Universidad del Valle de Guatemala
Data Science
Catedrática: Lynette García
Sergio Marchena - 16387
Pablo Viana - 16091

Laboratorio 1: Clustering, PCA y Reglas de Asociación

---

1. **Haga una exploración rápida de sus datos para eso haga un resumen de su dataset.**

Para hacer la exploración rápida de los datos lo que se hizo fue usar la función de summary() para ver los valores mínimos, máximos, la media, mediana, y cuartiles.

```
Id                MSSubClass       MSZoning     LotFrontage
Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00
1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00
Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00
Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05
3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00
Max.   :1460.0   Max.   :190.0                   Max.   :313.00
                                                 NA's   :259
LotArea          Street      Alley      LotShape  LandContour
Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63
1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50
Median :  9478               NA's:1369   IR3: 10   Low:  36
Mean   : 10517                           Reg:925   Lvl:1311
3rd Qu.: 11602
Max.   :215245


Utilities      LotConfig    LandSlope   Neighborhood   Condition1
AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260
NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81
              FR2    :  47   Sev:  13   OldTown:113   Artery :  48
```

```
              FR3    :   4              Edwards:100   RRAn   :  26
              Inside :1052              Somerst: 86   PosN   :  19
                                        Gilbert: 79   RRAe   :  11
                                        (Other):707   (Other):  15


Condition2     BldgType      HouseStyle   OverallQual
Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000
Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000
Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000
PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099
RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000
PosA   :   1                 1.5Unf : 14   Max.   :10.000
(Other):   2                 (Other): 19
OverallCond      YearBuilt     YearRemodAdd    RoofStyle
Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13
1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141
Median :5.000   Median :1973   Median :1994   Gambrel:  11
Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286
3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7
Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2


  RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0
Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0
WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0
WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.7
ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8   3rd Qu.: 166.0
Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0
(Other):   2   (Other):128   (Other):136                 NA's   :8


ExterQual ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure
Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
```

```
Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
          TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
                    Wood  :  3

BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
ALQ :220     Min.   :   0.0   ALQ :  19    Min.   :   0.00
BLQ :148     1st Qu.:   0.0   BLQ :  33    1st Qu.:   0.00
GLQ :418     Median : 383.5   GLQ :  14    Median :   0.00
LwQ : 74     Mean   : 443.6   LwQ :  46    Mean   :  46.55
Rec :133     3rd Qu.: 712.2   Rec :  54    3rd Qu.:   0.00
Unf :430     Max.   :5644.0   Unf :1256    Max.   :1474.00
NA's: 37                      NA's:  38

 BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
Median : 477.5   Median : 991.5   GasW :  18   Gd:241
Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
Max.   :2336.0   Max.   :6110.0   Wall :   4

Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
FuseP:   3   Median :1087   Median :   0   Median :  0.000
Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000

 GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
```

```
Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
Median :1464   Median :0.0000   Median :0.00000   Median :2.000
Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000

  HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual
Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
Median :0.0000   Median :3.000   Median :1.000   Gd:586
Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
Max.   :2.0000   Max.   :8.000   Max.   :3.000

TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu   GarageType
Min.   : 2.000   Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6
1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870
Median : 6.000   Min1:  31   Median :1.000   Gd  :380    Basment: 19
Mean   : 6.518   Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88
3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9
Max.   :14.000   Sev :   1   Max.   :3.000   NA's:690    Detchd :387
                 Typ :1360                               NA's   : 81

GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
Min.   :1900   Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3
1st Qu.:1961   RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48
Median :1980   Unf :605     Median :2.000   Median : 480.0   Gd  :  14
Mean   :1979   NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3
3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311
Max.   :2010                Max.   :4.000   Max.   :1418.0   NA's:  81
NA's   :81
```

```
GarageCond  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
Ex  :   2   N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
Fa  :  35   P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
Gd  :   9   Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
Po  :   7              Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
TA  :1326              3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
NA's:  81              Max.   :857.00   Max.   :547.00   Max.   :552.00

 X3SsnPorch      ScreenPorch        PoolArea        PoolQC
Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Ex  :   2
1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2
Median :  0.00   Median :  0.00   Median :  0.000   Gd  :   3
Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NA's:1453
3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
Max.   :508.00   Max.   :480.00   Max.   :738.000

  Fence      MiscFeature    MiscVal             MoSold
GdPrv:  59   Gar2:   2   Min.   :    0.00   Min.   : 1.000
GdWo :  54   Othr:   2   1st Qu.:    0.00   1st Qu.: 5.000
MnPrv: 157   Shed:  49   Median :    0.00   Median : 6.000
MnWw :  11   TenC:   1   Mean   :   43.49   Mean   : 6.322
NA's :1179   NA's:1406   3rd Qu.:    0.00   3rd Qu.: 8.000
                         Max.   :15500.00   Max.   :12.000

    YrSold        SaleType    SaleCondition    SalePrice
Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
Median :2008   COD    :  43   Alloca :  12   Median :163000
Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
               (Other):   9
```

Además, se utilizó la función de str() para ver la estructura de cada variable del dataset.

```
'data.frame':    1460 obs. of  81 variables:
 $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
 $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
 $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
 $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
 $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
 $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
 $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
 $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
 $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
 $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
 $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
 $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
 $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
 $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
 $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
 $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
 $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
 $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
 $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
 $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
 $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
 $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
 $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
 $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
 $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
```

```
$ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
$ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
$ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
$ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
$ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
$ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
$ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
$ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
$ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
$ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
$ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
$ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
$ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
$ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
$ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
$ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
$ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
$ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
$ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
$ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
$ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
$ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
$ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
$ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
$ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
$ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
$ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
$ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
$ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
$ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
$ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
$ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
```

```
$ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...

$ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...

$ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...

$ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...

$ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...

$ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...

$ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...

$ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...

$ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...

$ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...

$ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...

$ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...

$ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...

$ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...

$ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...

$ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...

$ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...

$ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...

$ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...

$ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

2. **Diga el tipo de cada una de las variables del dataset (cualitativa o categórica, cuantitativa, cuantitativa continua o cuantitativa discreta).**

Variable	**Descripción**	**Tipo**
Id	Identificador de los datos.	cuantitativa discreta
MSSubClass	Clase de la propiedad.	cuantitativa discreta
MSZoning	Clasificación de la zona general.	cuantitativa discreta
LotFrontage	Número de pies de la calle conectada a la propiedad.	cuantitativa discreta
LotArea	Tamaño del lote en pies cuadrados.	cuantitativa discreta
Street	Tipo de calle de acceso.	categórica
Alley	Tipo de callejón de acceso.	categórica

LotShape	Forma general del lote.	categórica
LandContour	Planitud del lote.	categórica
Utilities	Tipo de utilidades disponibles.	categórica
LotConfig	Configuración del lote,	categórica
LandSlope	Inclinación de la propiedad.	categórica
Neighborhood	Ubicaciones físicas dentro de los límites de la ciudad de Ames.	categórica
Condition1	Proximidad a la calle principal o líneas de tren.	categórica
Condition2	Proximidad a la calle principal o líneas de tren (si existe otra).	categórica
BldgType	Tipo de la vivienda.	categórica
HouseStyle	Estilo de la vivienda.	categórica
OverallQual	Calidad del material y acabados.	cuantitativa discreta
OverallCond	Rating de la condición de la propiedad.	cuantitativa discreta
YearBuilt	Año de construcción de la propiedad.	cuantitativa discreta
YearRemodAdd	Año de remodelación de la propiedad.	cuantitativa discreta
RoofStyle	Tipo del techo.	categórica
RoofMatl	Material del techo.	categórica
Exterior1st	Revestimiento exterior 1.	categórica
Exterior2nd	Revestimiento exterior 2 (si hay).	categórica
MasVnrType	Tipo de chapa de albañilería.	categórica
MasVnrArea	Area de chapa de albañilería.	categórica
ExterQual	Calidad del material exterior.	categórica
ExterCond	Condición presente del material exterior	categórica
Foundation	Tipo de cimiento	categórica
BsmtQual	Altura del sótano	categórica

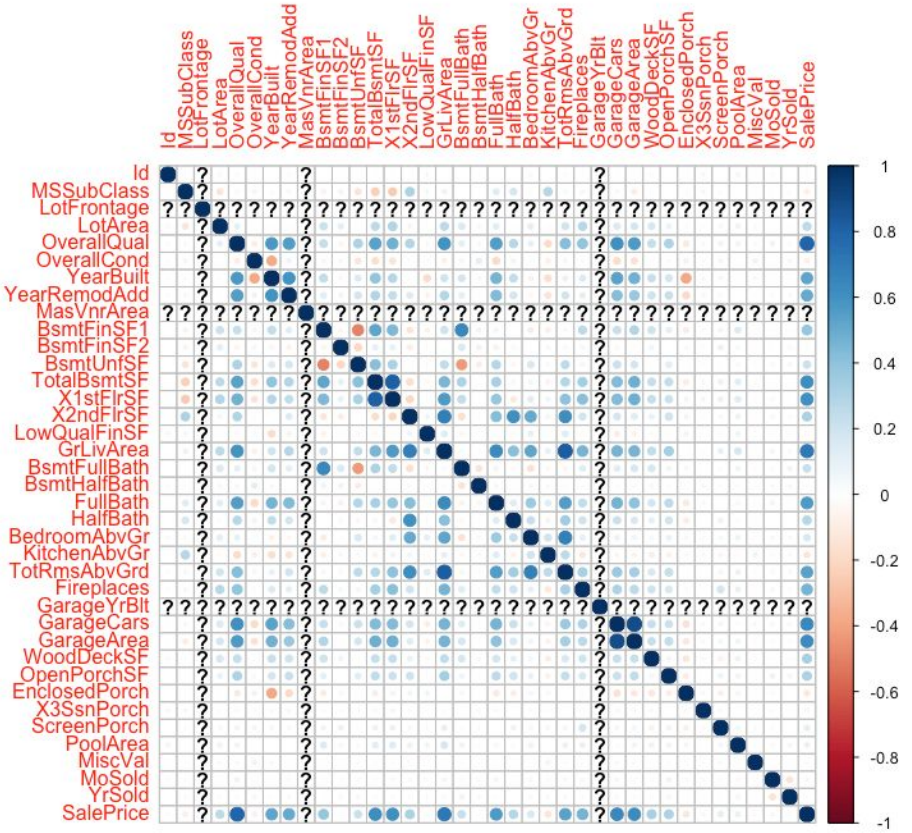
BsmtCond	Condiciones generales del sótano.	categórica
BsmtExposure	Muros de sótano a nivel de jardín o de salida	cualitativa
BsmtFinType1	Calidad del terminado del sótano	cualitativa
BsmtFinSF1	Terminado tipo 1 pies cuadrados.	cuantitativa discreta
BsmtFinType2	Cualidad de acabado segunda area (si existe)	cualitativa
BstmFinSF2	Pies cuadrados de terminado tipo 2	cuantitativa discreta
BsmtUnfSF	Pies cuadrados de área no terminada del sótano	cuantitativa discreta
TotalBsmtSF	Pies cuadrados totales del sótano	cuantitativa discreta
Heating	Tipo de calefacción	cualitativa
HeatingQC	Calidad y condición de la calefacción	cualitativa
