

Mark Ehler

King County Housing Data

About this Project

The task:

Given past house sales, predict what properties will be sell for high values when they come on the market.

The data:

21,597 samples of housing sales from May, 2014 until May, 2015 in the King County area.

About the Data

20 individual data points for each sample

6 statistics on square footage

3 statistics on geographical location

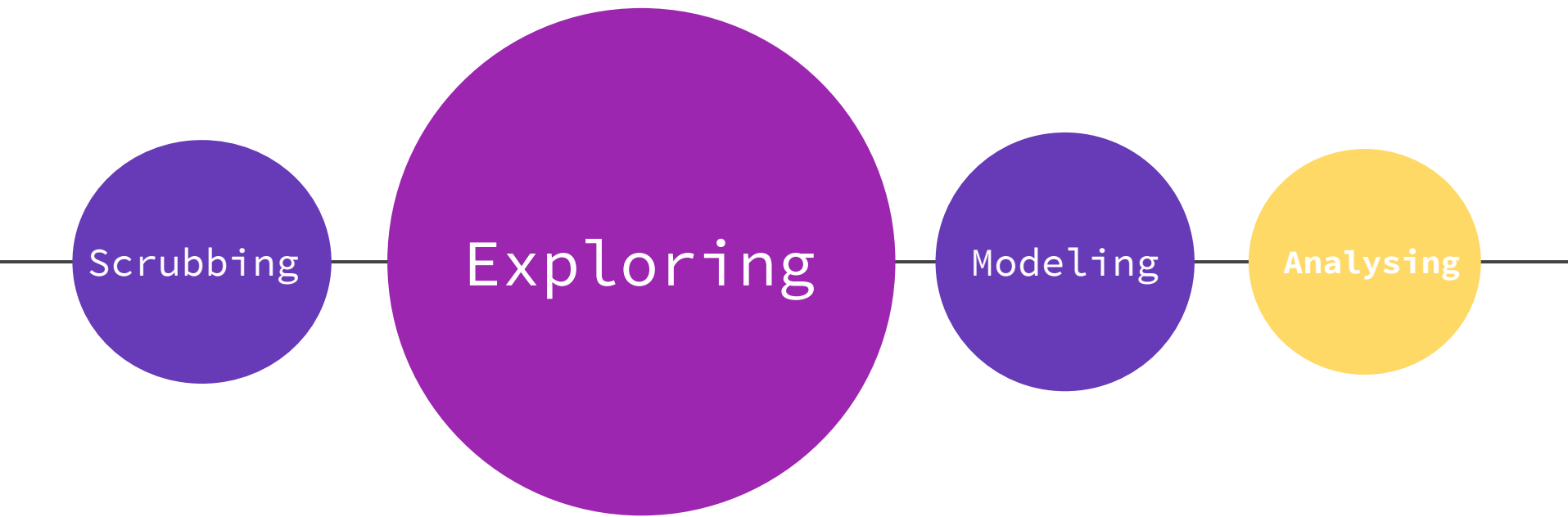
3 statistics dealing with dates

3 strictly quantifiable categories

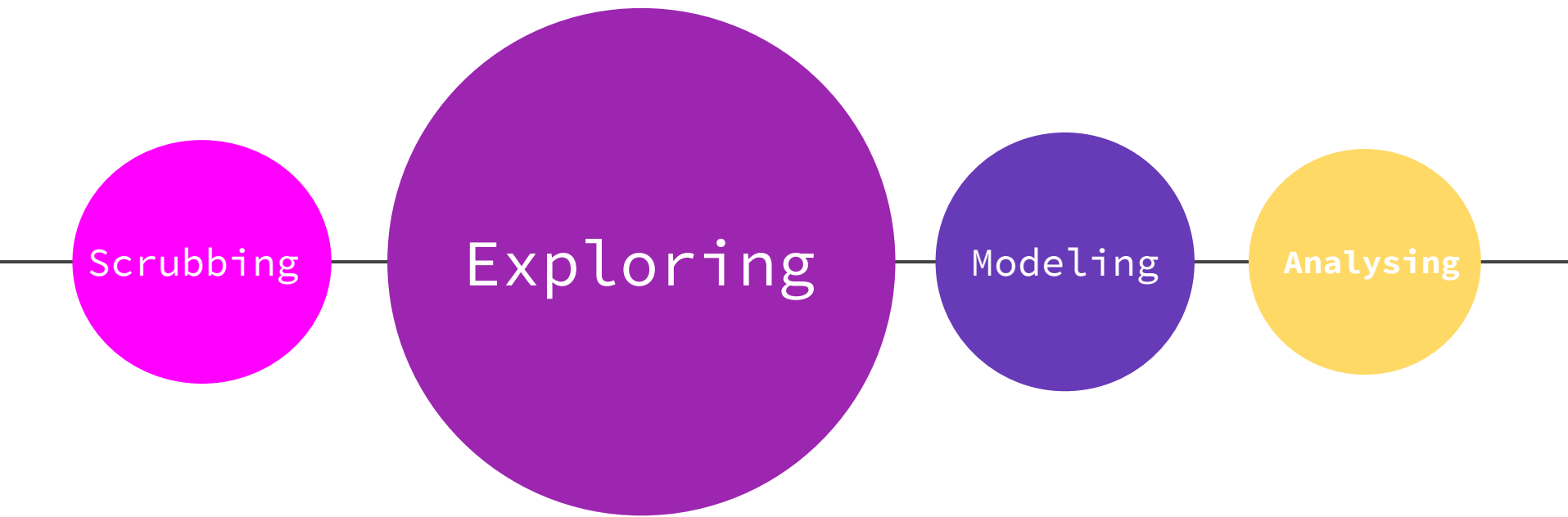
4 subjective values given by the researchers

