HOUSES IN USA – FIND GOOD DEALS AND AVOID SCAMS

NICOLAE MOROSAN

- DATA SCIENCE -

DATA

Lot Area

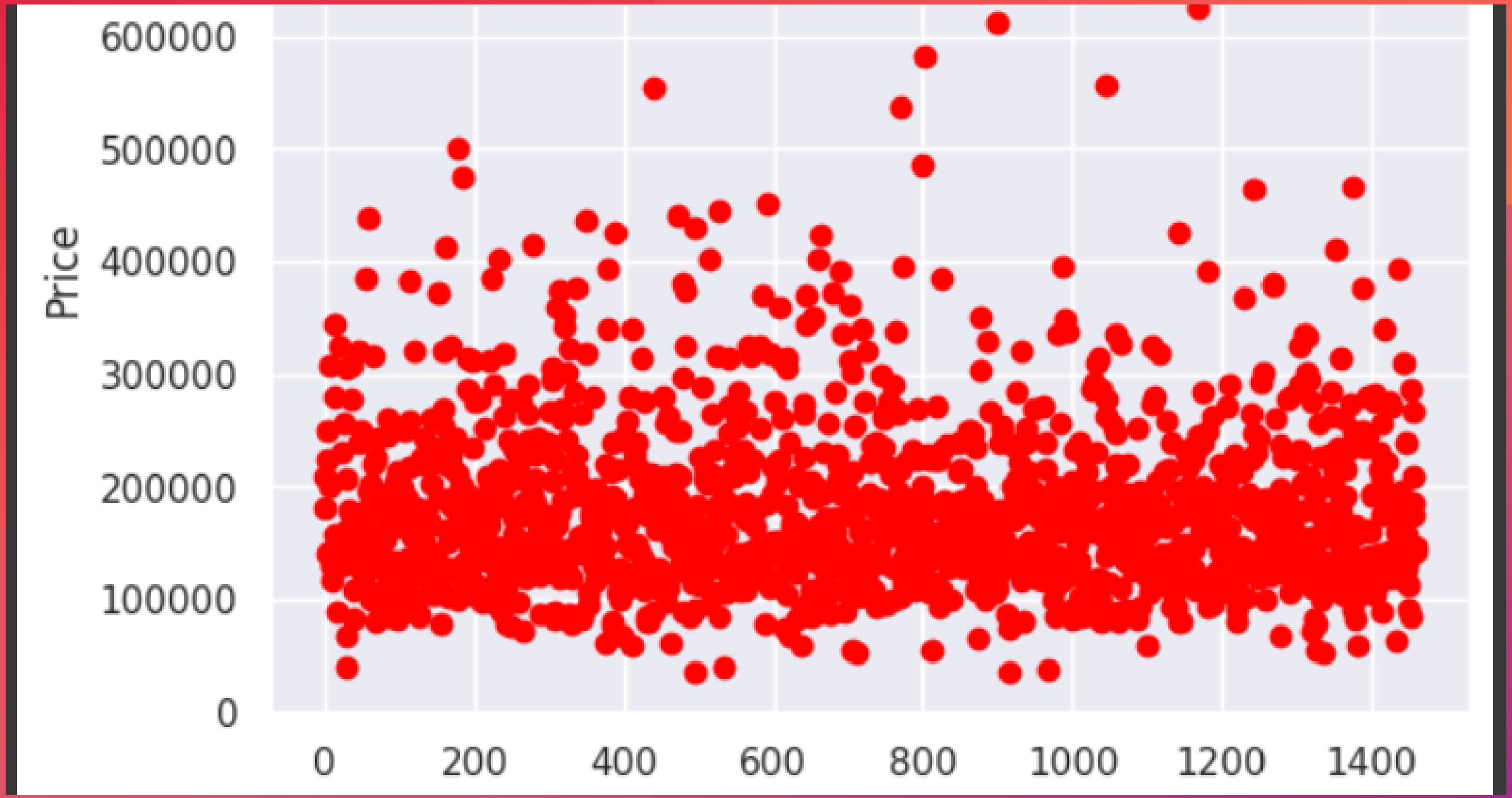
Garage

Year built

Materials

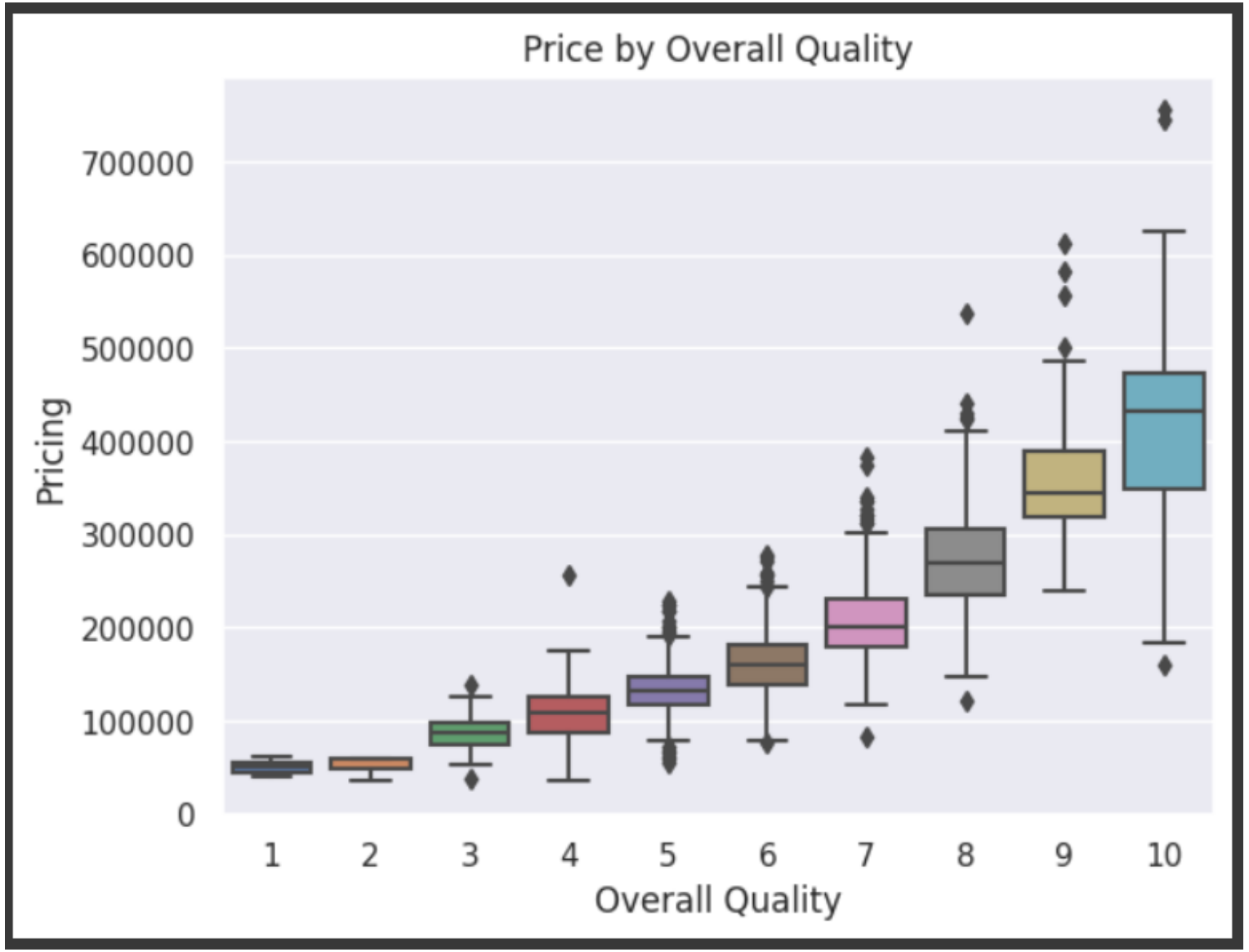
Alley

And so on