CentralAir	Aire acondicionado central	cualitativa
Electrical	Sistema eléctrico	categórica
1stFlrSF	Pies cuadrados primer piso	cuantitativa discreta
2ndFlrSF	Pies cuadrados segundo piso	cuantitativa discreta
LowQualFinSF	Acabados de baja calidad terminados (todos los pisos)	cuantitativa discreta
GrLivArea	Pies cuadrados de área habitable arriba del piso	cuantitativa discreta
BsmtFullBath	Baños completos en sótano	cuantitativa discreta
BsmtHalfBath	Baños a medias en el sótano	cuantitativa discreta
FullBath	Baños completos	cuantitativa discreta
HalfBath	Baños a medias	cuantitativa discreta
BedroomAbvGr	Número de cuartos arriba del nivel de sótano	cuantitativa discreta
KitchenAbvGr	Número de cocinas arriba del nivel de sótano	cuantitativa discreta
KitchenQual	Calidad de la cocina	categórica

TotRmsAbvGrd	Cantidad de cuartos arriba del nivel de sótano (no incluye baños)	cuantitativa discreta
Functional	Rating de funcionalidad de la casa	categórica
Fireplaces	Número de chimeneas	cuantitativa discreta
FireplacesQu	Calidad de chimeneas	categórica
GarageType	Localización del sótano	categórica
GarageYrBlt	Año de construcción del sótano	cuantitativa discreta
GarageFinish	Acabado interno del sótano	categórica
GarageCars	Capacidad para carros en el garage	cuantitativa discreta
GarageArea	Pies cuadrados del sótano	cuantitativa discreta
GarageQual	Calidad del garage	categórica
GarageCond	Condición del sótano	categórica
PavedDrive	Pavimento acera	categórica
WoodDeckSF	Pies cuadrados del pórtico de madera	cuantitativa discreta
OpenPorchSF	Pies cuadrados del pórtico	cuantitativa discreta
EnclosedPorch	Pies cuadrados del pórtico cerrado	cuantitativa discreta
3SsnPorch	Pies cuadrados “three season” pórtico	cuantitativa discreta
ScreenPorch	Área en pies cuadrados del “screen porch”	cuantitativa discreta
PoolArea	Pies cuadrados del área de la piscina	cuantitativa discreta
PoolQC	Calidad de la piscina	categórica
Fence	Calidad de cerca (barandilla)	categórica
MiscFeature	Características varias no cubiertas en otras categorías	categórica
MiscVal	Valor de características varias	cuantitativa discreta
MoSold	Mes vendido	cuantitativa discreta
YrSold	Año vendido	cuantitativa discreta

SaleType	Tipo de venta	categórica
SaleCondition	Condición de la venta	categórica
SalePrice	Precio de venta (variable a predecir)	cuantitativa discreta

3. **Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.**

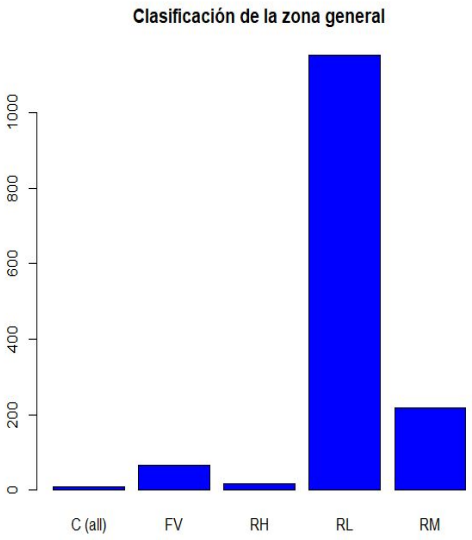
   Se utilizó la librería “corrplot” para poder ver la correlación en las variables numéricas de una forma más fácil y entendible para el lector. El gráfico que se obtuvo fue el siguiente:

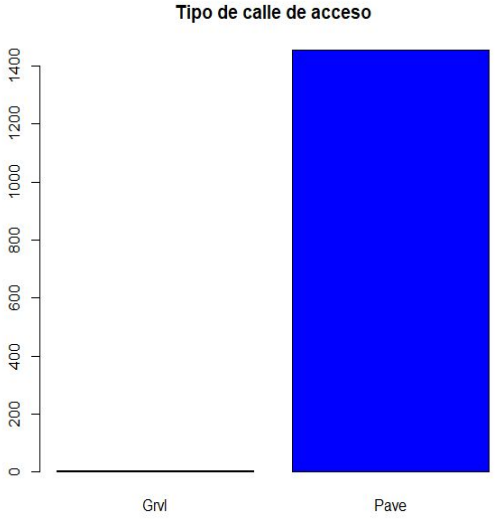


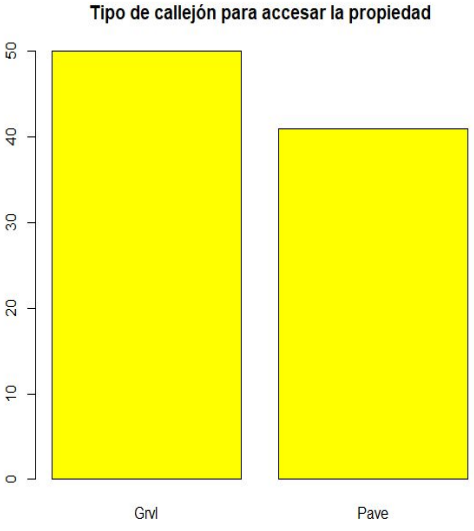
   Donde se puede ver que los puntos más azules, es donde hay una correlación fuerte entre las variables, mientras que en los puntos de color rojo y más claros, no hay correlación lineal entre esas variables.

