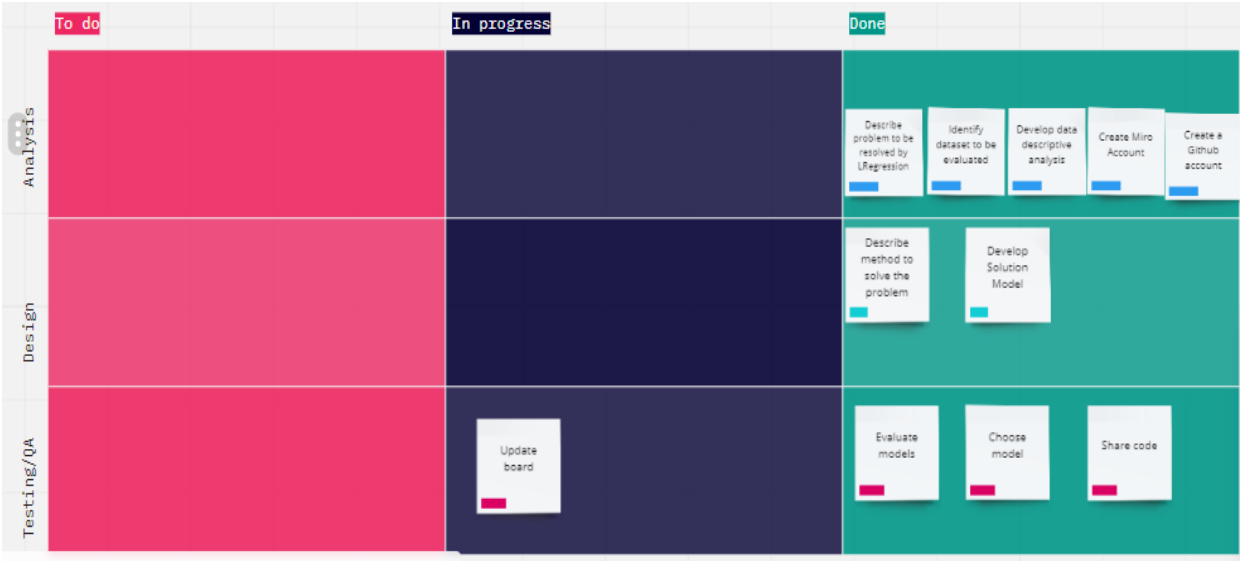
Ciencia de Datos
Analitica Descriptiva III

Asesor: Ana Gabriela Tavarez

Antonio Cabrera Diaz

https://github.com/acabreradiaz/Analitica_III_DataScience

https://miro.com/app/board/o9J_lu1BExU=/



Proyecto Matrizx para iniciar el análisis de datos

Usaremos como muestra la data de Aseguradoras de Salud

El problema consite en determinar el costo de los servicios de salud futuros y determinar los factores de riesgo que inciden en el costo facturado por servicios y su distribución ponderada.

Import dataset

```
print("Import dataset")
#path ='dataset/'
df = pd.read_csv('C:\\Users\\LENOVO\\Downloads\\insurance.csv')
print('\nNumber of rows and columns in the data set: ',df.shape)
print('')
```

Number of rows and columns in the data set: (1338, 7)

