```
import numpy as np
import pandas as pd
```

```
df = pd.read_csv('train_data.csv')
```

```
df_test = pd.read_csv('test_data.csv')
```

```
df_site_const = pd.read_csv('site_const_data-1.csv')
```

1 Dataset Preparation and Rationale

1. Why did Kathleen's team split the data into a training set (374 stores) and test set (85 stores)?

ANS: On dividing data into 374-train & 85-test, the model built by Kathleen's team achieved R-square of 0.786 on train & 0.719 on test, indicating strong predictive accuracy. Hence, they splitted data (81%-19%)

(a) What percentage of the total data is in the training set?

ANS: 18.51

(b) Explain what the training set will be used for, what the test set will be used for, and why it is important not to use the test set during model building.

ANS: The taining set is to train/build the model & test set is for validation of the model for accuracy.

If training set is used to build the model it will try to overfit and captures the noise from test data, model built in such a way will perform exceptionally well on the test data but poorly on new, truly unseen data because it has not learned to generalize.

2 Kathleen's Original Model

2. Fit a linear regression model using the training data with the four variables: agg.inc, sqft, col.grad, and com60. (a) Write the complete linear regression equation for predicting annual store profitability from these four predictors. Your equation should be in the form: annual.profit = β0 + β1 × agg.inc + β2 × sqft + β3 × col.grad + β4 × com60

```
import statsmodels.formula.api as smf
```

```
linear_regression_train = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60', data =df) # ols = or
linear_regression_result_train = linear_regression_train.fit()
coefficients = linear_regression_result_train.params
beta0=coefficients.iloc[0].round(4)
beta1=coefficients.iloc[1].round(4)
beta2=coefficients.iloc[2].round(4)
beta3=coefficients.iloc[3].round(4)
beta4=coefficients.iloc[4].round(4)
print("The equations is:")
print(f"annual.profit = {beta0} + {beta1} × agg.inc + {beta2} × sqft + {beta3} × col.grad + {beta4} × com60")
#
```

```
The equations is:
annual.profit = 83597.0925 + 0.0028 × agg.inc + 383.3631 × sqft + 346821.1387 × col.grad + 218265.7646 × com60
```

3. Using the estimated regression model, what annual profitability is predicted for a Milagro store located in an area with: • Aggregate income (agg.inc) of $100,000,000 • Store size (sqft) of 800 square feet • College graduate percentage (col.grad) of 0.30 (30%) • Long commute percentage (com60) of 0.10 (10%)

```
import pandas as pd

# Create a new DataFrame with the given values
new_data = pd.DataFrame({
    'agg.inc': [100000000],
    'sqft': [800],
    'col.grad': [0.30],
```

```
    'com60': [0.10]
})

predicted_profit = linear_regression_result_train.predict(new_data)

print(f"Predicted annual profitability: ${predicted_profit[0]:,.2f}")
```

```
Predicted annual profitability: $796,894.99
```

4. Evaluate the quality of the original model:

```
print("Linear Regression Summary for Train:")
print(linear_regression_result_train.summary())
```

```
Linear Regression Summary for Train:
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.784
Method:                 Least Squares   F-statistic:                     339.1
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          3.73e-122
Time:                        03:56:28   Log-Likelihood:                -5107.5
No. Observations:                 374   AIC:                         1.022e+04
Df Residuals:                     369   BIC:                         1.024e+04
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       8.36e+04   4.11e+04      2.035      0.043    2810.065    1.64e+05
Q("agg.inc")      0.0028      0.000     20.288      0.000       0.003       0.003
sqft            383.3631     61.230      6.261      0.000     262.960     503.766
Q("col.grad")  3.468e+05   1.13e+05      3.069      0.002    1.25e+05    5.69e+05
com60          2.183e+05   9.82e+04      2.222      0.027    2.51e+04    4.11e+05
==============================================================================
Omnibus:                       56.244   Durbin-Watson:                   2.108
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               96.491
Skew:                           0.880   Prob(JB):                     1.11e-21
Kurtosis:                       4.759   Cond. No.                     1.82e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.82e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(a) What is the R2 value on the training data?

ANS:

0.786

