Laboratorio 1

Oscar Godoy - Rafael Dubois

2022-07-24

Limpieza y exploración inicial

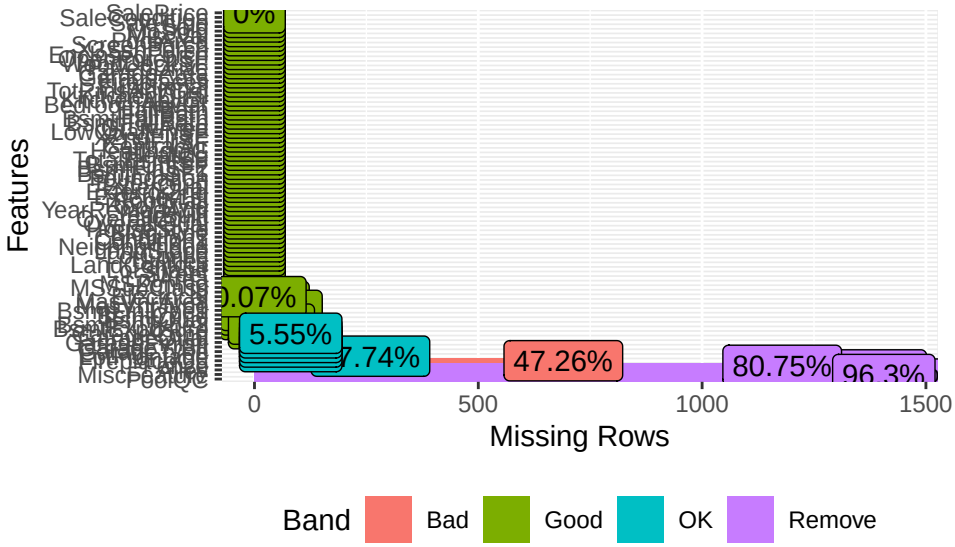
El data set contiene información sobre 1460 casas en la ciudad de Ames, Iowa. Para cada una de las casas, las cuales se identifican con un ID único, contamos con 79 variables exploratorias. Estas describen aspectos de la casa como zona, tamaño, forma, amenidades, entre otras. La variable objetivo, que se busca predecir, es el precio de venta de la casa.

```
introduce(dataOriginal)
```

```
##   rows columns discrete_columns continuous_columns all_missing_columns
## 1 1460      81               43                 38                   0
##   total_missing_values complete_rows total_observations memory_usage
## 1                 6965             0             118260       755280
```

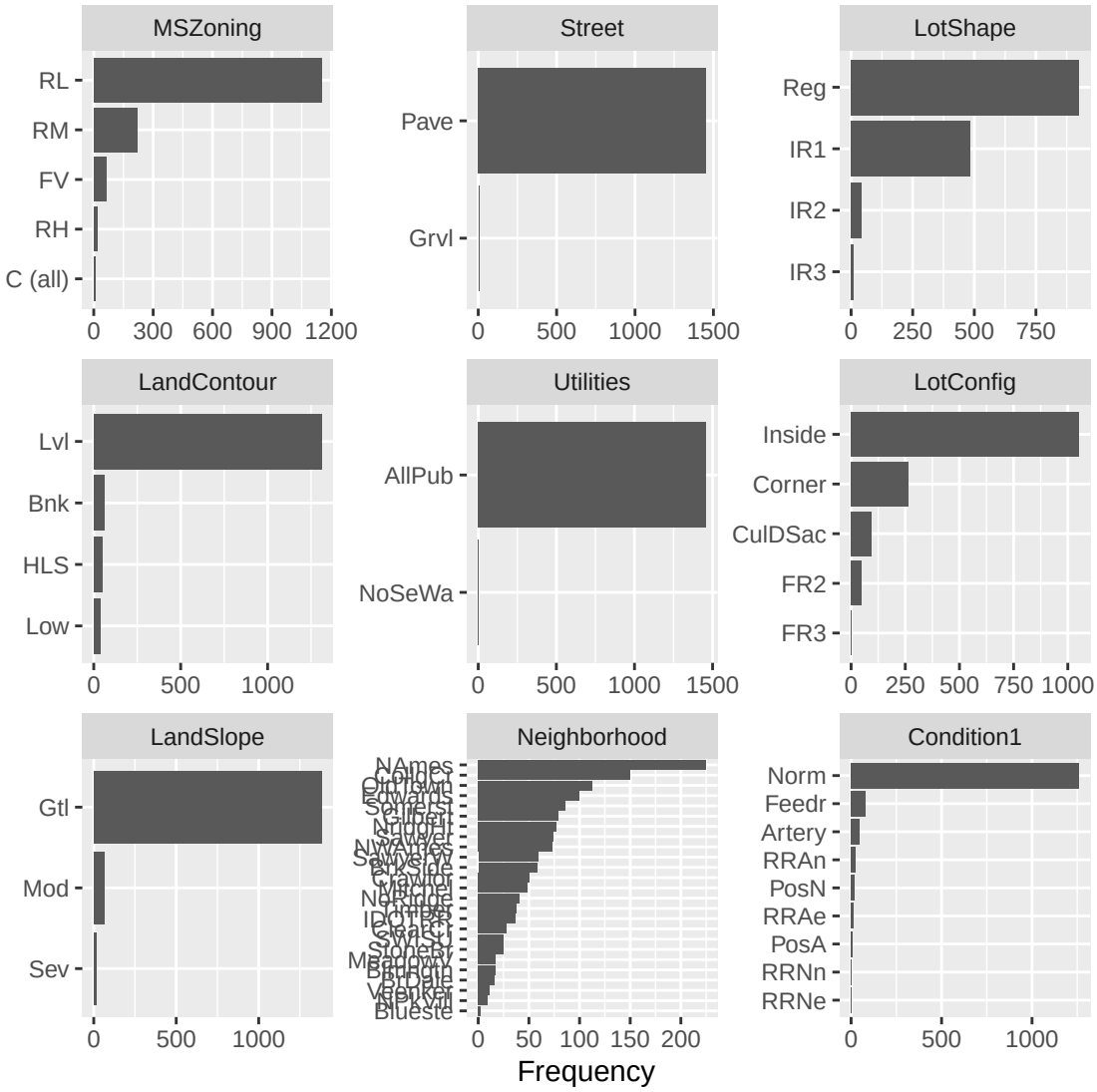
Procedemos a analizar la calidad de los datos. Vemos que hay variables con altos porcentajes de NA's, por lo que las descartamos. Estas variables tienen el factor en común de referirse a utilidades poco comunes que una casa pueda tener, como piscinas o cercas. La mayoría de las casas en el data set no posee estas características. En total, removemos las siguientes 5 variables:

