California Housing Market Analysis

Group 2: Tommy Greve, Erik Lapszynski, Ariel Stanalonis, & Zach Vozzo

Table of Contents

Project Motivation and Background

California population: 39 million (most in U.S.)

- 3 of the 10 largest U.S. cities (L.A., San Diego, San Jose)
- 5 more of the 50 largest U.S. cities (San Francisco, Fresno, Sacramento, Long Beach, Oakland)

California real estate market is the most chaotic in the U.S.!

Project Motivation and Background

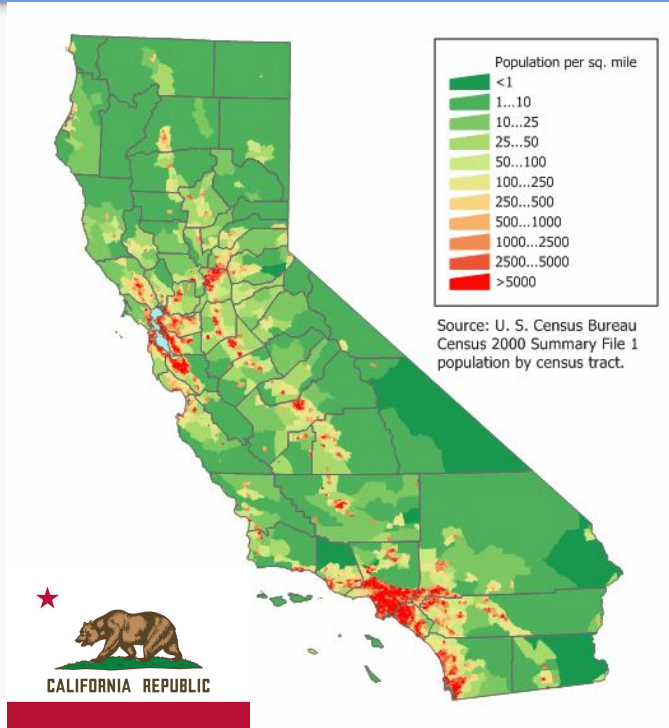
More problems:

- Not enough new properties built to satisfy demand
- Shortage constantly drives up costs for renters and homeowners in heavily populated areas
- Cali. has highest rate of functional poverty, second lowest rate of homeownership in U.S.

Project Motivation and Background

How can we discover the root causes of the California housing crisis and prevent it from continuing?

Are there any solutions to help prevent homeowners and renters in California from getting exploited by real estate agents and landlords?



Data Description

Initial Kaggle dataset:

- 35,390 rows/listings
- 39 columns/dimensions
- Mix of quant. & qual. data

Concerns:

- Derivative columns
- Pointless inputs
- Assigning roles/levels

Name	Role	Level
bathrooms	Input	Nominal
bedrooms	Input	Nominal
buildingArea	Input	Interval
city	Text Locatic	Nominal
cityId	Rejected	Interval
country	Rejected	Nominal
county	Text Locatic	Nominal
countyId	Rejected	Interval
currency	Rejected	Nominal
datePostedStri	Time ID	Interval
description	Rejected	Nominal
event	Rejected	Nominal
garageSpaces	Input	Nominal
hasBadGeocod	Rejected	Binary
hasGarage	Input	Binary
hasPetsAllowe	Input	Binary
homeType	Classificatio	Ordinal
id	ID	Nominal
isNewConstruc	Input	Binary
is_bankOwned	Input	Binary
is_forAuction	Input	Binary
latitude	Rejected	Interval
levels	Input	Nominal
livingArea	Input	Interval
livingAreaValue	Input	Interval
longitude	Rejected	Interval
lotAreaUnits	Rejected	Nominal
parking	Input	Binary
pool	Input	Binary
price	Target	Interval
pricePerSquare	Input	Interval
spa	Input	Binary
state	Rejected	Nominal
stateId	Rejected	Interval
streetAddress	Rejected	Nominal
time	Rejected	Interval
VAR1	Rejected	Interval
yearBuilt	Input	Nominal
zipcode	Rejected	Nominal