Exists m=1338 training examples and n=7 independent variables. –

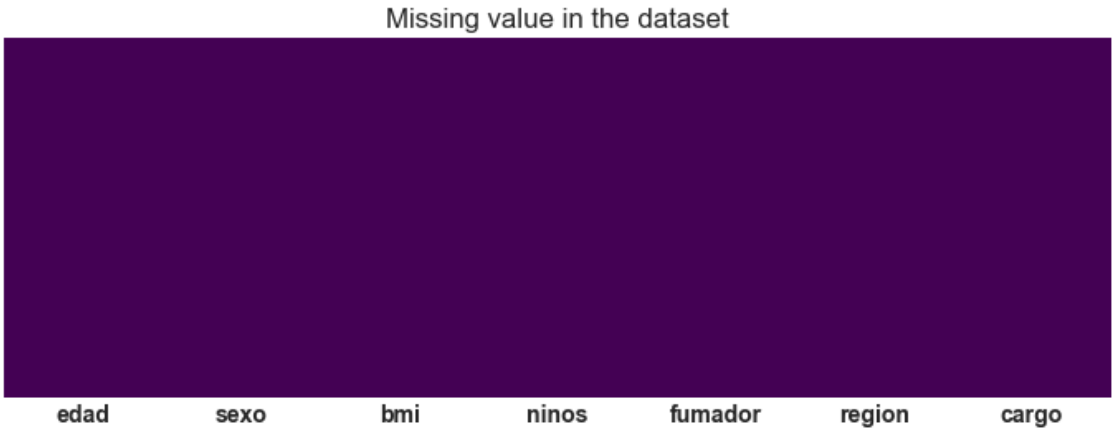
Conociendo la Data - Data Desc

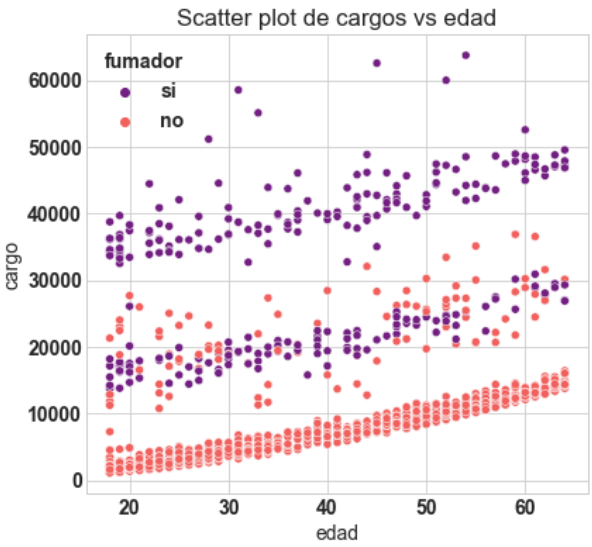
<class 'pandas.core.frame.DataFrame'>

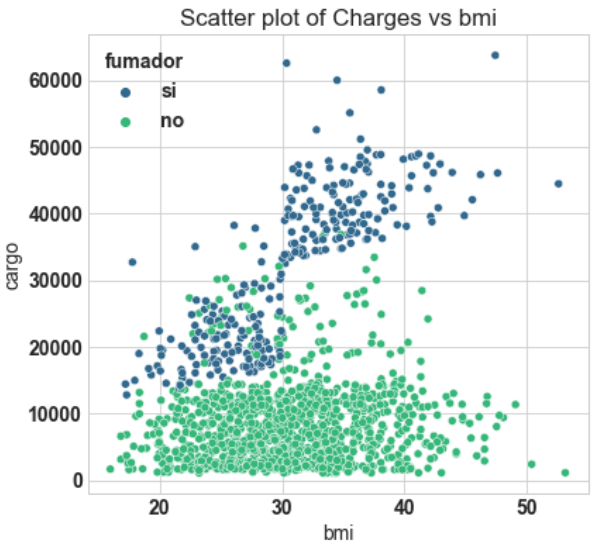
RangeIndex: 1338 entries, 0 to 1337

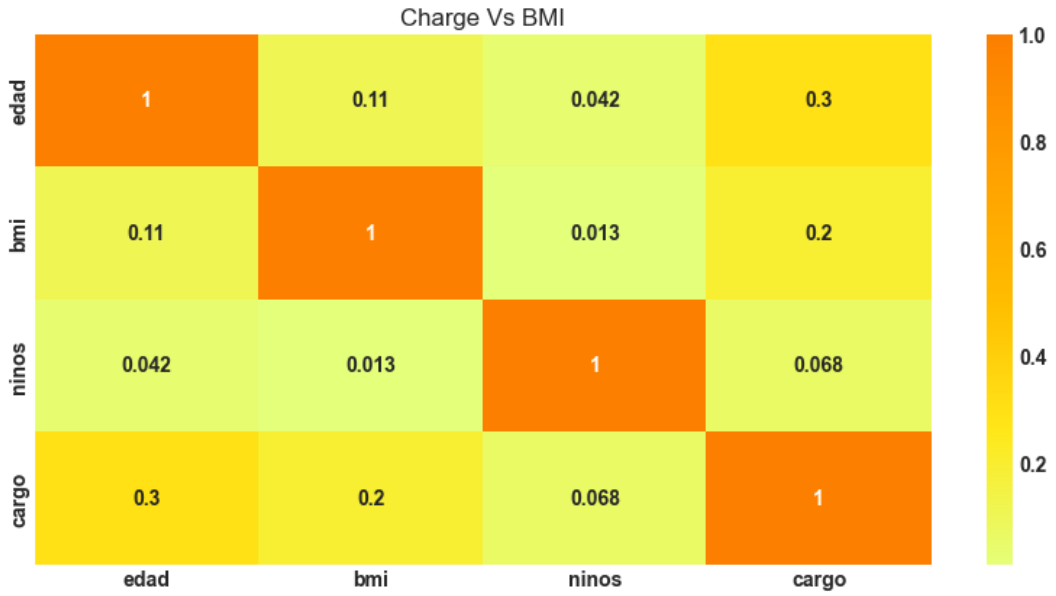
Data columns (total 7 columns):
edad 1338 non-null int64
sexo 1338 non-null object
bmi 1338 non-null float64
ninos 1338 non-null int64
fumador 1338 non-null object
region 1338 non-null object
cargo 1338 non-null float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB

La variable dependiente es el costo, supuesta a ser relacional con los demás argumentos.

