVA, ANALISIS ELECCIONES EUROPEAS

Alumna

Ivonne V. Yáñez Mendoza

Profesora

Aída Calviño Martínez

28 de junio de 2022

Índice

1 Introducción

En este ejercicio práctico del módulo de minería de datos se busca cumplir con los siguientes objetivos:

- Construir un modelo que permita predecir los resultados para las elecciones del periodo 2023.
- Tal y como se pide en el enunciado de esta práctica, se debe construir la variable objetivo a estudiar, dependiendo de lo que se desee investigar.
- Para esta práctica en particular se ha decidido estudiar como variable objetivo el porcentaje de votos a partidos de izquierda, sobre la cantidad de votos emitidos por municipio, los cuales se incluyen en el archivo *DatosEleccionesEuropeas2019.xlsx*. Para definir la variable objetivo se sumarán los votos contenidos en las columnas PSOE y Podemos y se calculará el porcentaje en relación a la columna *votosEmitidos*. Una vez construida la variable objetivo las columnas en amarillo de la tabla excel serán descartados para la práctica.
- Una vez que se ha definido la variable objetivo y realizada la eliminación de variables no útiles, el set de datos base a estudiar de forma preliminar contendrá los siguientes datos:

Variable	Explicación
CodigoINE	Código del municipio utilizado por el INE
Porcentaje (Objetivo)	% de votos a partidos de izquierda (PSOE y Podemos) sobre los votos emitidos
CCAA	Comunidad autónoma a la que pertenece el municipio
Censo	Población con derecho a votar
Population	Población del municipio en 2016
Age_under19_Ptge	Porcentaje de ciudadanos menores de edad
Age_over65_Ptge	Porcentaje de ciudadanos con más de 65 años
WomanPopulationPtge	Porcentaje de mujeres
ForeignersPtge	Porcentaje de extranjeros
UniversityPtge	Porcentaje de ciudadanos con estudios universitarios
Empresas	Número medio de empresas por cada 1000 habitantes en el municipio
IndustriaPtge	Porcentaje de empresas del sector industrial en el municipio
ConstruccionPtge	Porcentaje de empresas del sector de la construcción en el municipio
ComercTTEHosteleriaPtge	Porcentaje de empresas dedicadas a comercio, transporte u hostelería en el municipio
ServiciosPtge	Porcentaje de empresas del sector servicios en el municipio
Densidad	Densidad de población del municipio
PobChange_pct	Porcentaje de cambio en la población (valores negativos indican que ha disminuido)
PersonasInmueble	Número medio de personas que habita un inmueble
Explotaciones	Número medio de explotaciones agrícolas por cada 1000 habitantes en el municipio
PartidoCCAA	Partido Político que gobierna en la CCAA correspondiente en el momento de las elecciones
UnemploymentPtge	Tasa de paro registrado (en porcentaje)
WomenUnemploymentPtge	Porcentaje de parados del municipio que son mujeres
UnemployLess25_Ptge	Porcentaje de parados del municipio con menos de 25 años
UnemployMore40_Ptge	Porcentaje de parados del municipio con más de 45 años
AgricultureUnemploymentPtge	Porcentaje de parados del municipio en el sector de la agricultura
IndustryUnemploymentPtge	Porcentaje de parados del municipio en el sector de la industria
ConstructionUnemploymentPtge	Porcentaje de parados del municipio en el sector de la construcción
ServicesUnemploymentPtge	Porcentaje de parados del municipio en el sector servicios
AutonomosPtge	Porcentaje de profesionales autónomos

2 Análisis exploratorio y depuración de datos

2.1 Construcción de variable objetivo, lectura de datos y tipos de variables

La construcción de la variable objetivo se realiza directamente en R utilizando el código que se adjunta a continuación, realizando el calculo en la nueva columna Porcentaje.
Una vez declarada la variable objetivo se realiza una primera aproximación al set de datos y se comprueba si las variables han sido asignadas correctamente:

```
datos <- read_excel('DatosEleccionesEuropeas2019.xlsx')

# Crea columna Porcentaje la cual es el % de votos de partidos de izquierda sobre los votos emitidos
datos$Porcentaje <- (rowSums(cbind(datos$Podemos, datos$PSOE), na.rm = T ) * 100) / datos$VotosEmitidos

# Al crear la columna esta se agrega por defecto al final del set de datos, se relocaliza
datos <- datos %>%
  relocate(Porcentaje, .before = VotosEmitidos)

# Se descartan las variables que no serán utilizadas (en excel son las variables en amarillo)
datos<-data.frame(datos[,-c(3:13)])

# Primera lectura de datos
str(datos)
```

```
'data.frame':   8110 obs. of  29 variables:
 $ CodigoINE                  : chr  "01001" "01002" "01003" "01004" ...
 $ Porcentaje                 : num  32.6 22.3 9.1 19.6 34.7 ...
 $ CCAA                       : chr  "PaísVasco" "PaísVasco" "PaísVasco" "PaísVasco" ...
 $ Censo                      : num  2016 8047 1176 1354 176 ...
 $ Population                 : num  2760 9768 1503 1710 227 ...
 $ Age_under19_Ptge           : num  27 17.6 21.7 20.8 22.5 ...
 $ Age_over65_Ptge            : num  9.85 19.73 19.37 15.97 12.71 ...
 $ WomanPopulationPtge        : num  48.3 50.3 47.5 50.4 49.6 ...
 $ ForeignersPtge             : num  7.5 5.97 2.5 4.16 8.2 ...
 $ UniversityPtge             : num  22.7 17.6 23 17 10.8 ...
 $ Empresas                   : num  60.9 63.1 33.3 63.7 70.5 ...
 $ IndustriaPtge              : num  13.1 11.85 16 6.42 NA ...
 $ ConstruccionPtge           : num  17.9 15.1 16 16.5 NA ...
 $ ComercTTEHosteleriaPtge    : num  39.3 42.9 46 48.6 NA ...
 $ ServiciosPtge              : num  29.8 30.2 22 28.4 NA ...
 $ Densidad                   : num  138.3 101.4 20.5 62.3 17.5 ...
 $ PobChange_pct              : num  6.19 2.12 1.4 0.61 8.44 9.2 -1.96 3.54 -5.47 -0.55 ...
 $ PersonasInmueble           : num  2.14 1.99 2.15 1.58 1.59 2.06 1.58 2.23 1.19 1.77 ...
 $ Explotaciones              : num  7.25 22.01 67.2 33.92 52.86 ...
 $ PartidoCCAA                : chr  "Otro" "Otro" "Otro" "Otro" ...
 $ UnemploymentPtge           : num  6.89 8.67 1.19 6.79 5.11 4.23 5.13 4.7 3.95 2.53 ...
 $ WomenUnemploymentPtge      : num  55.4 61.5 42.9 71.7 44.4 ...
 $ UnemployLess25_Ptge        : num  5.76 7.16 0 3.26 11.11 ...
 $ UnemployMore40_Ptge        : num  48.2 47.7 64.3 50 44.4 ...
 $ AgricultureUnemploymentPtge: num  2.88 3.58 0 3.26 0 ...
 $ IndustryUnemploymentPtge   : num  12.2 12.5 42.9 16.3 0 ...
 $ ConstructionUnemploymentPtge: num  3.6 5.87 14.29 3.26 0 ...
 $ ServicesUnemploymentPtge   : num  64.8 62.6 42.9 59.8 100 ...
 $ AutonomosPtge              : num  10.27 7.84 4.59 8.05 11.36 ...
```