- PoolQC
- MiscFeature
- Alley
- Fence
- FireplaceQu

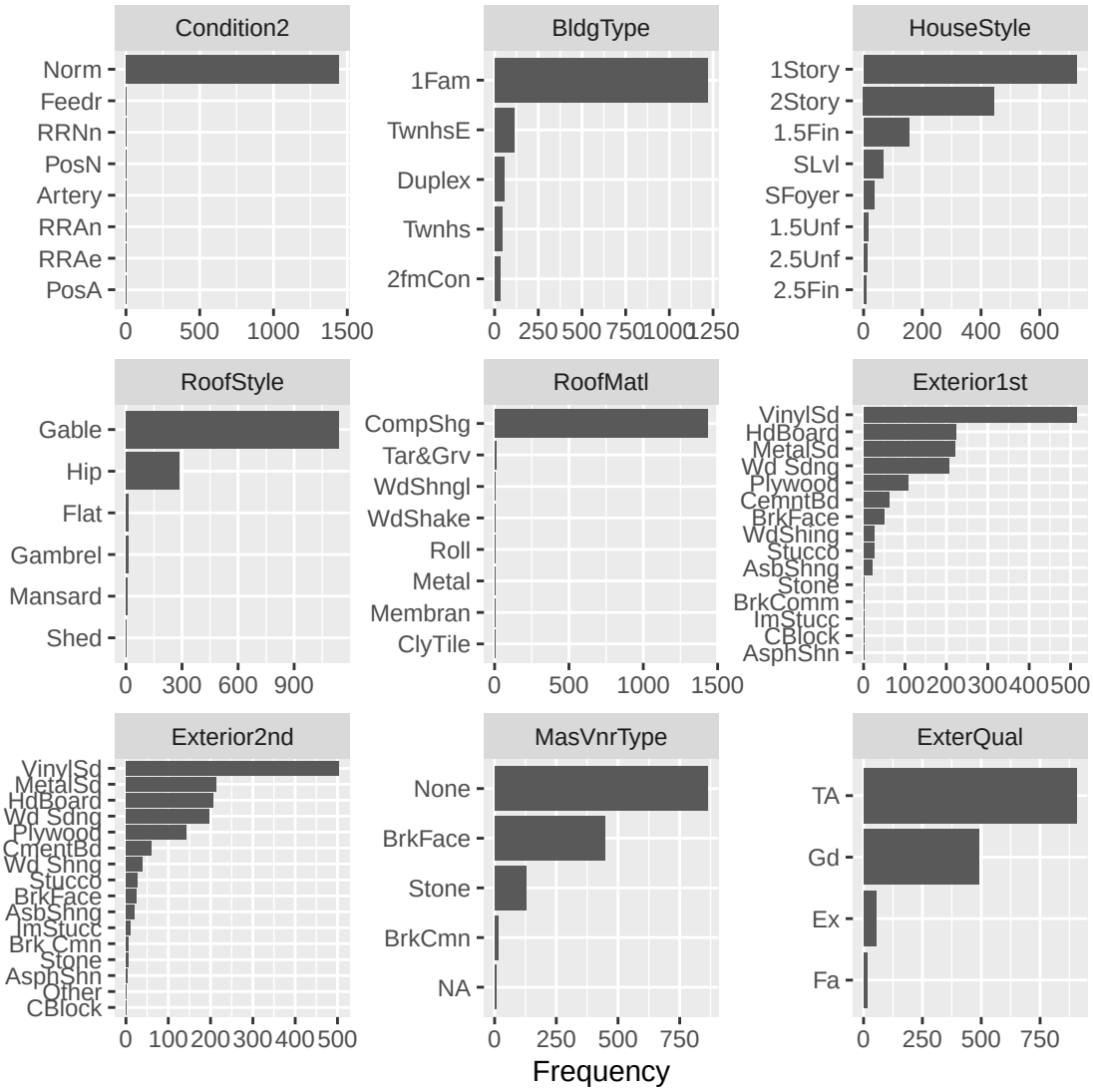


Ahora, analizamos la distribución de las variables cuantitativas y cualitativas. Descartamos las variables cualitativas que sean muy homogeneas en la poblacion. Se elige descartar aquellas variables donde una categoria concentra mas del 80% de las casas. Entonces, descartamos las siguientes 20 variables:

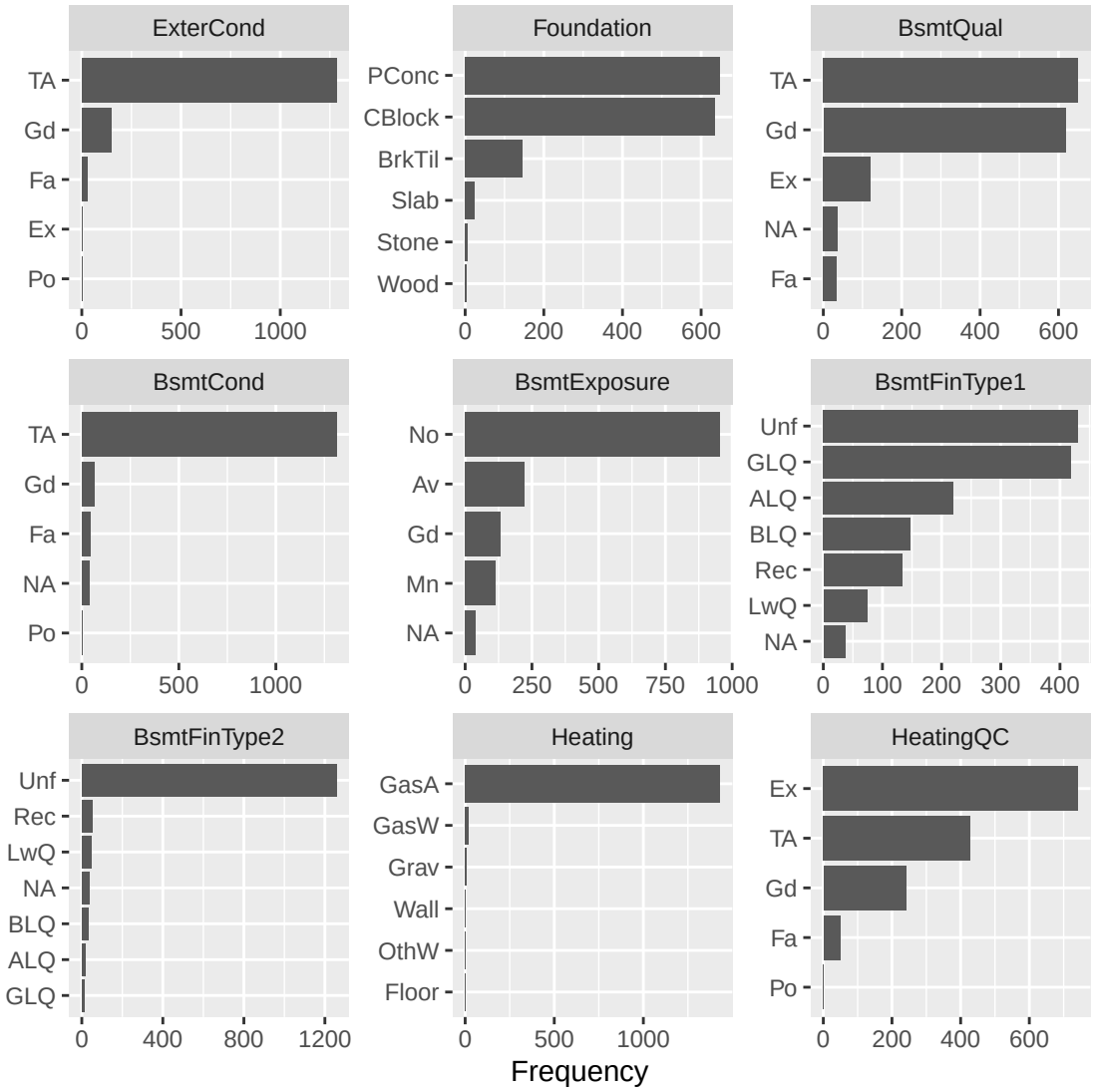
- Street
- LandContour
- Utilities
- LandSlope
- Condition1
- Condition2
- BldgType
- RoofMatl
- ExterCond
- BsmtCond
- BsmtFinType2
- Heating
- CentralAir
- Electrical
- Functional
- GarageQual
- GarageCond
- PavedDrive
- SaleType
- SaleCondition


MSZoning
RL
RM
FV
RH
C (all)
0 300 600 900 1200
Street
Pave
Grvl
0 500 1000 1500
LotShape
Reg
IR1
IR2
IR3
0 250 500 750
LandContour
Lvl
Bnk
HLS
Low
0 500 1000
Utilities
AllPub
NoSeWa
0 500 1000 1500
LotConfig
Inside
Corner
CulDSac
FR2
FR3
0 250 500 750 1000
LandSlope
Gtl
Mod
Sev
0 500 1000
Neighborhood
0 50 100 150 200
Condition1
Norm
Feedr
Artery
RRAn
PosN
RRAe
PosA
RRNn
RRNe
0 500 1000
Frequency

