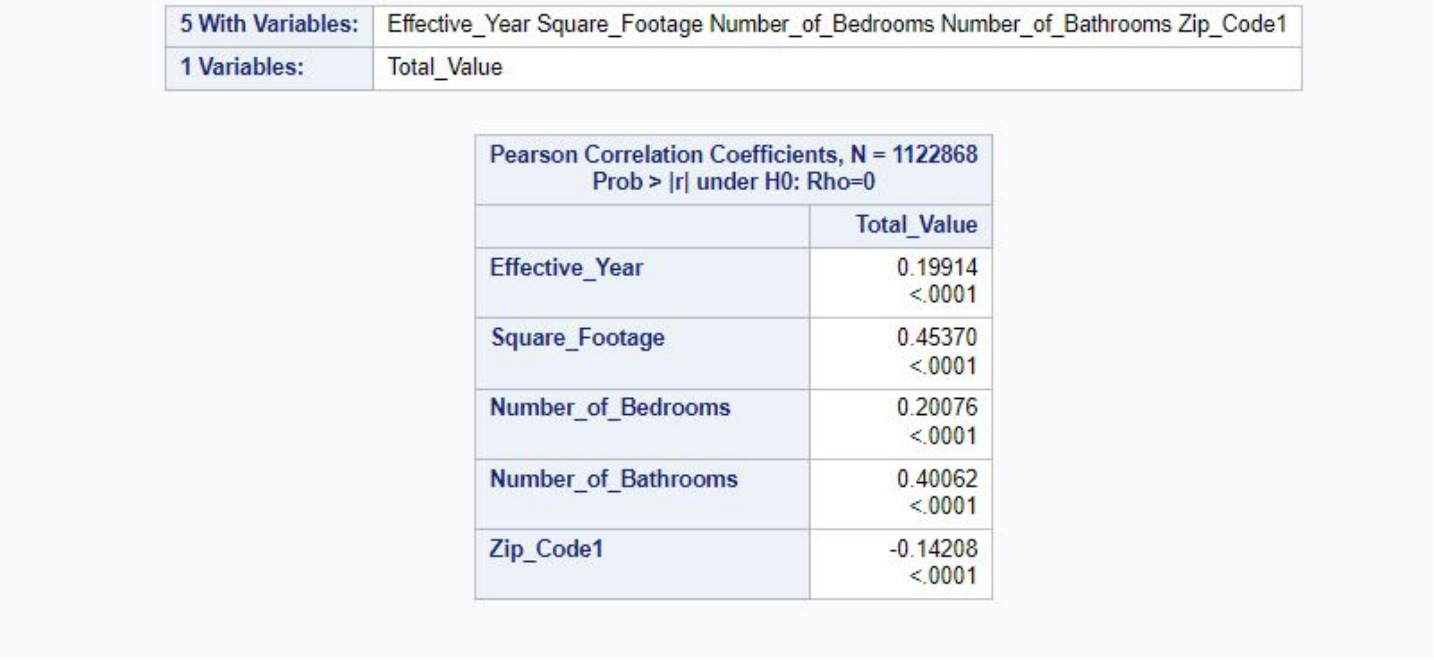
DSO 510: Business Analytics

Elmer, Issac, Wayne

Topic:
Project 2 Analysis

1. Executive Summary (WIP)
2. Goal & Variable Definition
3. Data Organization & Visualization
4. Analysis of Dependent Variable
5. Interpretation, Action, Adoption, Automation (WIP)

Correlations



5 With Variables:	Effective_Year Square_Footage Number_of_Bedrooms Number_of_Bathrooms Zip_Code1
1 Variables:	Total_Value

Pearson Correlation Coefficients, N = 1122868 Prob > |r| under H0: Rho=0	
	Total_Value
Effective_Year	0.19914 <.0001
Square_Footage	0.45370 <.0001
Number_of_Bedrooms	0.20076 <.0001
Number_of_Bathrooms	0.40062 <.0001
Zip_Code1	-0.14208 <.0001

Correlations

Effective Year (0.19914): This shows a weak positive correlation with the total value, suggesting that newer properties might have a slightly higher total value, but the relationship is not strong.