En la lectura inicial se observa que las variables CCAA y Partido CCAA que corresponden a las columnas 3 y 20 respectivamente no están correctamente codificadas. Se modifican y se reconvierten a factores con el siguiente código:

```
# Indicar numero de columna a convertir
datos[,c(3, 20)] <- lapply(datos[,c(3, 20)], as.factor)
```

Se verifica ademas que todas las variables numéricas tomen mas de 10 valores diferentes.

```
                porcentaje                      Censo
                      6324                       3287
                Population            Age_under19_Ptge
                      3489                       6002
           Age_over65_Ptge          WomanPopulationPtge
                      6773                       4519
            ForeignersPtge              UniversityPtge
                      5077                       1997
                 Empresas               IndustriaPtge
                      4241                       1328
          ConstruccionPtge      ComercTTEHosteleriaPtge
                      1452                       1796
             ServiciosPtge                   Densidad
                      1858                       4642
              PobChange_pct            PersonasInmueble
                      3048                        283
              Explotaciones            UnemploymentPtge
                      6003                       1443
       WomenUnemploymentPtge         UnemployLess25_Ptge
                      1785                       1174
         UnemployMore40_Ptge  AgricultureUnemploymentPtge
                      1914                       1540
    IndustryUnemploymentPtge ConstructionUnemploymentPtge
                      1622                       1322
    ServicesUnemploymentPtge               AutonomosPtge
                      2129                       1939
```

Están todas bien representadas, no se realizan cambios.

2.2 Análisis exploratorio y corrección de errores

Análisis descriptivo básico

```
summary(datos)
```

```
  CodigoINE           porcentaje               CCAA          Censo        
 Length:8110        Min.   : 0.00   CastillaLeón  :2248   Min.   :      3  
 Class :character   1st Qu.:27.40   Cataluña      : 947   1st Qu.:    136  
 Mode  :character   Median :40.00   CastillaMancha: 919   Median :    440  
                    Mean   :38.73   Andalucía     : 773   Mean   :   4331  
                    3rd Qu.:50.02   Aragón        : 731   3rd Qu.:   1855  
                    Max.   :94.12   ComValenciana : 542   Max.   :2391391  
                                    (Other)       :1950                    
   Population      Age_under19_Ptge Age_over65_Ptge WomanPopulationPtge
 Min.   :      3   Min.   : 0.000   Min.   : 0.00   Min.   :11.77      
 1st Qu.:    151   1st Qu.: 8.333   1st Qu.:19.82   1st Qu.:45.72      
 Median :    509   Median :13.879   Median :27.57   Median :48.48      
 Mean   :   5396   Mean   :13.562   Mean   :29.08   Mean   :47.30      
 3rd Qu.:   2290   3rd Qu.:19.059   3rd Qu.:36.91   3rd Qu.:50.00      
```

```
Max.   :2908032   Max.   :33.696   Max.   :76.47   Max.   :72.68

ForeignersPtge   UniversityPtge     Empresas        IndustriaPtge
Min.   : 0.000   Min.   : 0.00   Min.   :    0.00   Min.   : 1.050
1st Qu.: 1.621   1st Qu.: 5.56   1st Qu.:   37.50   1st Qu.: 5.750
Median : 4.144   Median : 9.09   Median :   56.41   Median : 8.450
Mean   : 6.380   Mean   :10.33   Mean   :   76.85   Mean   : 9.518
3rd Qu.: 8.824   3rd Qu.:13.87   3rd Qu.:   74.73   3rd Qu.:11.940
Max.   :71.468   Max.   :52.39   Max.   :44016.53   Max.   :44.680
                 NA's   :6                          NA's   :5109
ConstruccionPtge ComercTTEHosteleriaPtge ServiciosPtge      Densidad
Min.   : 4.32    Min.   :19.64           Min.   : 6.41   Min.   :    0.230
1st Qu.:12.20    1st Qu.:38.55           1st Qu.:23.98   1st Qu.:    4.642
Median :15.24    Median :43.73           Median :30.10   Median :   12.975
Mean   :16.02    Mean   :44.19           Mean   :30.93   Mean   :  171.396
3rd Qu.:19.23    3rd Qu.:49.25           3rd Qu.:36.95   3rd Qu.:   52.375
Max.   :40.80    Max.   :81.48           Max.   :71.67   Max.   :22366.670
NA's   :5060     NA's   :4931            NA's   :4984
PobChange_pct     PersonasInmueble Explotaciones     PartidoCCAA
Min.   :-52.270   Min.   :0.110    Min.   :   0.00   Otro:1652
1st Qu.:-10.405   1st Qu.:0.850    1st Qu.:  46.74   PP  :4121
Median : -4.970   Median :1.250    Median : 125.00   PSOE:2337
Mean   : -4.903   Mean   :1.295    Mean   : 160.65
3rd Qu.:  0.070   3rd Qu.:1.730    3rd Qu.: 229.48
Max.   :138.460   Max.   :3.330    Max.   :5333.33
NA's   :7         NA's   :5
UnemploymentPtge  WomenUnemploymentPtge UnemployLess25_Ptge
Min.   :  0.000   Min.   :  0.00        Min.   :  0.000
1st Qu.:  3.300   1st Qu.: 45.00        1st Qu.:  0.000
Median :  5.500   Median : 56.45        Median :  6.200
Mean   :  5.926   Mean   : 51.85        Mean   :  7.187
3rd Qu.:  8.030   3rd Qu.: 63.77        3rd Qu.: 10.000
Max.   :304.120   Max.   :100.00        Max.   :100.000
NA's   :1         NA's   :1             NA's   :1
UnemployMore40_Ptge AgricultureUnemploymentPtge IndustryUnemploymentPtge
Min.   :  0.00      Min.   :  0.000             Min.   :  0.000
1st Qu.: 45.45      1st Qu.:  0.000             1st Qu.:  0.000
Median : 53.85      Median :  3.030             Median :  6.340
Mean   : 53.51      Mean   :  8.102             Mean   :  9.354
3rd Qu.: 64.52      3rd Qu.: 11.110             3rd Qu.: 13.510
Max.   :100.00      Max.   :100.000             Max.   :100.000
NA's   :1           NA's   :1                   NA's   :1
ConstructionUnemploymentPtge ServicesUnemploymentPtge AutonomosPtge
Min.   :  0.000              Min.   :  0.00           Min.   :  0.00
1st Qu.:  0.000              1st Qu.: 53.05           1st Qu.:  7.98
Median :  5.710              Median : 66.67           Median : 10.53
Mean   :  8.074              Mean   : 62.59           Mean   : 11.46
3rd Qu.: 10.290              3rd Qu.: 76.82           3rd Qu.: 14.03
Max.   :100.000              Max.   :100.00           Max.   :359.79
NA's   :1                    NA's   :1                NA's   :14
```