```
linear_regression_test = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60', data =df_test) # ols
linear_regression_result_test = linear_regression_test.fit()
print(linear_regression_result_test.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.765
Model:                            OLS   Adj. R-squared:                  0.754
Method:                 Least Squares   F-statistic:                     65.23
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           2.08e-24
Time:                        04:53:58   Log-Likelihood:                -1144.9
No. Observations:                  85   AIC:                             2300.
Df Residuals:                      80   BIC:                             2312.
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept       3.65e+04   7.49e+04      0.487      0.627   -1.13e+05    1.86e+05
Q("agg.inc")      0.0025      0.000      8.719      0.000       0.002       0.003
sqft            439.4367     83.922      5.236      0.000     272.426     606.447
Q("col.grad")  1.425e+05   1.94e+05      0.735      0.464   -2.43e+05    5.28e+05
com60          4.839e+05   2.13e+05      2.276      0.026    6.08e+04    9.07e+05
==============================================================================
Omnibus:                        3.973   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.137   Jarque-Bera (JB):                3.487
```

```
Skew:                           0.492   Prob(JB):                        0.175
Kurtosis:                       3.121   Cond. No.                     1.62e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.62e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

(b) What is the R2 value on the test data?

ANS:**0.765**

```
Start coding or generate with AI.
```

5. Test the statistical significance of the predictors:

(a) Which independent variables are statistically significant at the 5% level (α = 0.05)?

ANS: These variables:(**agg.inc,sqft,col.grad,com60**) are statistically significant at 5% level

(b) Which variable has the smallest p-value (most statistically significant)?

ANS:**agg.inc** & **sqft** both have smallest p-value(0.000)

(c) Which variable has the largest p-value (least statistically significant, but still below 0.05)?

ANS:**com60** has largest p-value

3 Exploratory Correlation Analysis

6. Compute the correlation matrix for all numerical predictor variables (exclude store.number, annual.profit, and state). (a) The dataset now has 10 predictor variables: the 4 original variables plus 6 new variables. Identify the three pairs of variables with the strongest correlations (highest absolute values). Report the correlation coefficient for each pair.

```
excluded_col = ['store.number','annual.profit','state']
df_filtered = df.drop(columns=excluded_col)
correlation_matrix = df_filtered.corr()
correlation_matrix
```

	agg.inc	sqft	col.grad	com60	lci	nearcomp	nearmil	freestand	gini	housemed
agg.inc	1.000000	0.511834	0.670194	-0.238835	-0.314832	-0.147581	0.160179	0.173108	0.068613	0.486390
sqft	0.511834	1.000000	0.353035	-0.055354	-0.299658	-0.074148	0.121758	0.150469	0.110946	0.176864
col.grad	0.670194	0.353035	1.000000	-0.223868	-0.313549	-0.181825	0.065328	0.166153	0.001354	0.546932
com60	-0.238835	-0.055354	-0.223868	1.000000	-0.004081	0.024768	0.037709	-0.032012	-0.035073	-0.105689
lci	-0.314832	-0.299658	-0.313549	-0.004081	1.000000	0.071425	-0.114763	-0.219530	-0.104897	-0.151980
nearcomp	-0.147581	-0.074148	-0.181825	0.024768	0.071425	1.000000	-0.110449	0.104221	0.112272	-0.174588
nearmil	0.160179	0.121758	0.065328	0.037709	-0.114763	-0.110449	1.000000	0.211365	-0.012172	0.031857
freestand	0.173108	0.150469	0.166153	-0.032012	-0.219530	0.104221	0.211365	1.000000	0.020250	-0.011140
gini	0.068613	0.110946	0.001354	-0.035073	-0.104897	0.112272	-0.012172	0.020250	1.000000	-0.079713
housemed	0.486390	0.176864	0.546932	-0.105689	-0.151980	-0.174588	0.031857	-0.011140	-0.079713	1.000000

ANS:

Pair of variables with the strongest correlations:

1. aggr.inc & col.grad : 0.670194
2. col.grad & housemed : 0.546932
3. agg.inc & sqft : 0.511834

7. Statistical significance of new variables: Build a regression model using ALL 10 variables (the 4 original plus 6 new variables). Test the statistical significance of each variable at the 5% level (α = 0.05).

