﻿import pandas as pd

import seaborn as sns

df=pd.read\_csv("USA\_housing.csv")

df.info()

corr = df.corr()

x=df[["Avg. Area Income"]]

y=df["Price"]

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.3,random\_state=1)

from sklearn.linear\_model import LinearRegression

lr=LinearRegression()

lr.fit(x\_train, y\_train)

lr.coef\_

lr.intercept\_

### Price = -209867 + (Avg. Area Income) \* 21.0306331

y\_pred = lr.predict(x\_test)

from sklearn.metrics import mean\_squared\_error

mse=mean\_squared\_error(y\_test,y\_pred)

rmse=mse\*\*0.5

print("rmse is",rmse) # 269033.18772319076

## MULTIPLE LINEAR REGRESSION

x=df.drop(columns=['Price','Address'])

y=df["Price"]

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.3,random\_state=1)

from sklearn.linear\_model import LinearRegression

lr=LinearRegression()

lr.fit(x\_train, y\_train)

lr.coef\_

lr.intercept\_

### Price = -2645289 + (Avg. Area Income) \* 2.16398550e+01

### + (Avg. Area House Age) \* 1.65729214e+05 + (Avg. Area Number of Rooms) \* 1.20958349e+05

### + (Avg. Area Number of Bedrooms) \* 1.94909254e+03 + (Area Population) \* 1.52262240e+01

y\_pred = lr.predict(x\_test)

from sklearn.metrics import mean\_squared\_error

mse=mean\_squared\_error(y\_test,y\_pred)

rmse=mse\*\*0.5

print("rmse is",rmse) ## 102798.09614448597

The rmse of the simple linear regression ﻿with the “Avg. Area Income” (which had the highest correlation with the price) is 269033.19. This model can be improved by creating including all the variables (except for the address since there can’t be any real correlation between that and the price based on the representation of the address). This new improved model has a rmse of 102798.10.