Square Footage (0.45370): There is a moderate positive correlation with the total value, indicating that larger properties tend to have higher values. This is one of the more significant factors affecting property value in the data set.

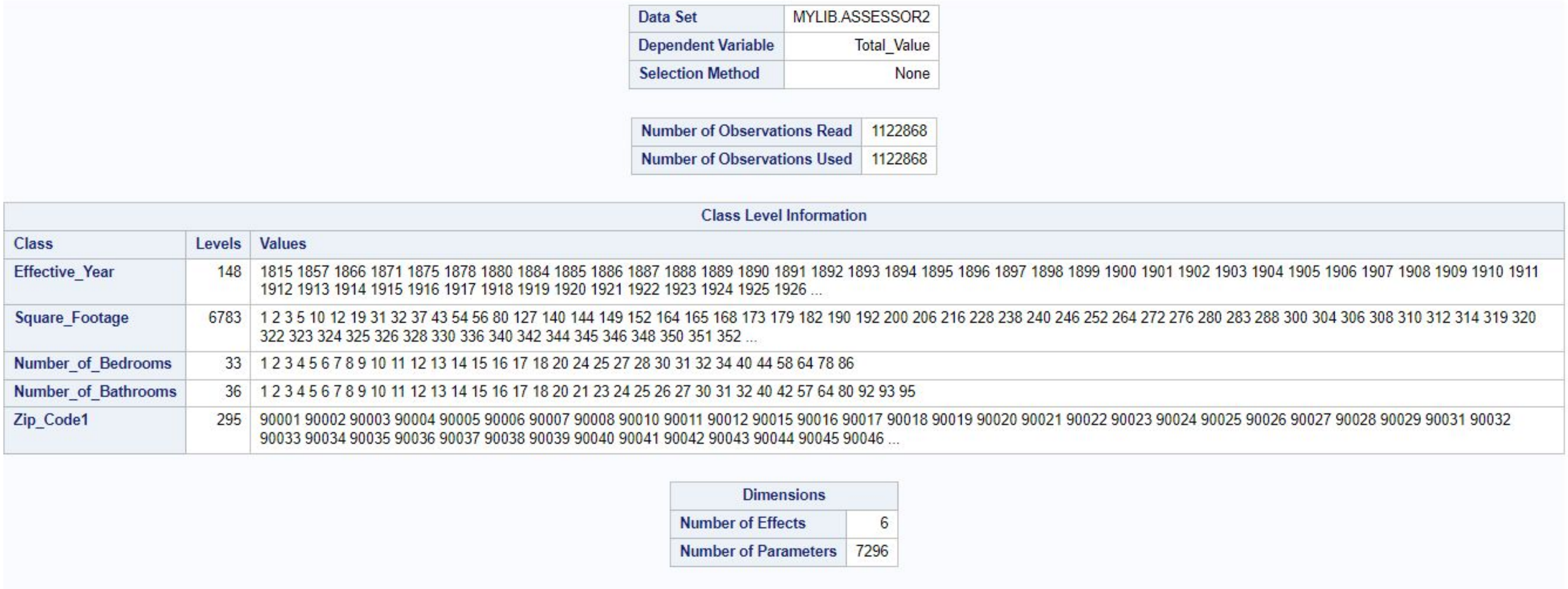
Number of Bedrooms (0.20076): Similar to the effective year, the number of bedrooms has a weak positive correlation with the total value. More bedrooms can slightly increase a property's value, but other factors might have more influence.

Correlations

Number of Bathrooms (0.40062): There's a moderate positive correlation, suggesting that properties with more bathrooms tend to have higher values. This is a significant factor, but not as strong as square footage.