1 ID tag

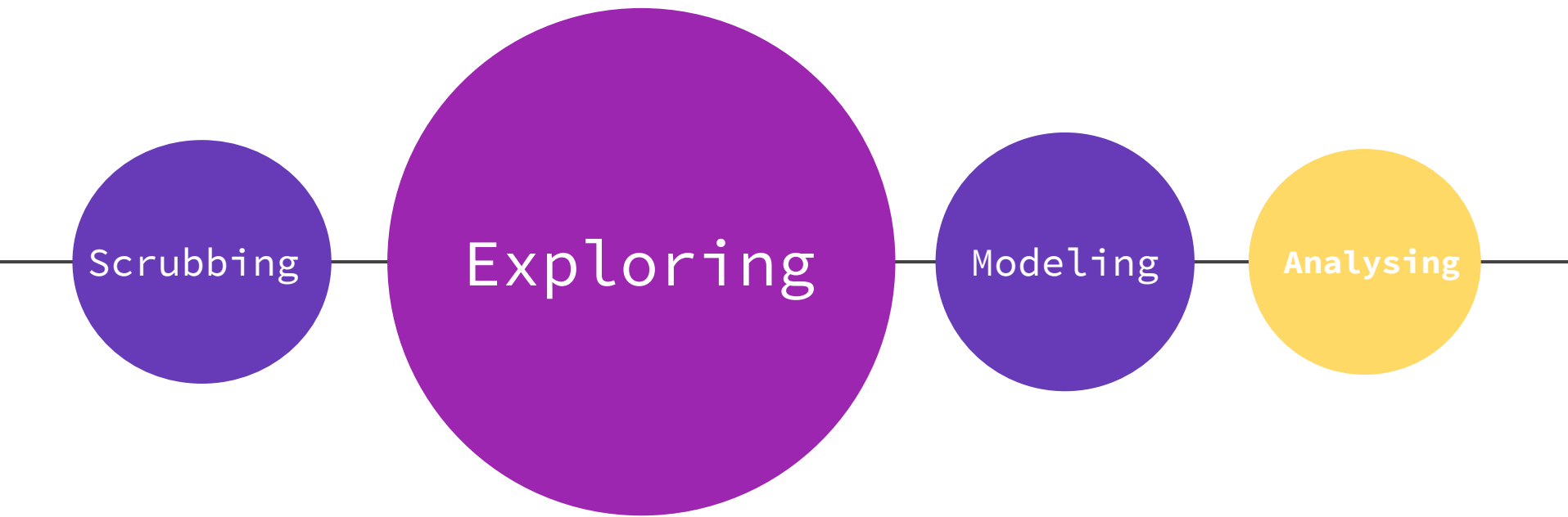
The Process



The Process



The Process

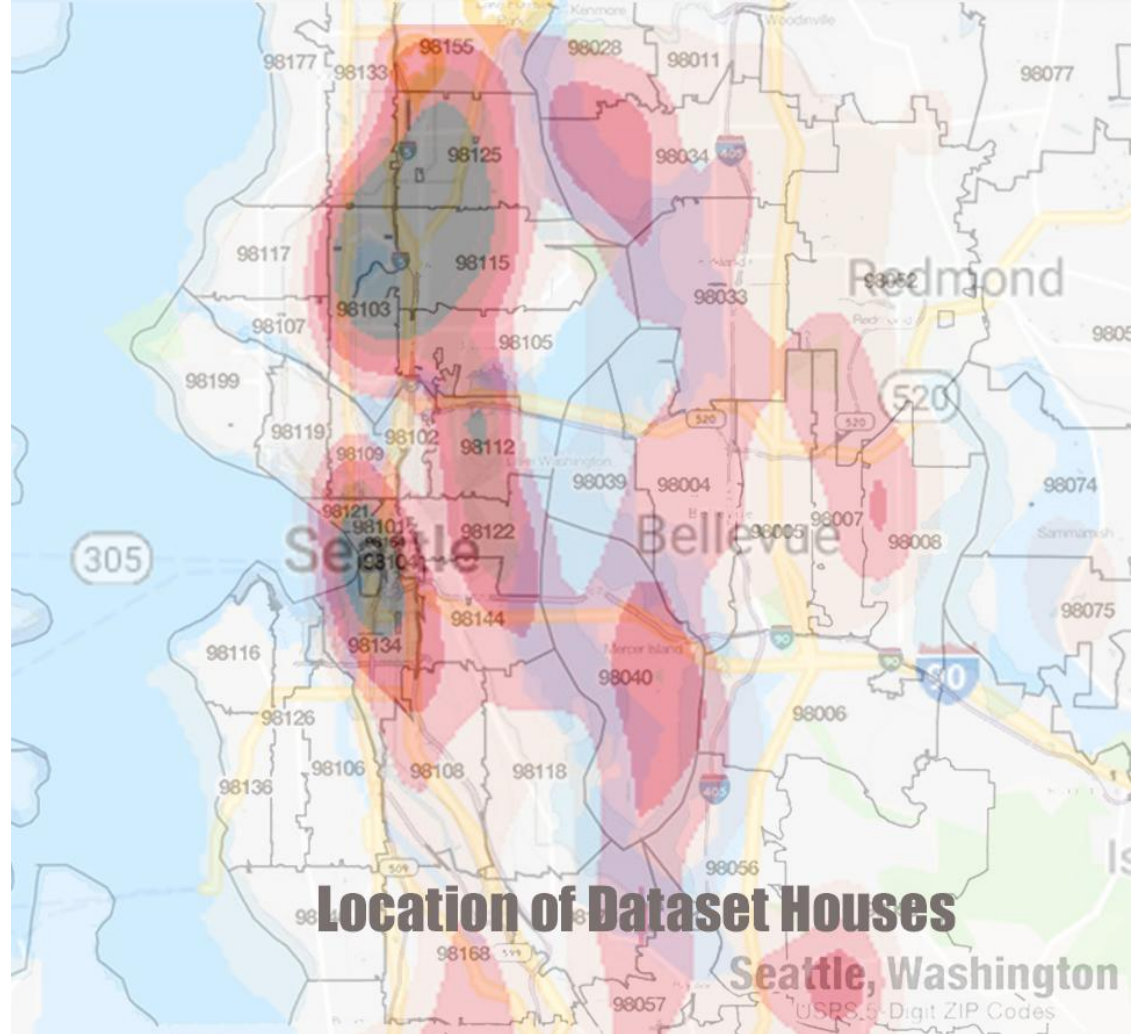


ubbing

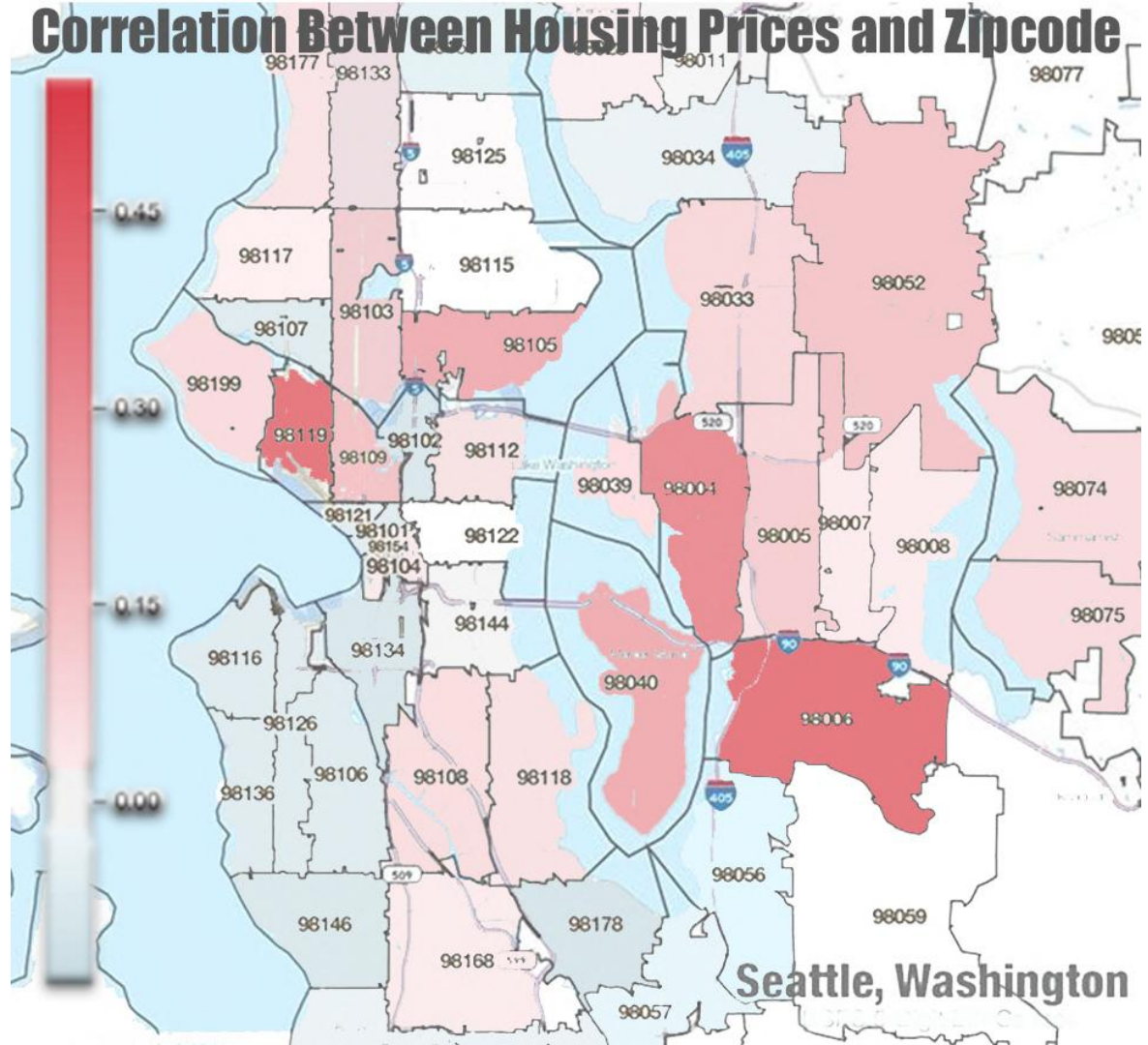
Exploring

Modeling

King County



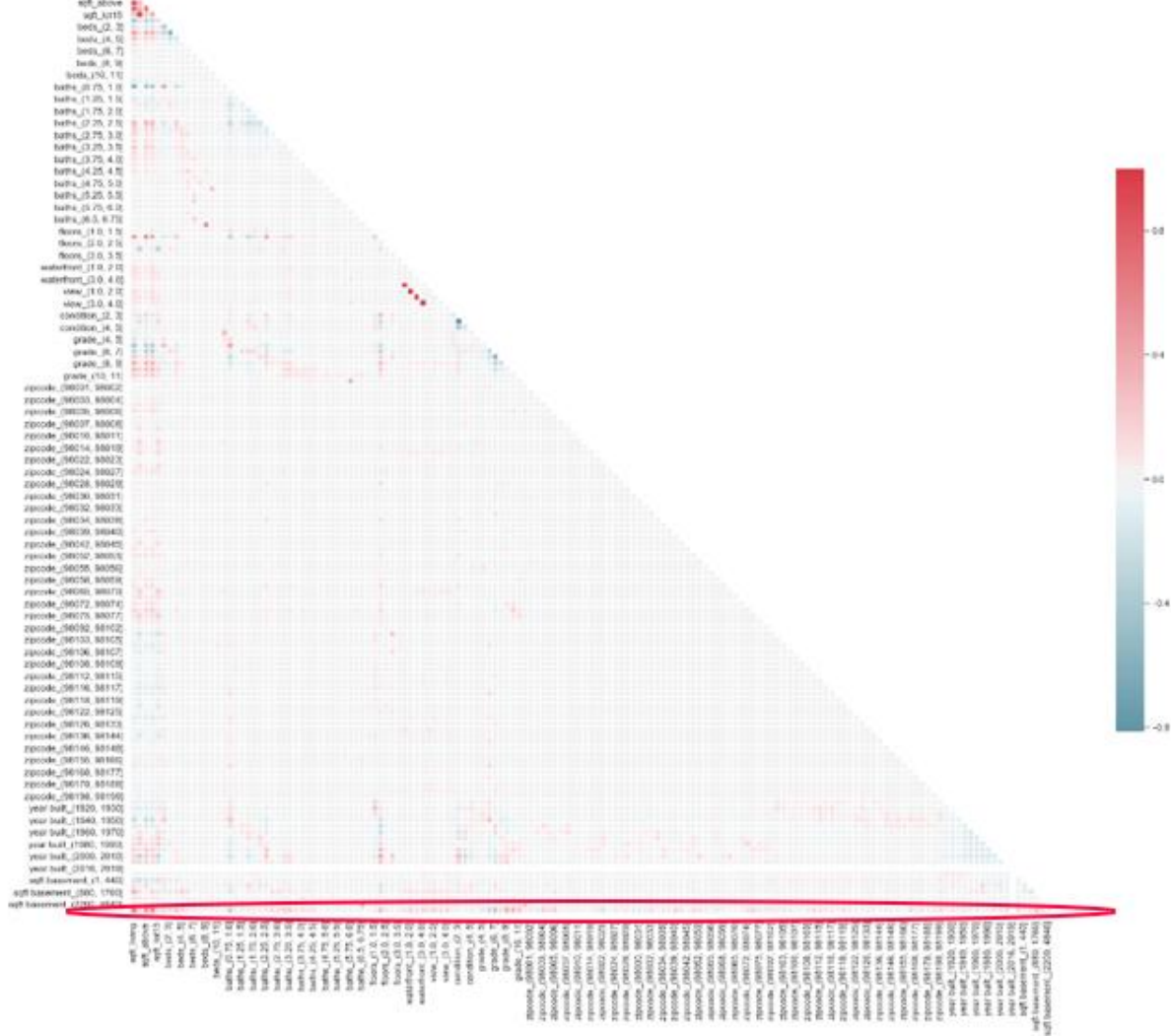
