Universidad del Valle de Guatemala
Data Science
Lynette García

Lab no.1 Data Science

Jorge Eduardo Súchite Martínez
Carnet 15293

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

Exploración Dataset train del laboratorio obtuvo 1460 objetos diferentes y 81 variables donde se tienen todas las especificaciones por la cual pasa una casa para asignarle un precio justo y predictivo. Este programa tiene en consideración cada detalle de la casa para luego crear otra dataset donde te guarda el precio que el algoritmo le dio el precio a la casa que estás considerando vender.

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

Cualitativa

Existen variables que muestran las cualidades de las casas, si tiene uno o dos patios; calle en la que está, garage, etc para poder venderlo.

Categórica

Existe una tabla donde se le asigna un número donde califica el acabado y los detalles de la casa al igual, otra donde se califica el estado, tipo y estilo de la casa.

Una variable donde asigna , zona urbana o rural; al igual dónde asigna calle, avenida, zona , ciudad y condado.

Si la casa queda cerca de diferentes puestos de negocios para las facilidades.

Otra donde define qué autopista tiene más cerca de ella

Además, otras variables donde se cualifica tipo de terraza; material de la terraza; si la terraza está cubierta o no; si está cubierta con dos materiales distintos.

También, existen categorías donde se califica a la baranda y su tipo de material junto con sus dimensionales.

Una categoría donde se califica qué tipo de material se tiene en los exteriores.

Una variable categórica donde se define qué tipo de fundición y la altura que esta tiene con fines de ponerle precio. al igual , la condición de la base y hasta donde termina en su base con respecto al área de la propiedad.

Una variable donde decide si el sótano está terminado o no, si está terminado lo agrega a la variable tipo 1 y decide ver en qué categoría está el sótano en calidad o confort, La variable tipo 2 es donde no está terminado el sótano y de qué material tiene hecho el sótano.

Una variable donde se define el tipo de calefacción y otra donde se define a calidad de la instalación de dicha calefacción.además, confiere si tiene un sistema de calefacción central o no.

Una variable donde define el tipo de instalaciones eléctricas y la calidad de la misma.

Luego, dos variables donde asigna si la casa tiene uno o más pisos; juntamente con la calidad que estos pisos tienen; si los pisos tienen uno o más baños privados o baños compartidos o si los cuartos no tienen baños.

Otra donde si se tiene una cocina en los pisos, al igual que su calidad; otra donde a alidad de los cuartos y la cantidad que se tienen en los diferentes pisos.

Si esta tiene chimenea o no y la calidad de la misma. Otra variable donde la locación del garage influye en el precio de la casa al igual, su calidad y la capacidad que tiene para resguardar automóviles y sus dimensiones de espacio, juntamente con el tiempo en el que el garage fue construido y si el interior del garaje está terminado o le faltan algunos acabados. Se considera también si la entrada al garaje está pavimentado o no.

Si se tiene un recubrimiento en la casa de madera y su tamaño. También, una variable donde le da una calificación de la entrada y salida de la casa y cuántas entradas se tiene en dicha casa y el tipo de fachada que tienen.

Se define si se tiene una piscina y si la tiene se le categoriza el tipo y las dimensiones que tiene al igual que su calidad.

Si se tiene cerca en la casa definir de qué material está hecha y la calidad que tiene de por sí.

Se tiene una variable categórica donde se agrega algo que al evaluador de casas no considera como prioritario y la agrega a la variable de " Miscelánea de agregados que tiene la casa que en otras categorías no pueden estar". POr ejemplo, si tiene elevador; dos garajes; si tiene pergola; cancha de tenis para luego evaluar la variable miscelánea.

Cuantitativa

Luego de categorizar en pocas palabras la casa existen variables como ID que se le asigna a cada casa que ingresa a la base de datos para que sí, se pueda tener un orden al igual que la variable precio la cual da el valor en dinero a la casa.

Una variable donde se define en qué forma fue vendida por ejemplo si con efectivo, crédito, por mensualidades.

Una variable donde define la condición de la venta de la casa por ejemplo, venta normal o si fue una casa que fue hipotecada y el nuevo integrante lo compra y si la casa es comprada por una o más personas, etc.