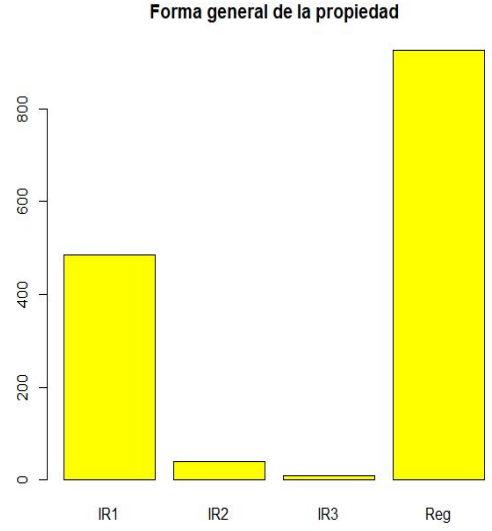
4. **Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos.**

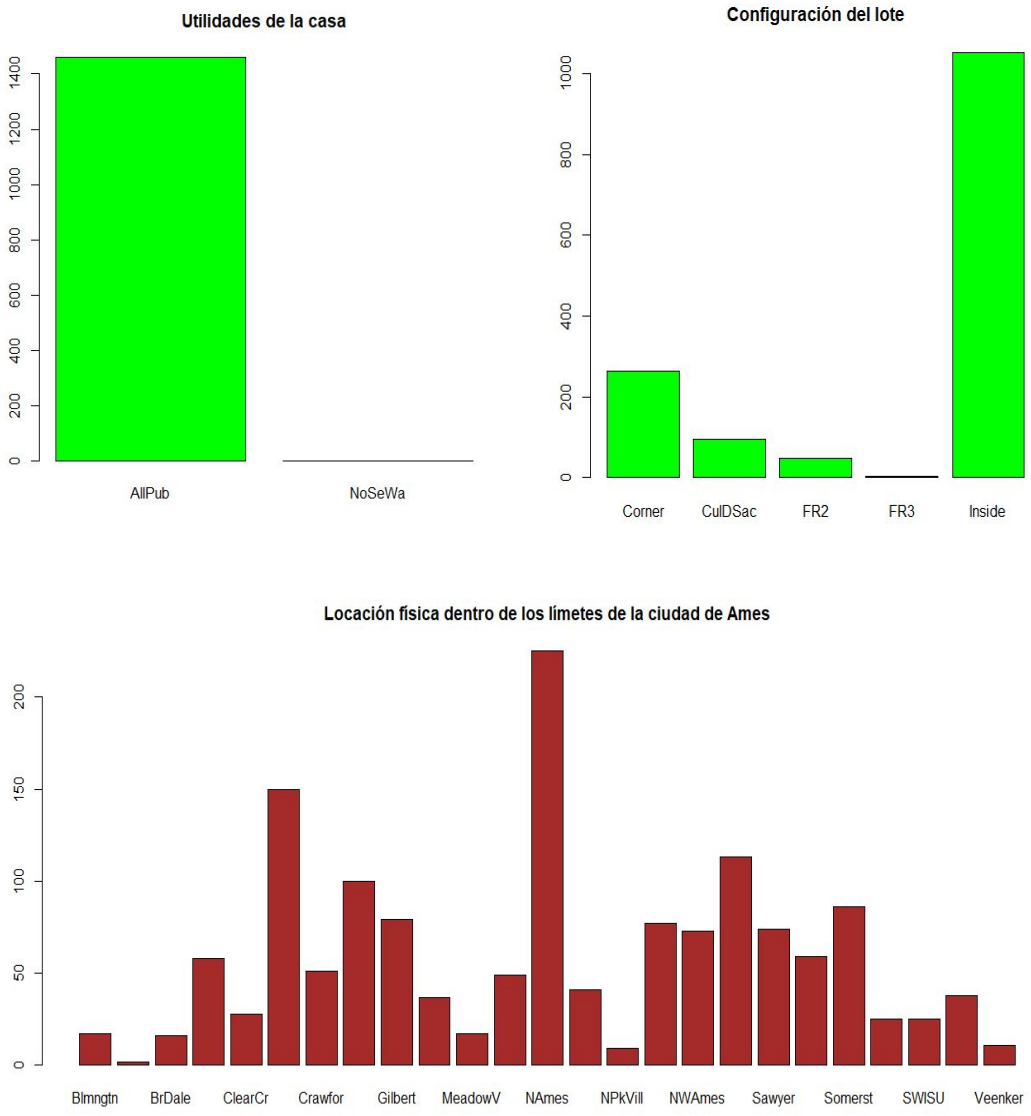
Se clasificaron 43 variables categóricas. Se procedió a crear las tablas de frecuencia de todas las variables. Se adjuntan algunas gráficas de barras obtenidas de las mismas.