Se observa que las variables ***IndustriaPtge, ConstruccionPtge, ComercTTEHosteleriaPtge y ServiciosPtge*** contienen un numero significativo de datos ausentes. Al explorar los registros la cantidad de datos NA supera el 50% por lo que han sido eliminados.

```
nrow(datos)
[1] 8110
```

```
datos$IndustriaPtge <- NULL
datos$ConstruccionPtge <- NULL
datos$ComercTTEHosteleriaPtge <- NULL
datos$ServiciosPtge <- NULL
```

Se revisan las frecuencias de las variables categóricas encontrándose lo siguiente:

Variable *PartidoCCAA*

```
freq(datos$PartidoCCAA)
        n    % val%
Otro 1652 20.4 20.4
PP   4121 50.8 50.8
PSOE 2337 28.8 28.8
```

Variable *CCAA*

```
freq(datos$CCAA)
                   n    % val%
Andalucía        773  9.5  9.5
Aragón           731  9.0  9.0
Asturias          78  1.0  1.0
Baleares          67  0.8  0.8
Canarias          88  1.1  1.1
Cantabria        102  1.3  1.3
CastillaLeón    2248 27.7 27.7
CastillaMancha   919 11.3 11.3
```

```
    Cataluña        947 11.7 11.7
    Ceuta             1  0.0  0.0
    ComValenciana   542  6.7  6.7
    Extremadura     387  4.8  4.8
    Galicia         313  3.9  3.9
    Madrid          179  2.2  2.2
    Melilla           1  0.0  0.0
    Murcia           45  0.6  0.6
    Navarra         272  3.4  3.4
    PaísVasco       243  3.0  3.0
    Rioja           174  2.1  2.1
```

Utilizando el código visto en clase, se dividen las comunidades autónomas en zonas, apelando a criterios geográficos:

```
datos$CCAA <- car::recode(datos$CCAA, "c('Galicia', 'Asturias', 'Cantabria',
'PaísVasco', 'Rioja', 'Navarra') = 'Zona1'; c('Aragón', 'Cataluña') = 'Zona2';
c('ComValenciana', 'Baleares') = 'Zona3'; c('CastillaLeón', 'Extremadura') = 'Zona4';
c('CastillaMancha', 'Madrid') = 'Zona5';
c('Andalucía', 'Murcia', 'Canarias', 'Ceuta', 'Melilla') = 'Zona6'")
```

Se realiza una comprobación del nuevo reparto de frecuencias de las CCAA

```
    freq(datos$CCAA)
             n    % val%
    Zona1 1182 14.6 14.6
    Zona2 1678 20.7 20.7
    Zona3  609  7.5  7.5
    Zona4 2635 32.5 32.5
    Zona5 1098 13.5 13.5
    Zona6  908 11.2 11.2
```

Las variables *UnemploymentPtge* y *AutonomosPtge* tienen valores fuera del intervalo 0-100, se recodifican:

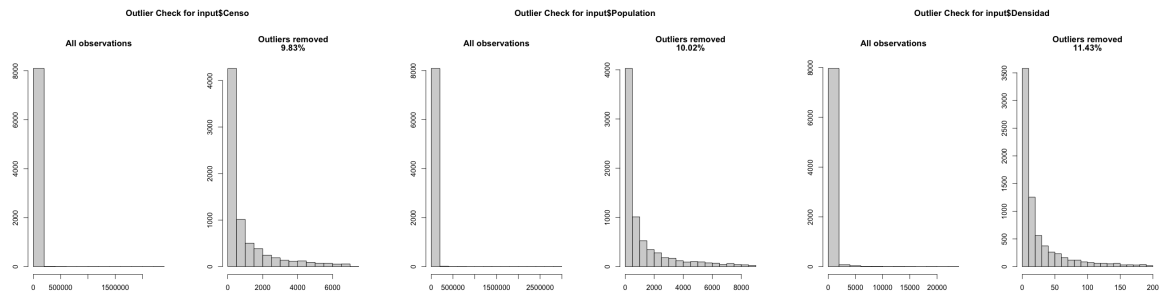
```
# Valores fuera de rango de las variables UnemploymentPtge y AutonomosPtge

datos$UnemploymentPtge <- replace(datos$UnemploymentPtge, which(datos$UnemploymentPtge > 100), NA)
datos$AutonomosPtge <- replace(datos$AutonomosPtge, which(datos$AutonomosPtge > 100), NA)
```

2.3 Detección de datos atípicos

El set de datos se ha generado con las correcciones correspondientes. A continuación se separa la variable objetivo *porcentaje* del resto de variables input. En el caso de la columna *CodigoINE* se indica que será utilizada como variable identificadora.

```
varObjCont<-datos$Porcentaje
input<-data.frame(datos[,-c(1:2)])
row.names(input) <- datos$CodigoINE
```

Haciendo uso del código entregado en clase, se realiza la detección de outliers: (Por espacio solo se muestran los gráficos de las variables que presentan un % considerable de atípicos (Censo, Population, Densidad), sin embargo estos **no han sido reemplazados** por valores NA debido a su porcentaje de importancia sobre un 5%)



A modo de resumen, se adjunta tabla con variables cuyo % de outliers si han sido reemplazadas por valores NA, estableciéndose como punto de corte un 5%:

Cuadro 2: Outliers removed

Variable	Outliers removed
WomanPopulationPtge	1.33%
ForeignersPtge	1.54%
UniversityPtge	0.31%
Empresas	0.79%
PobChange_pct	0.84%
Explotaciones	0.54%
UnemploymentPtge	0.09%
UnemployLess25_Ptge	1.01%
AgricultureUnemploymentPtge	2.2%
IndustryUnemploymentPtge	0.89%
ConstructionUnemploymentPtge	2.4%
AutonomosPtge	0.33%

2.4 Tratamiento de datos faltantes

Se realiza en primer lugar un análisis de la proporción del número de ausentes del set de datos, sin encontrar observaciones que contengan mas de 50% de datos ausentes. Observando la mediana en 0 la mitad de las observaciones están completas.

```
summary(input$prop_missings)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.000000 0.000000 0.005554 0.000000 0.434783
```

Se verifica por cada variable el número de ausentes.

```
(prop_missingsVars<-colMeans(is.na(input)))
                        CCAA                       Censo                  Population
                  0.0000000000                0.0000000000                0.0000000000
           Age_under19_Ptge             Age_over65_Ptge         WomanPopulationPtge
                  0.0000000000                0.0000000000                0.0133168927
             ForeignersPtge              UniversityPtge                   Empresas
                  0.0154130703                0.0038224414                0.0078914920
                  Densidad                PobChange_pct            PersonasInmueble
                  0.0000000000                0.0092478422                0.0006165228
              Explotaciones                 PartidoCCAA            UnemploymentPtge
                  0.0054254007                0.0000000000                0.0009864365
      WomenUnemploymentPtge         UnemployLess25_Ptge         UnemployMore40_Ptge
                  0.0001233046                0.0102342787                0.0001233046
AgricultureUnemploymentPtge     IndustryUnemploymentPtge ConstructionUnemploymentPtge
                  0.0220715166                0.0090012330                0.0241676942
   ServicesUnemploymentPtge               AutonomosPtge               prop_missings
                  0.0001233046                0.0051787916                0.0000000000
```

2.5 Eliminación

No se observan variables cuya % de ausentes supere el 50%, las que contenían % mayores fueron descartadas anteriormente. No se observan variables cuyo % de ausentes supere el 5% de los datos.