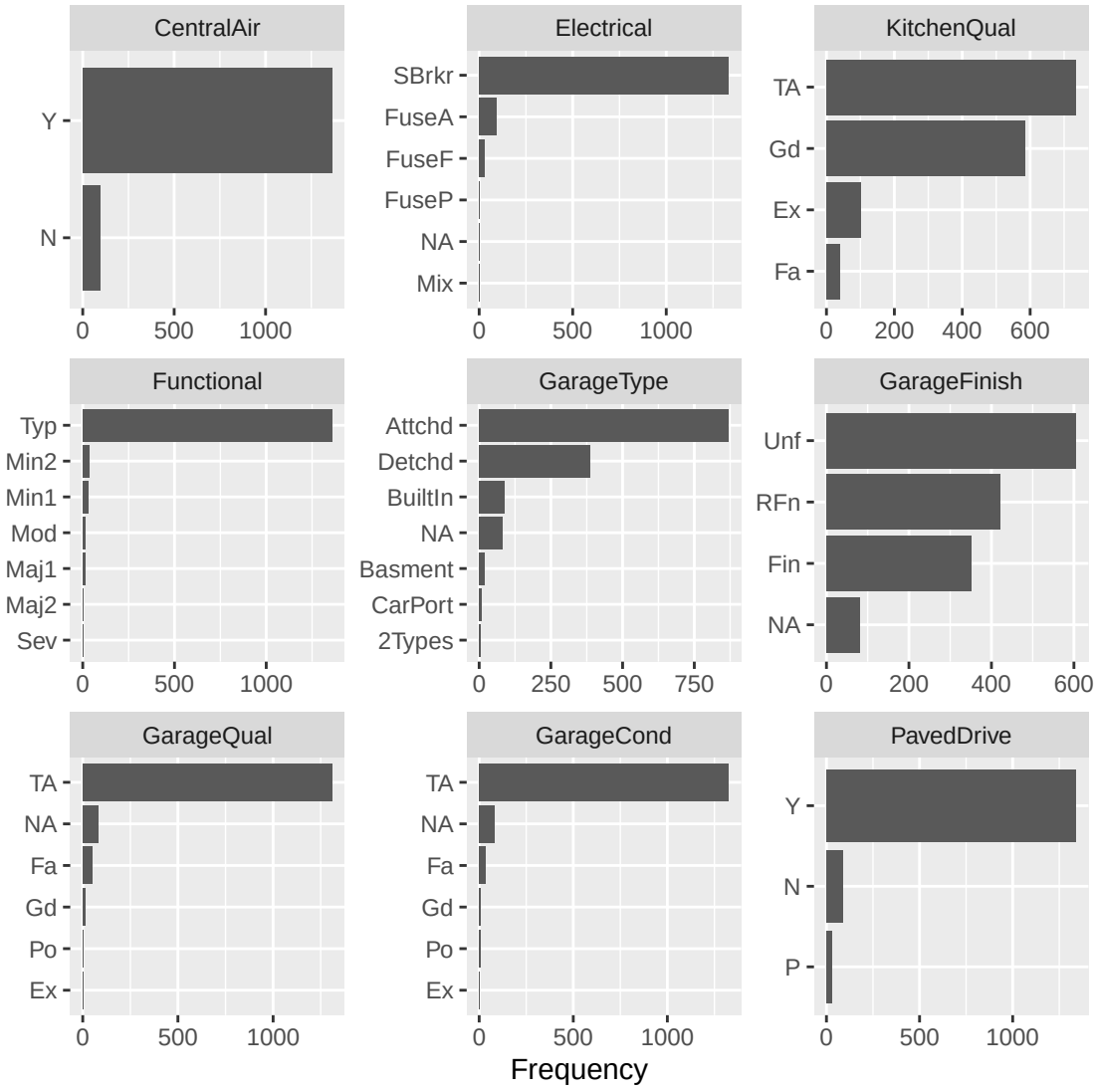
Page 1

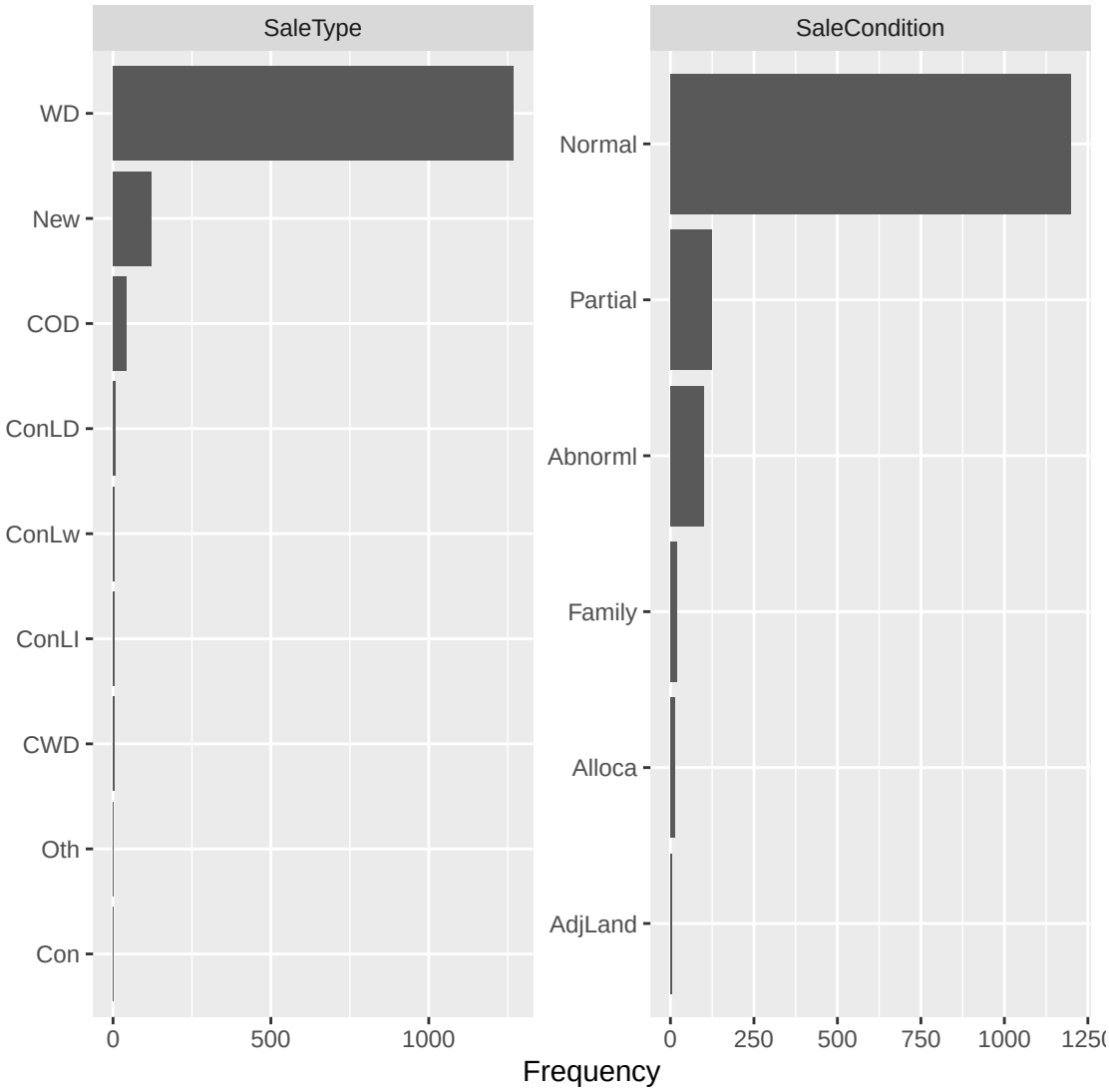

Condition2
Norm
Feedr
RRNn
PosN
Artery
RRAn
RRAe
PosA
0 500 1000 1500
BldgType
1Fam
TwnhsE
Duplex
Twnhs
2fmCon
0 250 500 750 1000 1250
HouseStyle
1Story
2Story
1.5Fin
SLvl
SFoyer
1.5Unf
2.5Unf
2.5Fin
0 200 400 600
RoofStyle
Gable
Hip
Flat
Gambrel
Mansard
Shed
0 300 600 900
RoofMatl
CompShg
Tar&Grv
WdShngl
WdShake
Roll
Metal
Membran
ClyTile
0 500 1000 1500
Exterior1st
VinylSd
HdBoard
MetalSd
Wd Sdng
Plywood
CemntBd
BrkFace
WdShing
Stucco
AsbShng
Stone
BrkComm
ImStucc
CBlock
AsphShn
0 100 200 300 400 500
Exterior2nd
VinylSd
MetalSd
HdBoard
Wd Sdng
Plywood
CmentBd
Wd Shng
Stucco
BrkFace
AsbShng
ImStucc
Brk Cmn
Stone
AsphShn
Other
CBlock
0 100 200 300 400 500
MasVnrType
None
BrkFace
Stone
BrkCmn
NA
0 250 500 750
ExterQual
TA
Gd
Ex
Fa
0 250 500 750
Frequency

Page 2


ExterCond
TA
Gd
Fa
Ex
Po
0
500
1000
Foundation
PConc
CBlock
BrkTil
Slab
Stone
Wood
0
200
400
600
BsmtQual
TA
Gd
Ex
NA
Fa
0
200
400
600
BsmtCond
TA
Gd
Fa
NA
Po
0
500
1000
BsmtExposure
No
Av
Gd
Mn
NA
0
250
500
750
1000
BsmtFinType1
Unf
GLQ
ALQ
BLQ
Rec
LwQ
NA
0
100
200
300
400
BsmtFinType2
Unf
Rec
LwQ
NA
BLQ
ALQ
GLQ
0
400
800
1200
Heating
GasA
GasW
Grav
Wall
OthW
Floor
0
500
1000
HeatingQC
Ex
TA
Gd
Fa
Po
0
200
400
600
Frequency