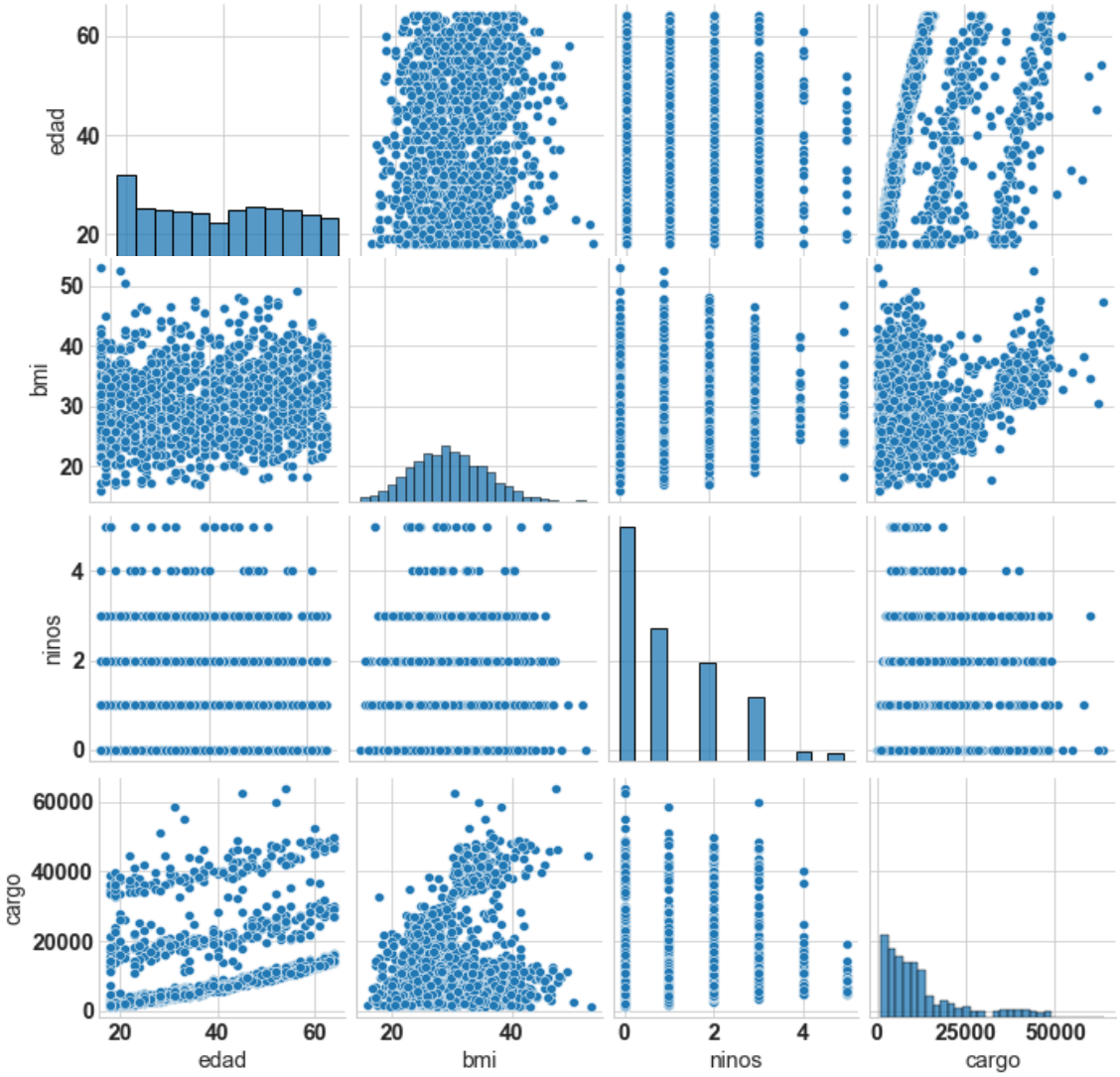


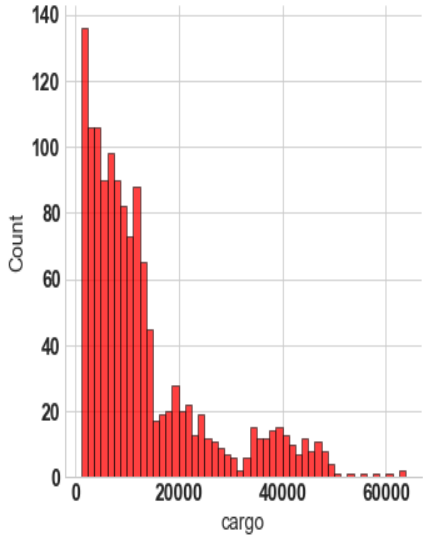


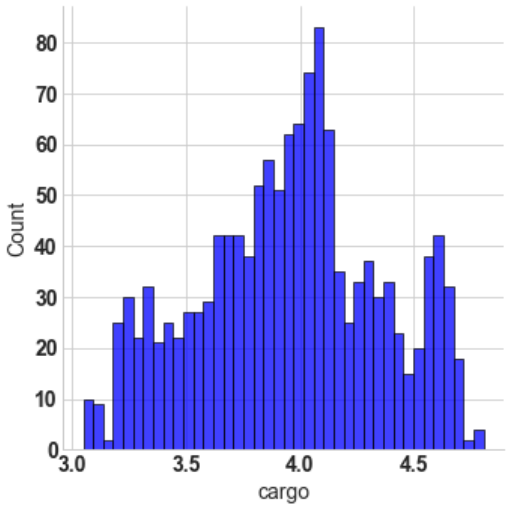




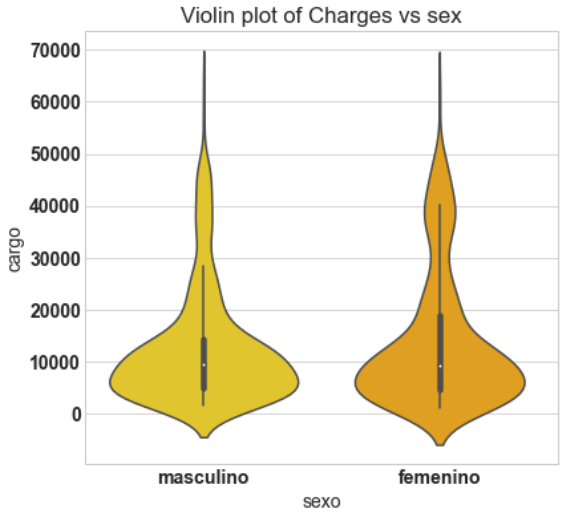
La grafica muestra que no hay correlacion entre las variables.