2.6 Imputación

Se lleva a cabo la imputación de los datos faltantes, diferenciando entre variables cualitativas y cuantitativas.

Se ha creado una variable input de nombre prop_missings y se verifica si cumple con las especificaciones (cantidad de valores diferentes) entregando el siguiente resultado:

```
length(unique(input$prop_missings))
[1] 6
```

Al tener un número bajo de observaciones, se decide convertir esta variable input en factor y después se comprueba que los niveles estén bien representados, en este caso se utiliza la función recode para agruparlos de forma que tengan una representación adecuada.

```
input$prop_missings<-as.factor(input$prop_missings)
freq(input$prop_missings)
                       n    % val%
0                   7215 89.0 89.0
0.0434782608695652   785  9.7  9.7
0.0869565217391304    91  1.1  1.1
0.130434782608696     13  0.2  0.2
0.173913043478261      5  0.1  0.1
0.434782608695652      1  0.0  0.0

input$prop_missings<-car::recode(input$prop_missings, "0='Ninguno';else='Alguno'")
freq(input$prop_missings)
           n  % val%
Alguno   895 11   11
Ninguno 7215 89   89
```

Como paso final se comprueba con summary el trabajo realizado anteriormente con la depuracion de los datos, el tratamiento de los ausentes y las re-codificaciones a categorias con baja representación:

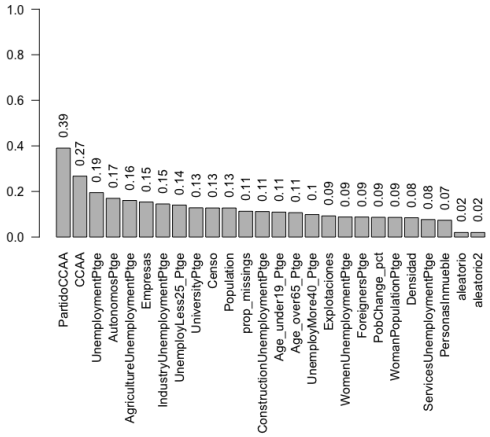
```
summary(input)
    CCAA          Censo           Population      Age_under19_Ptge Age_over65_Ptge
 Zona1:1182   Min.   :      3   Min.   :      3   Min.   : 0.000   Min.   : 0.00
 Zona2:1678   1st Qu.:    136   1st Qu.:    151   1st Qu.: 8.333   1st Qu.:19.82
 Zona3: 609   Median :    440   Median :    509   Median :13.879   Median :27.57
 Zona4:2635   Mean   :   4331   Mean   :   5396   Mean   :13.562   Mean   :29.08
 Zona5:1098   3rd Qu.:   1855   3rd Qu.:   2290   3rd Qu.:19.059   3rd Qu.:36.91
 Zona6: 908   Max.   :2391391   Max.   :2908032   Max.   :33.696   Max.   :76.47
 WomanPopulationPtge ForeignersPtge   UniversityPtge     Empresas         Densidad
 Min.   :32.91       Min.   : 0.000   Min.   : 0.00   Min.   :  0.00   Min.   :    0.230
 1st Qu.:45.89       1st Qu.: 1.598   1st Qu.: 5.56   1st Qu.: 37.28   1st Qu.:    4.642
 Median :48.53       Median : 4.040   Median : 9.09   Median : 56.05   Median :   12.975
 Mean   :47.51       Mean   : 5.846   Mean   :10.23   Mean   : 54.61   Mean   :  171.396
 3rd Qu.:50.00       3rd Qu.: 8.521   3rd Qu.:13.83   3rd Qu.: 74.08   3rd Qu.:   52.375
 Max.   :62.50       Max.   :30.416   Max.   :38.50   Max.   :180.85   Max.   :22366.670
 PobChange_pct     PersonasInmueble Explotaciones    PartidoCCAA UnemploymentPtge
 Min.   :-40.980   Min.   :0.110    Min.   :  0.00   Otro:1652   Min.   : 0.000
 1st Qu.:-10.447   1st Qu.:0.850    1st Qu.: 46.48   PP  :4121   1st Qu.: 3.300
 Median : -5.040   Median :1.250    Median :124.03   PSOE:2337   Median : 5.490
 Mean   : -5.263   Mean   :1.295    Mean   :155.03               Mean   : 5.874
 3rd Qu.:  0.000   3rd Qu.:1.730    3rd Qu.:227.52               3rd Qu.: 8.020
 Max.   : 31.190   Max.   :3.330    Max.   :775.00               Max.   :22.220
 WomenUnemploymentPtge UnemployLess25_Ptge UnemployMore40_Ptge AgricultureUnemploymentPtge
 Min.   :  0.00        Min.   : 0.000      Min.   :  0.00      Min.   : 0.000
 1st Qu.: 45.00        1st Qu.: 0.000      1st Qu.: 45.45      1st Qu.: 0.000
 Median : 56.45        Median : 6.100      Median : 53.85      Median : 2.830
 Mean   : 51.85        Mean   : 6.567      Mean   : 53.51      Mean   : 6.754
 3rd Qu.: 63.77        3rd Qu.: 9.930      3rd Qu.: 64.52      3rd Qu.:10.238
 Max.   :100.00        Max.   :40.000      Max.   :100.00      Max.   :44.440
 IndustryUnemploymentPtge ConstructionUnemploymentPtge ServicesUnemploymentPtge
 Min.   : 0.000           Min.   : 0.000               Min.   :  0.00
 1st Qu.: 0.000           1st Qu.: 0.000               1st Qu.: 53.05
 Median : 6.250           Median : 5.460               Median : 66.67
 Mean   : 8.643           Mean   : 6.711               Mean   : 62.59
 3rd Qu.:13.265           3rd Qu.:10.000               3rd Qu.: 76.82
 Max.   :50.000           Max.   :40.700               Max.   :100.00
 AutonomosPtge   prop_missings
 Min.   : 0.00   Alguno : 895
 1st Qu.: 7.98   Ninguno:7215
 Median :10.51
 Mean   :11.32
 3rd Qu.:13.95
 Max.   :32.20
```