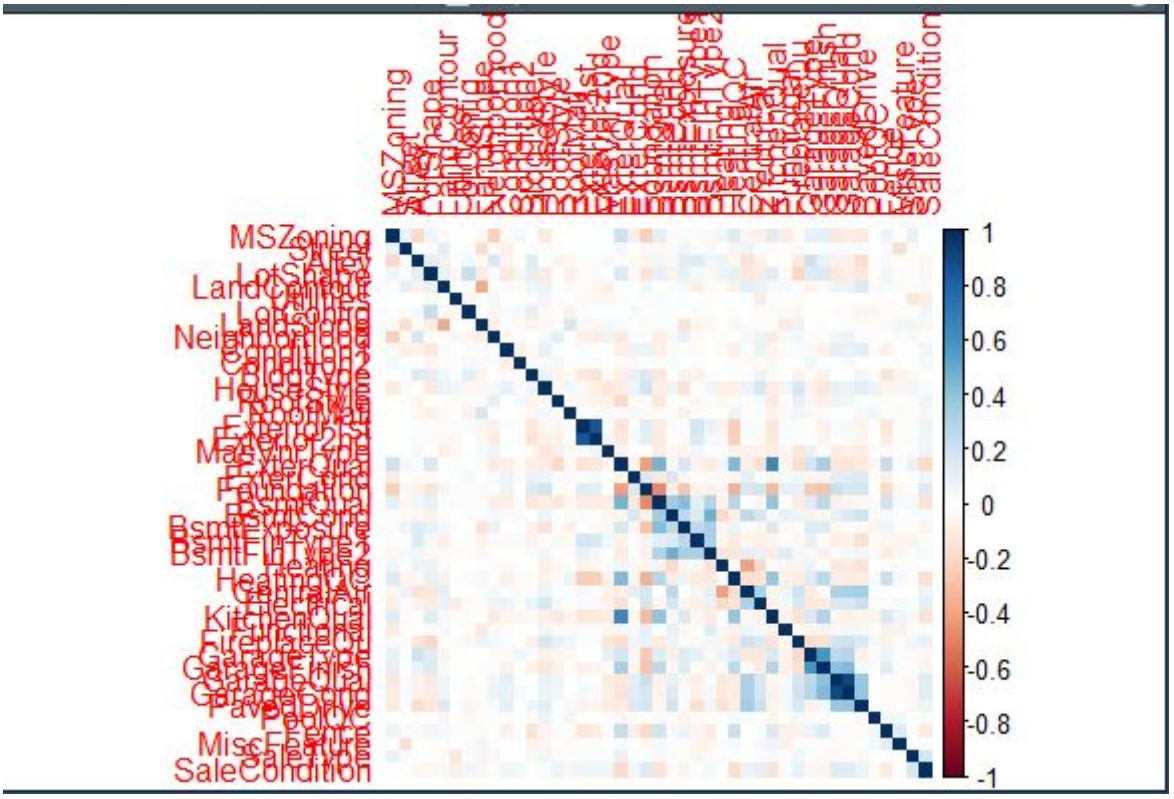
Continua

Existe una variable donde se define en qué año se vendió la casa
Existe una variable donde se define en qué mes fue vendida la casa para saber si fue en temporada de compras o no.

Cuantitativa Discreta

Número de cuartos
Número de baños
Número de cocinas
Número de pisos
Número de garages
Número de jardines