Data Preparation Activities

Cleaning/scrubbing data with Excel:

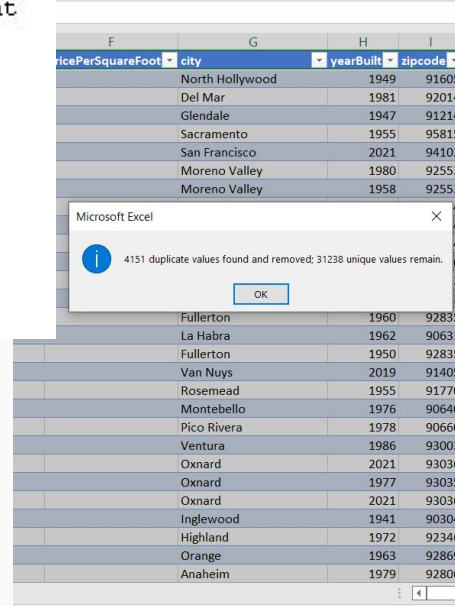
- Eliminated duplicate listings
- Imputed missing values
- Filtered outliers (i.e. lots)
- Standardizing levels

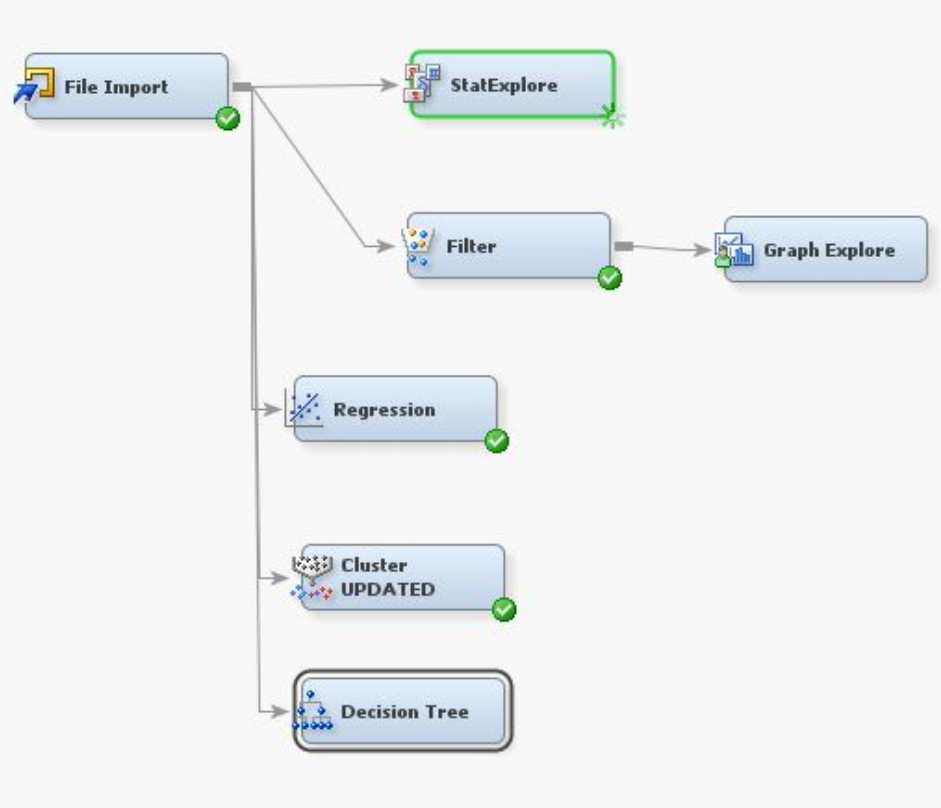
After cleaning:

- 3,152 unique rows/listings
- 17 columns/dimensions

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	1
INPUT	BINARY	8
INPUT	INTERVAL	4
INPUT	NOMINAL	5
REJECTED	BINARY	1
REJECTED	INTERVAL	7
REJECTED	NOMINAL	8
TARGET	INTERVAL	1