Page 3

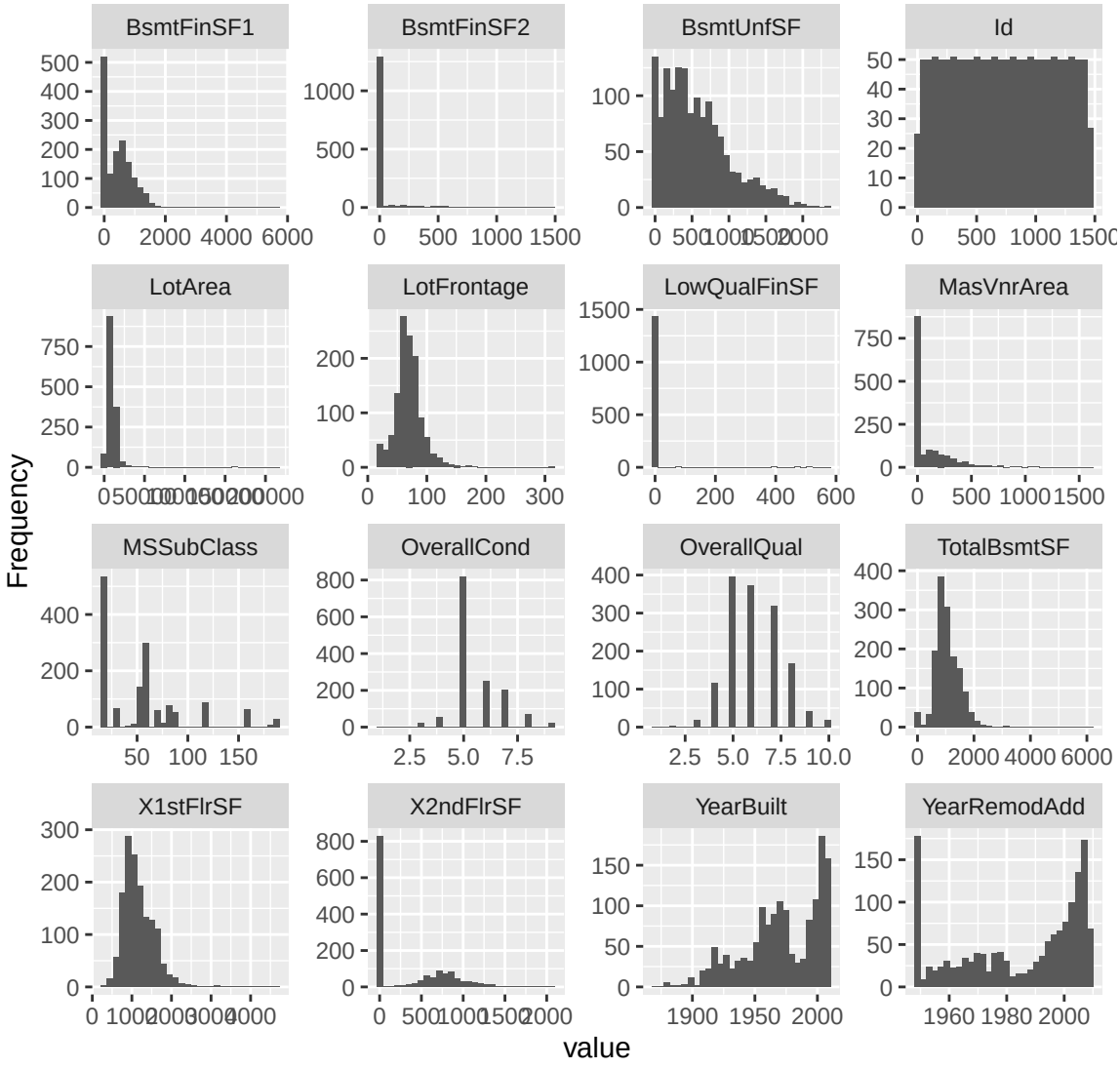

CentralAir
Y
N
0
500
1000
Electrical
SBrkr
FuseA
FuseF
FuseP
NA
Mix
0
500
1000
KitchenQual
TA
Gd
Ex
Fa
0
200
400
600
Functional
Typ
Min2
Min1
Mod
Maj1
Maj2
Sev
0
500
1000
GarageType
Attchd
Detchd
BuiltIn
NA
Basment
CarPort
2Types
0
250
500
750
GarageFinish
Unf
RFn
Fin
NA
0
200
400
600
GarageQual
TA
NA
Fa
Gd
Po
Ex
0
500
1000
GarageCond
TA
NA
Fa
Gd
Po
Ex
0
500
1000
PavedDrive
Y
N
P
0
500
1000
Frequency