(a) Which of the 6 new variables (lci, nearcomp, nearmil, freestand, gini, housemed) are statistically significant?

```
linear_regression_all = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") '
'+ sqft + Q("col.grad") + com60 + lci + nearcomp + nearmil + freestand + gini + housemed', data =df) # ols = ordinary least square
linear_regression_result_all = linear_regression_all.fit()
print(linear_regression_result_all.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                     406.1
Date:                Mon, 20 Oct 2025   Prob (F-statistic):         2.74e-190
Time:                        03:56:28   Log-Likelihood:                -4928.3
No. Observations:                 374   AIC:                             9879.
Df Residuals:                     363   BIC:                             9922.
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      1.279e+05    6.2e+04      2.062      0.040    5939.643     2.5e+05
Q("agg.inc")      0.0027   9.01e-05     29.854      0.000       0.003       0.003
sqft            294.6956     39.086      7.540      0.000     217.832     371.559
Q("col.grad")  3.401e+05    7.7e+04      4.414      0.000    1.89e+05    4.92e+05
com60          1.825e+05   6.19e+04      2.951      0.003    6.09e+04    3.04e+05
lci           -1.737e+04   4893.497     -3.549      0.000    -2.7e+04   -7742.203
nearcomp       3.131e+04   3360.382      9.317      0.000    2.47e+04    3.79e+04
nearmil        2642.5113    558.423      4.732      0.000    1544.361    3740.662
freestand      3.651e+05   2.07e+04     17.672      0.000    3.25e+05    4.06e+05
gini           1.819e+04   5.14e+04      0.354      0.724   -8.29e+04    1.19e+05
housemed        -15.0379     15.838     -0.949      0.343     -46.184      16.108
==============================================================================
Omnibus:                        7.389   Durbin-Watson:                   1.918
Prob(Omnibus):                  0.025   Jarque-Bera (JB):                9.371
Skew:                          -0.180   Prob(JB):                      0.00923
Kurtosis:                       3.687   Cond. No.                     2.07e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.07e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

ANS:

Statistically significant variables are: **lci, nearcomp, nearmil, freestand**

(b) Which of the new variables are NOT significant? What does this suggest about their usefulness in predicting store profitability?

ANS:

gini, housemed are not significant variables to predict the profitability

This suggest that these variables have no effect on outcome, so they are not useful in prediction of profitability

4. Model Comparison Now build and compare four different models.

8. Fit and evaluate four models using the training data:

Model A: Kathleen's Original Model Variables: agg.inc, sqft, col.grad, com60 (linear_regression_train)

Model B: Full Model Variables: All variables except store.number, annual.profit, and state

(linear_regression_all)

Model C: Parsimonious Model • Build this model by removing variables that meet either of these criteria: – Variables that are NOT statistically significant at the 5% level (from Question 7). – Variables involved in pairs with absolute correlation > 0.70 (from Question 6). For highly correlated pairs, keep the variable with stronger correlation to the outcome variable (annual.profit).

```
# running correlation again by including 'annual.profit' to check its correlation with other variables
excluded_col = ['store.number','state']
df_filtered2 = df.drop(columns=excluded_col)
correlation_matrix2 = df_filtered2.corr()
correlation_matrix2
```

	annual.profit	agg.inc	sqft	col.grad	com60	lci	nearcomp	nearmil	freestand	gini	housemed
annual.profit	1.000000	0.868277	0.579068	0.635193	-0.149642	-0.407641	0.038363	0.260600	0.485003	0.095546	0.374731
agg.inc	0.868277	1.000000	0.511834	0.670194	-0.238835	-0.314832	-0.147581	0.160179	0.173108	0.068613	0.486390
sqft	0.579068	0.511834	1.000000	0.353035	-0.055354	-0.299658	-0.074148	0.121758	0.150469	0.110946	0.176864
col.grad	0.635193	0.670194	0.353035	1.000000	-0.223868	-0.313549	-0.181825	0.065328	0.166153	0.001354	0.546932
com60	-0.149642	-0.238835	-0.055354	-0.223868	1.000000	-0.004081	0.024768	0.037709	-0.032012	-0.035073	-0.105689
lci	-0.407641	-0.314832	-0.299658	-0.313549	-0.004081	1.000000	0.071425	-0.114763	-0.219530	-0.104897	-0.151980
nearcomp	0.038363	-0.147581	-0.074148	-0.181825	0.024768	0.071425	1.000000	-0.110449	0.104221	0.112272	-0.174588
nearmil	0.260600	0.160179	0.121758	0.065328	0.037709	-0.114763	-0.110449	1.000000	0.211365	-0.012172	0.031857
freestand	0.485003	0.173108	0.150469	0.166153	-0.032012	-0.219530	0.104221	0.211365	1.000000	0.020250	-0.011140
gini	0.095546	0.068613	0.110946	0.001354	-0.035073	-0.104897	0.112272	-0.012172	0.020250	1.000000	-0.079713
housemed	0.374731	0.486390	0.176864	0.546932	-0.105689	-0.151980	-0.174588	0.031857	-0.011140	-0.079713	1.000000

