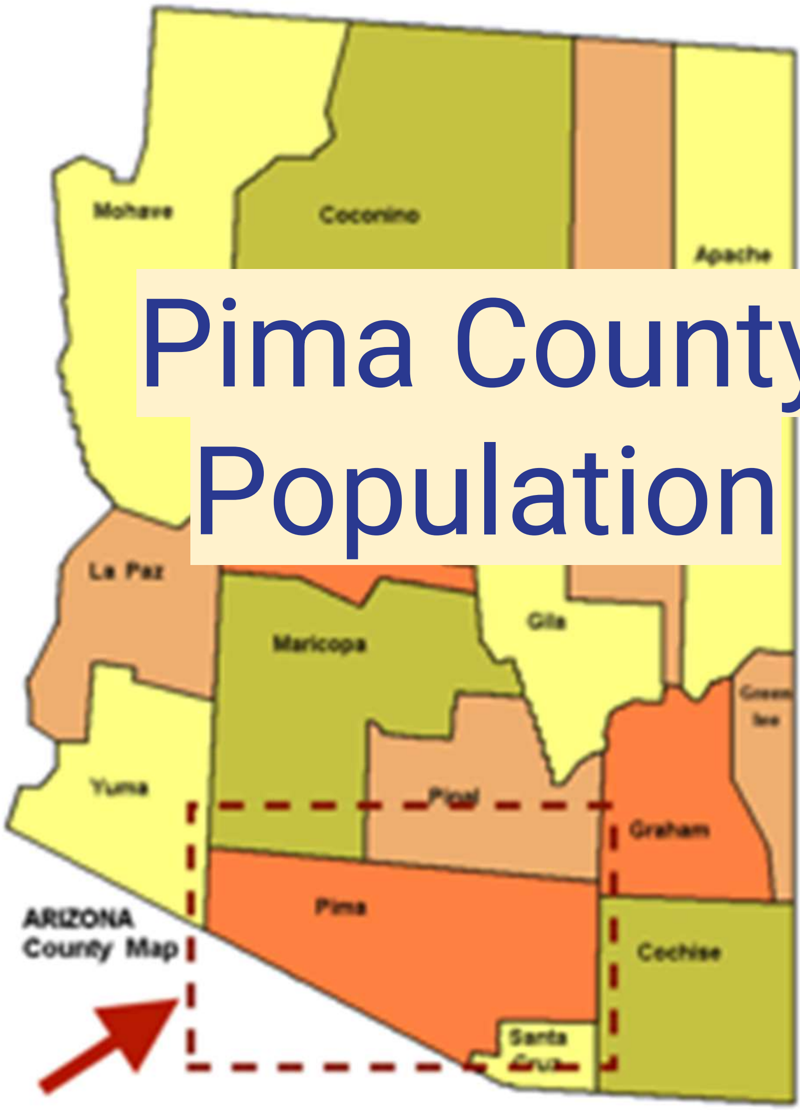
Pima County Housing Use Case

Stacking Models: Price/Sqft - Category Classification & Price Forecasting

David Gonzalez
John Kusmaul
Matthew Langschwager
Kevin Lee
Pablo Reynoso

Enhance IT

Pima County Population



- Population: 1,043,433
- Age Distribution: 18-65 (53.7%)
- Median Gross Rent: $907 USD
- Adults 25+ / Bachelor Diploma: 32.4%
- Median Wages Year: $53,379 USD
- Median Wages Year (per capita): $29,707 USD

Stacking Models for Sold Price Forecasting

Pima Dataset → Price/Sqft Category Classification → Sold Price Regression → Sold Price Forecasting for X location

- Dataset (LAT, LON, SoldPrice, SQFT):
  - Description
  - Statistics
- Cleaning Data:
  - Missing Data
  - Outliers
- Feature Engineering:
  - price/sqft category
  - floor_covering
- Feature Normalization:
  - min-max

- Cleaning Data:
  - Missing Data
  - Outliers
- Feature Engineering:
  - diff(price_sqft, min_category)
- KNN Classifier

- OLS Regressor

The Dataset for *Price-Sqft Category Classification*

Pima11B.csv

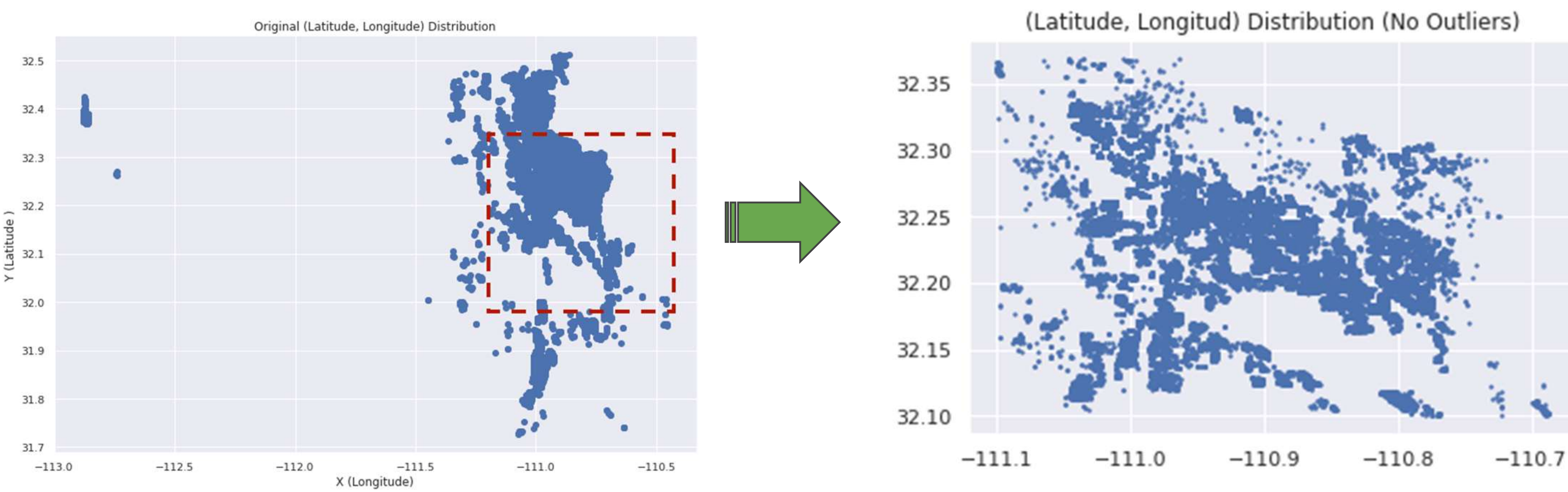
Dimensions:

