Real Estate Market Analysis

A Data-Driven Approach

Overview

• Overarching Goal: develop a data-centered application to automate specific components of housing market analysis

• Secondary Goal 1: illustrate process of developing novel and valid metrics (analysis of growth used for illustration - §1)

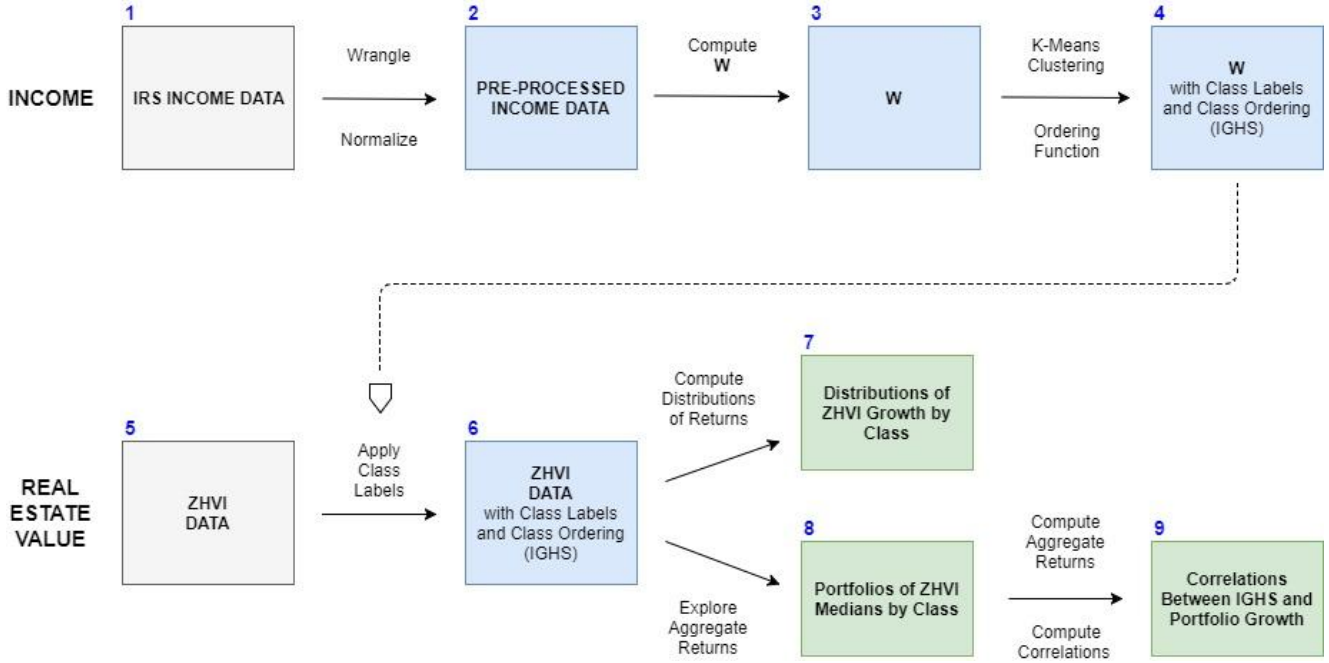
• Secondary Goal 2: illustrate approaches to building models upon established metrics (analysis of valuation used for illustration - §2)

• The systems developed are introductory to more robust metrics and models

§1: Growth

Exploratory cross-study between income levels and real estate value to develop a novel metric for growth

Overview: Growth



- Top half: development of novel metric for income growth
- Bottom half: testing income growth metric's relationship to real estate growth