```
#variables to remove:removing gini, housemed because they are not statistically significant
#there is no pair of variable having absolute correlation >0.7
linear_regression_Parsimonious = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60'
'+ lci + nearcomp + nearmil + freestand', data =df) # ols = ordinary least square
linear_regression_result_Parsimonious = linear_regression_Parsimonious.fit()
print(linear_regression_result_Parsimonious.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                     508.7
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          9.32e-193
Time:                        03:58:01   Log-Likelihood:                -4928.9
No. Observations:                 374   AIC:                             9876.
Df Residuals:                     365   BIC:                             9911.
Df Model:                           8
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      1.256e+05   5.35e+04      2.348      0.019    2.04e+04    2.31e+05
Q("agg.inc")      0.0027   8.74e-05     30.581      0.000       0.003       0.003
sqft            299.4467     38.756      7.726      0.000     223.233     375.660
Q("col.grad")    3.13e+05   7.22e+04      4.337      0.000    1.71e+05    4.55e+05
com60          1.783e+05   6.16e+04      2.893      0.004    5.71e+04       3e+05
lci           -1.763e+04   4865.956     -3.624      0.000   -2.72e+04   -8063.427
nearcomp       3.167e+04   3327.214      9.519      0.000    2.51e+04    3.82e+04
nearmil        2647.7553    557.527      4.749      0.000    1551.387    3744.124
freestand      3.673e+05   2.05e+04     17.927      0.000    3.27e+05    4.08e+05
==============================================================================
Omnibus:                        7.427   Durbin-Watson:                   1.921
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                9.432
Skew:                          -0.181   Prob(JB):                      0.00895
Kurtosis:                       3.689   Cond. No.                     1.96e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.96e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model D: Alternative Model • Start with the original 4 variables (agg.inc, sqft, col.grad, com60). • Add ONE variable from the 4 significant new variables identified in Question 7: lci, nearcomp, nearmil, freestand.

```
linear_regression_alt = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60'
'+ freestand ', data =df) # ols = ordinary least square
linear_regression_result_alt = linear_regression_alt.fit()
```

```
print(linear_regression_result_alt.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.892
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     605.5
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          4.27e-175
Time:                        04:04:19   Log-Likelihood:                -4980.4
No. Observations:                 374   AIC:                             9973.
Df Residuals:                     368   BIC:                             9996.
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      9.709e+04   2.93e+04      3.314      0.001    3.95e+04    1.55e+05
Q("agg.inc")      0.0027   9.88e-05     27.476      0.000       0.003       0.003
sqft            324.4291     43.758      7.414      0.000     238.383     410.476
Q("col.grad")  2.425e+05   8.08e+04      3.003      0.003    8.37e+04    4.01e+05
com60          2.049e+05      7e+04      2.927      0.004    6.72e+04    3.43e+05
freestand      4.241e+05   2.24e+04     18.926      0.000     3.8e+05    4.68e+05
==============================================================================
Omnibus:                        3.254   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.196   Jarque-Bera (JB):                3.126
Skew:                           0.148   Prob(JB):                        0.210
Kurtosis:                       3.336   Cond. No.                     1.82e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.82e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Test each of the 4 possible additions (one at a time) and choose the one that: – Improves test R2 compared to Model A, and – Maintains total profitability prediction ≥ $40 million

(a) For Model D report which variable you added

ANS: **freestand**

(b) For each model, report: i. Training R2 ii. Test R2 iii. Total predicted profitability for the 48 construction sites (in millions)

```
# Model A
selected_columns = ['agg.inc','sqft','col.grad','com60']
df_site_const['predicted_profit'] = linear_regression_result_train.predict(df_site_const[selected_columns])

