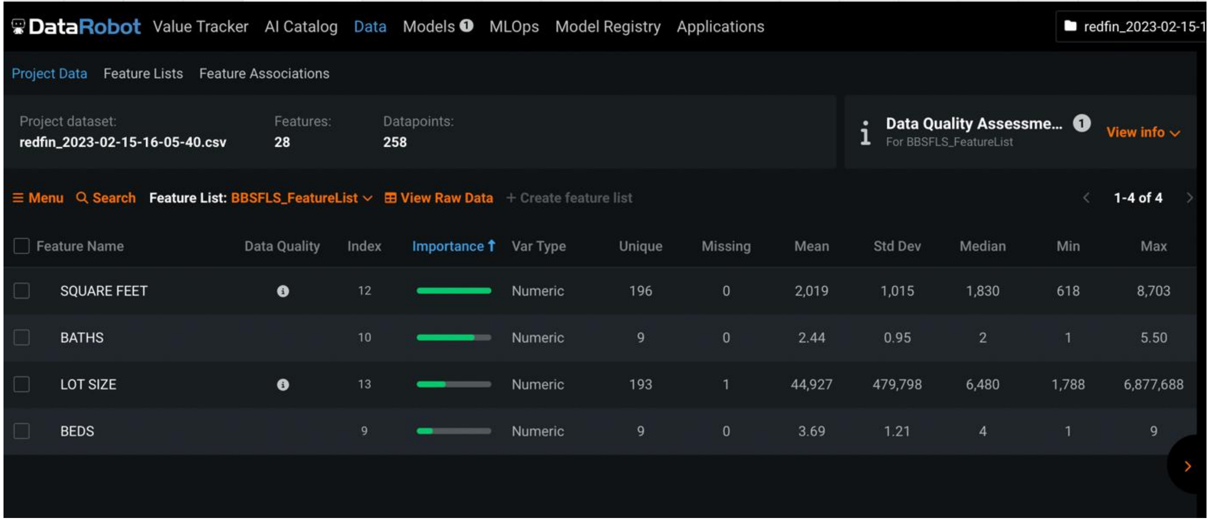
Assignment 2 - Real Estate Prices

Business Case:
Real Estate represents a single largest asset class (roughly around $217 trillion). People involved such as seller, buyer, brokers, lenders, insurers, investors, renters and other such market participants would be interested in a model to predict fair market value for residential properties.

To explore the potential for using MLS listings as a source of data to predict asking real estate prices , we will focus on single family homes located in San Jose, California. 258 properties that are currently in the market was sourced from redfin.com

Question 1

Feature set including bathrooms, bedrooms, lot size and square feet .
Property type is not included as we are working with one type of property that is single family homes and inclusion of the same will not have any effect on the model.



Project dataset: redfin_2023-02-15-16-05-40.csv

Features: 28

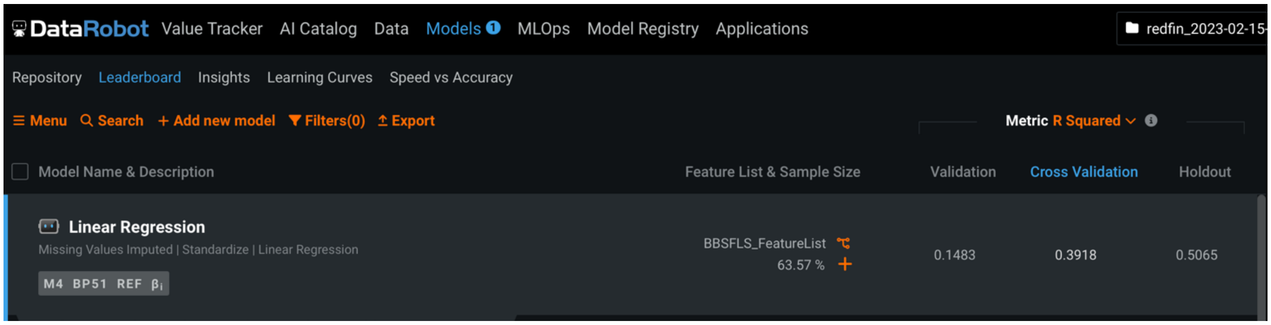
Datapoints: 258

Feature List: BBSFLS_FeatureList

□ Feature Name	Data Quality	Index	Importance ↑	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
□ SQUARE FEET	ℹ	12		Numeric	196	0	2,019	1,015	1,830	618	8,703
□ BATHS		10		Numeric	9	0	2.44	0.95	2	1	5.50
□ LOT SIZE	ℹ	13		Numeric	193	1	44,927	479,798	6,480	1,788	6,877,688
□ BEDS		9		Numeric	9	0	3.69	1.21	4	1	9