- Observations: 52918
- Features: 49
- Variables of Interest:
  - Latitude
  - Longitude
  - Sold Price
  - Sqrt_ft

SalePrice	LAT	LON	ZIP	ROOMS	...	SQFT
8062312	32.1681	-110.98	78746	6	...	1172
...	...	...	...	...	...	...
0	32.3163	-111.03	85741	5	...	1571

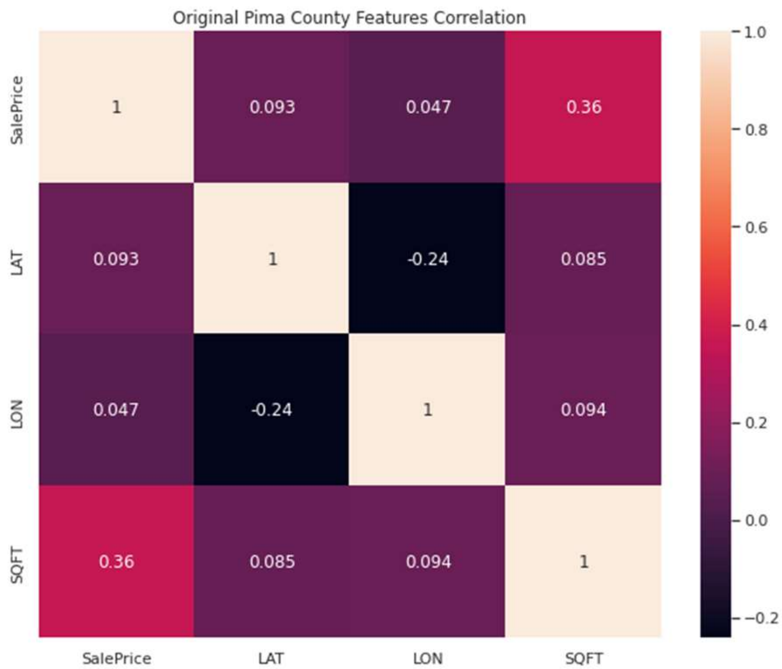
Data Cleaning: Outliers Removal

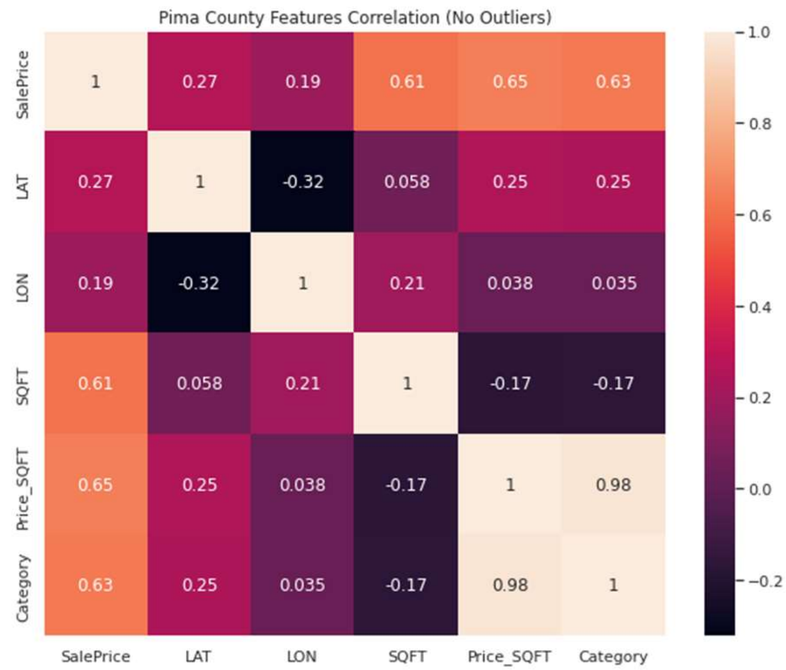
Pima County Map Reduction



Data Cleaning: Outliers Removal

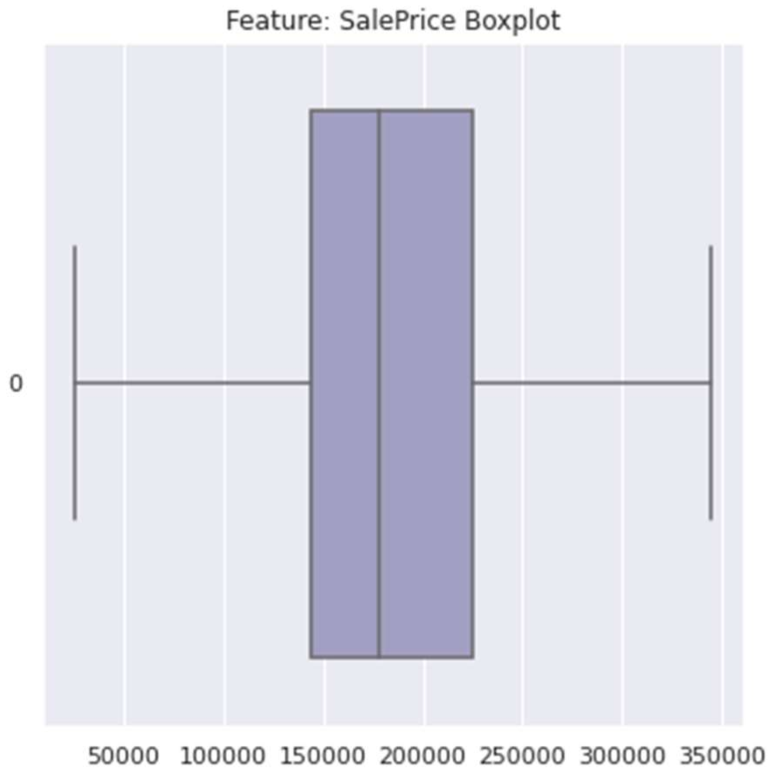