SAS Enterprise Miner Models

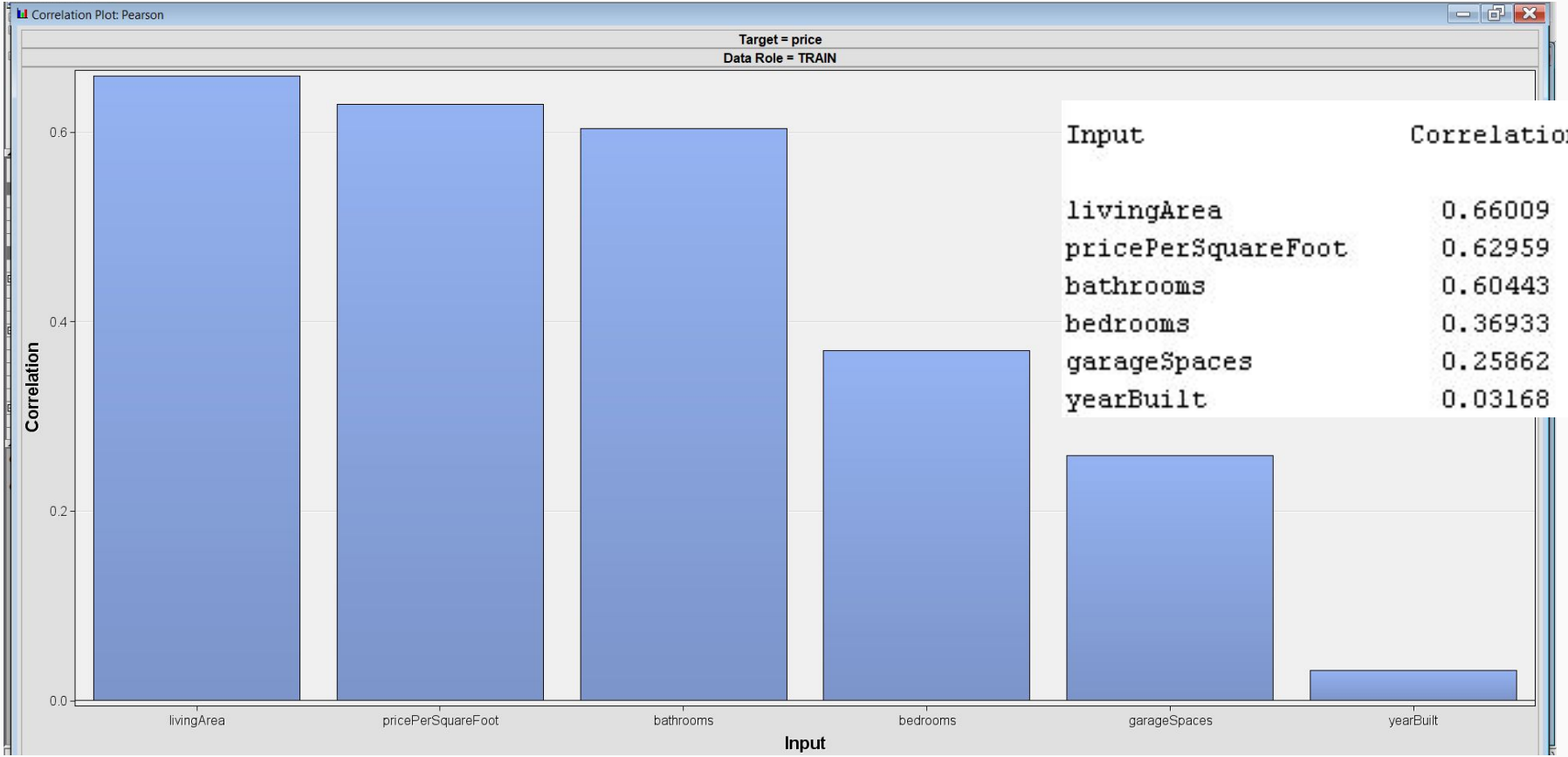
Exploring the data:

- StatExplore
- GraphExplore

Predictive models:

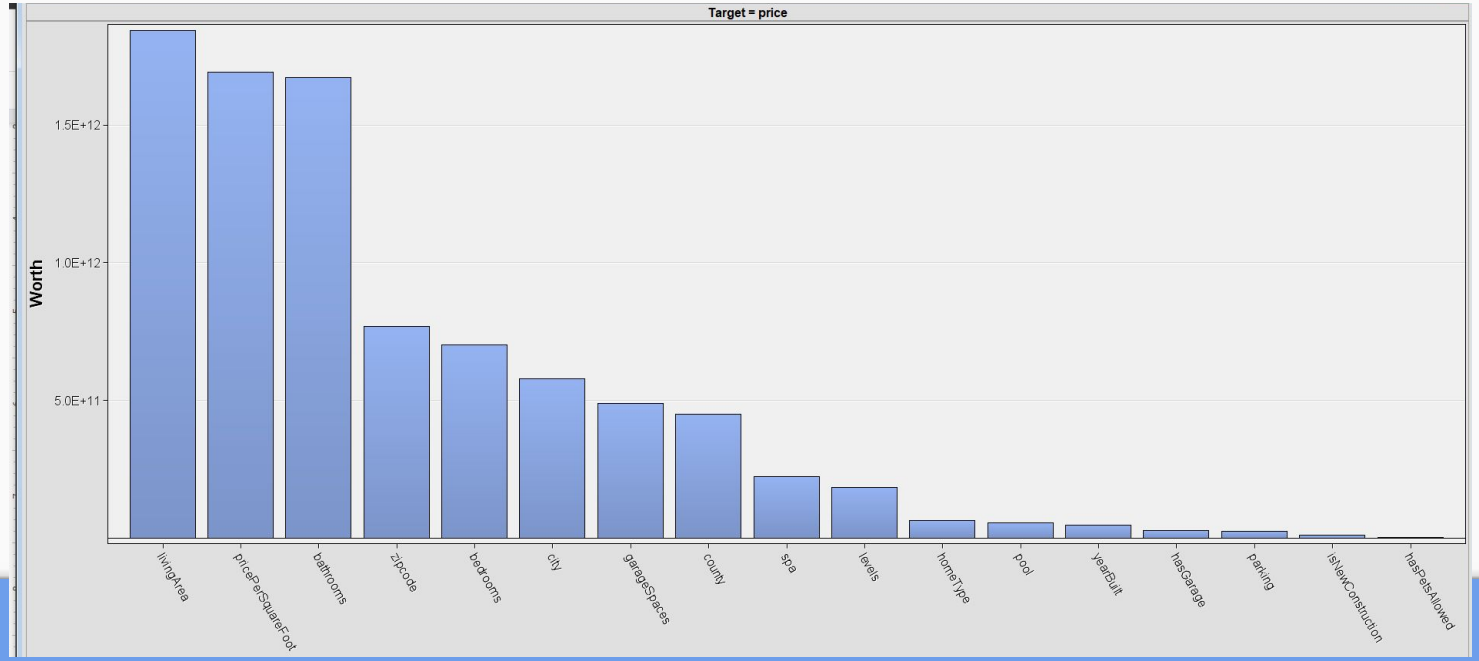
- Linear regression
- Cluster analysis
- Decision tree analysis

