UNIVERSIDAD DEL VALLE DE GUATEMALA
Data Science
Lynette García

Laboratorio 1

Análisis exploratorio, Clustering y PCA

Jonathan Álvarez, 15842

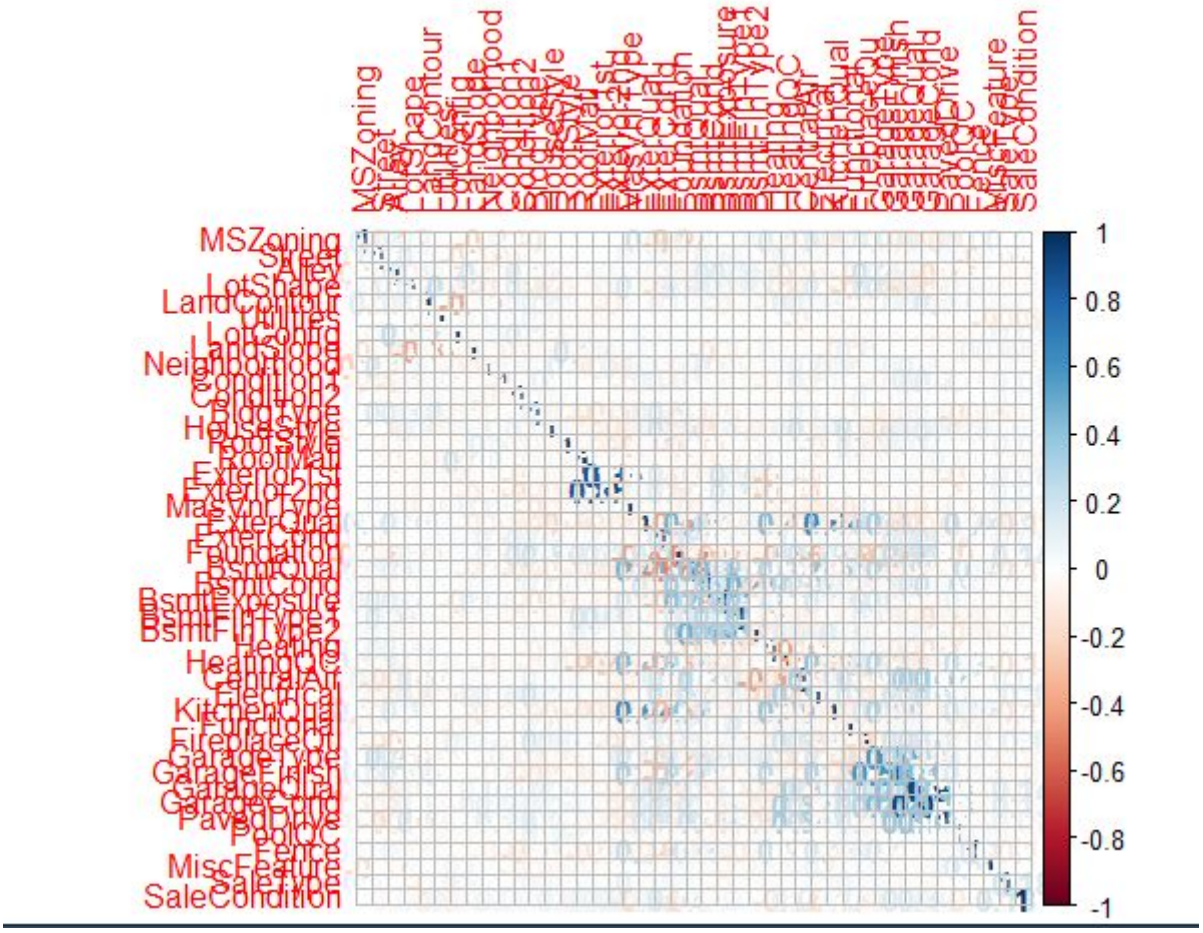
GUATEMALA, 8 de agosto de 2018

1) En el data encontramos un total de 1460 registros con 81 variables de diferentes tipos. En el cual se describen varias características que son tomadas en cuenta para colocar el precio de venta de las casas. Además, se cuenta con un dataset de testeo, el cual contiene el precio de venta de las casas. Con el fin de poder comparar el precio estimado contra el precio real.

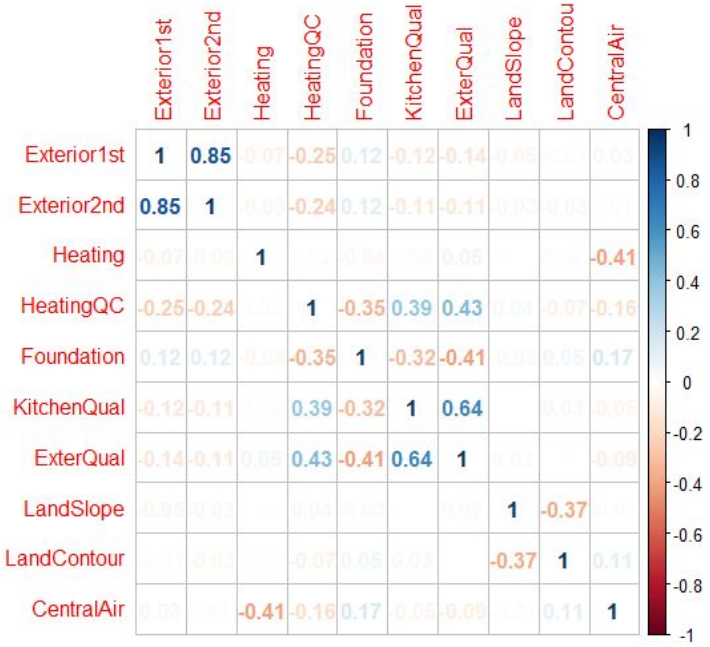
2) En el dataset encontramos variables cuantitativas y variables cualitativas. En las cuantitativas encontramos variables continuas, tales como las variables que nos indican el precio; y también variables discretas, como la cantidad de baños, pisos o el año de venta de la casa. En la siguiente tabla se puede observar el tipo de cada variable:

Cuantitativas	**Cualitativa**
"Id""MSSubClass" "LotFrontage" "LotArea" "OverallQual" "OverallCond" "YearBuilt" "YearRemodAdd" "MasVnrArea" "BsmtFinSF1" "BsmtFinSF2" "BsmtUnfSF" "TotalBsmtSF" "X1stFlrSF" "X2ndFlrSF" "LowQualFinSF" "GrLivArea" "BsmtFullBath" "BsmtHalfBath" "FullBath" "HalfBath" "BedroomAbvGr" "KitchenAbvGr" "TotRmsAbvGrd" "Fireplaces" "GarageYrBlt" "GarageCars" "GarageArea" "WoodDeckSF" "OpenPorchSF" "EnclosedPorch" "X3SsnPorch" "ScreenPorch" "PoolArea" "MiscVal" "MoSold" "YrSold" "SalePrice"	"MSZoning" "Street" "Alley" "LotShape" "LandContour" "Utilities" "LotConfig" "LandSlope" "Neighborhood" "Condition1" "Condition2" "BldgType" "HouseStyle" "RoofStyle" "RoofMatl" "Exterior1st" "Exterior2nd" "MasVnrType" "ExterQual" "ExterCond" "Foundation" "BsmtQual" "BsmtCond" "BsmtExposure" "BsmtFinType1" "BsmtFinType2" "Heating" "HeatingQC" "CentralAir" "Electrical" "KitchenQual" "Functional" "FireplaceQu" "GarageType" "GarageFinish" "GarageQual" "GarageCond" "PavedDrive" "PoolQC" "Fence" "MiscFeature" "SaleType" "SaleCondition"

3) Se realizó una correlación de “todas contra todas” las variables cuantitativas en el dataset (a excepción del Id), con el fin de tener un análisis preliminar de la posible relación entre cada una de las variables:



Hay que resaltar, que en la diagonal se encuentra la mayor correlación entre las variables, esto es debido a que las variables están siendo comparadas con ellas mismas, por lo que no se debe tomar en cuenta esta información. Luego podemos observar, gracias a la escala de colores, que las variables que presentan mayor correlación entre ellas son las que se representan en un tono azul.



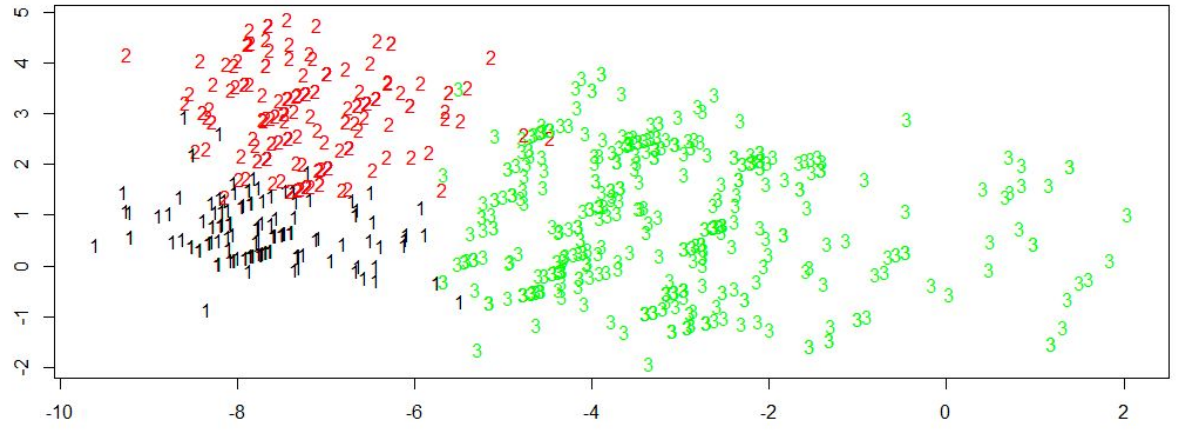
Posteriormente se seleccionó las variables que presentaran un coeficiente de correlación mayor a 0.50 (50%) en la matriz de correlación. Con lo que se obtuvo que existe una correlacion en:

- HeatingQC - Foundation
- Exterior1 - Exterior2
- Heating - Central air
- Foundation - ExterQual

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

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

6) Al realizar el clustering se pudieron identificar 3 grupos característicos, los cuales indican que el precio de una casa depende de factores como la calidad de la construcción, la calidad del aire, mejor calefacción, cuando se tiene dos jardines y la longitud de estos es amplia, también si la cocina presenta estar en buen estado.



7) Entre los hallazgos encontrados en el análisis exploratorio, se encontraron primeramente que existen muchas variables cualitativas que pueden presentar información valiosa para el modelo que se está buscando. Por lo que sería recomendable realizar un análisis con estadística no paramétrica para entender mejor la forma de los datos y cómo pueden afectar en el precio de compra de un casa.

Mientras que con las variables cuantitativas, se encontró que algunas de estas tienen un coeficiente de correlación bastante pequeño, a comparación del que se esperaba. Generalmente se espera que este sea mayor a un 80% para que el modelo sea representativo.