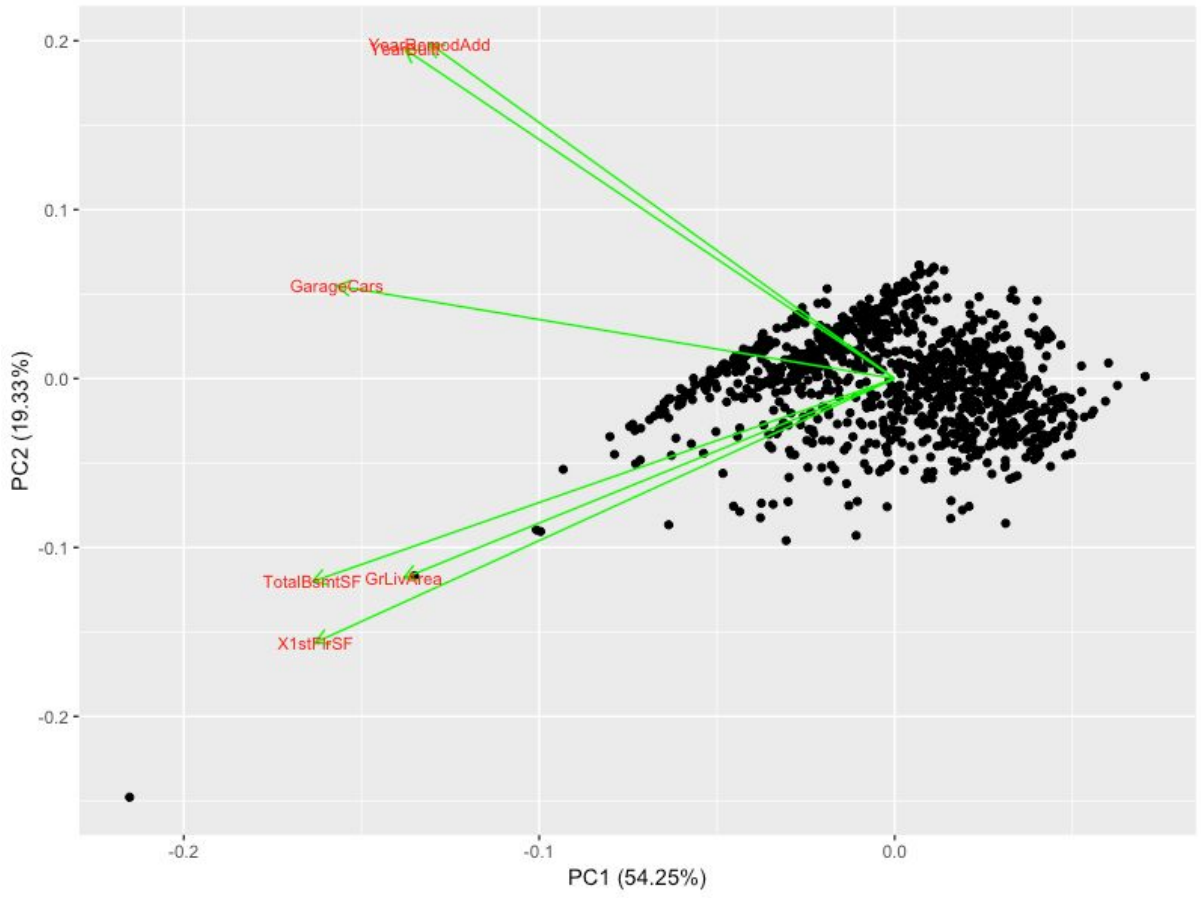






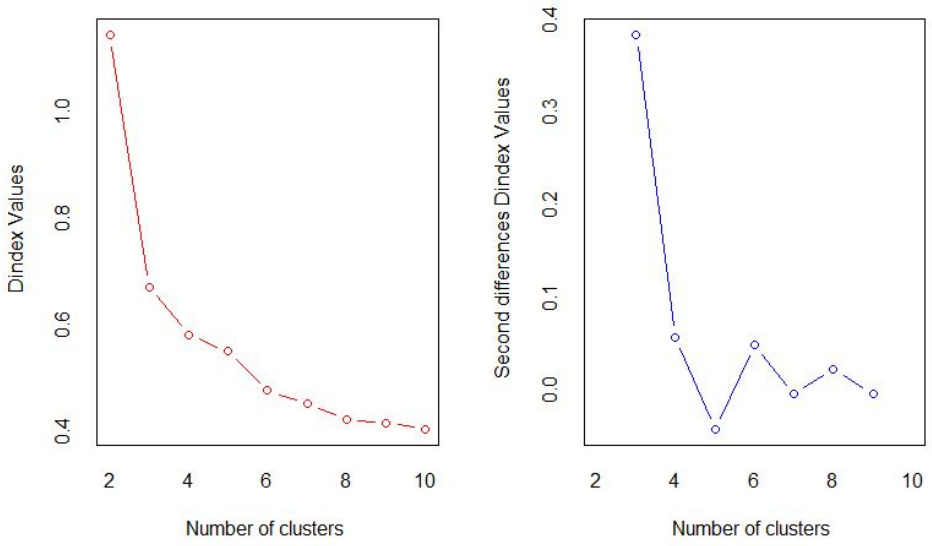
5. Haga un análisis de componentes principales, interprete los componentes.

Para hacer el análisis de componentes principales tuvimos que basarnos en la matriz de correlación de las variables, para ver qué variables estaban relacionadas entre sí. Luego, hicimos una prueba de esfericidad de Barlett, con el objetivo de determinar si era factible o no hacer un PCA. Luego de hacer esta prueba, según el algoritmo de PCA de R y la correlación entre las variables, estas variables fueron las que más explicaron la variabilidad de los datos. Dichas variables se pueden considerar como las más importantes según este estudio a la hora de comprar una casa. Las variables son: “YearBuilt”, “YearRemodAdd”, “TotalBsmtSF”, “X1stFlrSF”, “GrLivArea” y “GarageCars” respectivamente. En el siguiente gráfico se puede ver cómo se comportan con el resto de los datos.