StatExplore Outputs

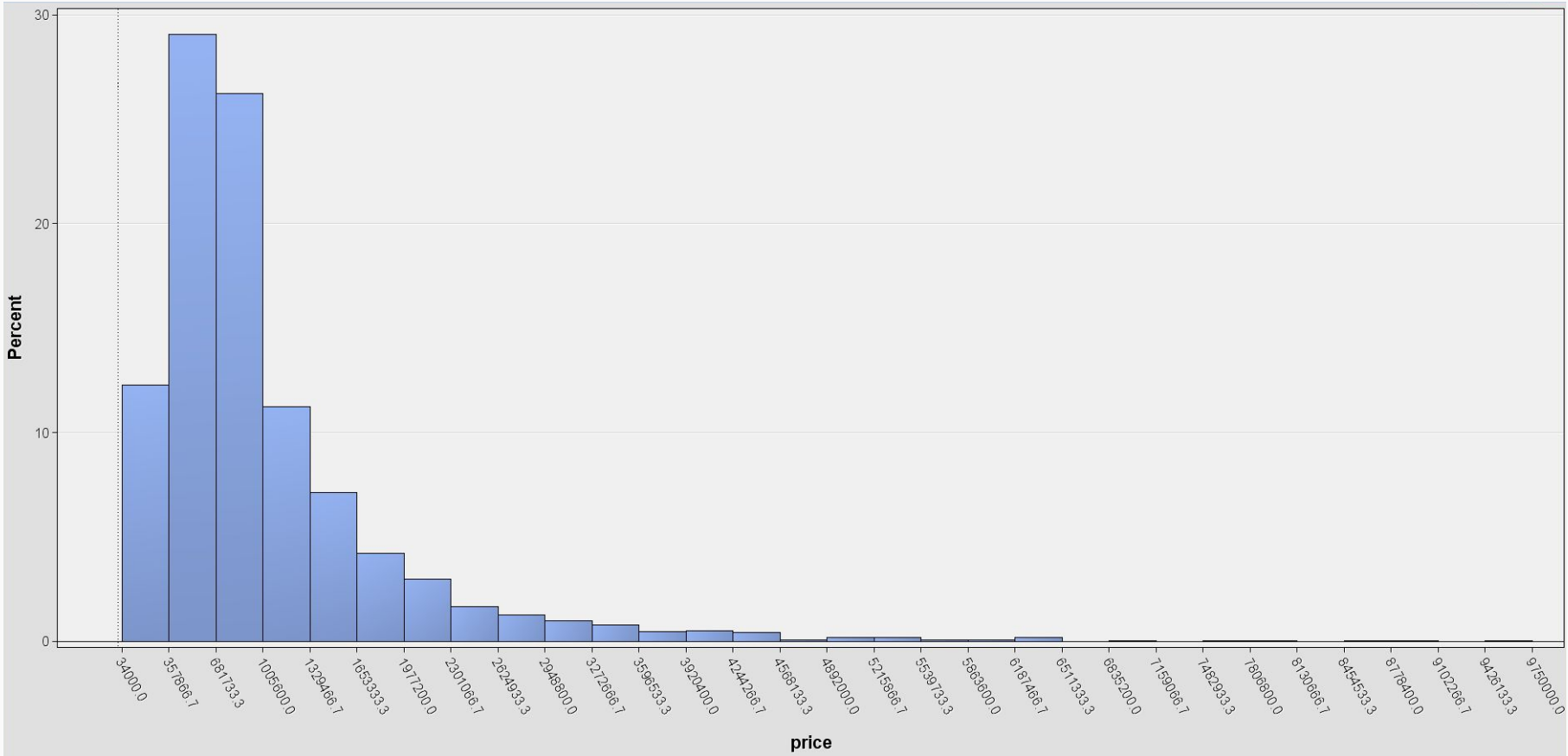


Input	Correlation
livingArea	0.66009
pricePerSquareFoot	0.62959
bathrooms	0.60443
bedrooms	0.36933
garageSpaces	0.25862
yearBuilt	0.03168

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
bathrooms	INPUT	2.537485	1.260895	3228	0	1	2	12	2.002661	8.082723
bedrooms	INPUT	3.284696	1.232421	3228	0	1	3	24	2.248262	26.70039
garageSpaces	INPUT	1.600372	1.241329	3228	0	0	2	18	1.952036	16.57831
livingArea	INPUT	2066.938	1445.596	3228	0	324	1701	20125	4.241803	31.96945
pricePerSquareFoot	INPUT	566.2261	388.0921	3228	0	14	515	5556	4.120227	36.11161
yearBuilt	INPUT	1971.964	29.51943	3228	0	1860	1977	2022	-0.63106	-0.03015
price	TARGET	1243947	1869448	3228	0	34000	799000	43000000	9.232655	138.1707