Price related to
Zipcode



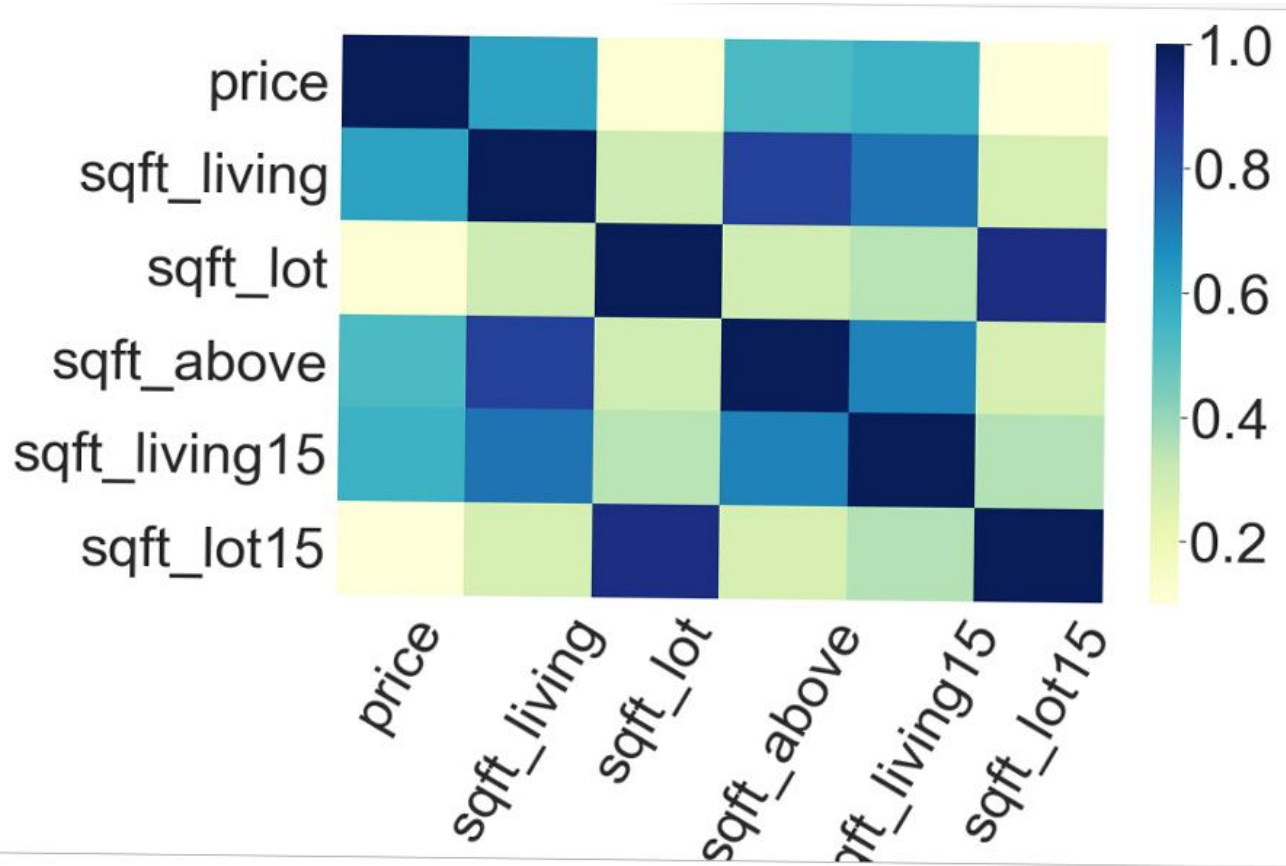
Heatmap showing the relationship between 1000 SNPs and 1000 genes. The color scale ranges from -0.8 (blue) to 0.8 (red). The x-axis lists 1000 SNPs, and the y-axis lists 1000 genes. The heatmap shows a dense pattern of red and blue dots, indicating strong correlations between many SNPs and genes. A color bar on the right indicates the correlation coefficient, ranging from -0.8 to 0.8.

SNP names (X-axis):