# Sum all predicted values
total_predicted_profit = df_site_const['predicted_profit'].sum()

print(total_predicted_profit)
```

```
40016174.42509793
```

ANS:

For "Model A:"

i. Training R2 = **0.786**

ii. Test R2 =**0.765**

iii. Total predicted profitability for the 48 construction sites (in millions)= **40016174.42509793**

```
# Model B
# R2 for train data
linear_regression_all = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") '
'+ sqft + Q("col.grad") + com60 + lci + nearcomp + nearmil + freestand + gini + housemed', data =df) # ols = ordinary least square
linear_regression_result_all = linear_regression_all.fit()
print(linear_regression_result_all.summary())


# R2 for Test data
linear_regression_all_test = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") '
'+ sqft + Q("col.grad") + com60 + lci + nearcomp + nearmil + freestand + gini + housemed', data =df_test)
```

```
linear_regression_all_test_result = linear_regression_all_test.fit()
print(linear_regression_all_test_result.summary())


# Prediction for 48 const sites
selected_columns = ['agg.inc','sqft','col.grad','com60','lci','nearcomp','nearmil','freestand','gini','housemed']
df_site_const['predicted_profit'] = linear_regression_result_all.predict(df_site_const[selected_columns])
total_predicted_profit = df_site_const['predicted_profit'].sum()

print("Total predicted profit for 48 const sites with Model B")
print(total_predicted_profit)
```

```
Dep. Variable:     Q("annual.profit")   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                     406.1
Date:                Mon, 20 Oct 2025   Prob (F-statistic):         2.74e-190
Time:                        05:05:32   Log-Likelihood:                -4928.3
No. Observations:                 374   AIC:                             9879.
Df Residuals:                     363   BIC:                             9922.
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      1.279e+05    6.2e+04      2.062      0.040    5939.643     2.5e+05
Q("agg.inc")      0.0027   9.01e-05     29.854      0.000       0.003       0.003
sqft            294.6956     39.086      7.540      0.000     217.832     371.559
Q("col.grad")  3.401e+05    7.7e+04      4.414      0.000    1.89e+05    4.92e+05
com60          1.825e+05   6.19e+04      2.951      0.003    6.09e+04    3.04e+05
lci           -1.737e+04   4893.497     -3.549      0.000    -2.7e+04   -7742.203
nearcomp       3.131e+04   3360.382      9.317      0.000    2.47e+04    3.79e+04
nearmil        2642.5113    558.423      4.732      0.000    1544.361    3740.662
freestand      3.651e+05   2.07e+04     17.672      0.000    3.25e+05    4.06e+05
gini           1.819e+04   5.14e+04      0.354      0.724   -8.29e+04    1.19e+05
housemed        -15.0379     15.838     -0.949      0.343     -46.184      16.108
==============================================================================
Omnibus:                        7.389   Durbin-Watson:                   1.918
Prob(Omnibus):                  0.025   Jarque-Bera (JB):                9.371
Skew:                          -0.180   Prob(JB):                      0.00923
Kurtosis:                       3.687   Cond. No.                     2.07e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.07e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.873
Model:                            OLS   Adj. R-squared:                  0.855
Method:                 Least Squares   F-statistic:                     50.70
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           4.61e-29
Time:                        05:05:32   Log-Likelihood:                -1119.0
No. Observations:                  85   AIC:                             2260.
Df Residuals:                      74   BIC:                             2287.
Df Model:                          10
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      2.074e+05   1.39e+05      1.496      0.139   -6.89e+04    4.84e+05
Q("agg.inc")      0.0025      0.000     10.822      0.000       0.002       0.003
sqft            345.8307     68.246      5.067      0.000     209.848     481.813
Q("col.grad")  1.767e+05   1.54e+05      1.146      0.256   -1.31e+05    4.84e+05
com60          2.404e+05   1.72e+05      1.399      0.166   -1.02e+05    5.83e+05
lci           -1.414e+04   1.09e+04     -1.303      0.197   -3.58e+04    7486.022
nearcomp       3.476e+04   9043.764      3.843      0.000    1.67e+04    5.28e+04
nearmil        2140.6479   1570.636      1.363      0.177    -988.912    5270.208
freestand      2.688e+05   5.53e+04      4.860      0.000    1.59e+05    3.79e+05
gini           -3.18e+04   1.19e+05     -0.267      0.790   -2.69e+05    2.05e+05
```

For "Model B:"

i. Training R2 = 0.918

ii. Test R2 =0.873

iii. Total predicted profitability for the 48 construction sites (in millions)= 36057340.47829171

