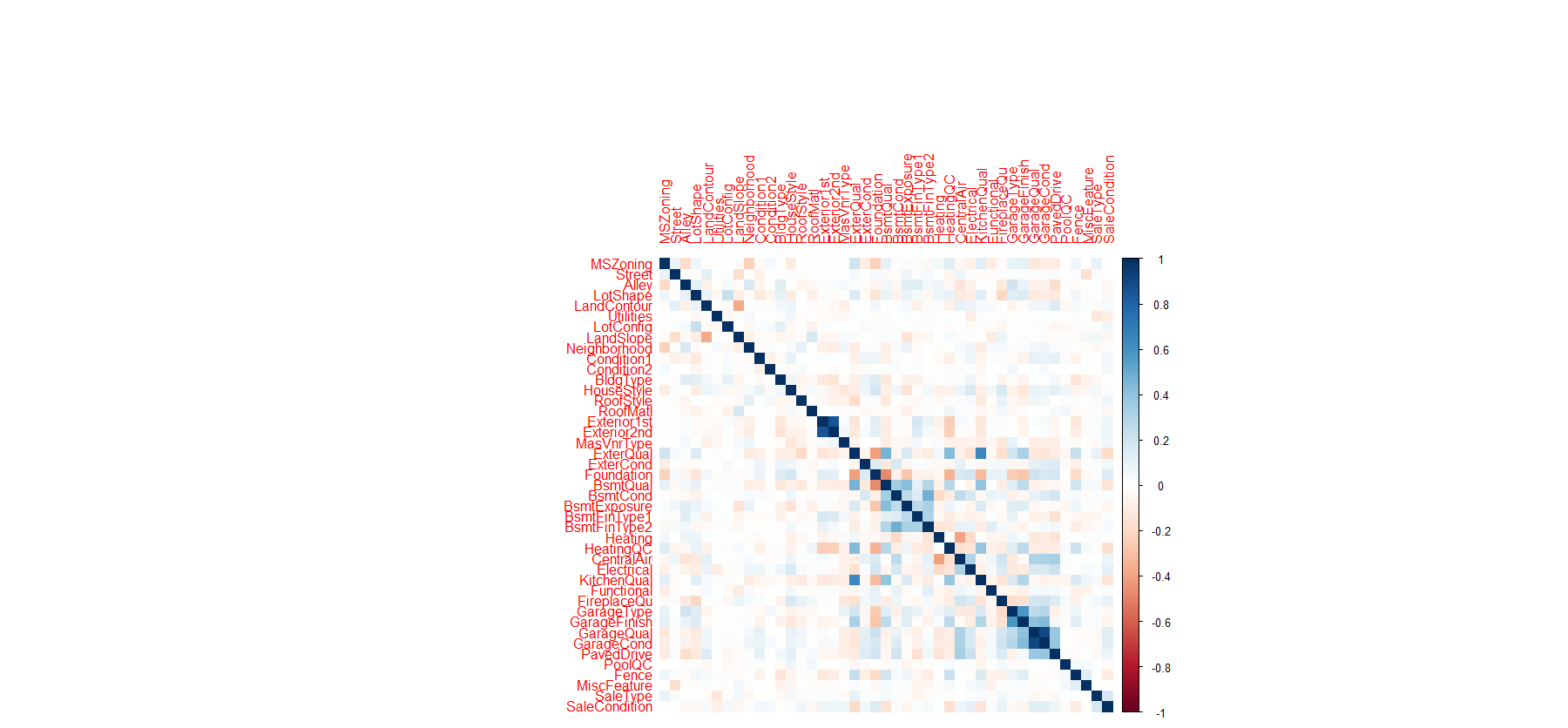
|  |  |
| --- | --- |
| Universidad del Valle de Guatemala | Julio Francisco Barahona Martínez |
| Data Science | 141206 |
| Lynette García | 8 de agosto de 2018 |

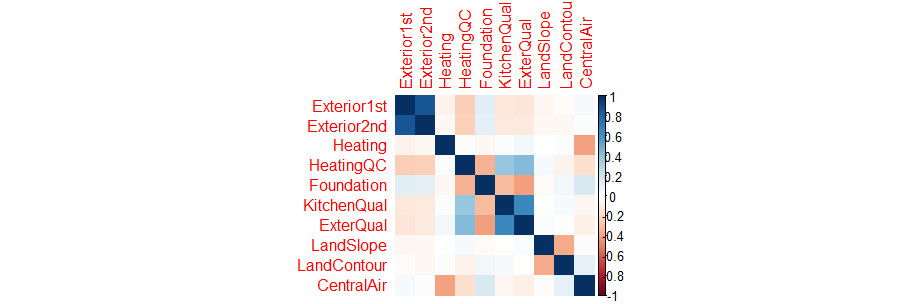
Laboratorio #1: Análisis exploratorio, clustering y PCA

1. Se tiene una base de datos para entrenamiento de 81 variables y 1500 observaciones (aproximadamente). Se tiene una base de datos para testeo de 1500 observaciones también, pero con la variable adicional de precio. No se tiene ningún caso completo
2. Tipos de variables:

|  |  |
| --- | --- |
| Tipo | Variable |
| Cualitativa o Categórica | ID, MSSubClass , MSSubClass, MSZoning , Street, Alley, LotShape, SaleCondition, SaleType, MiscFeature, PoolQC, MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, PavedDrive, GarageCond, GarageQual, GarageFinish, GarageType, FireplaceQu, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, Electrical, KitchenQual, Functional |
| Cuantitativa Discreta | LotFrontage, LotArea, MiscVal, LotFrontage, LotArea, PoolArea, ScreenPorch, 3SsnPorch, EnclosedPorch, OpenPorchSF, WoodDeckSF, GarageArea, GarageCars, YrSold, MoSold , Fence, GarageCars, GarageYrBlt, Fireplaces, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, TotRmsAbvGrd |

1. Análisis correlación variables categóricas:





* Como se puede ver hay alta correlación entre:
  + Exterior1 y Exterior2
  + Heating y Central air
  + Foundation y ExterQual
  + HeatingQC y Foundation
  + KitchenQual y ExterQual
  + LandSlope y LandContour

1. 4 Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o  
   cualquier otra técnica que le permita explorar los datos
2. Haga un análisis de componentes principales, interprete los componentes
3. Haga un análisis de clustering, describa los grupos.
4. Haga un resumen de los hallazgos más importantes encontrados al explorar los datos y llegue a  
   conclusiones sobre las posibles líneas de investigación.