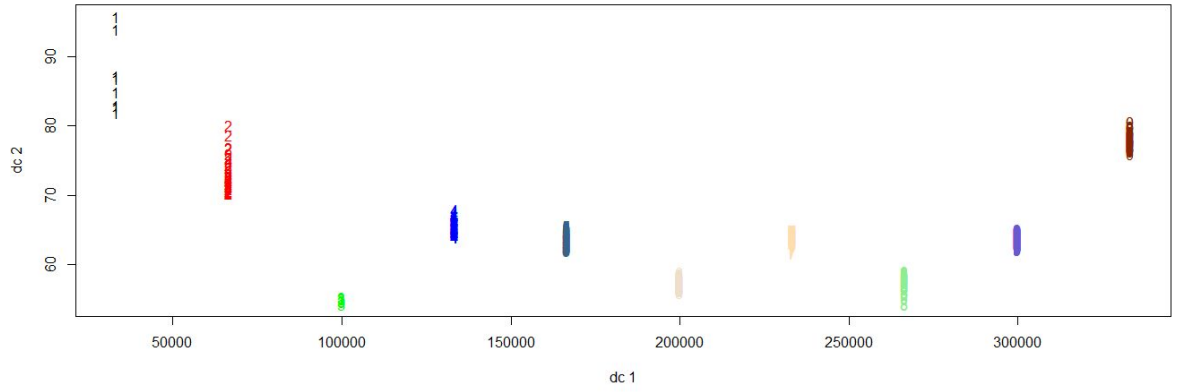


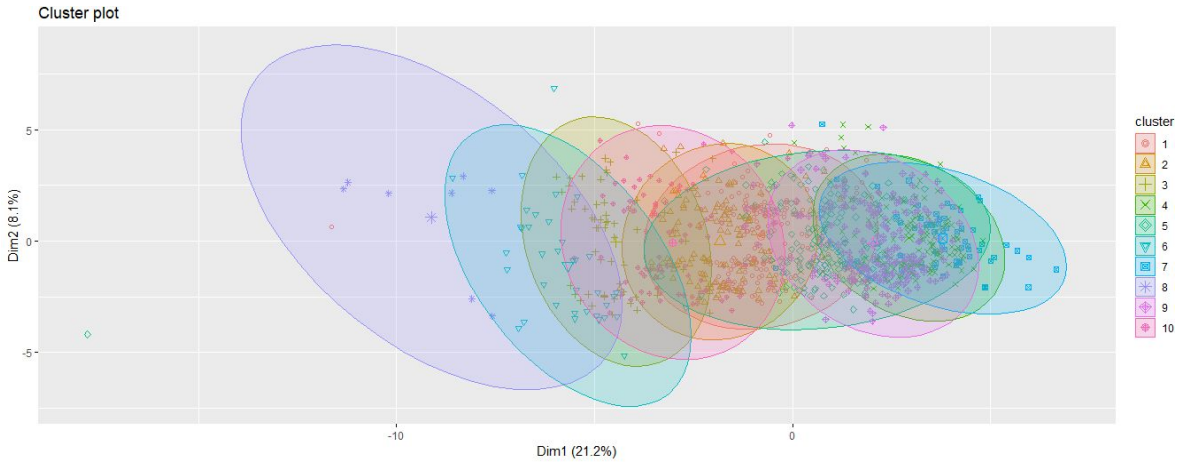
Se puede ver que las componentes principales explican aproximadamente un 73% de la variabilidad de los datos. (PC1+PC2)

6. **Haga un análisis de clustering, describa los grupos**.



Se utiliza la librería NbClust para encontrar un número óptimo de cluster; se determina que este es 10. El primer algoritmo de clustering a probar es k-medias con 10 grupos, del cual se generan las dos gráfica siguientes, con las librerías fpc y factoextra, respectivamente.





Con este método, como podemos observar, tenemos grupos transpuestos; esta observación nos hace probar como siguiente algoritmo fuzzy c-means, el cual arroja los siguientes porcentajes de pertenencia a cada grupo por fila.

```
> nums_completo[,51:60]
             1            2           3           4           5           6           7           8           9           10
1   2.155851e-03 5.320729e-04 5.963040e-03 0.0201569463 0.0601737769 0.0103196944 0.1299124062 5.564022e-03 0.7464406090 0.0187815817
2   1.344057e-03 3.761421e-04 3.185643e-03 0.0080568892 0.6546238269 0.0170667928 0.0246005493 7.405388e-03 0.2402515994 0.0430891115
3   2.426421e-03 5.516142e-04 7.519194e-03 0.0330588365 0.0292335689 0.0072514528 0.7815117078 4.226802e-03 0.1222499484 0.0119704545
4   8.935925e-05 2.916270e-05 1.800010e-04 0.0003555449 0.0062899261 0.0140613838 0.0007130077 2.186753e-03 0.0018466854 0.9742481764
5   6.532333e-03 1.250016e-03 2.676512e-02 0.3027260372 0.0250823273 0.0090288125 0.5523056911 5.807262e-03 0.0571814001 0.0133210047
6   3.888698e-04 1.256718e-04 7.907324e-04 0.0015821330 0.0318153897 0.0431763194 0.0032338446 8.227043e-03 0.0086693177 0.9019906791
7   2.616263e-02 2.834138e-03 6.629106e-01 0.2066979758 0.0134793769 0.0070062750 0.0458174922 5.098384e-03 0.0208792668 0.0091138390
9   8.365385e-04 2.818952e-04 1.633918e-03 0.0030956644 0.0361720039 0.5117139770 0.0058514376 3.362172e-02 0.0134988422 0.3932940037
10  2.986444e-05 1.042287e-05 5.645505e-05 0.0001025222 0.0008539963 0.9917347566 0.0001831614 2.486941e-03 0.0003809148 0.0041609654
11  8.430468e-04 2.843650e-04 1.644994e-03 0.0031129257 0.0360143006 0.5239887076 0.0058786840 3.432961e-02 0.0135000981 0.3804032643
12  8.123056e-02 4.299568e-03 7.979204e-01 0.0541460116 0.0094681590 0.0055824984 0.0229390410 4.270848e-03 0.0132123516 0.0069305206
14  3.142287e-03 4.685517e-04 2.241951e-02 0.9335016051 0.0041890874 0.0018980843 0.0230954763 1.314224e-03 0.0073787446 0.0025924250
16  8.320272e-04 2.785730e-04 1.635067e-03 0.0031230999 0.0392598597 0.3575993322 0.0059695433 2.997079e-02 0.0140702165 0.5472614884
18  2.455844e-04 9.231360e-05 4.347285e-04 0.0007282749 0.0036498022 0.0284296994 0.0011769002 9.542852e-01 0.0020734340 0.0088840851
19  1.544088e-03 4.720242e-04 3.322057e-03 0.0072081326 0.5334655726 0.0568851234 0.0166965027 1.758814e-02 0.0599755871 0.3028427760
20  1.710291e-04 5.600270e-05 3.433911e-04 0.0006752215 0.0113589222 0.0300993401 0.0013448303 4.389654e-03 0.0034388435 0.9481227652
21  9.061655e-04 7.340585e-05 9.953553e-01 0.0020111716 0.0002380561 0.0001323903 0.0006685176 9.891209e-05 0.0003480696 0.0001680000
22  1.487992e-04 4.866200e-05 2.991265e-04 0.0005891742 0.0100914549 0.0250947736 0.0011762569 3.750981e-03 0.0030228655 0.9557779058