```
# Model C
# R2 for train data
linear_regression_Parsimonious = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60'
'+ lci + nearcomp + nearmil + freestand', data =df) # ols = ordinary least square
linear_regression_result_Parsimonious = linear_regression_Parsimonious.fit()
print(linear_regression_result_Parsimonious.summary())


# R2 for Test data
linear_regression_Parsimonious_test = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60'
'+ lci + nearcomp + nearmil + freestand', data =df_test)
linear_regression_Parsimonious_test_result = linear_regression_Parsimonious_test.fit()
print(linear_regression_Parsimonious_test_result.summary())


# Prediction for 48 const sites
selected_columns = ['agg.inc','sqft','col.grad','com60','lci','nearcomp','nearmil','freestand']
df_site_const['predicted_profit'] = linear_regression_result_Parsimonious.predict(df_site_const[selected_columns])
total_predicted_profit = df_site_const['predicted_profit'].sum()

print("Total predicted profit for 48 const sites with Model C")
print(total_predicted_profit)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.918
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                     508.7
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          9.32e-193
Time:                        05:14:17   Log-Likelihood:                -4928.9
No. Observations:                 374   AIC:                             9876.
Df Residuals:                     365   BIC:                             9911.
Df Model:                           8
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      1.256e+05   5.35e+04      2.348      0.019    2.04e+04    2.31e+05
Q("agg.inc")      0.0027   8.74e-05     30.581      0.000       0.003       0.003
sqft            299.4467     38.756      7.726      0.000     223.233     375.660
Q("col.grad")   3.13e+05   7.22e+04      4.337      0.000    1.71e+05    4.55e+05
com60          1.783e+05   6.16e+04      2.893      0.004    5.71e+04       3e+05
lci           -1.763e+04   4865.956     -3.624      0.000   -2.72e+04   -8063.427
nearcomp       3.167e+04   3327.214      9.519      0.000    2.51e+04    3.82e+04
nearmil        2647.7553    557.527      4.749      0.000    1551.387    3744.124
freestand      3.673e+05   2.05e+04     17.927      0.000    3.27e+05    4.08e+05
==============================================================================
Omnibus:                        7.427   Durbin-Watson:                   1.921
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                9.432
Skew:                          -0.181   Prob(JB):                      0.00895
Kurtosis:                       3.689   Cond. No.                     1.96e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.96e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.871
Model:                            OLS   Adj. R-squared:                  0.858
Method:                 Least Squares   F-statistic:                     64.24
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           1.08e-30
Time:                        05:14:17   Log-Likelihood:                -1119.5
No. Observations:                  85   AIC:                             2257.
Df Residuals:                      76   BIC:                             2279.
Df Model:                           8
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      1.613e+05   1.21e+05      1.332      0.187   -7.98e+04    4.02e+05
Q("agg.inc")      0.0025      0.000     11.160      0.000       0.002       0.003
sqft            348.7931     67.018      5.204      0.000     215.315     482.272
Q("col.grad")  1.875e+05   1.52e+05      1.230      0.223   -1.16e+05    4.91e+05
com60           2.82e+05   1.64e+05      1.717      0.090    -4.5e+04    6.09e+05
lci           -1.542e+04   1.07e+04     -1.447      0.152   -3.67e+04    5804.950
nearcomp       3.519e+04   8955.664      3.929      0.000    1.74e+04     5.3e+04
nearmil        1965.0242   1533.122      1.282      0.204   -1088.453    5018.502
freestand      2.703e+05   5.47e+04      4.942      0.000    1.61e+05    3.79e+05
==============================================================================
```

For "Model C:"

i. Training R2 = **0.918**

ii. Test R2 =**0.871**

iii. Total predicted profitability for the 48 construction sites (in millions)= **36292492.40864575**

```
# Model D
# R2 for train data
linear_regression_alt = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60'
'+ freestand ', data =df) # ols = ordinary least square
linear_regression_result_alt = linear_regression_alt.fit()
print(linear_regression_result_alt.summary())