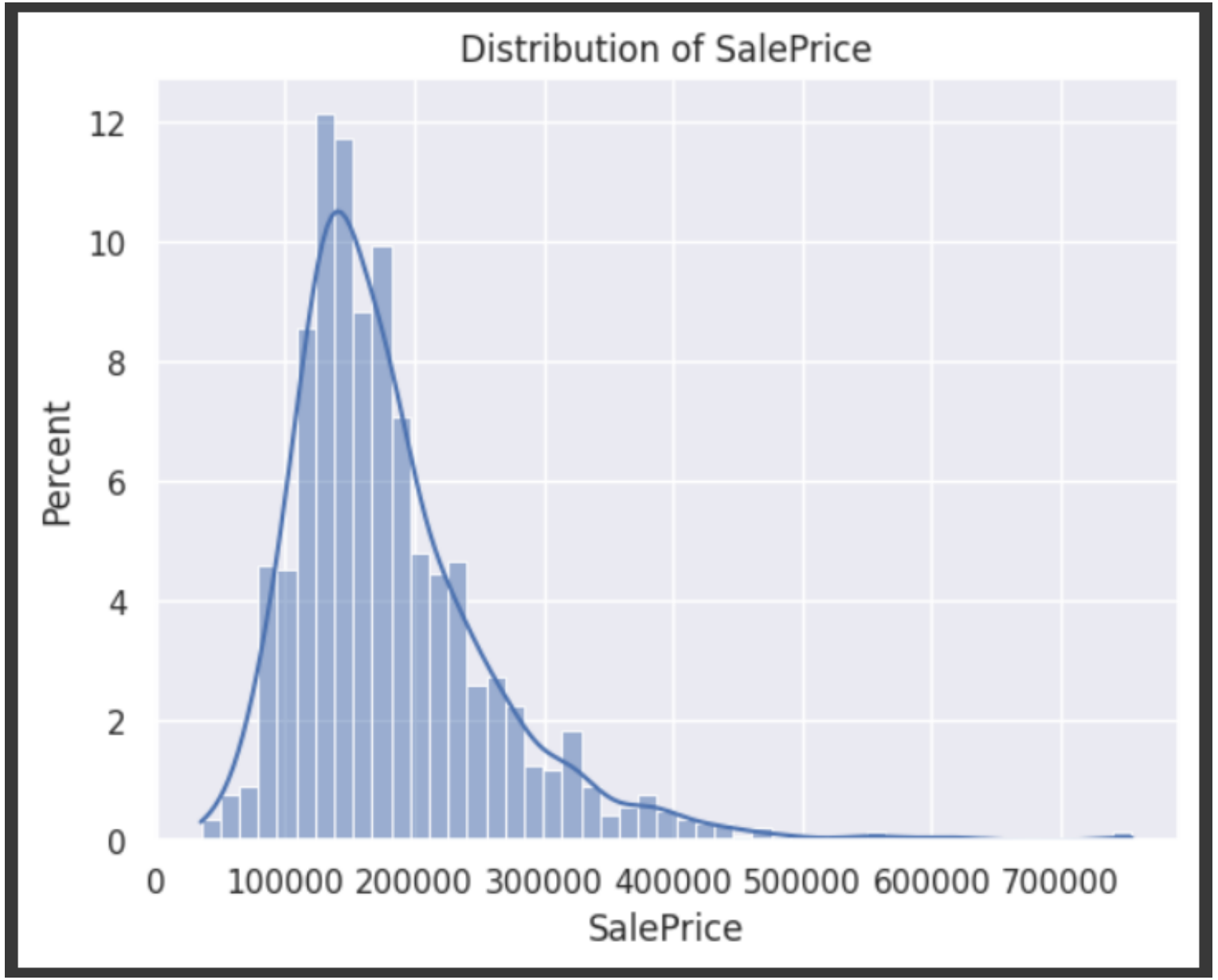

Price
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400

PRICE BY QUALITY

AS EXPECTED



PRICES DISTRIBUTION

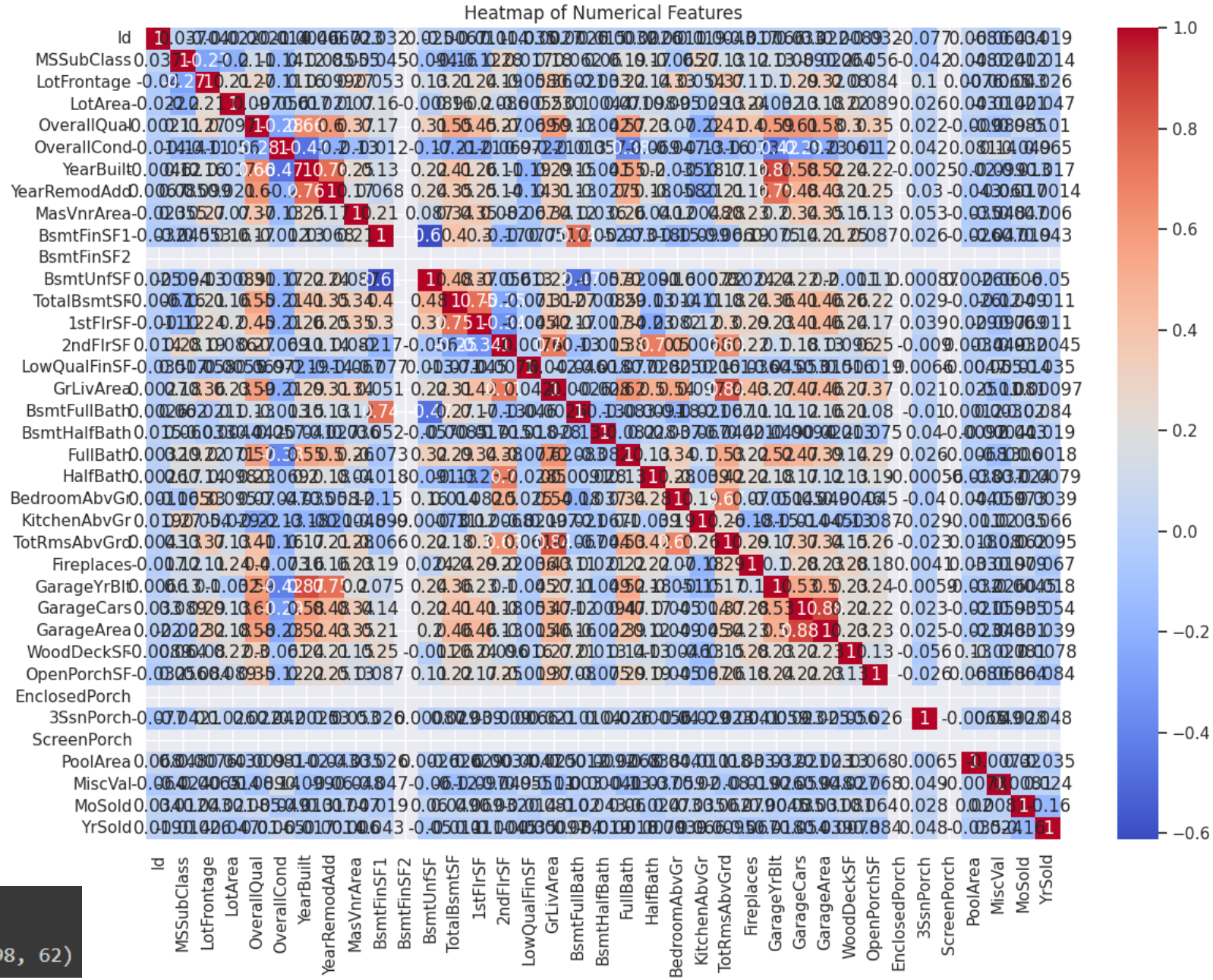


CLEANING