Features Correlation Optimization

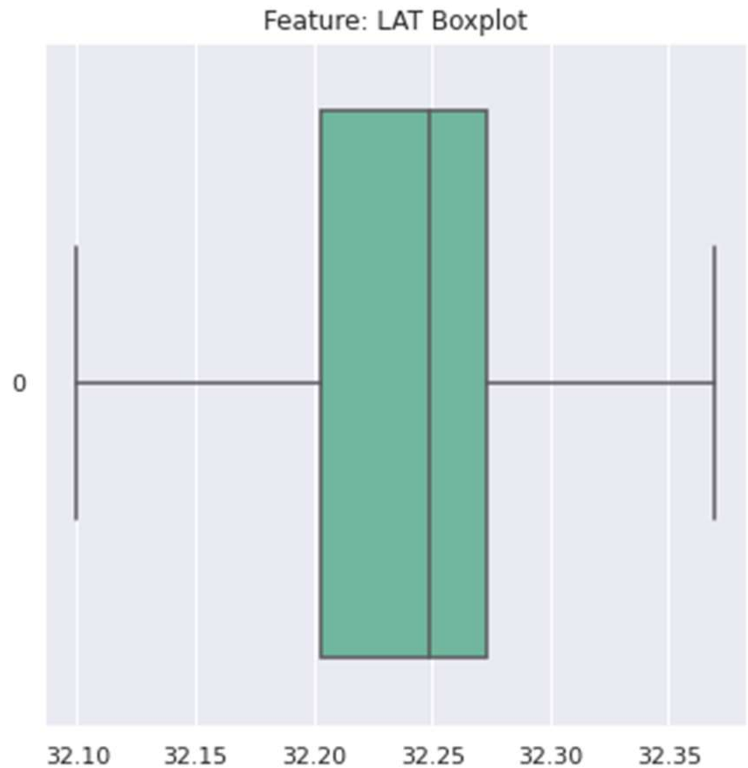




Data Cleaning: Outliers Removal

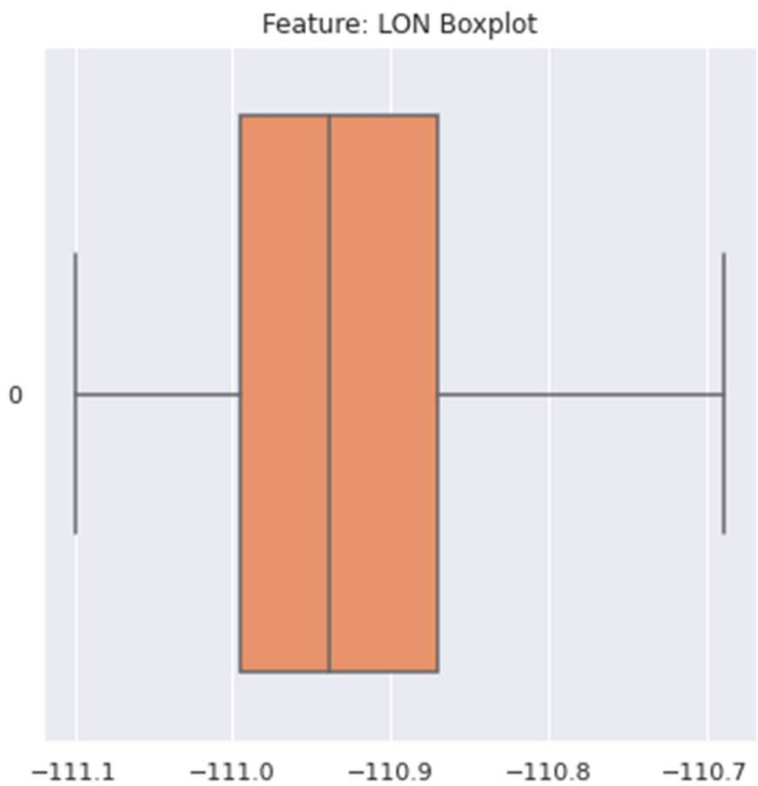
Features Whisker Boxplots (Sale Price, Latitude)

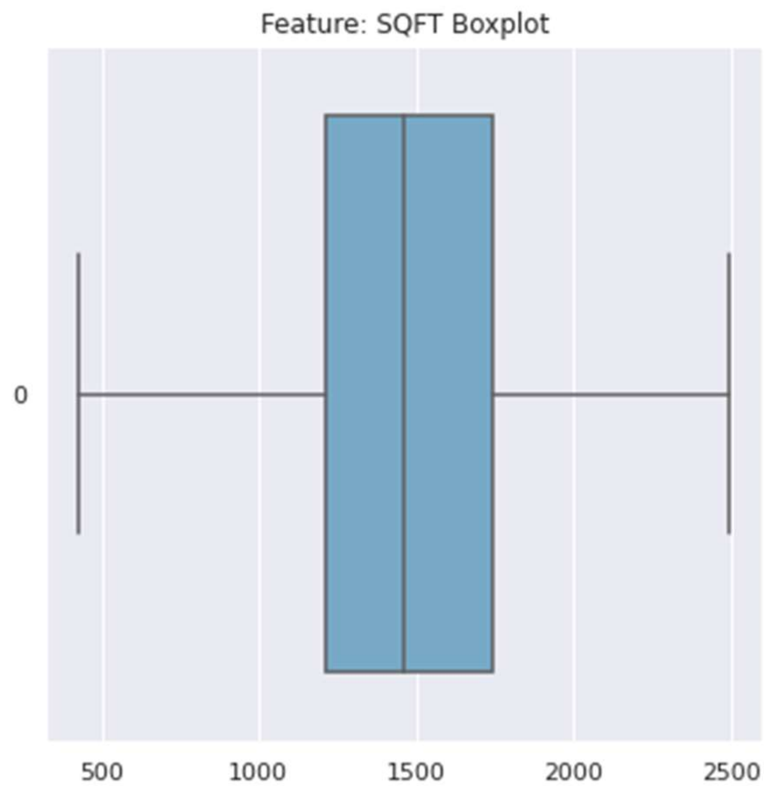




Data Cleaning: Outliers Removal

Features Whisker Boxplots (Longitude, SQFT)





Feature Engineering

Data Preprocessing

- Add Columns

  (‘ Price_SQFT ’)

- Generate a Category

  (‘ Category ‘)