3 Análisis de las relaciones entre variables

3.1 Detección de relaciones entre inputs y variable objetivo

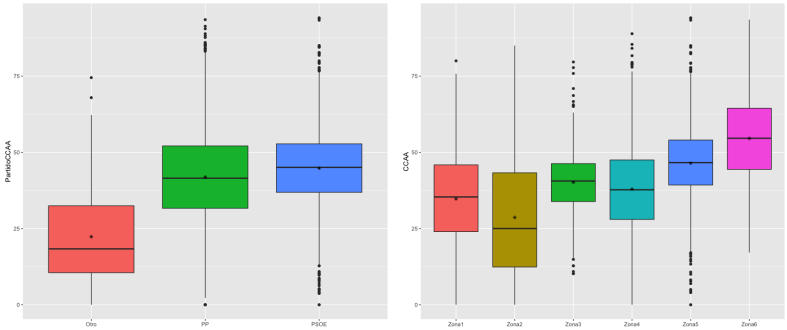
En el siguiente apartado se evaluarán las relaciones entre las variables input y la variable objetivo *Porcentaje*

En primer lugar se revisa el efecto de las variables más importantes sobre la variable objetivo utilizando el gráfico de la V de cramer. No se observa ninguna variable que esté por debajo de las variables *aleatorio* y *aleatorio2*.

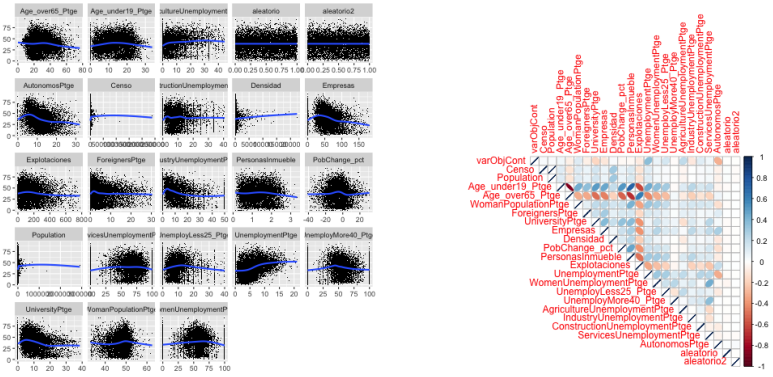


Las variables mas importantes según el gráfico son: *PartidoCCAA*, *CCAA*, *UnemployementPtge*, *AutonomosPtge*.

Utilizando un gráfico de caja y bigote se revisa el efecto de los dos factores mas importantes, PartidoCCAA y CCAA:



En el primer gráfico para el caso de *PartidoCCAA* se observa que las medias entre PP y PSOE son bastante parejas con una mínima diferencia a favor del PSOE (caja a mayor altura) y un aumento de % votos en caso que ese partido gobierne al compararlo con el PP o con Otro. En el segundo gráfico, se observa que la variable *CCAA* si tiene influencia sobre el *Porcentaje*, (en la zona 6 dan lugar a mayor % de votos).



En el caso del gráfico de dispersión no se observa que alguna de las variables input tengan una

influencia significativa sobre el porcentaje, esto al no evidenciarse una relación lineal. Mismo caso para el gráfico del coeficiente de relación, no se evidencia una relación directa entre la variable objetivo y alguna de las inputs.

4 Regresión lineal

4.1 Construcción de modelos de regresión lineal

Una vez se han depurado los datos y realizada la correspondiente partición *train-test* se crean los siguientes modelos de regresión lineal. *Nota:* Al final de este apartado se adjunta una tabla resumen con los valores r2 train y r2 test para cada modelo probado.

4.1.1 Modelo 1

Este modelo ha sido construido incluyendo todas las variables input

```
modelo1<-lm(varObjCont~.,data=data_train)
```

4.1.2 Modelo 2

Segundo modelo eliminando variables con escasa relevancia predictiva

```
modelo2<-lm(varObjCont~PartidoCCAA + CCAA + AutonomosPtge + UnemploymentPtge +
              AgricultureUnemploymentPtge + Age_over65_Ptge + Densidad + UniversityPtge +
              ForeignersPtge + Explotaciones + PersonasInmueble + WomanPopulationPtge, data =
              data_train)
```

4.1.3 Modelo 3

El modelo anterior contiene variables que podrían no ser decisivas al momento de realizar una predicción por lo que se prueba a quitarlas en este modelo.

```
modelo3<-lm(varObjCont~PartidoCCAA + CCAA + AutonomosPtge + UnemploymentPtge +
              Age_over65_Ptge + AgricultureUnemploymentPtge + Explotaciones,
            data = data_train)
```

4.1.4 Modelo 4

Teniendo en cuenta los resultados de los modelos anteriores se decide probar con un enfoque mas acotado con solo 4 variables:

```
modelo4<-lm(varObjCont~PartidoCCAA + CCAA + AutonomosPtge + UnemploymentPtge, data = data_train)
```

4.1.5 Modelo 5

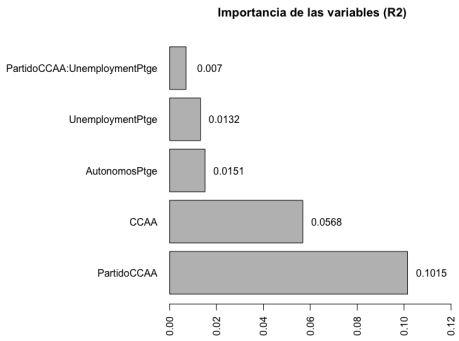
En este modelo se ha decidido probar con una interacción entre *PartidoCCAA* y el porcentaje de desempleados para ver si existe alguna incidencia o efecto.

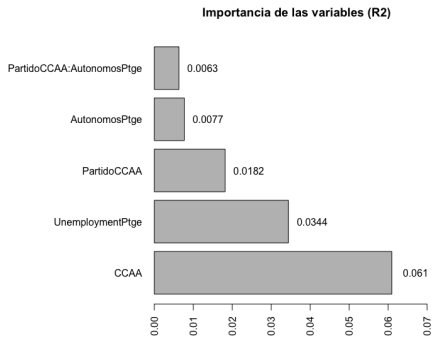
```
modelo5<-lm(varObjCont~PartidoCCAA + CCAA + AutonomosPtge + UnemploymentPtge +
              PartidoCCAA:UnemploymentPtge, data = data_train)
```

4.1.6 Modelo 6

En esta segunda interacción se ha querido estudiar si existe algún efecto en la interacción PartidoCCAA y AutonomosPtge observándose lo siguiente:

```
modelo6 <- lm(varObjCont~PartidoCCAA + CCAA + AutonomosPtge + UnemploymentPtge+
              PartidoCCAA:AutonomosPtge, data = data_train)
```





Notas: En cuanto al gráfico, no se observa que las interacciones creadas tanto en el modelo 5 y 6 tengan algún impacto significativo, de hecho es el efecto menos importante en ambos modelos.