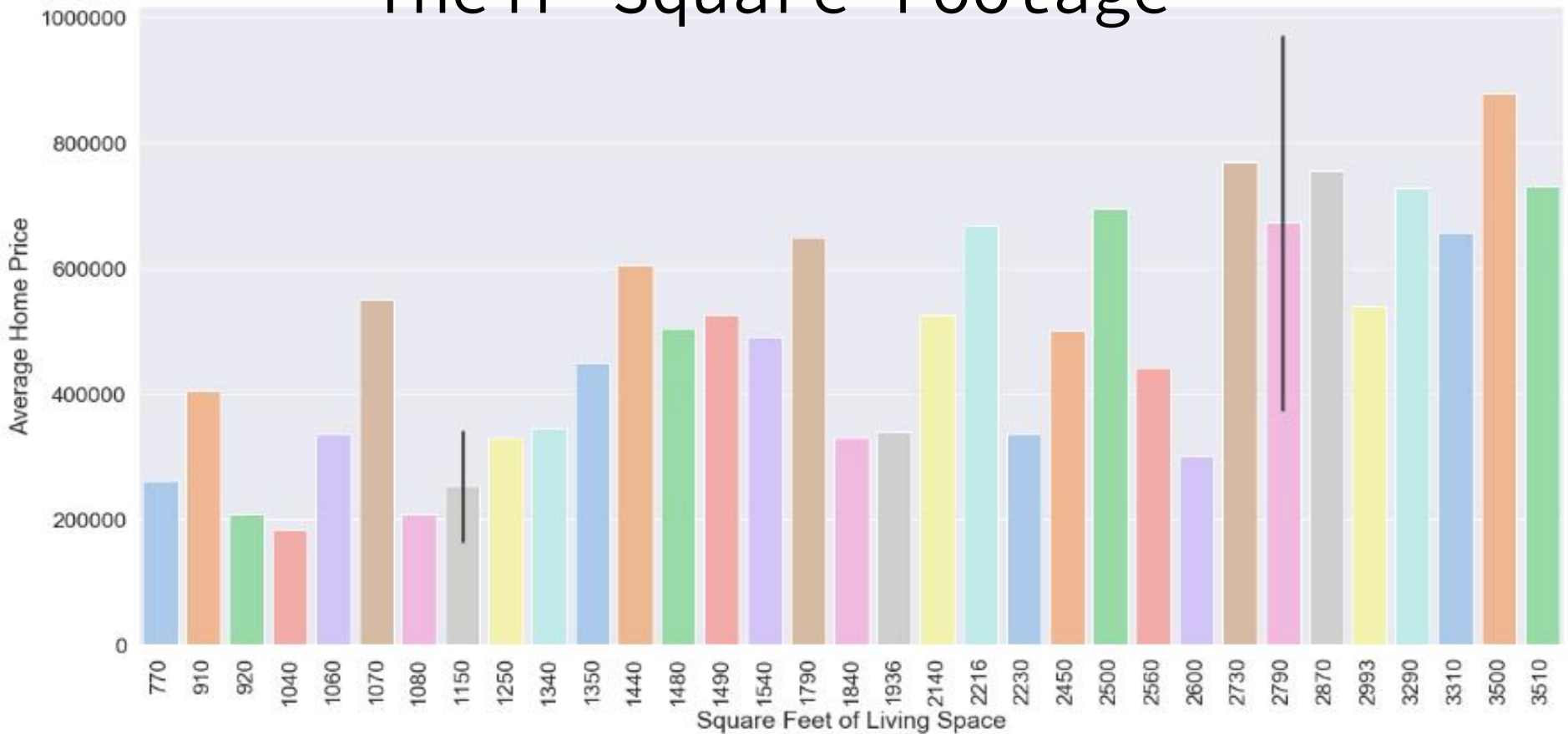
- rs11234567
- rs11234568
- rs11234569
- rs11234570
- rs11234571
- rs11234572
- rs11234573
- rs11234574
- rs11234575
- rs11234576
- rs11234577
- rs11234578
- rs11234579
- rs11234580
- rs11234581
- rs11234582
- rs11234583
- rs11234584
- rs11234585
- rs11234586
- rs11234587
- rs11234588
- rs11234589
- rs11234590
- rs11234591
- rs11234592
- rs11234593
- rs11234594
- rs11234595
- rs11234596
- rs11234597
- rs11234598
- rs11234599
- rs11234600
- rs11234601
- rs11234602
- rs11234603
- rs11234604
- rs11234605
- rs11234606
- rs11234607
- rs11234608
- rs11234609
- rs11234610
- rs11234611
- rs11234612
- rs11234613
- rs11234614
- rs11234615
- rs11234616
- rs11234617
- rs11234618
- rs11234619
- rs11234620
- rs11234621
- rs11234622
- rs11234623
- rs11234624
- rs11234625
- rs11234626
- rs11234627
- rs11234628
- rs11234629
- rs11234630
- rs11234631
- rs11234632
- rs11234633
- rs11234634
- rs11234635
- rs11234636
- rs11234637
- rs11234638
- rs11234639
- rs11234640
- rs11234641
- rs11234642
- rs11234643
- rs11234644
- rs11234645
- rs11234646
- rs11234647
- rs11234648
- rs11234649
- rs11234650
- rs11234651
- rs11234652
- rs11234653
- rs11234654
- rs11234655
- rs11234656
- rs11234657
- rs11234658
- rs11234659
- rs11234660
- rs11234661
- rs11234662
- rs11234663
- rs11234664
- rs11234665
- rs11234666
- rs11234667
- rs11234668
- rs11234669
- rs11234670
- rs11234671
- rs11234672
- rs11234673
- rs11234674
- rs11234675
- rs11234676
- rs11234677
- rs11234678
- rs11234679
- rs11234680
- rs11234681
- rs11234682
- rs11234683
- rs11234684
- rs11234685
- rs11234686
- rs11234687
- rs11234688
- rs11234689
- rs11234690
- rs11234691
- rs11234692
- rs11234693
- rs11234694
- rs11234695
- rs11234696
- rs11234697
- rs11234698
- rs11234699
- rs11234700
- rs11234701
- rs11234702
- rs11234703
- rs11234704
- rs11234705
- rs11234706
- rs11234707
- rs11234708
- rs11234709
- rs11234710
- rs11234711
- rs11234712
- rs11234713
- rs11234714
- rs11234715
- rs11234716
- rs11234717
- rs11234718
- rs11234719
- rs11234720
- rs11234721
- rs11234722
- rs11234723
- rs11234724
- rs11234725
- rs11234726
- rs11234727
- rs11234728
- rs11234729
- rs11234730
- rs11234731
- rs11234732
- rs11234733
- rs11234734
- rs11234735
- rs11234736
- rs11234737
- rs11234738
- rs11234739
- rs11234740
- rs11234741
- rs11234742
- rs11234743
- rs11234744
- rs11234745
- rs11234746
- rs11234747
- rs11234748
- rs11234749
- rs11234750
- rs11234751
- rs11234752
- rs11234753
- rs11234754
- rs11234755
- rs11234756
- rs11234757
- rs11234758
- rs11234759
- rs11234760
- rs11234761
- rs11234762
- rs11234763
- rs11234764
- rs11234765
- rs11234766
- rs11234767
- rs11234768
- rs11234769
- rs11234770
- rs11234771
- rs11234772
- rs11234773
- rs11234774
- rs11234775
- rs11234776
- rs11234777
- rs11234778
- rs11234779
- rs11234780
- rs11234781
- rs11234782
- rs11234783
- rs11234784
- rs11234785
- rs11234786
- rs11234787
- rs11234788
- rs11234789
- rs11234790
- rs11234791
- rs11234792
- rs11234793
- rs11234794
- rs11234795
- rs11234796
- rs11234797
- rs11234798
- rs11234799
- rs11234800
- rs11234801
- rs11234802
- rs11234803
- rs11234804
- rs11234805
- rs11234806
- rs11234807
- rs11234808
- rs11234809
- rs11234810
- rs11234811
- rs11234812
- rs11234813
- rs11234814
- rs11234815
- rs11234816
- rs11234817
- rs11234818
- rs11234819
- rs11234820
- rs11234821
- rs11234822
- rs1123482