Data

	A	B	C	D	E
1	CALIFORNIA				
2	Individual Income Tax Returns: Selected Income and Tax Items by State, ZIP Code, and Size of Adjusted Gross Income, Tax Year 2014				
3	[Money amounts are in thousands of dollars]				
4	ZIP code [1]	Size of adjusted gross income	Number of returns	Number of single returns	Number of joint returns
5					
6			(1)	(2)	(3)
7	**00000**	**Total**	**17,113,670**	**8,156,450**	**6,155,880**
8	00000	$1 under $25,000	6,336,730	4,292,060	868,910
9	00000	$25,000 under $50,000	3,965,570	1,890,290	1,142,610
10	00000	$50,000 under $75,000	2,207,030	919,390	930,030
11	00000	$75,000 under $100,000	1,411,040	458,010	796,580
12	00000	$100,000 under $200,000	2,214,810	462,630	1,611,640
13	00000	$200,000 or more	978,490	134,070	806,110
14					
15	90001		21,370	9,410	5,080
16	90001	$1 under $25,000	13,020	7,130	1,880
17	90001	$25,000 under $50,000	6,110	1,810	2,010
18	90001	$50,000 under $75,000	1,630	370	810
19	90001	$75,000 under $100,000	420	70	260
20	90001	$100,000 under $200,000	170	30	120
21	90001	$200,000 or more	20	**	**

Analysis of Income:
U.S. Internal Revenue Service - CA Income Taxes

	A	B	C	D	E	F	G	H
1	RegionID	RegionNa	State	Metro	1996-04	1996-05	1996-06	1996-07
2	61639	10025	NY	New York				
3	84654	60657	IL	Chicago	146700	146500	146300	146300
4	84616	60614	IL	Chicago	198000	195500	194200	193800
5	93144	79936	TX	El Paso	70800	71000	71000	71400
6	61616	10002	NY	New York				
7	84640	60640	IL	Chicago	102300	101300	100700	100600
8	91733	77084	TX	Houston	75600	75400	75100	75100
9	97564	94109	CA	San Franci	298200	295700	296400	298500
10	90668	75070	TX	Dallas-Fort Worth				
11	62037	11226	NY	New York				
12	91940	77449	TX	Houston	72100	72300	72100	71500
13	61630	10016	NY	New York				
14	71831	32162	FL	The Villag	91800	92100	91800	90800
15	84646	60647	IL	Chicago	124500	124000	123600	124100
16	74242	37211	TN	Nashville	80900	81600	82500	83600
17	96107	90250	CA	Los Angeles-Long Beach-Anaheim				
18	74101	37013	TN	Nashville	89300	90200	91100	92000
19	84620	60618	IL	Chicago	140900	140300	139500	139000
20	61703	10128	NY	New York				
21	61625	10011	NY	New York				
22	69816	28269	NC	Charlotte	103300	104000	104700	105700

Analysis of Real Estate Value:
Zillow Home Value Index - All Medians

Analysis of Income: Data Pre-Processing

- Study gross incomes above $25,000 annually as representation of participants in housing market

- Normalize income brackets to study percentage distributions rather than counts

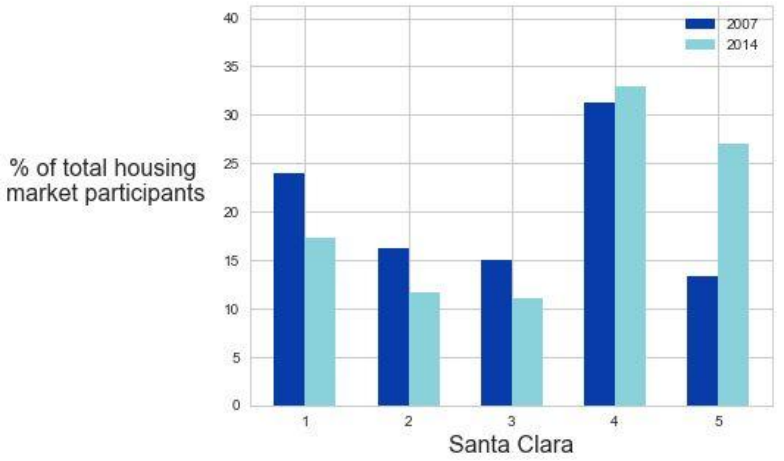
- Compute **W,** table of net changes in percentages from 2007 to 2014 of each income bracket