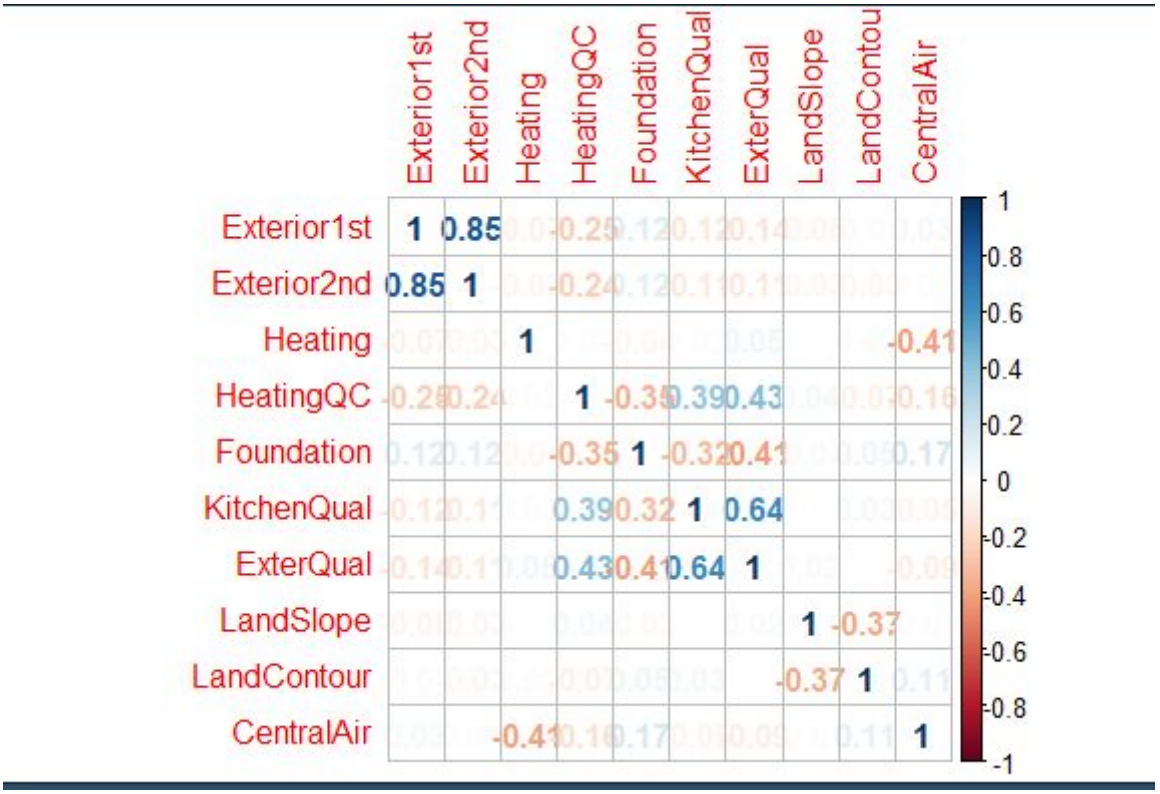
3. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.



	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope
MSZoning	1.0000000000	0.087653854	-0.191736477	6.188727e-02	-0.017854136	-0.0011920336	-9.895037e-03	-0.02205503
Street	0.0876538539	1.000000000	0.015611498	-1.022399e-02	0.115995172	0.0016817668	1.396030e-02	-0.17935952
Alley	-0.1917364773	0.015611498	1.000000000	1.001070e-01	-0.077078021	-0.0063624371	4.706776e-02	-0.02151808
LotShape	0.0618872737	-0.010223991	0.100107029	1.000000e+00	0.085434489	-0.0361006797	2.211018e-01	-0.09995091
LandContour	-0.0178541358	0.115995172	-0.077078021	8.543449e-02	1.000000000	0.0082380300	-2.552735e-02	-0.37426717
Utilities	-0.0011920336	0.001681767	-0.006362437	-3.610068e-02	0.008238030	1.0000000000	-3.258930e-02	-0.00590928
LotConfig	-0.0098950374	0.013960299	0.047067764	2.211018e-01	-0.025527354	-0.0325893034	1.000000e+00	-0.00725611
LandSlope	-0.0220550393	-0.179359521	-0.021518089	-9.995092e-02	-0.374267174	-0.0059092853	-7.256118e-03	1.00000000
Neighborhood	-0.2333799147	-0.012902538	0.166000694	-2.602994e-02	-0.002013160	0.0482143941	-3.327017e-02	-0.08026618
Condition1	-0.0278738156	-0.071657455	-0.057591828	-1.150033e-01	0.024800960	-0.0009500550	2.145658e-02	-0.01676169
Condition2	0.0446060724	0.002038937	-0.007713679	-4.376767e-02	-0.016185045	-0.0008309651	3.386811e-02	-0.02632154
BldgType	0.0056904708	-0.018243035	0.139717963	1.162622e-01	0.051142773	-0.0107781323	1.072295e-01	-0.05358235

Se eligió entre todas las variables las que fueran mayores o que estuvieran en el rango de <-0.4 y >0.4 puesto que esas son las que tienen un mayor porcentaje.