Zip Code (-0.14208): This shows a weak negative correlation with the total value, implying that certain zip codes might be associated with slightly lower property values. However, the relationship is weak, indicating that the zipcode impact on property value might be less significant compared to other factors or could be influenced by other variables not captured in this analysis.

ANOVA



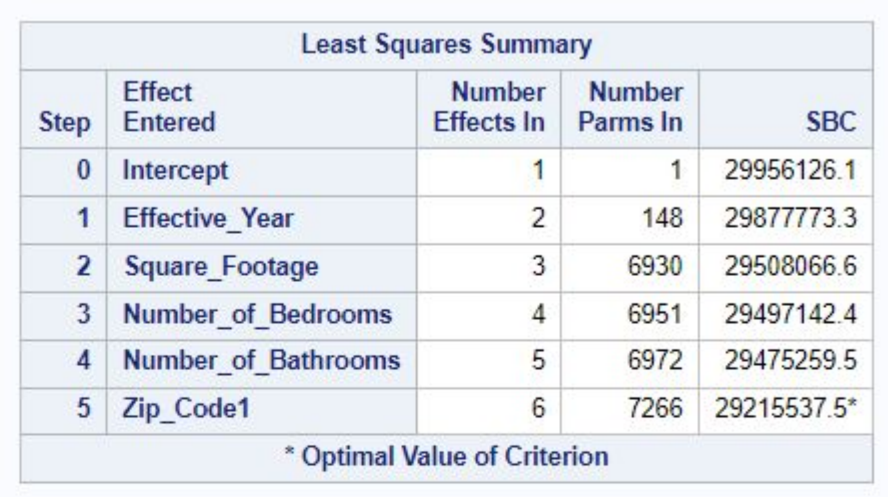
Data Set	MYLIB.ASSESSOR2
Dependent Variable	Total_Value
Selection Method	None

Number of Observations Read	1122868
Number of Observations Used	1122868

Class Level Information		
Class	Levels	Values
Effective_Year	148	1815 1857 1866 1871 1875 1878 1880 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 ...
Square_Footage	6783	1 2 3 5 10 12 19 31 32 37 43 54 56 80 127 140 144 149 152 164 165 168 173 179 182 190 192 200 206 216 228 238 240 246 252 264 272 276 280 283 288 300 304 306 308 310 312 314 319 320 322 323 324 325 326 328 330 336 340 342 344 345 346 348 350 351 352 ...
Number_of_Bedrooms	33	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 24 25 27 28 30 31 32 34 40 44 58 64 78 86
Number_of_Bathrooms	36	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 23 24 25 26 27 30 31 32 40 42 57 64 80 92 93 95
Zip_Code1	295	90001 90002 90003 90004 90005 90006 90007 90008 90010 90011 90012 90015 90016 90017 90018 90019 90020 90021 90022 90023 90024 90025 90026 90027 90028 90029 90031 90032 90033 90034 90035 90036 90037 90038 90039 90040 90041 90042 90043 90044 90045 90046 ...