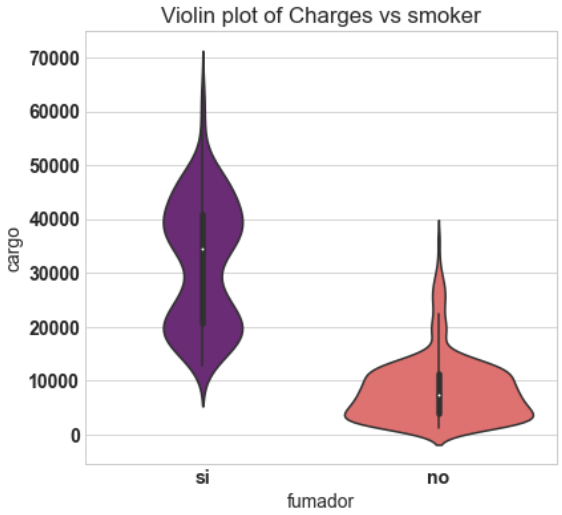


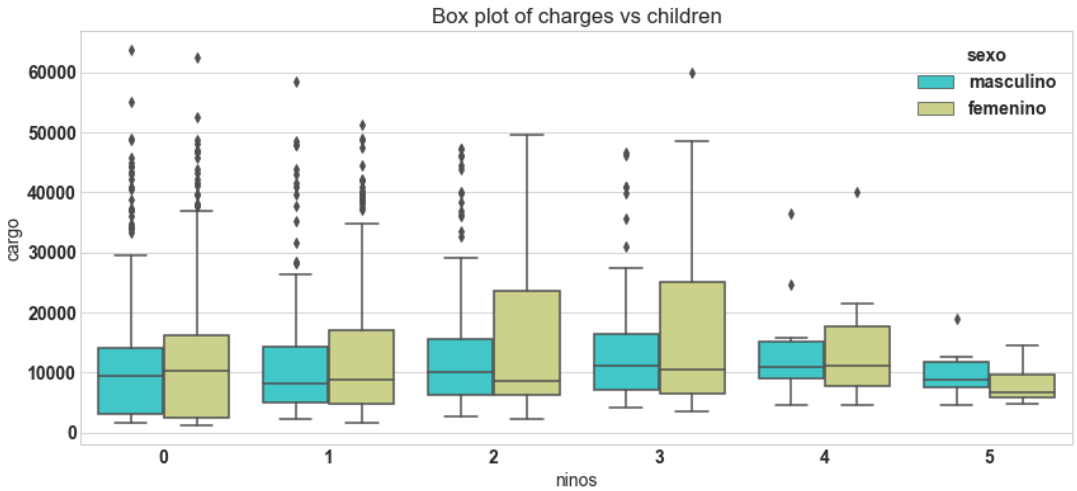


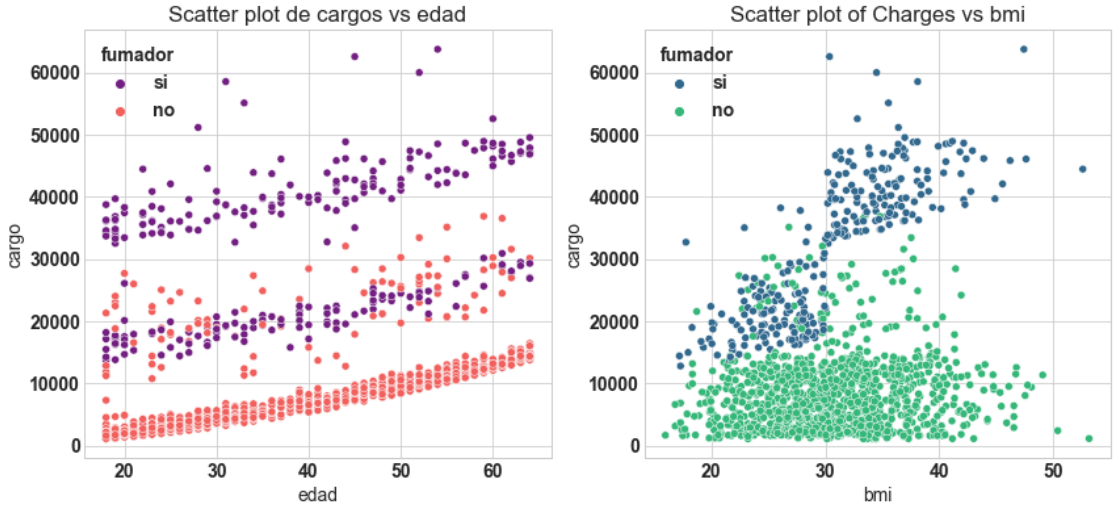


Notamos que los valores varian entre 1120 y 63500 el plot esta sesgado a la derecha.









Columns in data frame after encoding dummy variable:
['edad' 'bmi' 'cargo' 'OHE_masculino' 'OHE_1' 'OHE_2' 'OHE_3' 'OHE_4'
'OHE_5' 'OHE_si' 'OHE_este' 'OHE_oeste' 'OHE_sur']

Creando el modelo:

 # R_square Regression Linear
The Mean Square Error(MSE) or J(theta) is: 0.18729622322981587
R square obtain for normal equation method is : 0.7795687545055354