Feature Engineering

`preprocess_data()`

Calculate Bins of evenly intervals using [Max, Min] of

‘ Price_SQRT ‘

Bins = [Max, Min] / N

N = 20

Feature Engineering

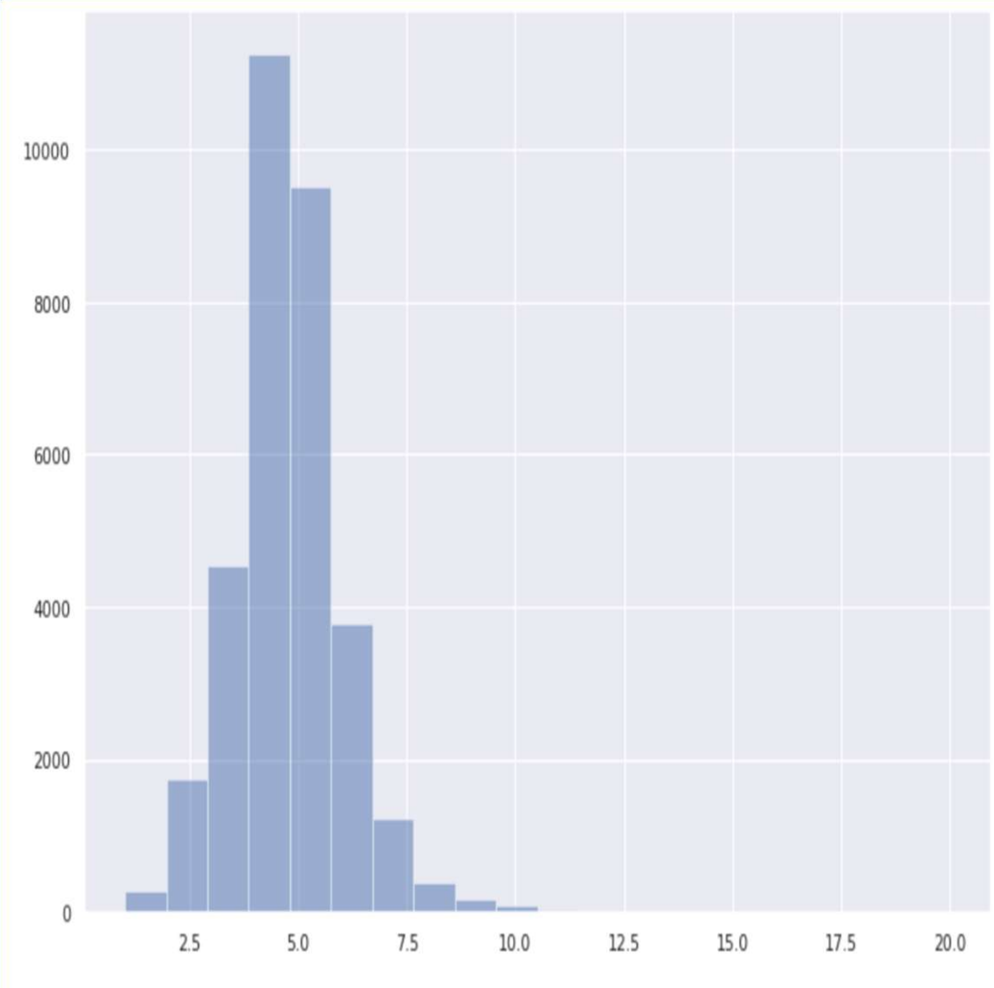
`preprocess_data()`

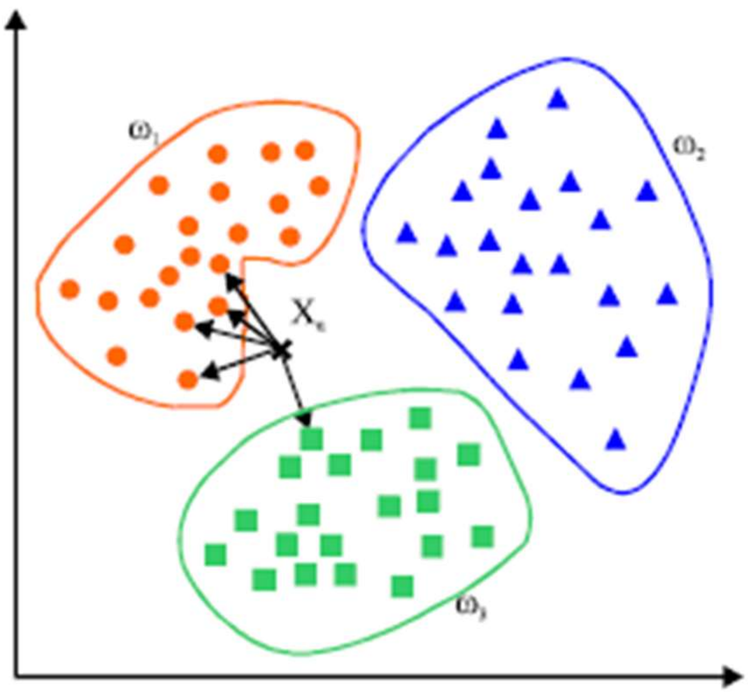
	SalePrice	LAT	LON	SQFT	Price_SQFT	Category
0	325000	32.251658	-110.954721	590	550.85	20
1	318889	32.235314	-110.961662	640	498.26	18
2	318889	32.235314	-110.961662	640	498.26	18
3	315000	32.239954	-110.937697	688	457.85	16
4	315000	32.239954	-110.937697	688	457.85	16
...	...	...	...	...	...	...
33066	37000	32.174740	-110.974629	1896	19.51	1
33067	42000	32.204042	-110.795308	2196	19.13	1
33068	26500	32.165369	-110.975436	1582	16.75	1
33069	26500	32.165369	-110.975436	1582	16.75	1
33070	28538	32.232718	-110.880543	2079	13.73	1

33071 rows × 6 columns

Feature Engineering

Distribution of Category





KNN Classifier

John

Splitting the Data

Split the data into two separate datasets: train and test

The dataset was sorted randomly then split

The train dataset contained 70% of the data

While the test set contained the remaining 30%

Finding Ideal K Value

Through trial and error determined that a K value of 2 returned the highest accuracy