23  5.581725e-04 1.221039e-04 1.831517e-03 0.0093802764 0.0049431292 0.0013752971 0.9621475590 8.242400e-04 0.0166214999 0.0021962049
24  8.989379e-04 3.029844e-04 1.755362e-03 0.0033241993 0.0385189287 0.5088357801 0.0062766925 3.600570e-02 0.0144585496 0.3896228659
26  7.102737e-03 1.296305e-03 3.180696e-02 0.5737608779 0.0215118975 0.0082233970 0.2934575024 5.386405e-03 0.0455564566 0.0118974592
27  6.072698e-04 2.015343e-04 1.203527e-03 0.0023251370 0.0326135537 0.1762076026 0.0045162343 1.901444e-02 0.0109798419 0.7523308638
28  2.676751e-02 2.937098e-03 6.333371e-01 0.2290119301 0.0142660894 0.0073842047 0.0491339159 5.365071e-03 0.0221768701 0.0096202271
29  2.501831e-03 6.205700e-04 6.867559e-03 0.0228046873 0.0718965448 0.0122495407 0.1377108968 6.587391e-03 0.7163773540 0.0223836252
30  2.312227e-03 9.139705e-04 3.924977e-03 0.0062633329 0.0245550591 0.1019459832 0.0095637398 7.870744e-01 0.0155054524 0.0479408803
31  8.558683e-03 3.588268e-03 1.386414e-02 0.0210098863 0.0656728102 0.1817482818 0.0303110251 5.211139e-01 0.0454109005 0.1087221387
33  1.103394e-03 3.108557e-04 2.595260e-03 0.0064804538 0.7443304634 0.0149375253 0.0193017773 6.372393e-03 0.1657083759 0.0388595019
34  5.897978e-04 1.759829e-04 1.302100e-03 0.0029379227 0.8739982180 0.0154968468 0.0072616118 5.393354e-03 0.0315233948 0.0613207711
35  2.573853e-03 3.919428e-04 1.731257e-02 0.9432021784 0.0036844696 0.0016500233 0.0212163483 1.137677e-03 0.0065682053 0.0022627339
36  2.383395e-02 2.507534e-03 7.278886e-01 0.1608653716 0.0114081016 0.0059784846 0.0378596310 4.364439e-03 0.0175398953 0.0077539546
37  1.861791e-04 5.976532e-05 3.811624e-04 0.0007700250 0.0177638250 0.0182786963 0.0015961372 3.658402e-03 0.0044295217 0.9528762863
38  1.203381e-03 3.758854e-04 2.532569e-03 0.0053205803 0.2204815986 0.0656744221 0.0116991500 1.723837e-02 0.0372599368 0.6382141083
```

Tras analizar los dos métodos por medio de la silueta obtenemos valores muy cercanos entre ambos. Con k-medias la media de la silueta es 0.48 y con fuzzy c-means el valor es de 0.477; estos número no representan la mejor distribución de grupos pero están en el rango aceptable.

7. **Obtenga reglas de asociación más interesantes del dataset. Discuta sobre el nivel de confianza y soporte.**

Se procede a utilizar el algoritmo apriori con la división de la muestra categórica, obtenida en ejercicios anteriores. Debido a la gran cantidad de datos que tenemos, hacer reglas de asociación podría tener un tiempo computacional muy alto si se utilizan los parámetros por defecto. Por esta razón, se utilizan los siguientes:

- confidence: 0.9
- support: 0.9
- maxlen: 5
- maxtime: 3
- minlen: 2
- target: “rules”

Se utilizan valores altos de confianza y soporte para obtener reglas muy específicas con bastante ocurrencia en el dataset y así disminuir el número de datos. Los demás argumentos se utilizan para reducir el tiempo computacional poniendo límites al tiempo de búsqueda y cantidad de items en la regla. Minlen igual a dos se utiliza para que no existan reglas del tipo $\{\} \Rightarrow \{item\}$ en el set de reglas.

Una vez que se corre el algoritmo con estas especificaciones, procedemos a eliminar las reglas redundantes para solo quedarnos con aquellas reglas específicas que contienen a reglas más generales.