Correlation Between Continuous Variables of our Data



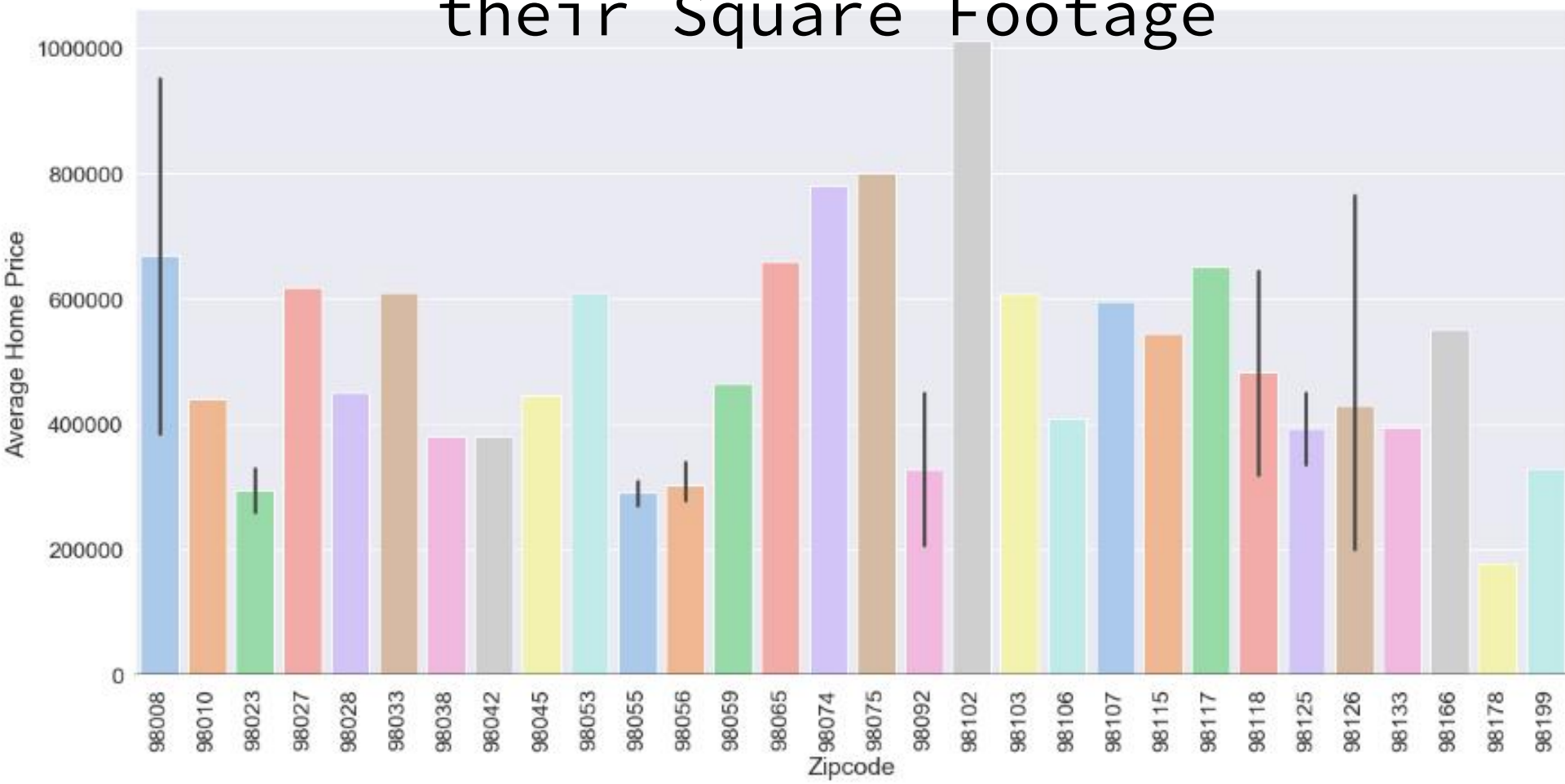
Samples of Home Prices and Their Square Footage



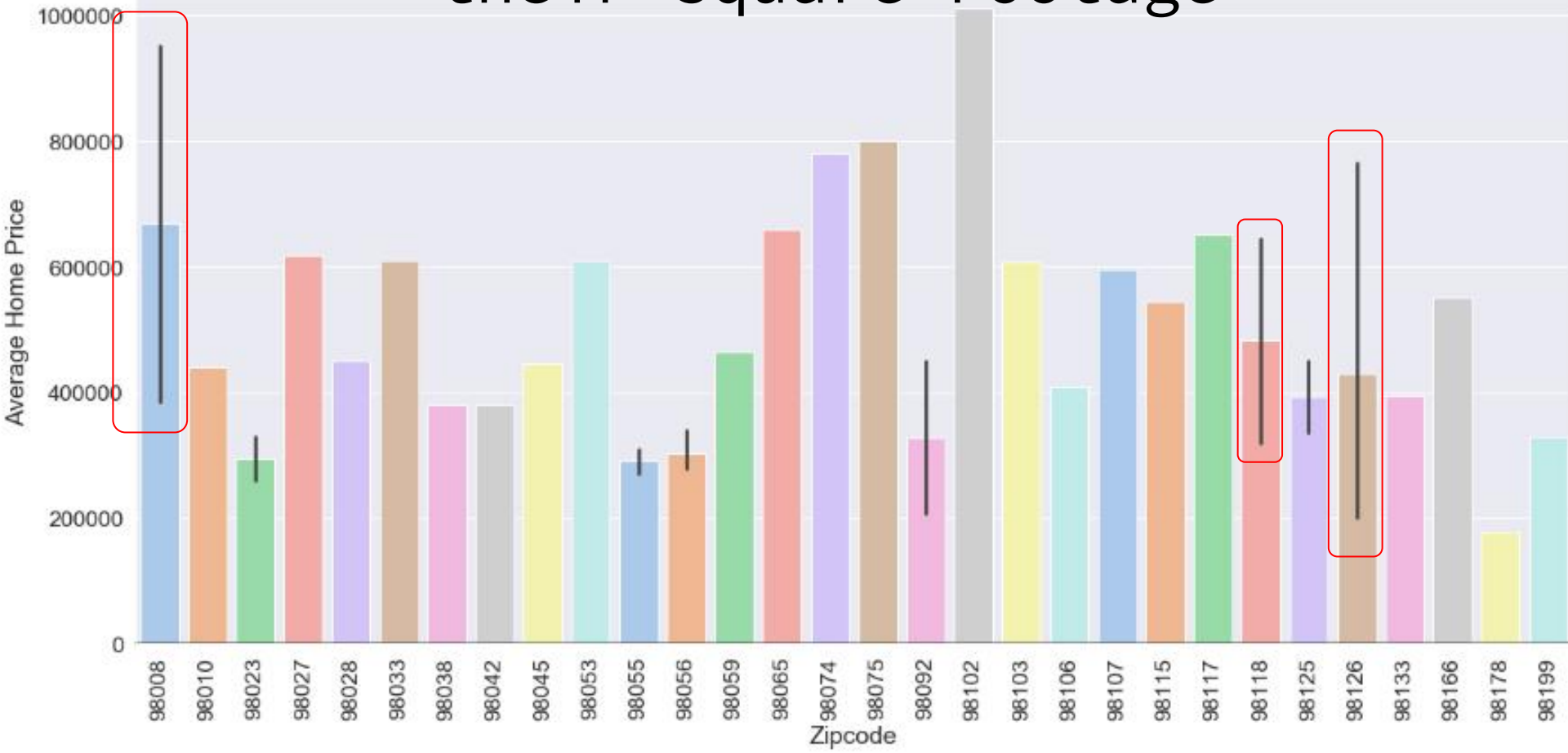
Samples of Home Prices and Their Square Footage