- IQR – top 7 features with most outliers
  i. 798/1460 entries left
- Correlation (62 out of 80 features left)
  i. Higher Bound: +0.6
  ii. Lower Bound: -0.36
- One-hot encoding
  i. Heavy interference with normalization
- Removal of non-numerical features
  i. Heavily decreases performance

CORRELATION

- YEAR BUILT-GARAGE BUILT
- QUALITY-SIZE OF GARAGE
- BASEMENT (FINISHED OR NOT) – BASEMENT QUALITY



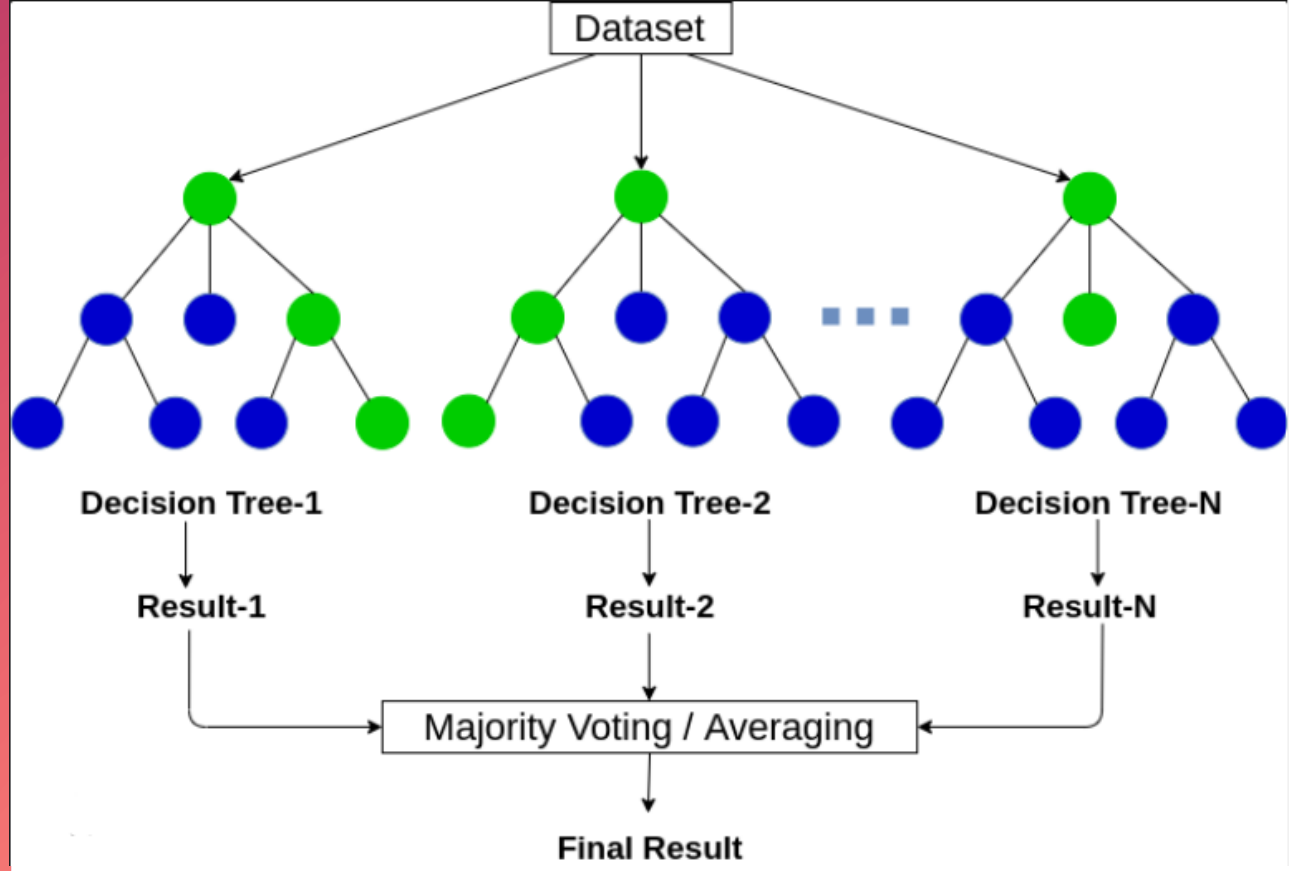

```
Shape of original dataset: (1460, 81)
Shape of x_75 before correlation removal: (798, 80)
Shape of x_75_final (after correlation removal): (798, 62)
```

```
MSE: 2736943208.96
RMSE: 52315.80266955674
Mean: Predicted Sale Price    183393.805
dtype: float64
```