El resultado es el siguiente:

```
> inspect(reglas[!is.redundant(reglas)])
     lhs                    rhs                    support   confidence      lift count
[1]  {GarageCond=TA}     => {Condition2=Norm}   0.9000000  0.9909502 1.0012369  1314
[2]  {Condition2=Norm}   => {GarageCond=TA}     0.9000000  0.9093426 1.0012369  1314
[3]  {GarageCond=TA}     => {Street=Pave}       0.9047945  0.9962293 1.0003402  1321
[4]  {Street=Pave}       => {GarageCond=TA}     0.9047945  0.9085282 1.0003402  1321
[5]  {GarageCond=TA}     => {Utilities=AllPub}  0.9075342  0.9992459 0.9999307  1325
[6]  {Utilities=AllPub}  => {GarageCond=TA}     0.9075342  0.9081563 0.9999307  1325
[7]  {Electrical=SBrkr}  => {Heating=GasA}      0.9006849  0.9857571 1.0078469  1315
[8]  {Heating=GasA}      => {Electrical=SBrkr}  0.9006849  0.9208683 1.0078469  1315
[9]  {Electrical=SBrkr}  => {Condition2=Norm}    0.9041096  0.9895052 0.9997769  1320
[10] {Condition2=Norm}   => {Electrical=SBrkr}  0.9041096  0.9134948 0.9997769  1320
[11] {Electrical=SBrkr}  => {Street=Pave}       0.9102740  0.9962519 1.0003630  1329
[12] {Street=Pave}       => {Electrical=SBrkr}  0.9102740  0.9140303 1.0003630  1329
[13] {Electrical=SBrkr}  => {Utilities=AllPub}  0.9136986  1.0000000 1.0006854  1334
[14] {Utilities=AllPub}  => {Electrical=SBrkr}  0.9136986  0.9143249 1.0006854  1334
[15] {PavedDrive=Y}      => {Heating=GasA}      0.9047945  0.9858209 1.0079121  1321
[16] {Heating=GasA}      => {PavedDrive=Y}      0.9047945  0.9250700 1.0079121  1321
[17] {PavedDrive=Y}      => {RoofMatl=CompShg}  0.9000000  0.9805970 0.9983763  1314
[18] {RoofMatl=CompShg}  => {PavedDrive=Y}      0.9000000  0.9163180 0.9983763  1314
[19] {PavedDrive=Y}      => {Condition2=Norm}    0.9095890  0.9910448 1.0013324  1328
[20] {Condition2=Norm}   => {PavedDrive=Y}      0.9095890  0.9190311 1.0013324  1328
[21] {PavedDrive=Y}      => {Street=Pave}       0.9143836  0.9962687 1.0003798  1335
[22] {Street=Pave}       => {PavedDrive=Y}      0.9143836  0.9181568 1.0003798  1335
[23] {PavedDrive=Y}      => {Utilities=AllPub}  0.9171233  0.9992537 0.9999386  1339
[24] {Utilities=AllPub}  => {PavedDrive=Y}      0.9171233  0.9177519 0.9999386  1339
[25] {Functional=Typ}    => {Heating=GasA}      0.9136986  0.9808824 1.0028629  1334
[26] {Heating=GasA}      => {Functional=Typ}    0.9136986  0.9341737 1.0028629  1334
[27] {Functional=Typ}    => {RoofMatl=CompShg}  0.9171233  0.9845588 1.0024100  1339
[28] {RoofMatl=CompShg}  => {Functional=Typ}    0.9171233  0.9337517 1.0024100  1339
[29] {Functional=Typ}    => {Condition2=Norm}    0.9212329  0.9889706 0.9992367  1345
[30] {Condition2=Norm}   => {Functional=Typ}    0.9212329  0.9307958 0.9992367  1345
[31] {Functional=Typ}    => {Street=Pave}       0.9273973  0.9955882 0.9996966  1354
```

Un total de 95 reglas de asociación con soporte y confianza mayor o igual a 0.9 y conteo de apariciones en el dataset mayor a mil. Entre ellas están por ejemplo:

- Si la pendiente de la propiedad es tiene poca inclinación (Gentle slope) y el tipo de calefacción es horno de aire caliente forzado a gas (gas forced warm air furnace) entonces el material del techo es teja estándar (Standard Composite Shingle)

- Si la casa tiene un camino pavimentado entonces el techo es de teja estándar.

Reduciendo los valores de confianza y soporte a 0.75 se encuentran las siguientes:

- SI la calle está pavimentada y la funcionalidad de la casa es típica entonces la clasificación general de la zona de la venta es residencial con poca densidad.

- Si la calle esta pavimentada y se tienen todas las utilidades públicas (electricidad, gas, agua, …) entonces los acabados del sótano no están terminados.

8. **Haga un resumen de los hallazgos más importantes encontrados al explorar los datos y llegue a conclusiones sobre las posibles líneas de investigación**.

Lo primero que pudimos al realizar la exploración de los datos es la cantidad de variables numéricas y categóricas, existen 38 variables numéricas y 43 variables numéricas. Nos dimos cuenta también que los datos ya estaban limpios y consistentes.

Después, analizando las variables categóricas con gráficas de barras, frecuencias, etc, nos dimos cuenta que todas las propiedades en promedio tienen 1.57 baños completos y que el rating de la condición general promedio de las propiedades es de 5.58. Y así con muchas variables categóricas más.

Ahora bien, de las variables numéricas, pudimos encontrar que los datos se pueden separar en 10 grupos (Clustering).

Sabiendo que el objetivo final es predecir los precios de las casas utilizando el dataset, encontramos que al ver la correlaciones de SalePrice contra todas las demás variables, las variables que más influyen son: La Condición General con 0.79, El Año de Construcción con 0.52, Año de Remodelación con 0.52, Área Total del Sótano en Pies Cuadrados con 0.61, Metros Cuadrados del Primer Piso con 0.60, Superficie Habitable Por Encima del Nivel del Suelo (en pies cuadrados) con 0.70, Baños completos en el Segundo Nivel con 0.56, Habitaciones En el Segundo Nivel con 0.547, Capacidad de Carros en el Garaje con 0.64 y Tamaño del Garaje en Pies Cuadrados con 0.61.De estas 10 variables según el análisis de componentes principales, sólo 6 podrían explicar el 73% de la variabilidad de los datos.

Sabemos que el hecho de que haya correlación no indica la causalidad, es por eso que vimos también las reglas de asociación. Otro hallazgo encontrado que nos llamó mucho la atención fueron las reglas de asociación. Encontramos reglas muy interesantes y que no se pueden saber o ver a simple vista, como por ejemplo si la pendiente de la propiedad es tiene poca inclinación (Gentle slope) y el tipo de calefacción es horno de aire caliente forzado a gas (gas forced warm air furnace) entonces el material del techo es teja estándar (Standard Composite Shingle), entre muchas otras. Estás varían según el soporte y el nivel de confianza.