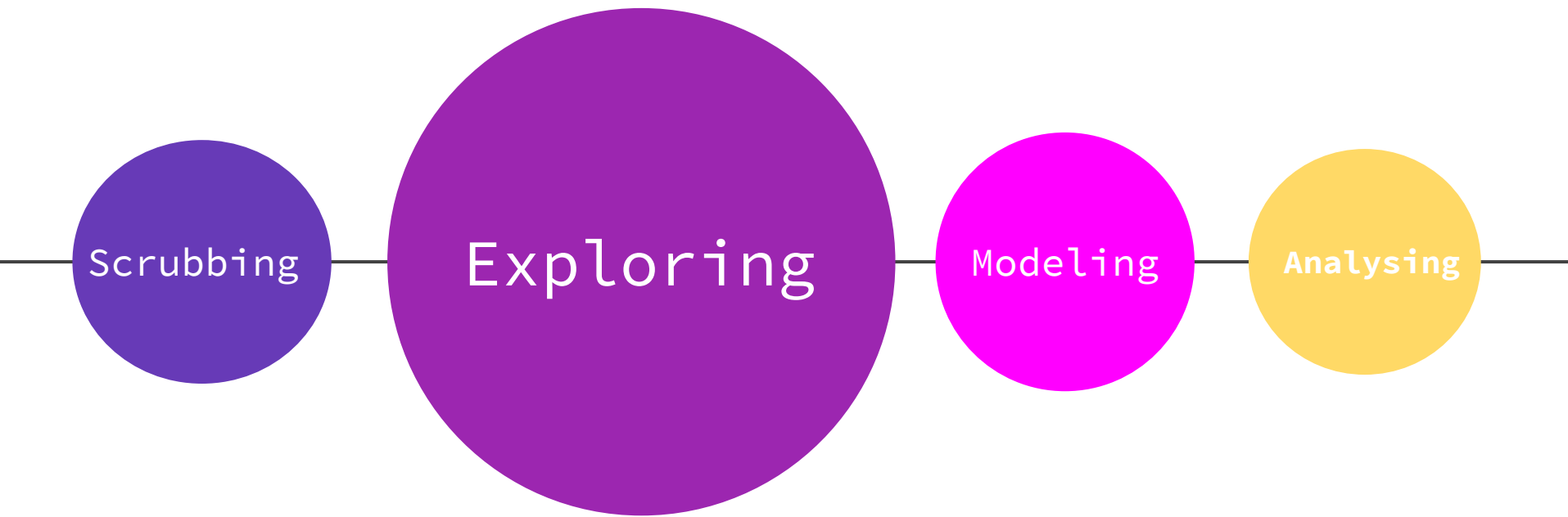
Samples of Zip Codes and their Square Footage



Samples of Zip Codes and their Square Footage



The Process



OLS Regression Results

In [130]: `model.summary()`

Out[130]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.844
Model:	OLS	Adj. R-squared:	0.843
Method:	Least Squares	F-statistic:	754.4
Date:	Fri, 25 Jan 2019	Prob (F-statistic):	0.00
Time:	18:19:05	Log-Likelihood:	-2.6590e+05
No. Observations:	20756	AIC:	5.321e+05
Df Residuals:	20607	BIC:	5.333e+05
Df Model:	148		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-1.439e+05	1.02e+05	-1.411	0.158	-3.44e+05	5.61e+04
sqft_living	1.866e+05	2.73e+04	6.833	0.000	1.33e+05	2.4e+05
sqft_lot	2.767e+05	1.5e+04	18.418	0.000	2.47e+05	3.06e+05
sqft_above	3.797e+05	2.53e+04	15.004	0.000	3.3e+05	4.29e+05
sqft_living15	1.35e+05	9319.504	14.490	0.000	1.17e+05	1.53e+05
sqft_lot15	-5.82e+04	1.47e+04	-3.968	0.000	-8.7e+04	-2.95e+04
beds_(1, 2]	-8281.9848	6901.719	-1.200	0.230	-2.18e+04	5245.931
beds_(2, 3]	-1.088e+04	6984.881	-1.557	0.119	-2.46e+04	2812.882
beds_(3, 4]	-1.119e+04	7186.644	-1.557	0.120	-2.53e+04	2898.758
beds_(4, 5]	-2.06e+04	7612.035	-2.707	0.007	-3.55e+04	-5683.931
beds_(5, 6]	-2.711e+04	9431.195	-2.874	0.004	-4.56e+04	-8619.441
beds_(6, 7]	-6.831e+04	1.22e+04	-5.627	0.000	-1.54e+05	-3.05e+04

OLS Regression Results

In [130]: `model.summary()`

Out[130]:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.844
Model:	OLS	Adj. R-squared:	0.843
Method:	Least Squares	F-statistic:	754.4
Date:	Fri, 25 Jan 2019	Prob (F-statistic):	0.00
Time:	18:19:05	Log-Likelihood:	-2.6590e+05
No. Observations:	20756	AIC:	5.321e+05
Df Residuals:	20607	BIC:	5.333e+05
Df Model:	148		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-1.439e+05	1.02e+05	-1.411	0.158	-3.44e+05	5.61e+04
sqft_living	1.866e+05	2.73e+04	6.833	0.000	1.33e+05	2.4e+05
sqft_lot	2.767e+05	1.5e+04	18.418	0.000	2.47e+05	3.06e+05
sqft_above	3.797e+05	2.53e+04	15.004	0.000	3.3e+05	4.29e+05
sqft_living15	1.35e+05	9319.504	14.490	0.000	1.17e+05	1.53e+05
sqft_lot15	-5.82e+04	1.47e+04	-3.968	0.000	-8.7e+04	-2.95e+04
beds_(1, 2]	-8281.9848	6901.719	-1.200	0.230	-2.18e+04	5245.931
beds_(2, 3]	-1.088e+04	6984.881	-1.557	0.119	-2.46e+04	2812.882
beds_(3, 4]	-1.119e+04	7186.644	-1.557	0.120	-2.53e+04	2898.758
beds_(4, 5]	-2.06e+04	7612.035	-2.707	0.007	-3.55e+04	-5883.931
beds_(5, 6]	-2.711e+04	9431.195	-2.874	0.004	-4.56e+04	-8619.441
beds_(6, 7]	-6.831e+04	1.22e+04	-5.627	0.000	-1.04e+05	-3.76e+04

Smallest Predictive Errors

```
In [194]: top_ten = X.columns[selector.support_ == True]
list(top_ten)
```