4.1.7 Evaluación de modelos

Como apartado final y para determinar cual de los modelos es preferible se comparan la cantidad de variables de cada modelo más los valores de r^2 para train y test.

Cuadro 3: Comparación de modelos de regresión lineal

Modelo	Rank	r2 Train	r2 Test
Modelo1	32	0.4783307	0.5037294
Modelo2	18	0.4728199	0.4951773
Modelo3	13	0.4619070	0.4883348
Modelo4	10	0.4438234	0.4718783
Modelo5	12	0.4509770	0.4787307
Modelo6	12	0.4498981	0.4785064

Por principio de parsimonia el modelo 4 es el más sencillo al tener 10 variables y aunque tiene un r^2 un poco más pequeño en prueba y test en relación a los modelos anteriores no es una diferencia significativa.

Se interpretan algunos parámetros del modelo preferido:

```
coef(modelo4)
     (Intercept)    PartidoCCAAPP  PartidoCCAAPSOE        CCAAZona2        CCAAZona3
       27.172941        14.623328        27.060110        -8.241076       -16.295382
       CCAAZona4        CCAAZona5        CCAAZona6    AutonomosPtge UnemploymentPtge
       -3.381762        -9.031131         8.054304        -0.470340         1.022337
```

El modelo4 consta de 4 variables donde las mas relevantes son *PartidoCCAA* seguido de *CCAA* y donde el r^2 explica en un 44% la variable objetivo.

En cuanto a las variables mas importantes se puede interpretar lo siguiente:

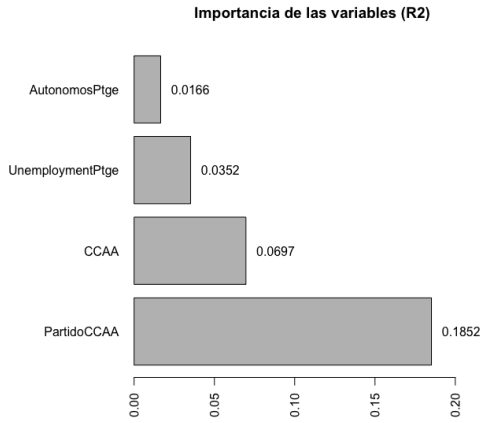
1. PartidoCCAAPSOE tiene un parámetro de 27 lo que indica que si la comunidad autónoma es gobernada por ese partido existe la posibilidad de que el % de votos aumentará en 27 puntos.

2. CCAAZona6 tiene un parámetro de 8 lo que indica que el % de votos hacia la izquierda podría subir 8 puntos en cambio en la CCAAZona3 el % disminuye en 16 puntos. y asi en las otras zonas geográficas.

Estabilidad del modelo, se puede decir que es un modelo estable.

```
# Estabilidad del modelo
> Rsq(modelo4,"varObjCont",data_train)
[1] 0.4438234
> Rsq(modelo4,"varObjCont",data_test)
[1] 0.4718783
```

En el gráfico que muestra la distribución de variables en el modelo elegido, siguen siendo PartidoCCAA y CCAA las variables más importantes.



5 Regresión lineal, modelos automáticos

Antes de transformar y discretizar variables input se ha separado la variable objetivo *Porcentaje* del conjunto de datos.

5.1 Transformación de variables input cuantitativas

Se analiza si es posible aplicar una transformación a las variables input, utilizando la función *Transf_Auto*, separando del conjunto de datos las variables aleatorias creadas anteriormente (contenidas en las columnas 25 y 26).

De 21 variables numéricas 18 han sido transformadas:

```
names(TransfCont) #veo las transformaciones que se han aplicado
 [1] "log_Censo"                          "log_Population"
 [3] "log_Age_under19_Ptge"               "cuarta_Age_over65_Ptge"
 [5] "inv_ForeignersPtge"                 "sqr_UniversityPtge"
 [7] "sqr_Empresas"                       "inv_Densidad"
 [9] "cuarta_PobChange_pct"               "cuarta_PersonasInmueble"
[11] "sqr_Explotaciones"                  "inv_WomenUnemploymentPtge"
[13] "log_UnemployLess25_Ptge"            "inv_UnemployMore40_Ptge"
[15] "raiz4_AgricultureUnemploymentPtge" "inv_IndustryUnemploymentPtge"
[17] "inv_ConstructionUnemploymentPtge"   "inv_ServicesUnemploymentPtge"
```

5.2 Discretización de variables input cuantitativas

Se indica que se discreticen todas las variables input numéricas salvo las aleatorias. Al discretizar las variables, se debe revisar que los niveles estén bien representados y en caso de no ser así, reagrupar

categorías con baja frecuencia.

De las variables discretizadas las siguientes debieron ser re categorizadas al tener categorias infrarrepresentadas (menor a 5%)

disc_Censo, disc_Population, disc_Age_over65_Ptge, disc_WomanPopulationPtge, disc_ForeignersPtge, disc_UniversityPtge, disc_Densidad, disc_PersonasInmueble, disc_Explotaciones, disc_WomenUnemploymentPtge, disc_UnemployLess25_Ptge, disc_UnemployMore40_Ptge, disc_AgricultureUnemploymentPtge, disc_ConstructionUnemploymentPtge, disc_ServicesUnemploymentPtge y $disc_AutonomosPtge.

Es decir, se tuvo que realizar re categorizaciones en casi todas las variables discretizadas.

5.3 Selección de variables automáticas

Una vez unidos en un mismo set de datos la variable objetivo, las input originales mas las transformaciones y discretizaciones, se realiza la partición *train-test* para evaluar los modelos. Con la partición ya hecha, se seleccionan las variables para formar parte del modelo.

Se comparan ademas los modelos obtenidos con el modelo manual que resulto ser el preferible en el apartado de regresión lineal manual, el modelo 4:

```
> #Modelo 1
> modeloManual<-lm(varObjCont~PartidoCCAA + CCAA + AutonomosPtge + UnemploymentPtge, data = data_train)
> modeloManual$rank
[1] 10
> Rsq(modeloManual,"varObjCont",data_train)
[1] 0.4439145
> Rsq(modeloManual,"varObjCont",data_test)
[1] 0.4729172
```

5.3.1 Selección de variables con input originales

Nota: A continuación se adjunta el código para los modelos ejecutados y probados y al final de este apartado en el momento del análisis de los modelos, se agrega una tabla resumen con los valores de r^2 para train y test por cada modelo manual y automático.

```
#Modelo 2
modeloStepAIC<-step(null, scope=list(lower=null, upper=full), direction="both", trace = F)

# Modelo 3
modeloBackAIC<-step(full, scope=list(lower=null, upper=full), direction="backward", trace = F)

# Modelo 4
modeloForwAIC<-step(null, scope=list(lower=null, upper=full), direction="forward", trace = F)

# Modelo 5
modeloStepBIC<-step(null, scope=list(lower=null, upper=full), direction="both",
                    trace = F, k=log(nrow(data_train)))

# Modelo 6
modeloBackBIC<-step(full, scope=list(lower=null, upper=full),direction="backward",
                    trace = F, k=log(nrow(data_train)))

# Modelo 7
modeloForwBIC<-step(null, scope=list(lower=null, upper=full), direction="forward",
                    trace = F, k=log(nrow(data_train)))
```