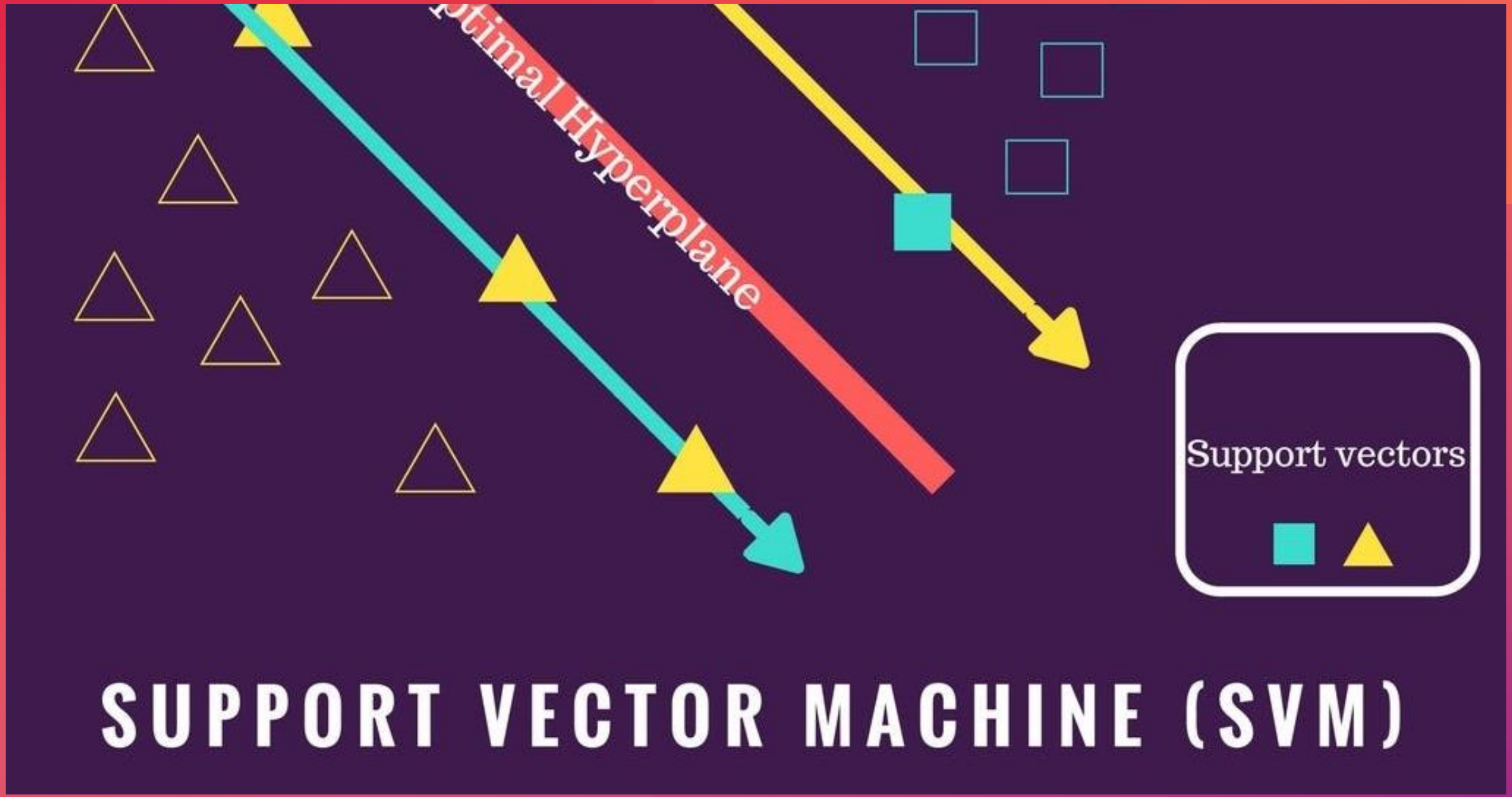
```
Normalized RMSE: 14.552379045773781 %
```