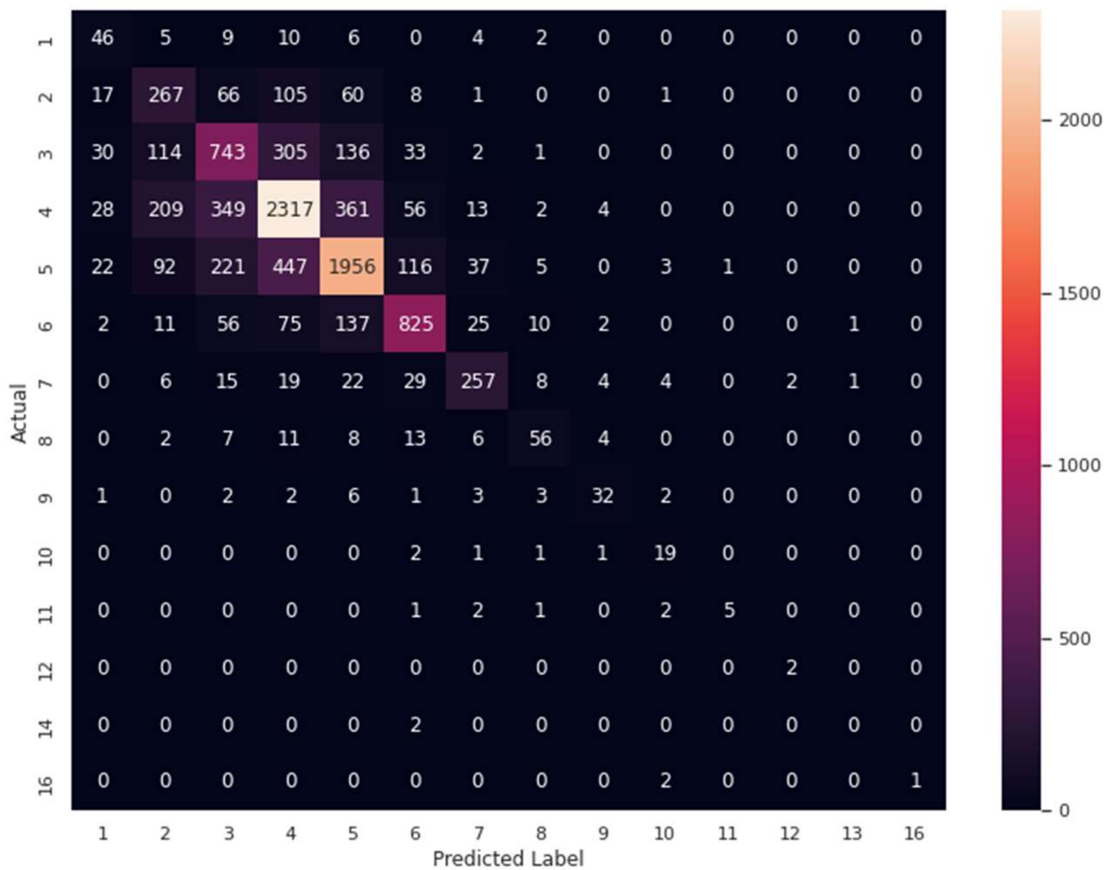
Tried different values for K such as 10, 7, 5 and 4

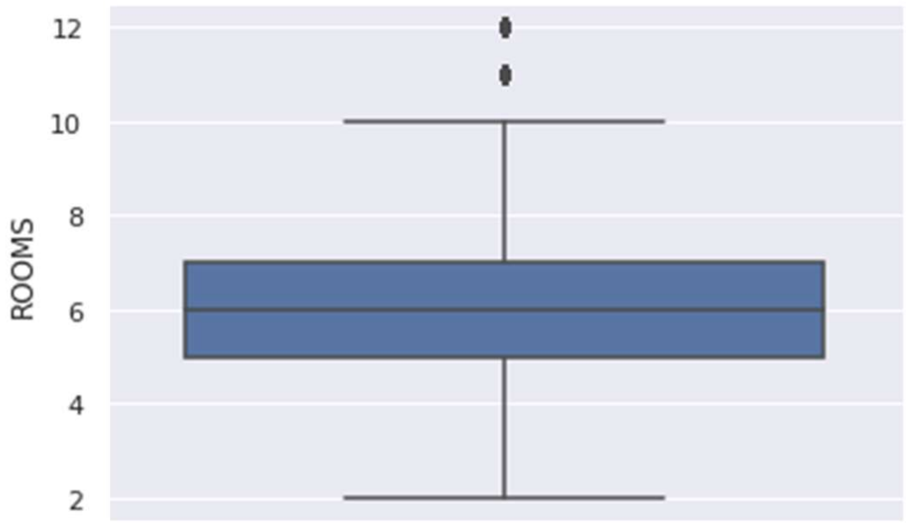
The results I ended up with for Train results were 92%

While the Test results were 65%

Given the limited amount of data entries and the fact that we were constrained to only using latitude longitude and price per square foot I believe that these figures are adequate

Confusion Matrix





Data Cleaning:

Regression Features

Matthew

Additional Feature Outliers

- The provided Pima County housing dataset contains 49 features
- Four were used in the classification model:
  - Sales Price
  - Square Footage
  - Longitude
  - Latitude
- An additional nine features were chosen for potential use in the regression model:
  - Sale Date
  - ZIP Code
  - Lot Size
  - Year Built
  - Number of Rooms
  - Number of Bathrooms
  - Number of Garages
  - Garage Capacity
  - Pool Size

Additional Feature Outliers

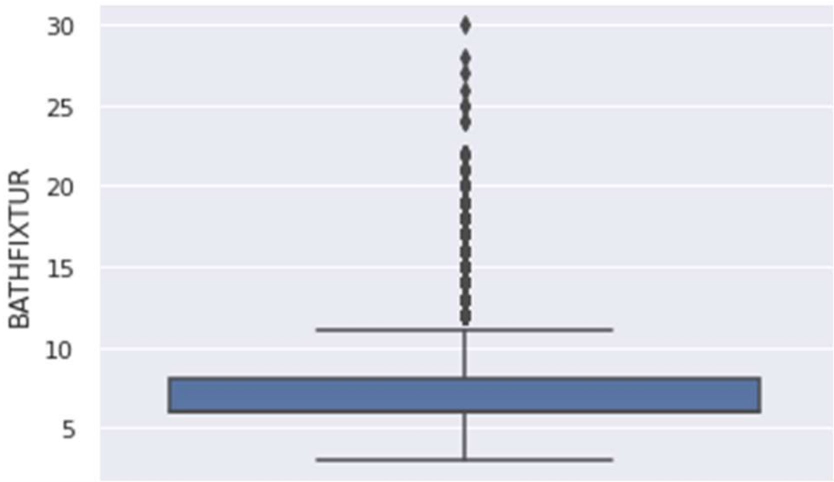
- For classification, the selected features were subjected to iterations of outlier removal, to ensure robust data
- A notable number of houses were cast off to achieve this