Dimensions	
Number of Effects	6
Number of Parameters	7296

ANOVA

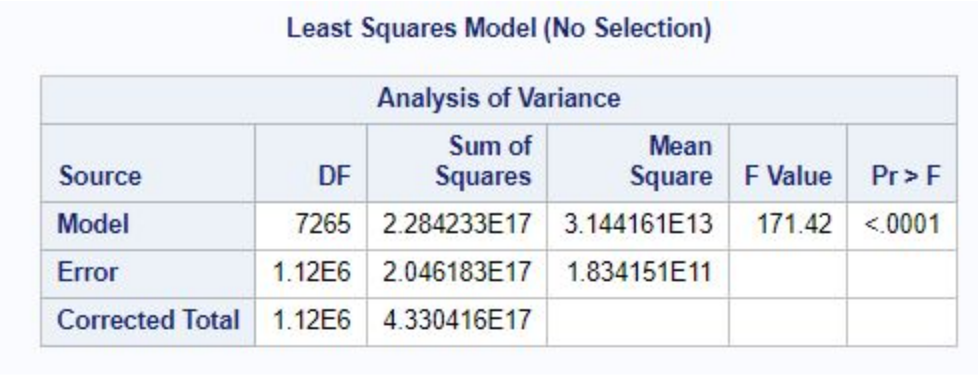


Least Squares Summary

Step	Effect Entered	Number Effects In	Number Parms In	SBC
0	Intercept	1	1	29956126.1
1	Effective_Year	2	148	29877773.3
2	Square_Footage	3	6930	29508066.6
3	Number_of_Bedrooms	4	6951	29497142.4
4	Number_of_Bathrooms	5	6972	29475259.5
5	Zip_Code1	6	7266	29215537.5*

* Optimal Value of Criterion

ANOVA



Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7265	2.284233E17	3.144161E13	171.42	<.0001
Error	1.12E6	2.046183E17	1.834151E11		
Corrected Total	1.12E6	4.330416E17			

Regression

- Model was selected using a backwards elimination process starting with all variables until only variables with significant p-values were retained

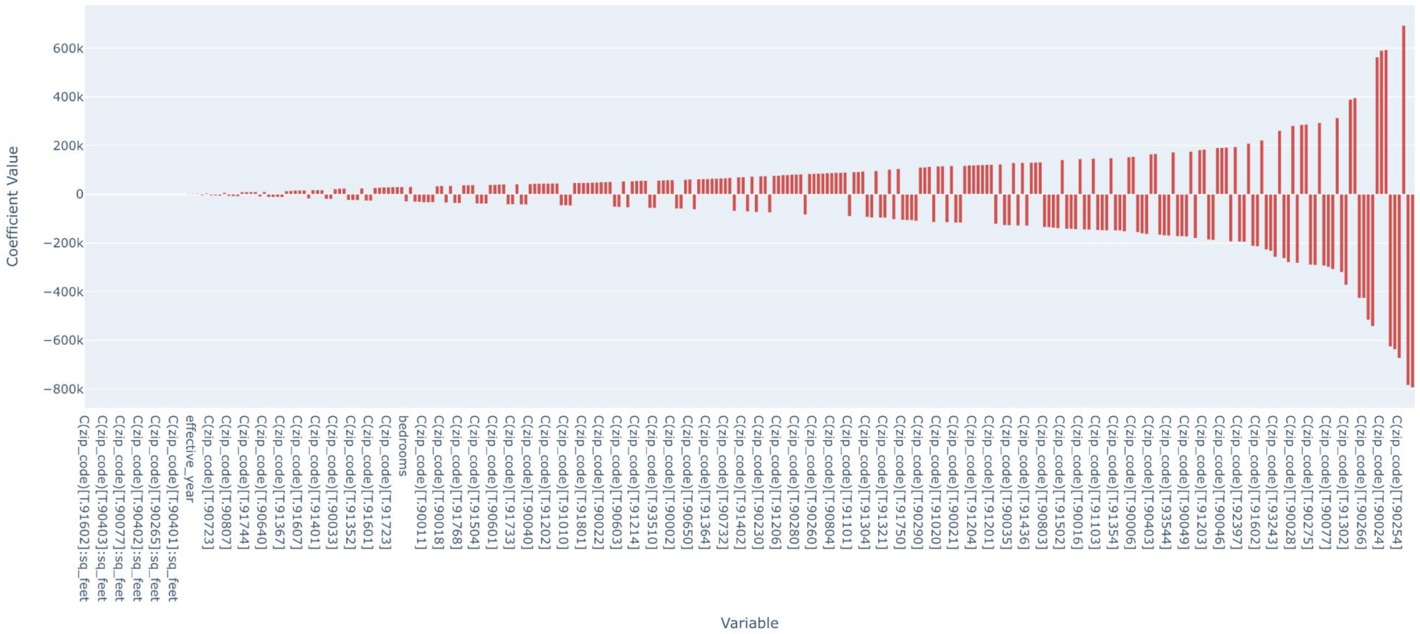
- Final model formula
  total_value ~ C(zip_code) * sq_feet + bedrooms + bathrooms + effective_year