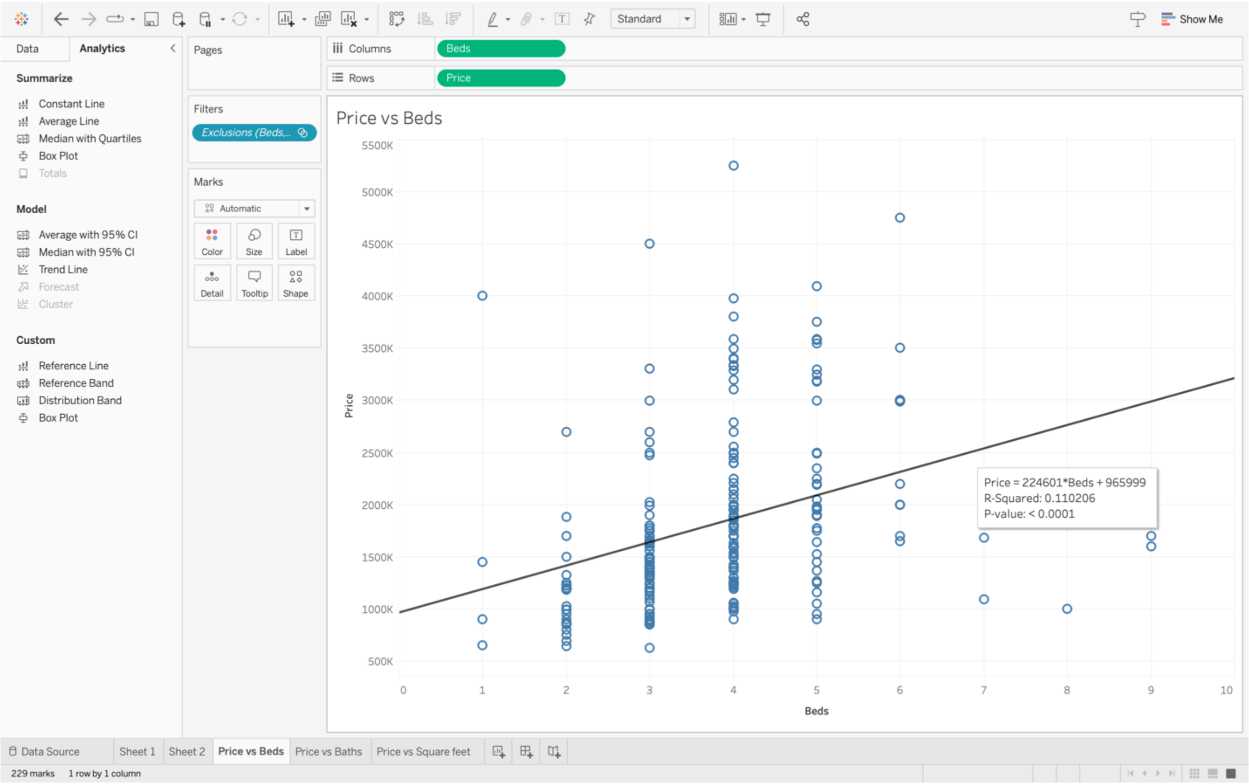
Question 2

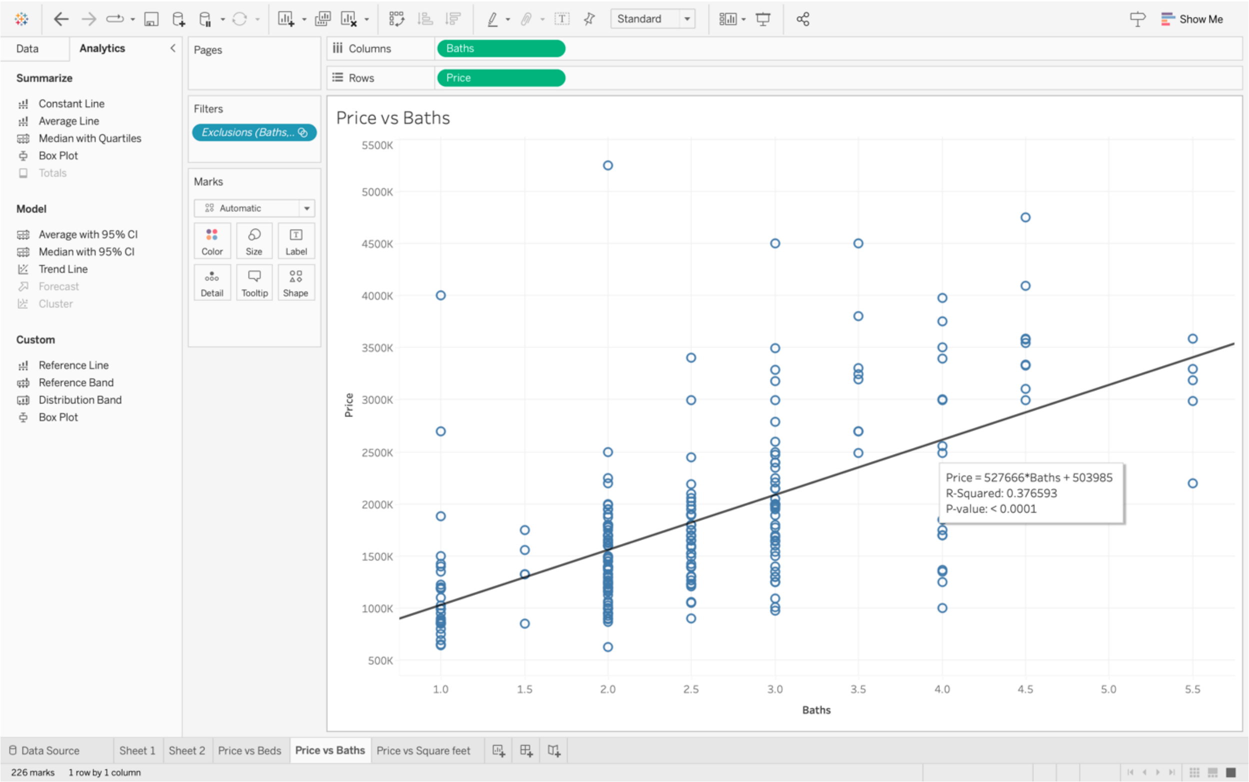
Data	R2	MAPE	MAE	RMSE
Cross-Validation	0.39	22.40%	$403,626	$639,783
Holdout	0.51	26.05%	$388,047	$504,901



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Linear Regression Missing Values Imputed | Standardize | Linear Regression M4 BP51 REF βi	BBSFLS_FeatureList 63.57 %	0.1483	0.3918	0.5065