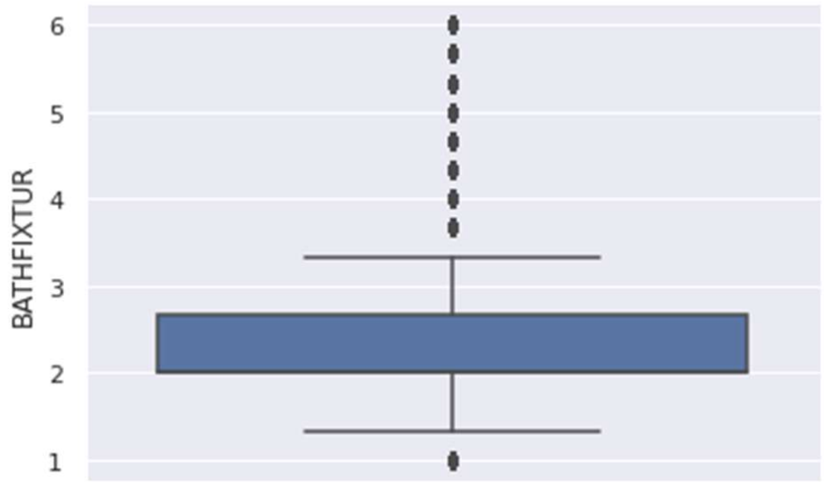
- For regression, outlier removal was also required, but the focus is minimizing additional loss of categories
- More data => better regression model

Additional Feature Outliers

	Original Range	Edited Range
Sales Price	201701 - 201908	(no change)
ZIP Code	0 - 99925	85600 - 85800
Lot Size	0.0315 - 74.27	0 - 6
Year Built	1898 - 2019	1902 - present
# of Rooms	1 - 99	1 - 13
# of Bathrooms*	3 - 30*	0 - 6
# of Garages	1 - 9	0 - 9
Garage Capacity	0 - 9	0 - 7
Pool Size	0 - 1100	(0 or 1)

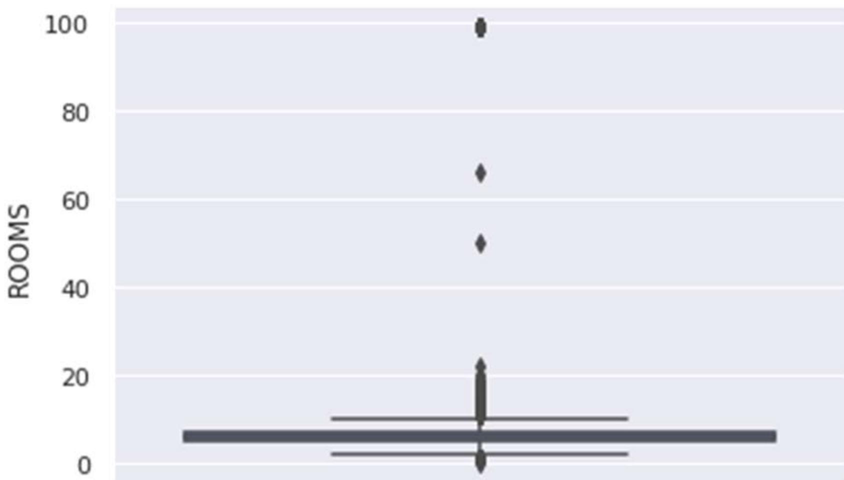
Additional Feature Outliers: Bath Fixtures

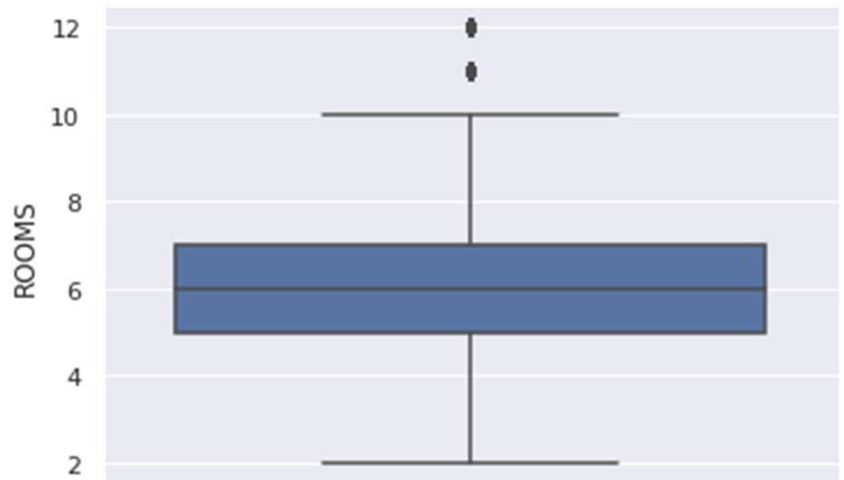




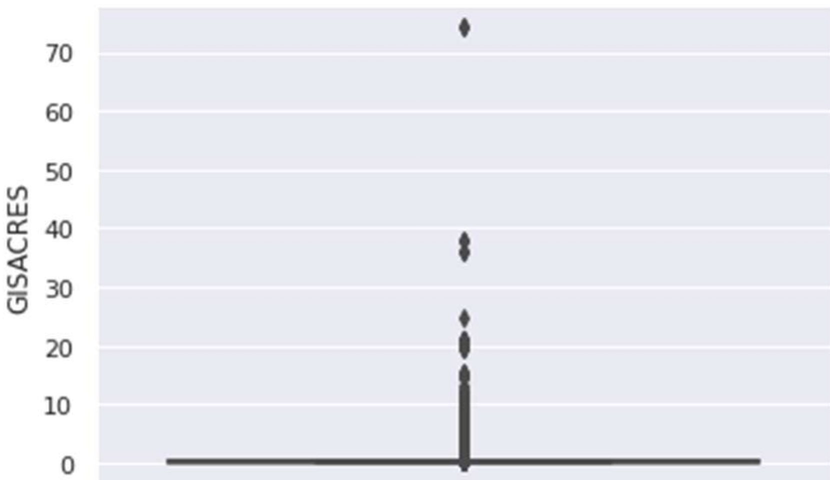
In order to convert “Bath Fixtures” to “Bathrooms”, an assumption of 3 fixtures per bathroom was made (i.e. sink, bathtub, toilet). Any non-whole number of bathrooms should be considered half-baths.