# R2 for Test data
linear_regression_alt_test = smf.ols(formula = 'Q("annual.profit") ~ Q("agg.inc") + sqft+ Q("col.grad") + com60'
'+ freestand ', data =df_test)
linear_regression_alt_result_test = linear_regression_alt_test.fit()
print(linear_regression_alt_result_test.summary())


# Prediction for 48 const sites
selected_columns = ['agg.inc','sqft','col.grad','com60','freestand']
df_site_const['predicted_profit'] = linear_regression_result_alt.predict(df_site_const[selected_columns])
total_predicted_profit = df_site_const['predicted_profit'].sum()

print("Total predicted profit for 48 const sites with Model D")
print(total_predicted_profit)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.892
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     605.5
Date:                Mon, 20 Oct 2025   Prob (F-statistic):          4.27e-175
Time:                        05:14:21   Log-Likelihood:                -4980.4
No. Observations:                 374   AIC:                             9973.
Df Residuals:                     368   BIC:                             9996.
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept       9.709e+04   2.93e+04      3.314      0.001    3.95e+04    1.55e+05
Q("agg.inc")       0.0027   9.88e-05     27.476      0.000       0.003       0.003
sqft             324.4291     43.758      7.414      0.000     238.383     410.476
Q("col.grad")   2.425e+05   8.08e+04      3.003      0.003    8.37e+04    4.01e+05
com60           2.049e+05      7e+04      2.927      0.004    6.72e+04    3.43e+05
freestand       4.241e+05   2.24e+04     18.926      0.000     3.8e+05    4.68e+05
==============================================================================
Omnibus:                        3.254   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.196   Jarque-Bera (JB):                3.126
Skew:                           0.148   Prob(JB):                        0.210
Kurtosis:                       3.336   Cond. No.                     1.82e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.82e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("annual.profit")   R-squared:                       0.838
Model:                            OLS   Adj. R-squared:                  0.827
Method:                 Least Squares   F-statistic:                     81.46
Date:                Mon, 20 Oct 2025   Prob (F-statistic):           1.01e-29
Time:                        05:14:21   Log-Likelihood:                -1129.3
No. Observations:                  85   AIC:                             2271.
Df Residuals:                      79   BIC:                             2285.
Df Model:                           5
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept       1.168e+05   6.42e+04      1.821      0.072   -1.09e+04    2.45e+05
Q("agg.inc")       0.0026      0.000     10.847      0.000       0.002       0.003
sqft             320.1889     73.093      4.381      0.000     174.701     465.677
Q("col.grad")   1.225e+05   1.62e+05      0.755      0.453   -2.01e+05    4.46e+05
```

```
com60         3.255e+05    1.8e+05      1.809      0.074   -3.27e+04    6.84e+05
freestand     3.252e+05   5.49e+04      5.925      0.000    2.16e+05    4.34e+05
==============================================================================
Omnibus:                        1.896   Durbin-Watson:                   1.887
Prob(Omnibus):                  0.387   Jarque-Bera (JB):                1.638
Skew:                          -0.204   Prob(JB):                        0.441
Kurtosis:                       2.456   Cond. No.                     1.65e+09
==============================================================================
```

For "Model D:"

i. Training R2 = **0.892**

ii. Test R2 =**0.838**

iii. Total predicted profitability for the 48 construction sites (in millions)= **37333040.45001802**

9. Model recommendation and the dilemma: Which model would you recommend to Harriman Capital? In your answer, discuss whether you should prioritize statistical performance (higher test R2) even if it means revising the $40Mprofitabilityestimatedownward, orprioritizemeetingthebusinessrequirement($40M target) even with lower predictive accuracy. What are the business risks of each choice?

ANS:

- I recommend using Model C because its R-squared values for both training (0.918) and testing (0.871) data are strong compared to Models A and D. It is also more optimized than Model B, as it excludes insignificant variables. As noted in the case, "the fewer variables used, the better."
- While it's possible that Harriman might reject the proposal due to the lower profitability predicted by Model C, three out of four models with strong predictive accuracy also estimate profits below $40M. Therefore, prioritizing statistical performance over optimistic profitability demonstrates transparency and can help build Harriman Capital's trust in BVA and Milagro.
- If they recommend Model A, it may show higher profitability. However, if Harriman Capital consults other advisors and discovers that these sites actually have lower profitability, they will likely reject the offer.