GraphExplore Output



Linear Regression Analysis

Significant Variables

1. Bathrooms
2. City
3. Home type
4. Living area
5. Parking
6. Pool
7. Price per square foot
8. Spa

Linear Regression Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	499	9.1787701E15	1.8394329E13	23.91	<.0001
Error	2728	2.0990689E15	769453426564		
Corrected Total	3227	1.1277839E16			

Model Fit Statistics

R-Square	0.8139	Adj R-Sq	0.7798
AIC	88803.7055	BIC	88988.9227
SBC	91843.5145	C(p)	500.0000

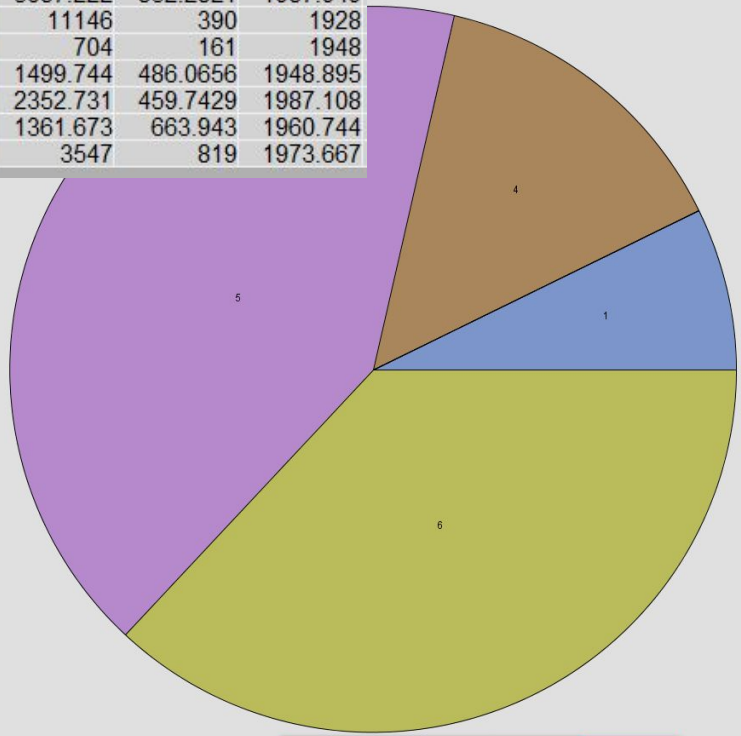
Effect	DF	Sum of Squares	F Value	Pr > F
bathrooms	1	4.62965E13	60.17	<.0001
bedrooms	1	2.47667E11	0.32	0.5705
city	316	4.99812E14	2.06	<.0001
county	3	4.45938E12	1.93	0.1223
garageSpaces	1	1.11778E12	1.45	0.2282
hasGarage	1	6.21301E11	0.81	0.3690
hasPetsAllowed	1	4.73829E11	0.62	0.4327
homeType	3	2.38725E13	10.34	<.0001
isNewConstruction	1	8.64859E12	11.24	0.0008
is_bankOwned	1	4.79058E11	0.62	0.4302
is_forAuction	1	2.12226E10	0.03	0.8681
levels	5	9.70317E12	2.52	0.0276
livingArea	1	6.8352E14	888.32	<.0001
parking	1	1.05549E13	13.72	0.0002
pool	1	4.03381E12	5.24	0.0221
pricePerSquareFoot	1	1.54997E15	2014.38	<.0001
spa	1	5.48883E12	7.13	0.0076
yearBuilt	1	7.29569E11	0.95	0.3303
zipcode	158	1.35252E14	1.11	0.1660

Cluster Analysis

- Most expensive characteristics:
  - 4 bed/3 bath, garage accessibility, built in 1987+
- 4 reasonable unique clusters generated
- Most prominent housing available, most expensive