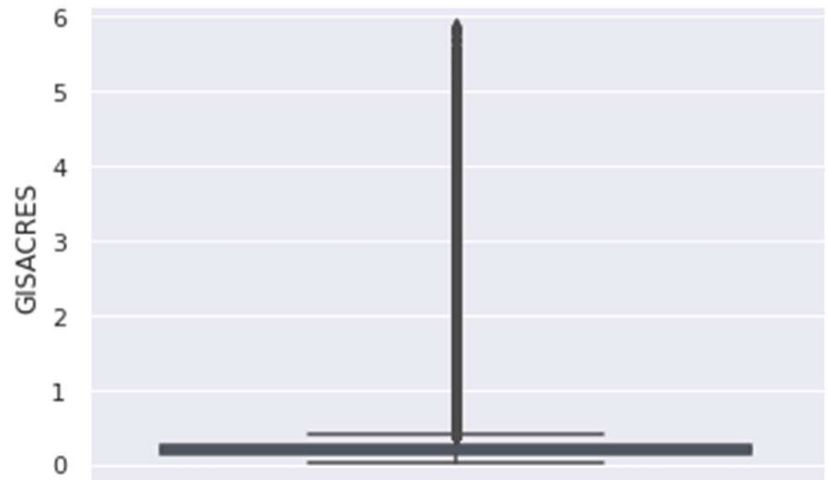
Additional Feature Outliers: Rooms

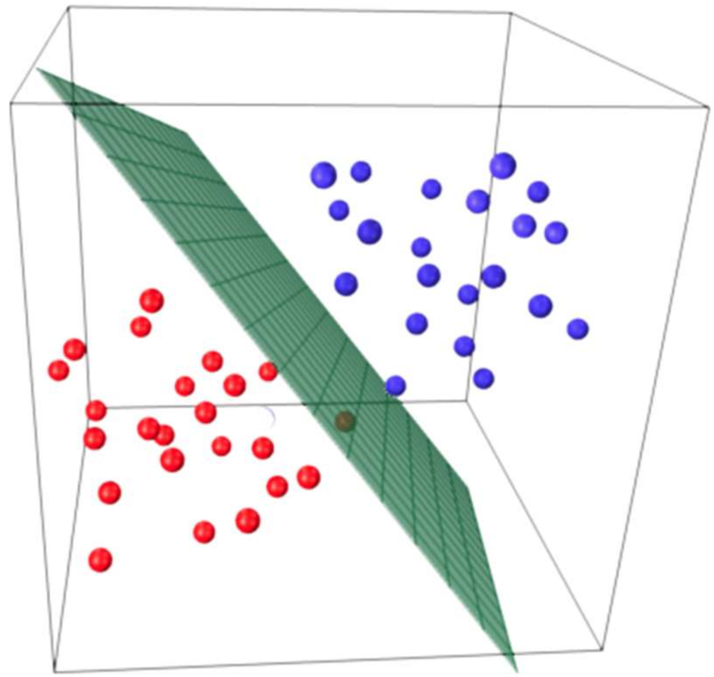




Additional Feature Outliers: Lot Size







Logistic Regression

(David)

Data PreProcessing (Adding PRICE/SQRT_FT)

In this regression model, the predicted variables in KNN Classifier are the interest values.

For this reason we concatenate them to our main database. (y_hat_train - y_hat_test).

PRICE_SQFT_CAT = y_hat_train (add) y_hat_test

Shape = (25564,)

Final Data PreProcessing

We take back the clean database with the addition of our new feature..

- Lon
- Lat
- Sale Price
- Sqft
- Sale Date
- Zip
- Gis Acres
- Year
- Rooms
- Bathrooms
- Garage
- Garage Capa
- Pool Area
- Price Sqft
- Category
- Price Sqft Cat

Shape = (25564, 16)

Obtaining the y_train (PRICE SQFT)

bins = (min [Price_SQFT], max [Price_SQFT], 20)

New column: **MIN_CAT (array(bins[min_indices])**

New variable

to predict

y_train = Price_Sqft - Min_Cat

Logistic Regression Model

```
X=d1_[['ROOMS', 'BATHFIXTUR', 'GARAGE', 'Category']]
```

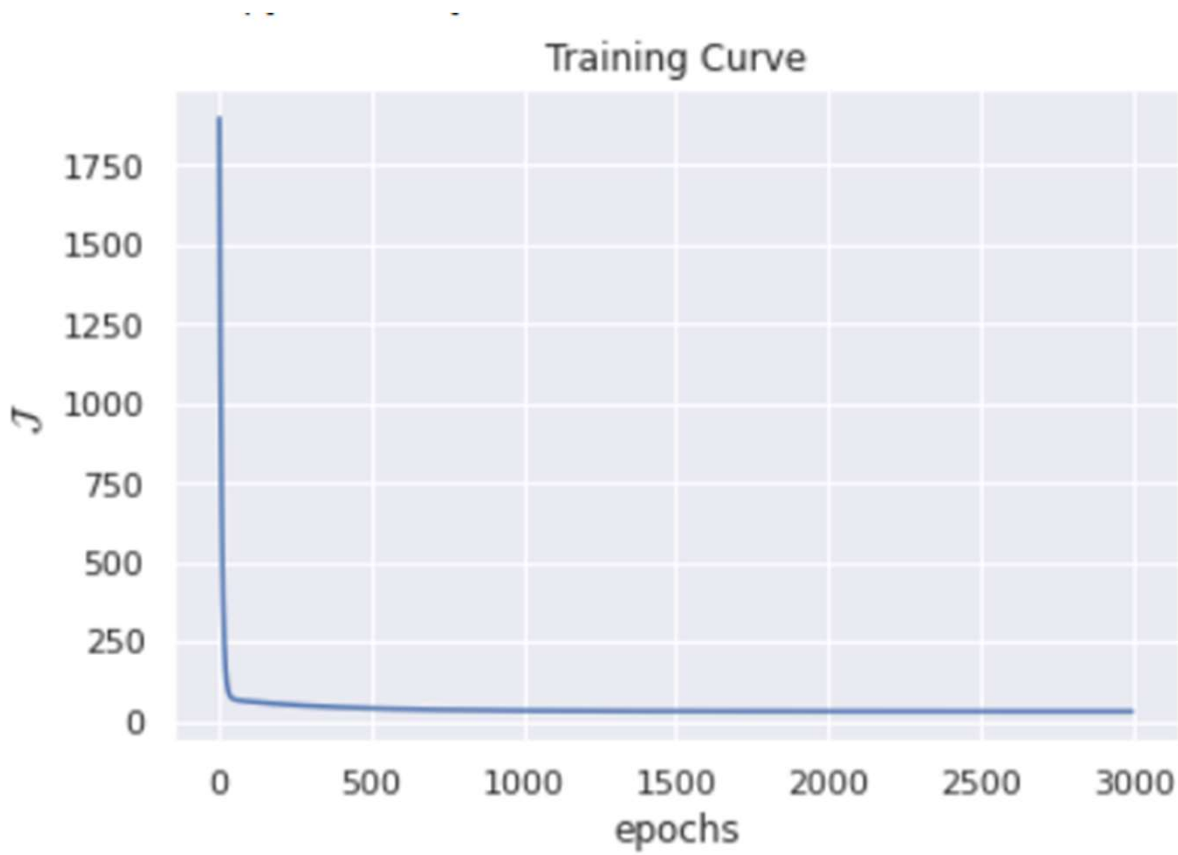
ROOMS	BATHFIXTUR	GARAGE	GARAGECAPA	POOLAREA	Price_SQFT	Category	PRICE_SQFT_CAT	MIN_CAT	DIFF_PRICESQFT_MIN_CAT
4	2.3333333333333335	1	2	0	186.06	7	5	105.0378947368421	81.0221052631579
4	2.3333333333333335	1	2	0	186.75	7	5	105.0378947368421	81.7121052631579
4	2.0	1	3	0	186.83	7	2	105.0378947368421	81.79210526315791
4	2.3333333333333335	1	2	0	187.45	7	5	105.0378947368421	82.41210526315788
4	2.0	1	2	0	192.31	7	5	105.0378947368421	87.2721052631579
4	2.0	1	2	0	193.61	7	6	105.0378947368421	88.57210526315791
4	2.0	1	2	0	193.61	7	4	105.0378947368421	88.57210526315791
4	2.0	1	2	0	193.61	7	10	105.0378947368421	88.57210526315791
4	2.0	1	2	0	193.61	7	5	105.0378947368421	88.57210526315791