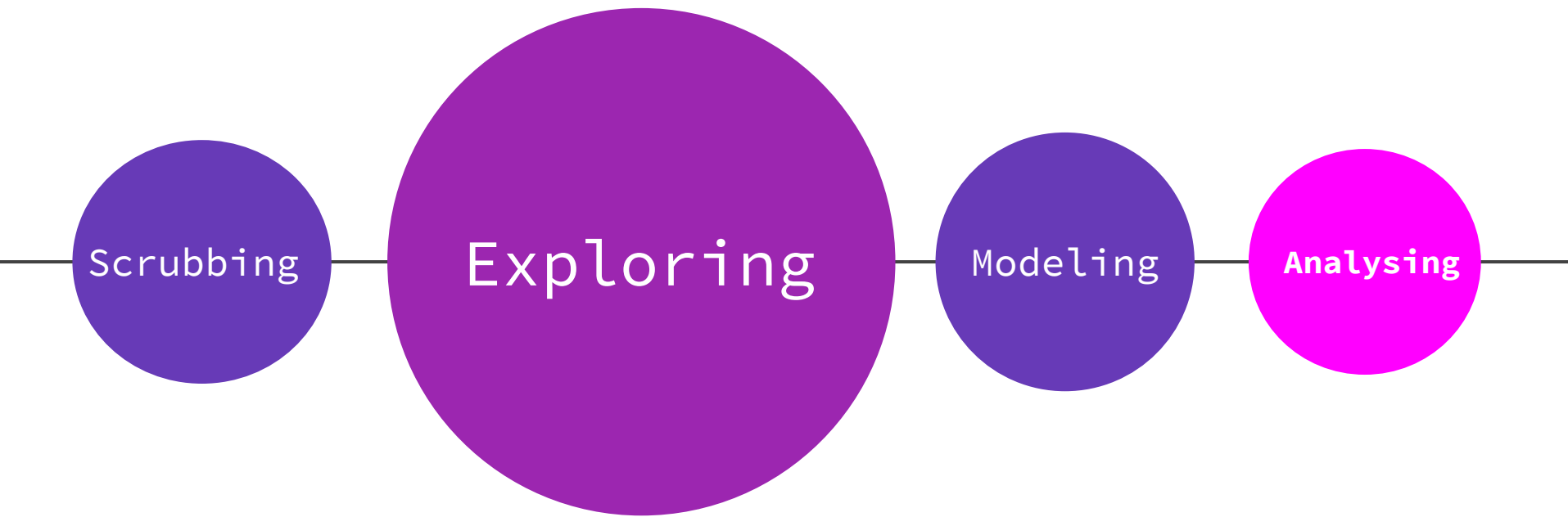
```
Out[194]: ['sqft_living',
'sqft_living15',
'baths_(5.75, 6.0]',
'grade_(11, 12]',
'zipcode_          98004]',
'zipcode_          98039]',
'zipcode_          98040]',
'zipcode_          98109]',
'zipcode_          98112]',
'zipcode_          98119]']
```

Least Accurate Variables

```
In [266]: dropped = X.drop(returns, axis=1)
          dropped.columns
```

```
Out[266]: Index(['beds_(1, 2]', 'beds_(2, 3]', 'beds_(3, 4]', 'beds_(5, 6]',
                'beds_(7, 8]', 'beds_(8, 9]', 'beds_(9, 10]', 'beds_(10, 11]',
                'baths_(0.5, 0.75]', 'baths_(0.75, 1.0]', 'baths_(1.0, 1.25]',
                'baths_(2.0, 2.25]', 'baths_(2.25, 2.5]', 'baths_(4.25, 4.5]',
                'baths_(4.5, 4.75]', 'baths_(4.75, 5.0]', 'baths_(5.0, 5.25]',
                'baths_(5.25, 5.5]', 'baths_(5.5, 5.75]', 'baths_(6.0, 6.5]',
                'baths_(6.5, 6.75]', 'baths_(6.75, 7.5]', 'floors_(1.0, 1.5]',
                'floors_(2.0, 2.5]', 'floors_(3.0, 3.5]', 'grade_(3, 4]',
                'grade_(4, 5]', 'grade_(6, 7]', 'zipcode_(98034, 98038]',
                'zipcode_(98053, 98055]', 'zipcode_(98056, 98058]',
                'zipcode_(98166, 98168]', 'zipcode_(98178, 98188]',
                'zipcode_(98188, 98198]', 'year_built_(1950, 1960]',
                'year_built_(1960, 1970]', 'year_built_(1980, 1990]',
                'year_built_(1990, 2000]', 'year_built_(2016, 2019]',
                'sqft_basement_(1, 440]', 'sqft_basement_(440, 880]',
                'sqft_basement_(880, 1760]', 'sqft_basement_(1760, 2200]',
                'sqft_basement_(2200, 4840]'],
              dtype='object')
```

The Process



R Squared After Dropping Variables

```
In [173]: y = y  
X = stepwise_X  
  
preds_int = sm.add_constant(X)  
model = sm.OLS(y, preds_int).fit()  
model.summary()
```

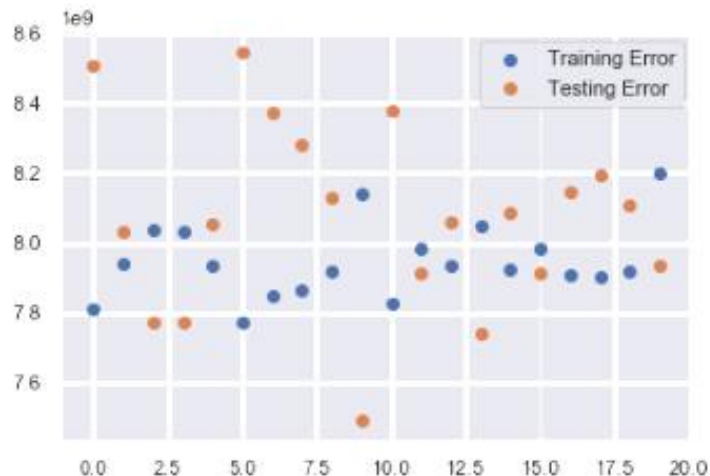
Out[173]: OLS Regression Results

Dep. Variable:	price	R-squared:	0.842
Model:	OLS	Adj. R-squared:	0.841
Method:	Least Squares	F-statistic:	1050.
Date:	Sun, 27 Jan 2019	Prob (F-statistic):	0.00
Time:	09:41:27	Log-Likelihood:	-2.6603e+05
No. Observations:	20756	AIC:	5.323e+05
Df Residuals:	20650	BIC:	5.331e+05
Df Model:	105		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-2.782e+05	1.79e+04	-15.500	0.000	-3.13e+05	-2.43e+05
grade_[8, 9]	1.2e+05	2744.073	43.747	0.000	1.15e+05	1.25e+05
sqft_living	3.84e+05	1.09e+04	35.093	0.000	3.63e+05	4.05e+05
grade_[9, 10]	1.946e+05	4037.820	48.187	0.000	1.87e+05	2.02e+05

MSE After Dropping Ineffective Variables

```
train_test_split(stepwise_X, y, test_size = 0.25)
```



Price Predictor Model v1.0

```
In [122]: ###MSE is the squared price error  
from math import sqrt  
pos_MSE = -1*cv_20_results  
Median_Error = sqrt(pos_MSE)  
print(f'Average error of prediction model: ${round(Median_Error, 2)}')
```

Average error of prediction model: \$89977.49

Highlights

Get the Square Footage

It is highly correlated with price

Zip Code Hotspots

Consider properties in the top 5 correlated zip codes to be more accurate predictions

Accuracy of Model

Plus or minus \$90,000

— — —

Further Findings

Geo Map

Better visualization and more accurate geographical correlations

Adjusted for inflation

A huge variable not addressed by the current model

More Widespread

More samples from under represented neighborhoods

— — —

Thanks for your time