- R Squared of 0.44 with a total 563 variables
  - 322/563 variables had statistical significance

- RMSE for the fitted values was 463,297.21

Regression

Pareto Plot of Coefficients from Regression



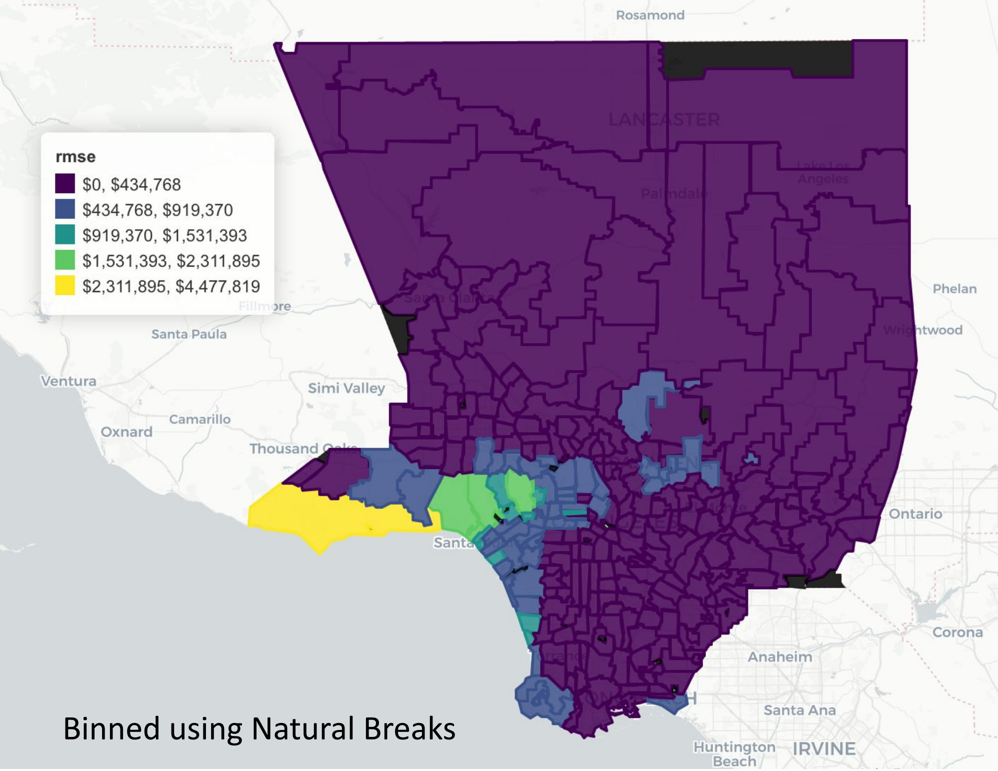
Regression

Coefficient value for non Zip Code variables (all significant)

- Bedrooms: -30,520
- Bathrooms: 47,750
- Effective Year: 2,899.25
- Sq Feet: 54.65

Regression

- RMSE by Zip Code
- Calculated by squaring the error and averaging by zip code then taking the square root



Binned using Natural Breaks

Analytical Interpretation

- Given that the models most contributing variables were categorical and sparse some clustering or dimensionality reduction may help to condense zip code information
- RMSE by Zip Code follows a similar pattern as home price by zip code, given that the homes in the most expensive zip codes have the highest RMSE as well. Homes in areas with high priced zip codes have higher square footage.

Decision Points

- A working idea we have is to identify the homes in the zipcodes who have the largest coefficient values from the regression and use that in assessing a home's value
- An additional idea is to identify the square footage of homes in the zipcodes and determine the price per square foot for a given zip code

Q&A (QUESTIONS & ANSWERS)

Questions?

Thank you!