LINEAR REGRESSION

RANDOM FORESTS




```
MSE Random Forests: 2410110614.83
MAE Random Forests: 34009.12
RMSE Random Forests: 49092.87743481736
Mean Random Forests: 182755.265
Normalized RMSE Random Forests: 13.655876894246832 %
```


ptimal Hyperplane
Support vectors
SUPPORT VECTOR MACHINE (SVM)

TESTS & METRICS

	Predicted values (SVM)	Actual/real values
0	189407	260000
1	197265	143000
2	224325	180000
3	154804	255900
4	239544	128000
..	...	...
195	221820	269500
196	138816	176000
197	167817	202500
198	230674	310000
199	160834	194000

```
# Change in C (regularization parameter) - RESULTS
# NOTE: R SQUARED MUST BE POSITIVE, otherwise we have a problem

# EPSILON = 0.1
# 0.1: r squared is -0.0222
# 1 / 2: r squared is -0.0217
# 10: r squared is -0.0195
# 50: r squared is -0.0079
# 100: r squared is 0.008969
# 500: r is 0.1
# 1.000: r is 0.1863
# 5.000: r is 0.33966
# 10.000: r is 0.37119
# 50.000: r is 0.38363
# 100.000: r is 0.3857

# Interpretation: An R^2 of 0.3857 means that about 38.57% of the total
# The rest (very large - 61.43%) cannot be explained using our model
```

TESTS & METRICS

	Gap (Regression)	Gap (Random Forests)	Gap (SVM)
0	-55779	-97678	-70593
1	69664	46626	54265
2	65186	28815	44325
3	-90098	-66820	-101096
4	104378	64990	111544
..	...	...	...
195	-41696	-20296	-47680
196	-39303	-12423	-37184
197	-31486	-10415	-34683
198	-70495	-48505	-79326
199	-35120	-5001	-33166

SO, YAY OR NAY?

Good deals / Steals:

	Gap (Regression)	Gap (Random Forests)	Gap (SVM)
103	83979	120357	87172
98	71888	110836	68684
18	56382	93166	55626
163	61363	83022	56396
87	63330	80764	47335
19	65849	78289	59631
188	27402	75069	47973
142	95203	73093	88745
151	28795	71203	22266
164	75603	66426	89225

Bad deals / Scams:

	Gap (Regression)	Gap (Random Forests)	Gap (SVM)
38	-85067	-99937	-93085
186	-83412	-100009	-90934
180	-108675	-107187	-116784
143	-113067	-113345	-118025
67	-124503	-115925	-152268
105	-130736	-123457	-134925
133	-170143	-128091	-178954
42	-144406	-133586	-152032
32	-124225	-133918	-125245
25	-78026	-161012	-100378
24	-237245	-258982	-241774

- Thank you