Cluster Analysis Output

Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	bathrooms	bedrooms	garageSpaces	livingArea	pricePerSquareFoot	yearBuilt
1	234	0.211298	13.09613	5	3.882629	4.931624	4.824786	2.92735	5087.222	832.2821	1987.949
2	1	.	0	1	20.23652	9	24	2	11146	390	1928
3	1	.	0	4	56.85803	1	2	0	704	161	1948
4	457	0.12698	10.98534	6	3.577526	1.857768	2.80744	0.004376	1499.744	486.0656	1948.895
5	1338	0.105418	13.16659	6	2.13627	2.949178	3.804933	2.224963	2352.731	459.7429	1987.108
6	1194	0.114606	13.28522	5	2.13627	1.859296	2.562814	1.249581	1361.673	663.943	1960.744
7	3	0.131439	4.157642	5	32.88732	4	4.666667	2.666667	3547	819	1973.667



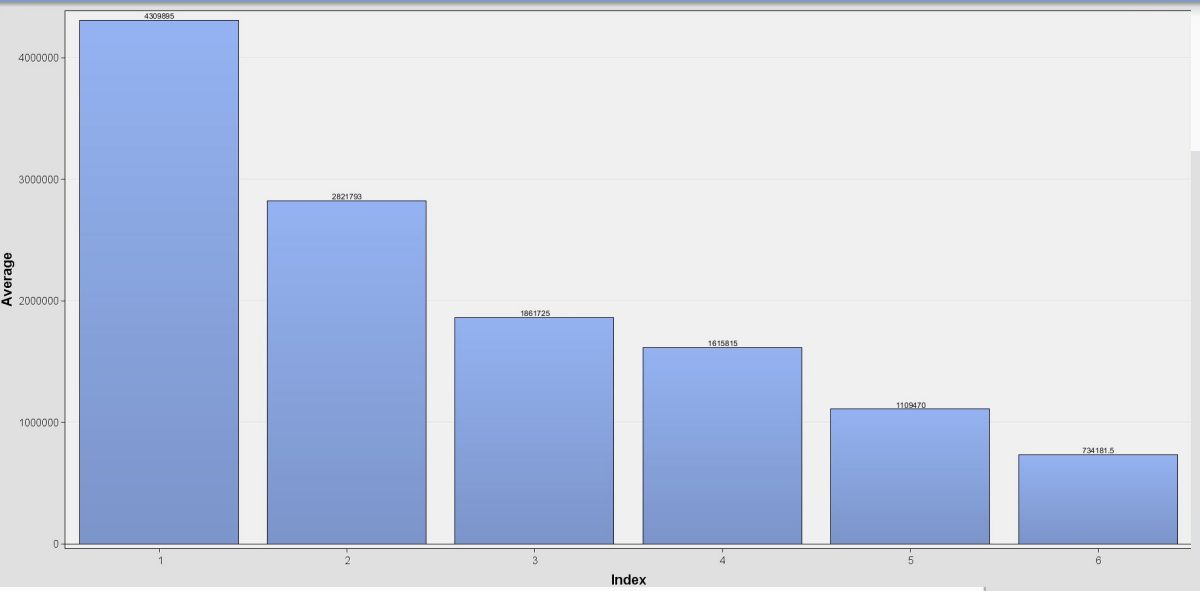
Centroid clustering method:

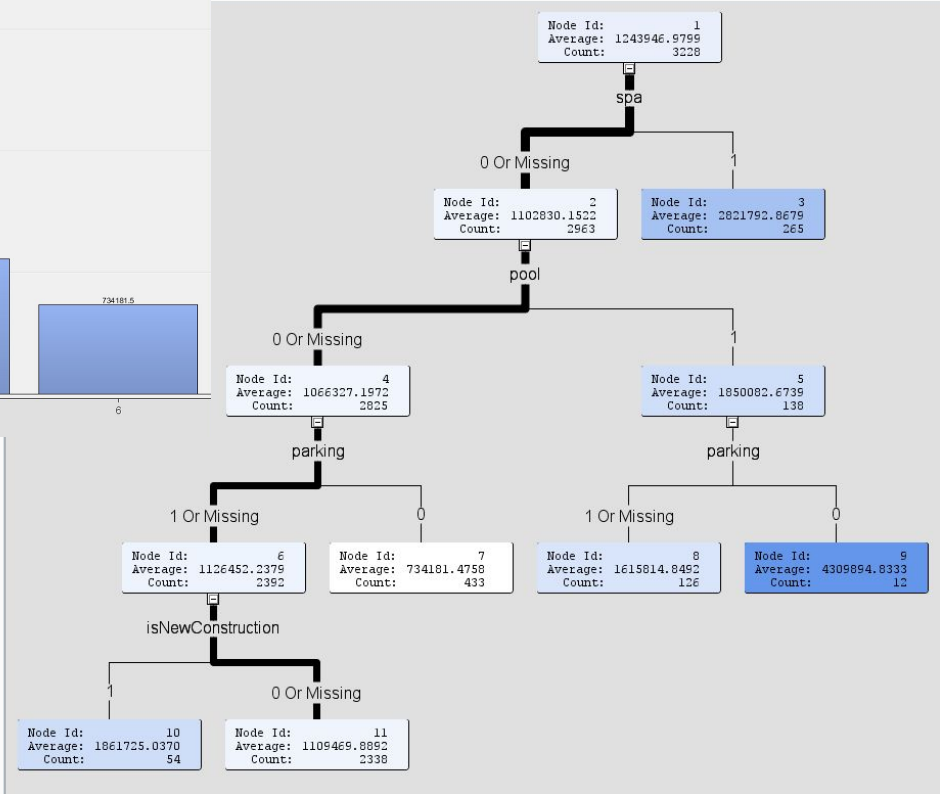
- **Cluster 1**: newer houses with 5+ bed/bath
- **Cluster 4**: oldest/smallest living spaces
- **Cluster 5:** newer/moderately sized houses
- **Cluster 6:** newer and smaller living spaces

Decision Tree Analysis

- Utilized binary variables to predict house price ranges
  - 4 significant variables:
    - Spa
    - Pool
    - Parking
    - isNewConstruction
  - 6 tree nodes produced

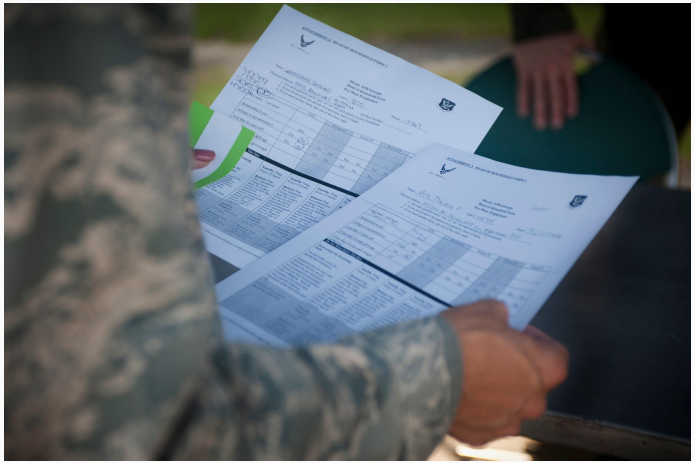
Decision Tree Output





Results and Findings

1. Expensive homes include certain amenities that drive up their price.
2. Properties tend to be small and expensive, compared to Pennsylvania. (Price per square foot in the data is quite high at $566 versus $151)
3. Certain amenities are more valuable and have a greater indication of the price than others



Managerial Implications and Conclusion

1. Eliminate the “extras” when possible.
2. Home buyers might want to purchase a “fixer upper”
3. Understand that there are more individuals than homes
4. More developers need to be drawn to the market

Further Research:

- Looking into different pricing based on location (diagnose areas where the inflation of houses is greater)

- Analysis of California housing prices over time