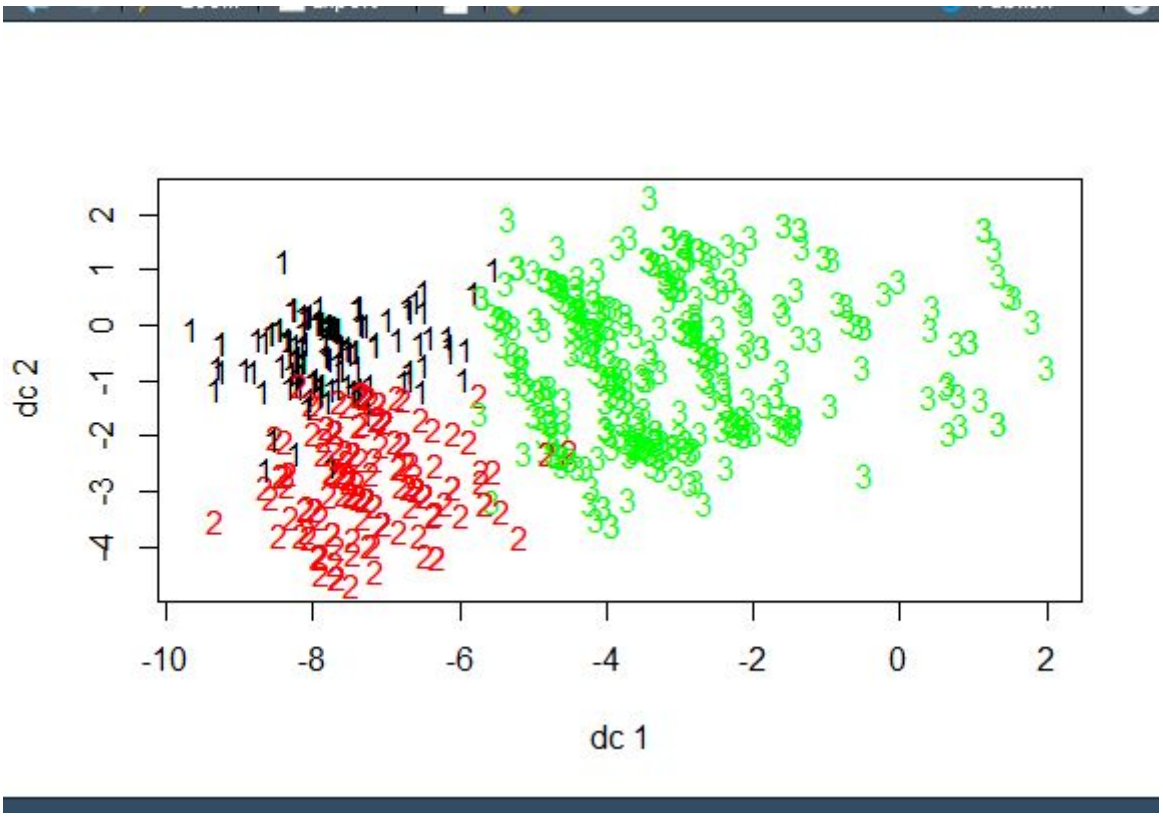
para luego plotearlas y tener una mejor vista de los datos



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. Haga un análisis de clustering, describa los grupos.



Se consideró toda las tablas que tengan un porcentaje mayor a 0.40 y -0.40

Las personas que quieren comprar casas de esta base de datos tienen como prioridades:

- La casa tiene un peso mayor cuando se tiene dos jardines
- La casa tiene un peso mayor cuando la calidad de los jardines o exteriores es muy buena juntamente con que la cocina sea de buena calidad.
- La casa tiene un peso mayor con la calidad de calefacción
- La casa tiene un peso mayor si esta está en un espacio plano
- La casa tiene un peso mayor si esta tiene una calefacción central y que tenga buena calidad
- La casa tiene un peso mayor según la calidad de la fundición

Nótese que los resultados son una predicción y se asume que este caso es un caso controlado

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