Question 3





Columns: Baths

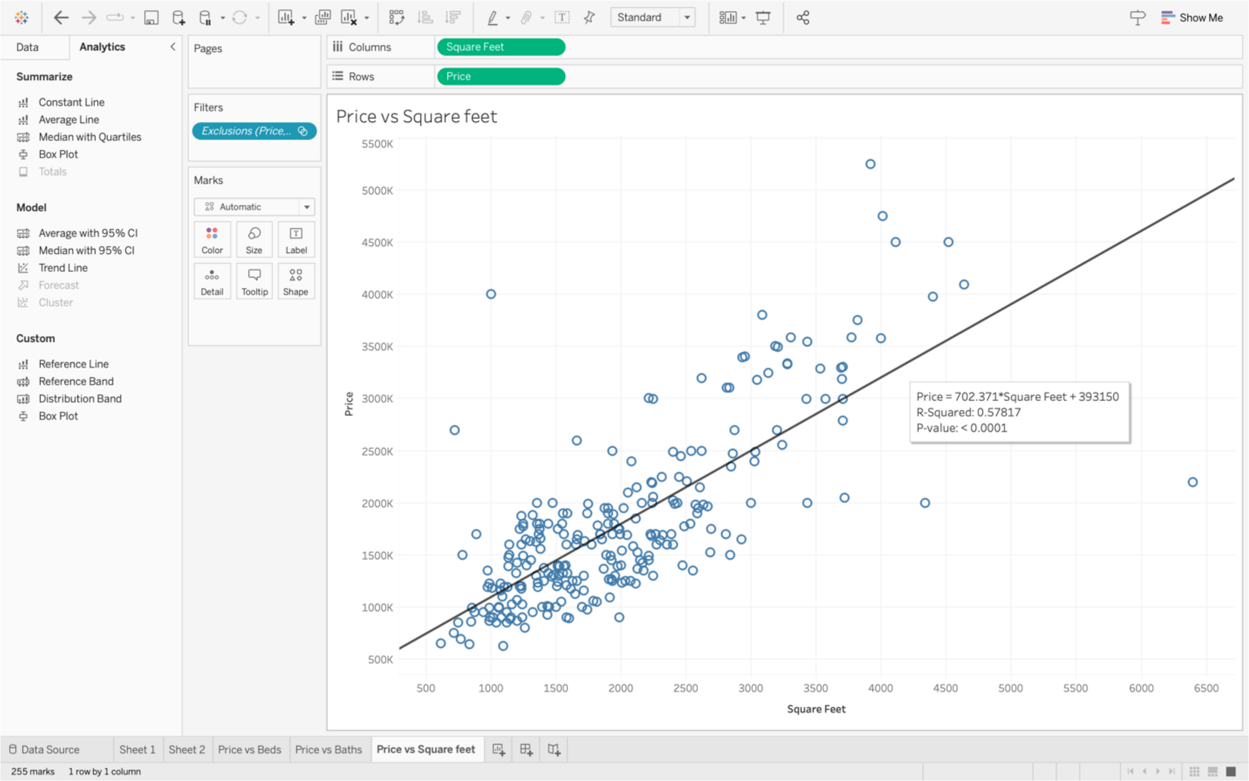
Rows: Price

Filters: Exclusions (Baths,...

Price vs Baths

Price = 527666*Baths + 503985
R-Squared: 0.376593
P-value: < 0.0001

226 marks 1 row by 1 column



Columns: Square Feet

Rows: Price

Filters: Exclusions (Price,...

Price vs Square feet

Price = 702.371*Square Feet + 393150
R-Squared: 0.57817
P-value: < 0.0001

255 marks 1 row by 1 column

Visualization excludes 1 outlier property priced at $7,495,000 having 5 beds and 5.5 baths. Square feet is selected for this visualization as it had no missing values.

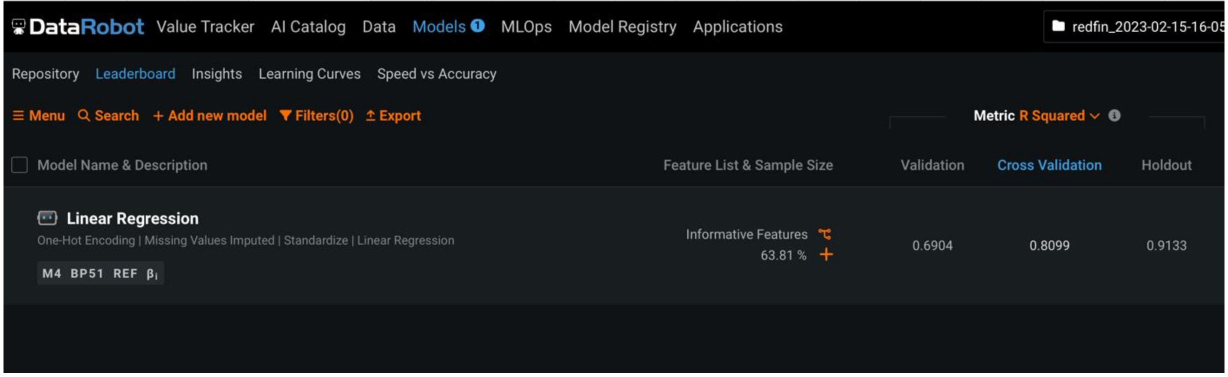
Predictor	R2
Beds	0.11
Baths	0.377
Square Feet	0.578

Question 4

Based on R2 error, square feet is the best predictor for real estate asking price for homes in San Jose, California as of February 2023. That means the area of the house decides the price of the property.

Additional Note:

When the outlier property priced at $7,495,000 is removed and the model is rerun it gives a better performance:



Metric R Squared

□ Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Linear Regression One-Hot Encoding | Missing Values Imputed | Standardize | Linear Regression M4 BP51 REF βᵢ	Informative Features 63.81 %	0.6904	0.8099	0.9133

Data	R2	MAPE	MAE	RMSE
Cross-Validation	0.81	9.36%	$180,934	$328,332
Holdout	0.91	7.62%	$150,372	$272,170