5.3.2 Selección de variables con input originales e interacciones

```
# Modelo 8
# Este modelo se descarta por demorar 8 min en ejecutar
modeloStepAIC_int<-step(null, scope=list(lower=null, upper=fullInt), direction="both", trace = F)

# Modelo 9
modeloBackAIC_int<-step(full, scope=list(lower=null, upper=fullInt),
                        direction="backward", trace = F)
```

```
# Modelo 10
# Este modelo también se descarta por excesivo tiempo de ejecución
modeloForwAIC_int<-step(null, scope=list(lower=null, upper=fullInt), direction="forward", trace = F)

# Modelo 11
modeloStepBIC_int<-step(null, scope=list(lower=null, upper=fullInt), direction="both",
                        trace = F, k=log(nrow(data_train)))

# Modelo 12
modeloBackBIC_int<-step(full, scope=list(lower=null, upper=fullInt), direction="backward",
                        trace = F, k=log(nrow(data_train)))

# Modelo 13
# También se descarta por tiempo de ejecución
modeloForwBIC_int<-step(null, scope=list(lower=null, upper=fullInt), direction="forward",
                        trace = F, k=log(nrow(data_train)))
```

5.3.3 Selección de variables con input originales y transformadas

```
# Modelo 14
modeloStepAIC_trans<-step(null, scope=list(lower=null, upper=fullT), trace = F, direction="both")

# Modelo 15
modeloBackAIC_trans<-step(full, scope=list(lower=null, upper=fullT),
                          trace = F, direction="backward")

# Modelo 16
modeloForwAIC_trans<-step(null, scope=list(lower=null, upper=fullT), trace = F,
                          direction="forward")

# Modelos BIC
# Modelo 17
modeloStepBIC_trans<-step(null, scope=list(lower=null, upper=fullT),
                          trace = F, direction="both",k=log(nrow(data_train)))

# Modelo 18
modeloBackBIC_trans<-step(full, scope=list(lower=null, upper=fullT),
                          trace = F, direction="backward",k=log(nrow(data_train)))

# Modelo 19
modeloForwBIC_trans<-step(null, scope=list(lower=null, upper=fullT),
                          trace = F, direction="forward",k=log(nrow(data_train)))
```

5.3.4 Selección de variables con input originales, transformaciones y discretizaciones

```
# Modelo 20
modeloStepAIC_todo<-step(null, scope=list(lower=null, upper=fulltodo), trace = F, direction="both")

# Modelo 21
modeloBackAIC_todo<-step(full, scope=list(lower=null, upper=fulltodo), direction="backward")

# Modelo 22
modeloForwkAIC_todo<-step(null, scope=list(lower=null, upper=fulltodo),
                          trace = F, direction="forward")

# Modelo 23
modeloStepBIC_todo<-step(null, scope=list(lower=null, upper=fulltodo),
                         trace = F, direction="both",k=log(nrow(data_train)))

# Modelo 24
modeloBackBIC_todo<-step(full, scope=list(lower=null, upper=fulltodo),
                         trace = F, direction="backward",k=log(nrow(data_train)))

# Modelo 25
modeloForwBIC_todo<-step(null, scope=list(lower=null, upper=fulltodo),
                         trace = F, direction="forward",k=log(nrow(data_train)))
```

5.3.5 Selección de variables incluyendo input, transformaciones, discretizaciones e interacciones

```
# Modelo 26 Stepwise
modeloStepBIC_todoInt<-step(null, scope=list(lower=null, upper=fullIntT),
                            direction="both",trace =F, k=log(nrow(data_train)))

# Modelo 27 Backwise
modeloBackBIC_todoInt<-step(full, scope=list(lower=null, upper=fullIntT),
                            direction="backward",trace = F, k=log(nrow(data_train)))

# Modelo 28 Forward
modeloForwBIC_todoInt<-step(null, scope=list(lower=null, upper=fullIntT),
                            direction="forward",trace = F, k=log(nrow(data_train)))

# Modelo 29 AIC Forward

modeloForwkAIC_todoInt<-step(null, scope=list(lower=null, upper=fulltodo),
                             trace = F, direction="forward")

# Modelo 30 AIC Back
modeloBackAIC_todoInt<-step(full, scope=list(lower=null, upper=fulltodo),
                            trace = F, direction="backward")

# Modelo 31 AIC Both
modeloStepAIC_todoInt<-step(null, scope=list(lower=null, upper=fulltodo),
                            trace = F, direction="both")
```

5.3.6 Comparación de modelos

En la siguiente tabla se pueden observar los resultados de todos los modelos tanto manuales como automáticos que han sido ejecutados:

Esta tabla no incluye los modelos *modeloStepAIC_int*, *modeloForwAIC_int* y *modeloForwBIC_int* pues fueron descartados por tiempo de ejecución.

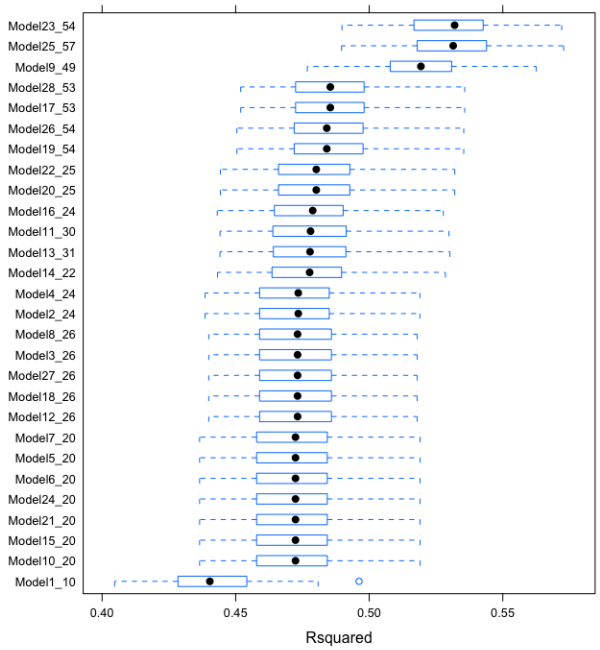
Cuadro 4: Comparación de modelos de regresión lineal