R_square Sklearn
The Mean Square Error(MSE) or J(theta) is: 0.18729622322981895
R square obtain for scikit learn library is : 0.7795687545055319

Los valores obtenidos son los mismos en nuestro modelo usando la ecuacion de regresion normal y verificandola con el modulo sklearn

Validacion del modelo de regresion linear
6.991404120436252

Test set evaluation:

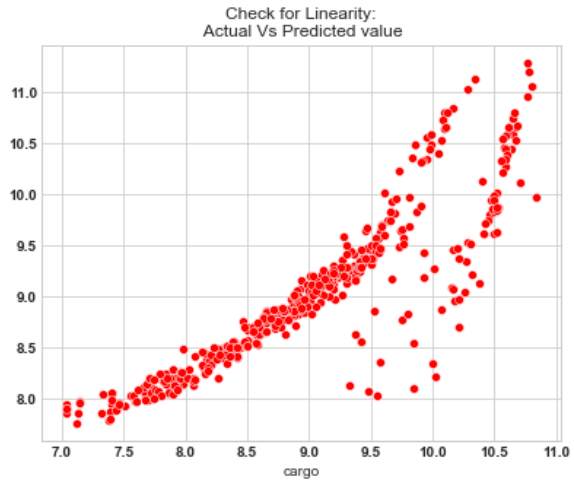
_____
MAE: 2.2005759824668525
MSE: 6.0105070273690595
RMSE: 2.451633542634188
R2 Square -6.073840183475098

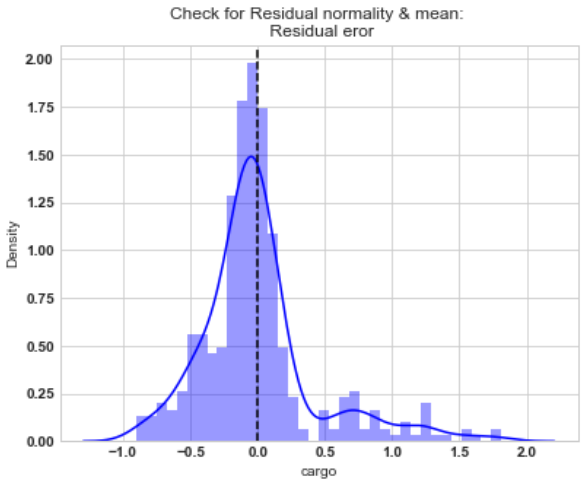
_____
Train set evaluation:

_____
MAE: 2.2056163703196585
MSE: 6.132075523111328
RMSE: 2.4763027930992867
R2 Square -6.302012624566442

#Resultados del modelo lineal simple
****

#En nuestro modelo el plot del valor actual vs la prediccion es curva de modo que la asuncion linear usada falla.
#La media residual es cero y el error residual en el plot esta sesgado a la derecha.
#Q-Q plot muestra como valor el log value mayor a 1.5 tendente a incrementar.
#el grafico muestra heterocedasticidad, el error incrementara despues de cierto punto.
#El valor del factor de varianza de inflacion es menor de 5, no tiene multicolinealidad.





##Robust Regression
Test set evaluation:

_____
MAE: 2.2005759824668525
MSE: 6.0105070273690595

RMSE: 2.451633542634188
R2 Square -6.073840183475098

_____
Train set evaluation:

_____
MAE: 2.2056163703196585
MSE: 6.132075523111328
RMSE: 2.4763027930992867
R2 Square -6.302012624566442

_____
Test set evaluation:

_____
MAE: 0.25762642106842076
MSE: 0.21036522115384151
RMSE: 0.45865588533653584
R2 Square 0.7524185650515673

_____
==========
Train set evaluation:

_____
MAE: 0.2674621665182601
MSE: 0.24215572710183542
RMSE: 0.4920932097700957
R2 Square 0.7116434444187184

##Ridge Regression
Test set evaluation:

_____
MAE: 0.309605840076702
MSE: 0.2007728534823148
RMSE: 0.4480768388148564
R2 Square 0.7637079414021039

_____
==========
Train set evaluation:

_____
MAE: 0.29699944388575056
MSE: 0.20436855525643116
RMSE: 0.45207140504176013
R2 Square 0.756640020997378

Lasso Regression
Test set evaluation:

_____
MAE: 0.3421881384272713
MSE: 0.24621476004853263
RMSE: 0.4962032249942425

R2 Square 0.7102267985936679

_____
==========
Train set evaluation:

_____
MAE: 0.33209119253687025
MSE: 0.24479679406040653
RMSE: 0.49476943525283223
R2 Square 0.7084984889788959

Elastic Net
Test set evaluation:

_____
MAE: 0.3388182950369143
MSE: 0.24328829808425767
RMSE: 0.49324263611761876
R2 Square 0.7136709879347727

_____
==========
Train set evaluation:

_____
MAE: 0.3290279847848223
MSE: 0.24264830617940852
RMSE: 0.4925934491844248
R2 Square 0.7110568863064642

Polinomial Regression

Test set evaluation:

_____
MAE: 42784539701.62653
MSE: 7.358677686597079e+23
RMSE: 857827353643.9065
R2 Square -8.660518934536093e+23

_____
==========
Train set evaluation:

_____
MAE: 0.20079792217650938
MSE: 0.13143923980495928
RMSE: 0.3625455003236963
R2 Square 0.8434835016623777

 # Stochastic Gradient Descent
Test set evaluation:

_____
MAE: 0.3436525244229725
MSE: 0.24524770626486245
RMSE: 0.4952249047300251
R2 Square 0.7113649361723045

_____
==========
Train set evaluation:

_____
MAE: 0.34472640209697425
MSE: 0.2567997784642071
RMSE: 0.5067541597897417
R2 Square 0.6942054582882766

Random Forest Regressor
Test set evaluation:

_____
MAE: 0.22374523345408992
MSE: 0.15367840018305812
RMSE: 0.3920183671501351
R2 Square 0.8191339869337229

_____
Train set evaluation:

_____
MAE: 0.08283868998376569
MSE: 0.02676433217329005
RMSE: 0.16359808120295927
R2 Square 0.9681293078206751

De acuerdo a los resultados obtenidos en los diferentes modelos el que nos ofrece mejores resultados es el Random Forest Regressor con un R2 Square de 96%.