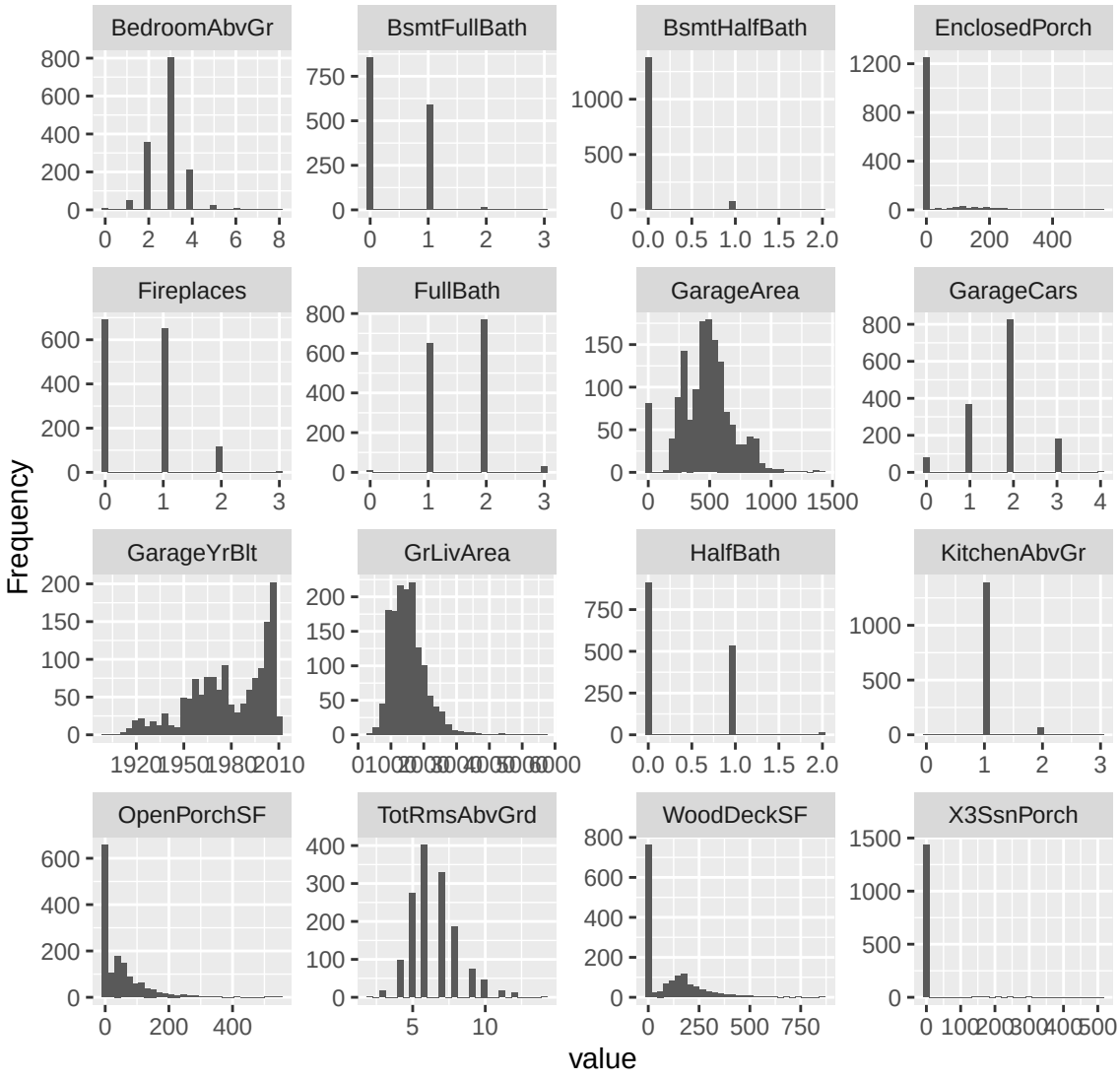

SaleType
WD
New
COD
ConLD
ConLw
ConLI
CWD
Oth
Con
0
500
1000
SaleCondition
Normal
Partial
Abnorml
Family
Alloca
AdjLand
0
250
500
750
1000
Frequency

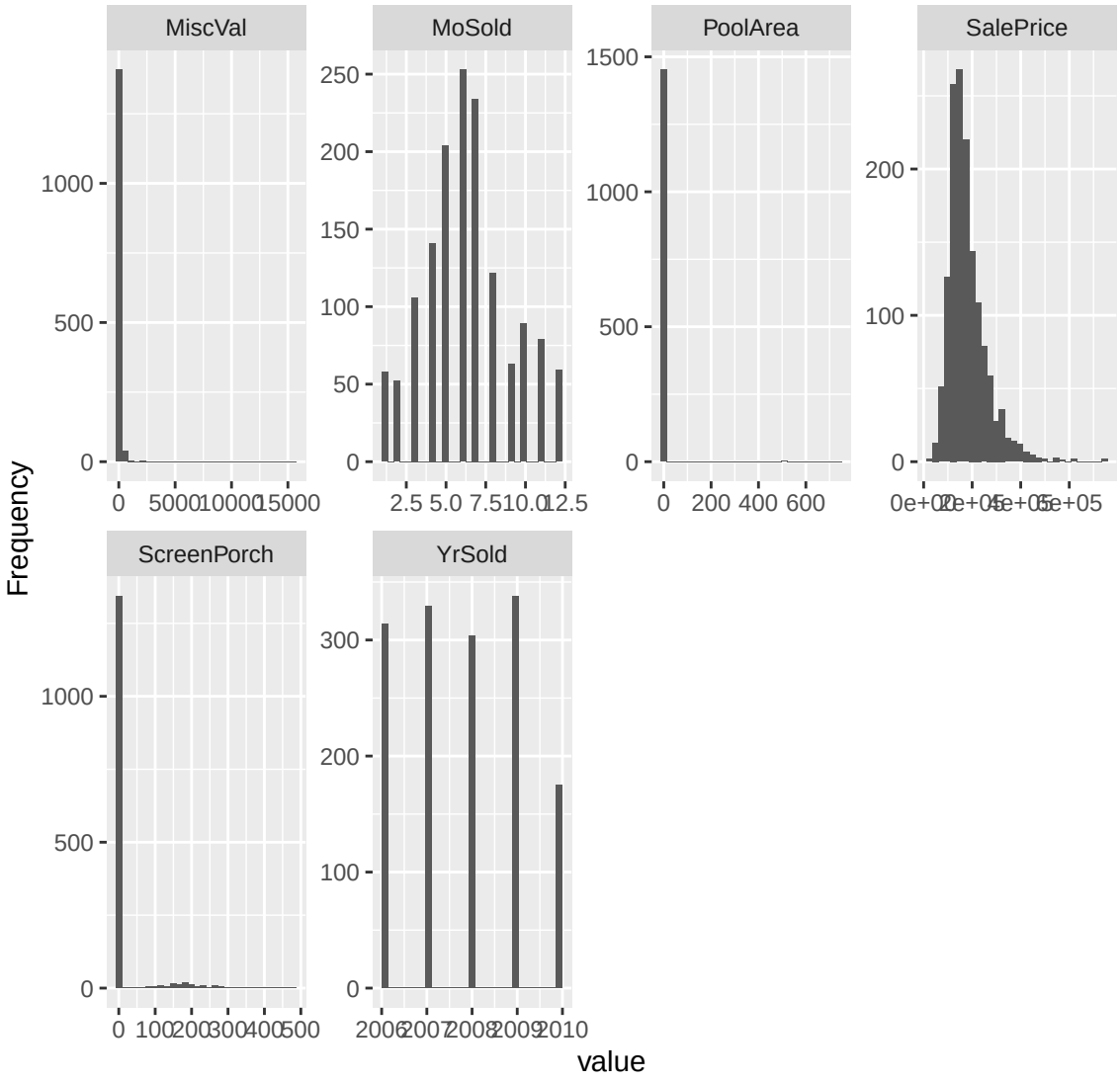
Page 5

De igual manera, descartamos las variables cuantitativas que esten muy sesgadas. Descartamos aquellas variables donde al menos el 80% de los datos asuman un mismo valor. Encontramos 9 de estas variables:

- BsmtFinSF2
- LotArea
- LowQualFinSF
- BsmtHalfBath
- EnclosedPorch
- KitchenAbvGr
- X3SsnPorch
- MiscVal
- PoolArea
- ScreenPorch




BedroomAbvGr
BsmtFullBath
BsmtHalfBath
EnclosedPorch
Fireplaces
FullBath
GarageArea
GarageCars
GarageYrBlt
GrLivArea
HalfBath
KitchenAbvGr
OpenPorchSF
TotRmsAbvGrd
WoodDeckSF
X3SsnPorch
Frequency
value



Finalmente, contamos con 46 variables cualitativas y cuantitativas que nos proponemos a investigar.

```
introduce(dataOriginal)
```

```
##   rows columns discrete_columns continuous_columns all_missing_columns
## 1 1460      46               18                 28                   0
##   total_missing_values complete_rows total_observations memory_usage
## 1                  630          1096              67160       391160
```

Análisis de Componentes Principales

Filtramos las variables numéricas para realizarles un PCA. Luego, de las variables restantes descartamos aquellas que no sean continuas. Variables discretas y ordinales se evitan. Además, eliminamos la variable SalePrice para solo analizar las variables exploratorias. Las variables descartadas son las siguientes:

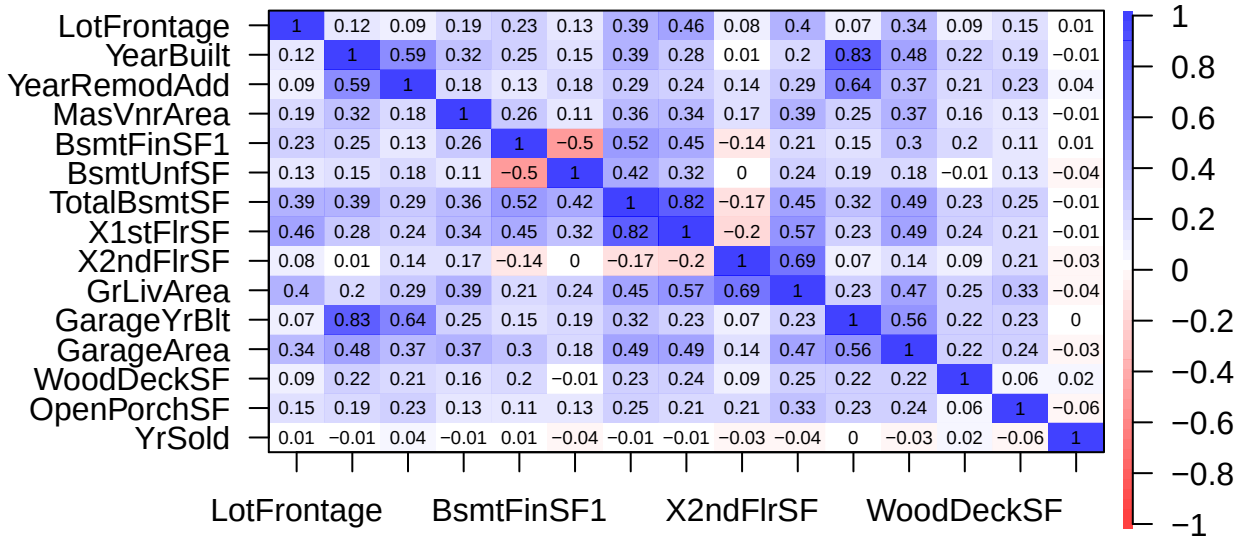
- Id
- MSSubClass
- OverallQual
- OverallCond
- BsmtFullBath
- FullBath
- HalfBath
- BedroomAbvGr
- TotRmsAbvGrd
- Fireplaces
- GarageCars
- MoSold

Después de la limpieza, quedamos con 16 variables continuas. Deseamos reducir la dimensionalidad de este data set empleando PCA. Primero, creamos la matriz de correlación de este set de datos, y notamos que su determinante es muy cercano a 0. Esto puede evidenciarse que hay bastante correlación entre las matrices, es decir, dependencia lineal entre las variables originales.

```
rcor <- cor(dataPCA, use = "pairwise.complete.obs")
det(rcor)
```

```
## [1] 6.016125e-06
```

```
cor.plot(rcor)
```



Para asegurarnos de que PCA es aplicable, realizamos el test de esfericidad de Bartlett. El p-valor es prácticamente 0, por lo que descartamos la hipótesis nula del test y concluímos que la matriz de correlación de las variables es distinta a la identidad.

```
cortest.bartlett(dataPCA)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 17468.61
##
## $p.value
## [1] 0
##
## $df
## [1] 105
```

Con estos requisitos cumplidos, procedemos a realizar un PCA:

```
compPrinc <- prcomp(na.omit(dataPCA[,-1]), scale = T)
summary(compPrinc)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.1079 1.3313 1.2652 1.2097 1.00692 0.94551 0.91814
## Proportion of Variance 0.3174 0.1266 0.1143 0.1045 0.07242 0.06386 0.06021
## Cumulative Proportion  0.3174 0.4440 0.5583 0.6628 0.73526 0.79912 0.85933
##                           PC8     PC9    PC10    PC11    PC12   PC13    PC14
## Standard deviation     0.80057 0.72468 0.61581 0.48451 0.38070 0.2048 0.04973
## Proportion of Variance 0.04578 0.03751 0.02709 0.01677 0.01035 0.0030 0.00018
## Cumulative Proportion  0.90511 0.94262 0.96971 0.98648 0.99683 0.9998 1.00000
```