Modelo	Modelos	ranking	rtest	rtrain
1	modeloManual	10	0.4718783	0.4438234
2	modeloStepAIC	24	0.5029228	0.4773708
3	modeloBackAIC	26	0.5028253	0.4780461
4	modeloForwAIC	24	0.5029228	0.4773708
5	modeloStepBIC	20	0.4986513	0.4758112
6	modeloBackBIC	20	0.4986513	0.4758112
7	modeloForwBIC	20	0.4986513	0.4758112
8	modeloBackAIC_int	26	0.5028253	0.4780461
9	modeloStepBIC_int	49	0.5472125	0.5273960
10	modeloBackBIC_int	20	0.4986513	0.4758112
11	modeloStepAIC_trans	30	0.5085348	0.4847584
12	modeloBackAIC_trans	26	0.5028253	0.4780461
13	modeloForwAIC_trans	31	0.5086244	0.4848856
14	modeloStepBIC_trans	22	0.5055675	0.4816614
15	modeloBackBIC_trans	20	0.4986513	0.4758112
16	modeloForwBIC_trans	24	0.5060944	0.4826059
17	modeloStepAIC_todo	53	0.5153569	0.4967459
18	modeloBackAIC_todo	26	0.5028253	0.4780461
19	modeloForwkAIC_todo	54	0.5153391	0.4967461
20	modeloStepBIC_todo	25	0.5080702	0.4852266
21	modeloBackBIC_todo	20	0.4986513	0.4758112
22	modeloForwBIC_todo	25	0.5080702	0.4852266
23	modeloStepBIC_todoInt	54	0.5583550	0.5398435
24	modeloBackBIC_todoInt	20	0.4986513	0.4758112
25	modeloForwBIC_todoInt	57	0.5583375	0.5409726
26	modeloForwkAIC_todoInt	54	0.5153391	0.4967461
27	modeloBackAIC_todoInt	26	0.5028253	0.4780461

Modelo	Modelos	ranking	rtest	rtrain
28	modeloStepAIC_todoInt	53	0.5153569	0.4967459

De todos los modelos tanto por equilibrio como capacidad predictiva el modelo *modeloStepBIC_trans* es el que mejor resultados muestra.

Se lleva a cabo la validación cruzada respectiva para evaluar los modelos, obteniéndose lo siguiente:




```
summary(resamples(vcrTodosModelos),metric=c("Rsquared"))

Call:
summary.resamples(object = resamples(vcrTodosModelos), metric = c("Rsquared"))

Models: Model1_10, Model2_24, Model3_26, Model4_24, Model5_20, Model6_20, Model7_20, Model8_26,
Model9_49, Model10_20, Model11_30, Model12_26, Model13_31, Model14_22, Model15_20, Model16_24,
Model17_53, Model18_26, Model19_54, Model20_25, Model21_20, Model22_25, Model23_54, Model24_20,
Model25_57, Model26_54, Model27_26, Model28_53
Number of resamples: 100

Rsquared
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
Model1_10  0.4046811 0.4284404 0.4403327 0.4425415 0.4541025 0.4961934    0
Model2_24  0.4385406 0.4593744 0.4734374 0.4731663 0.4850413 0.5189712    0
Model3_26  0.4399262 0.4591736 0.4731721 0.4735598 0.4857633 0.5179647    0
Model4_24  0.4385406 0.4593744 0.4734374 0.4731663 0.4850413 0.5189712    0
Model5_20  0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model6_20  0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model7_20  0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model8_26  0.4399262 0.4591736 0.4731721 0.4735598 0.4857633 0.5179647    0
Model9_49  0.4767975 0.5079858 0.5193470 0.5188826 0.5307347 0.5624985    0
Model10_20 0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model11_30 0.4441478 0.4640511 0.4780294 0.4790206 0.4913558 0.5298315    0
Model12_26 0.4399262 0.4591736 0.4731721 0.4735598 0.4857633 0.5179647    0
Model13_31 0.4441187 0.4641894 0.4778412 0.4789987 0.4912224 0.5302009    0
Model14_22 0.4432132 0.4636780 0.4776930 0.4777272 0.4894813 0.5285194    0
Model15_20 0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model16_24 0.4431031 0.4651417 0.4788425 0.4784617 0.4901613 0.5277266    0
Model17_53 0.4518793 0.4725308 0.4854342 0.4862493 0.4980582 0.5357292    0
Model18_26 0.4399262 0.4591736 0.4731721 0.4735598 0.4857633 0.5179647    0
Model19_54 0.4504438 0.4719724 0.4841087 0.4858726 0.4976385 0.5353822    0
Model20_25 0.4442847 0.4661767 0.4801762 0.4807759 0.4926687 0.5320197    0
Model21_20 0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model22_25 0.4442847 0.4661767 0.4801762 0.4807759 0.4926687 0.5320197    0
Model23_54 0.4897969 0.5168412 0.5319637 0.5302055 0.5426430 0.5720512    0
Model24_20 0.4364895 0.4579774 0.4723954 0.4725615 0.4842249 0.5190414    0
Model25_57 0.4896558 0.5179645 0.5314214 0.5309875 0.5438914 0.5728014    0
Model26_54 0.4504438 0.4719724 0.4841087 0.4858726 0.4976385 0.5353822    0
Model27_26 0.4399262 0.4591736 0.4731721 0.4735598 0.4857633 0.5179647    0
Model28_53 0.4518793 0.4725308 0.4854342 0.4862493 0.4980582 0.5357292    0
```

Revisando la información que entrega el gráfico y la tabla comparativa de los modelos, los modelos

generados con AIC y BIC tienen similitudes en capacidad predictiva pero siempre es preferible considerar los que posean menor numero de parámetros.

En el gráfico los modelos con mejor R2 son ademas los que poseen un número de parámetros muy elevado por ende, son descartados de inmediato. Se busca un equilibrio entre predicción y simpleza por lo que se decide por el modelo14 *modeloStepBIC_trans* por mostrar equilibrio entre el número de parámetros y un R2 aceptable.

5.3.7 Análisis del modelo ganador

El modelo ganador para esta practica es el *modeloStepBIC_trans*

```
> coef(modeloStepBIC_trans)
                  (Intercept)                PartidoCCAAPP              PartidoCCAAPSOE
                41.960845092                13.934154204                27.505885002
                    CCAAZona2                    CCAAZona3                    CCAAZona4
                -7.784425915                -16.973102308                -1.284287035
                    CCAAZona5                    CCAAZona6           UnemploymentPtge
                -7.248196699                  8.081398805                 0.416472160
              AutonomosPtge      cuarta_Age_over65_Ptge           sqr_Explotaciones
                -0.471325919                -0.006918045                -0.268685767
                  Densidad          sqr_UniversityPtge             ForeignersPtge
                 0.001005086                -0.278274067                -0.190765506
    log_UnemployLess25_Ptge           PersonasInmueble              log_Population
                 0.192013888                -3.206055223                 0.787394931
AgricultureUnemploymentPtge    ServicesUnemploymentPtge             Age_over65_Ptge
                 0.152171947                 0.023313836                -0.116466604
```

Conclusiones:

El r^2 explica el modelo en un 0.5055675, no se observa una diferencia significativa en train y test.

El % de votación aumentaría en 27 puntos cuando el partido que gobierna la comunidad autónoma al momento de la elección es del PSOE.

El % de votación aumentaría en 8 puntos en caso de que la comunidad autónoma sea de la zona 6 (sur de España).

Las variables que explican en su mayoría el modelo son PartidoCCAA, CCAA y AutonomosPtge.