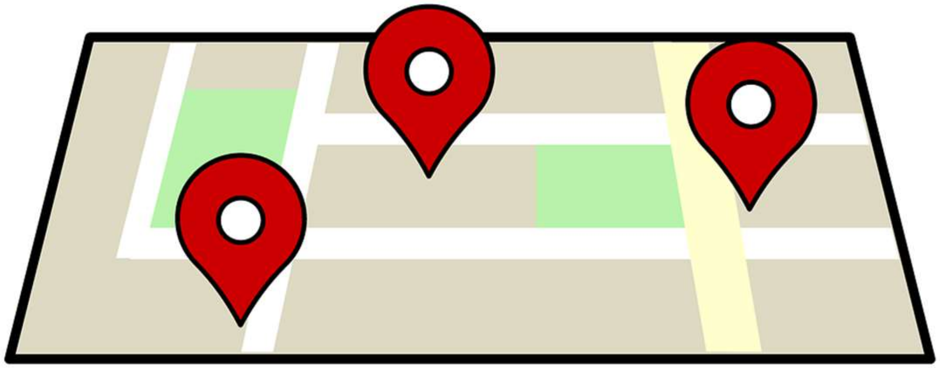
```
XTest = np.array([[4,3,1,8]])
```

```
myReg.predict(XTest)
```

```
array([85.83893189])
```

Gradient Descent





```
1785 The Exchange SE Atlanta GA
120 Madrid Valle Dorado Tlalnepantla
2274 Hidden Glen Drive Marietta GA
(3, 2)
[[ 33.90915665 -84.4791487 ]
 [ 19.5480573  -99.2118424 ]
 [ 33.9201813  -84.4999197 ]]
```

Address to Lat Long

(David)

Conclusions

- In Classification, data cleaning (outlier & missing data) improved the correlation of features (latitude, longitude) towards sold_price and some to sqrt_ft.
- Sequential individual removal of outliers of features and predicted variable does not imply overall outliers removal of features overall.
- Normalization Min-Max of features improved accuracy ~[0.2-0.4] in both Training/Testing.

Q&A

Appendix

Feature Normalization

Min-Max Approach

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Feature Engineering

floor_covering to One-Hot-Encoding

```
Other: concrete tile
Other: Italian tile
Other: gray saltillo
Other: Tile
Other: 20 x 20 on Diagonal
Other: travertine/flagstone
Other: porclain tile
Other: Dyed Concrete
Other: acrylic overlay
Other: Saltillo tile
Other: Travertine tile
Other: Italian Tile
Other: 100% Porcelain Tile
Other: Pergo
Other: Slate
Other: Lime Stone
Laminate
Other: Brick Floor
Other: carpet- guest house
Other: Flagstone
Other: Lux Vinyl
None
Other: TBD
Other: Porcelain-wood
Other: Rojo Concrete Overla
Other: Polish concrete
Mexican Tile
Other: Porcelain Tile
Other: Wood like
Other: Concrete tile
Natural Stone
Vinyl
Other: Egytian sandstone
```

Diff Values: 120

```
flagstone
pergo
marble
20x20_on_diagonal
mesquite_wood_floor
acrylic_overlay
brick
multiple_type
porcelain_wood
porcelain_tile_24x24
new_plank_tile
porcelain_plank_tile
slate_tile
vinyl_plank
canterra_stone
porcelain
granite
talavera_floor
custom_saltillo
travertine_tile
plank_tile
egyptian_sandstone
travertine_marble
carpet
studio_laminate
organic_wool_carpet
luxury_vinyl
brick_floor
polished_brick
polished_concrete
saltillo_tile
indoor_outdoor
wood_laminate
```

Diff Values: 92

flagstone
pergo
egyptian_sandstone
concrete
marble
tile
20x20_on_diagonal
carpet
upg_flooring
throughout_home
travertine
laminate
tbd
saltillo
other
wood
~~none~~
acrylic_overlay
brick
multiple_type
slate
indoor_outdoor
quartzite
natural_stone
canterra_stone
porcelain
bamboo
limestone
granite
talavera_floor
parquet
vinyl
real_polished_aggregated
terrazzo
cork
brazilian_pergo

Diff Values: 36 (incl. 'none')

flagstone	pergo	egyptian_sandstone	concrete	marble	tile	...	wood	...	natural_stone	brazilian_pergo
1	0	0	1	1	1	...	0	...	1	0
0	0	0	0	0	1	...	1	...	1	0
...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	...	0	...	0	0

Diff Values: 35