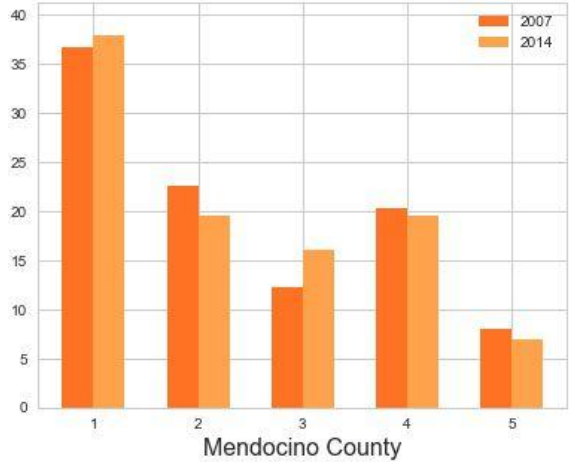
		%distribution2007	%distribution2008	%distribution2009
zipcode	gross_income_bracket			
96161	1: 25,000-50,000	28.821266	30.669399	32.037534
	2: 50,000-75,000	20.286814	19.808743	19.638070
	3: 75,000-100,000	16.124519	15.437158	15.080429
	4: 100,000-200,000	24.309199	24.863388	25.268097
	5: > 200,000	10.458202	9.221311	7.975871

	1: 25k-50k	2: 50k-75k	3: 75k-100k	4: 100k-200k	5: >200k
zipcode					
90001	-4.674271	2.391328	1.298178	0.983740	0.001025
90002	-3.706238	1.639324	1.139695	0.927219	0.000000
90003	-3.370728	2.144693	0.718696	0.667275	-0.159936
90004	-3.772951	1.277511	0.392643	1.393961	0.708835
90005	-5.056257	1.734964	1.056504	1.597609	0.667180

Top: e.g. subset of income distributions; Bottom: e.g. subset of **W**

Analysis of Income:
Changes in Income Brackets





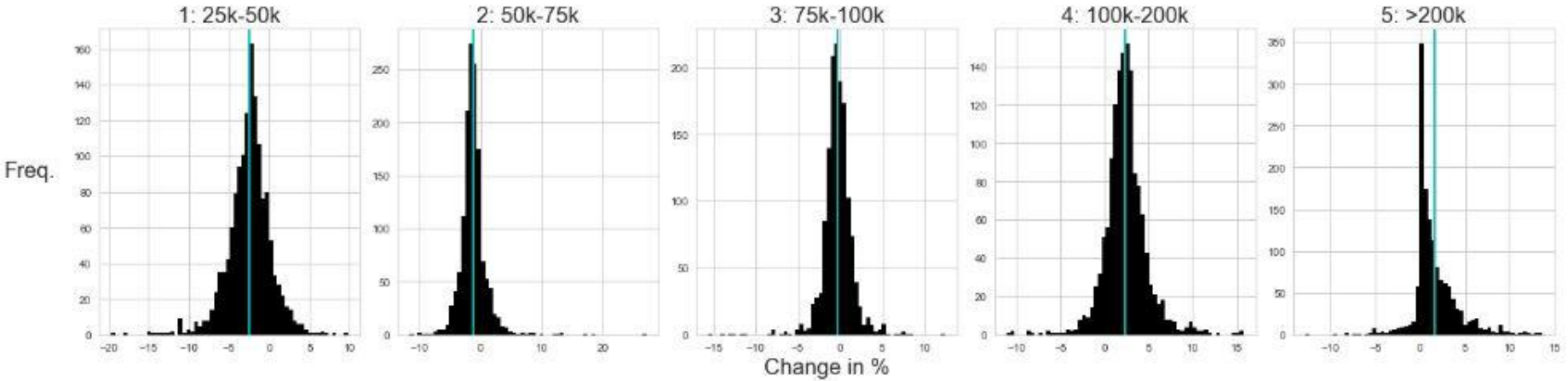
- 1: $25k - $50k
- 2: $50k - $75k
- 3: $75k - $100k
- 4: $100k - $200k
- 5: $200k +

95054	-6.732233	-4.556195	-3.918435	1.568799	13.638064

95460	1.200911	-3.080151	3.773925	-0.840509	-1.054176

- Example plots above represent the income distributions in 2007 and 2014
- Respective w vectors in **W** represent the net changes in income distributions

Analysis of Income:
Changes in Income Brackets



- Statistical summary of income brackets' % changes for zip codes in **W** between 2007 and 2014
- Respective means of each income bracket marked in cyan

Analysis of Income:
Clustering Changes in Income Brackets

- Clustering regions allows us to later measure changes in real estate value over samples rather than independent regions

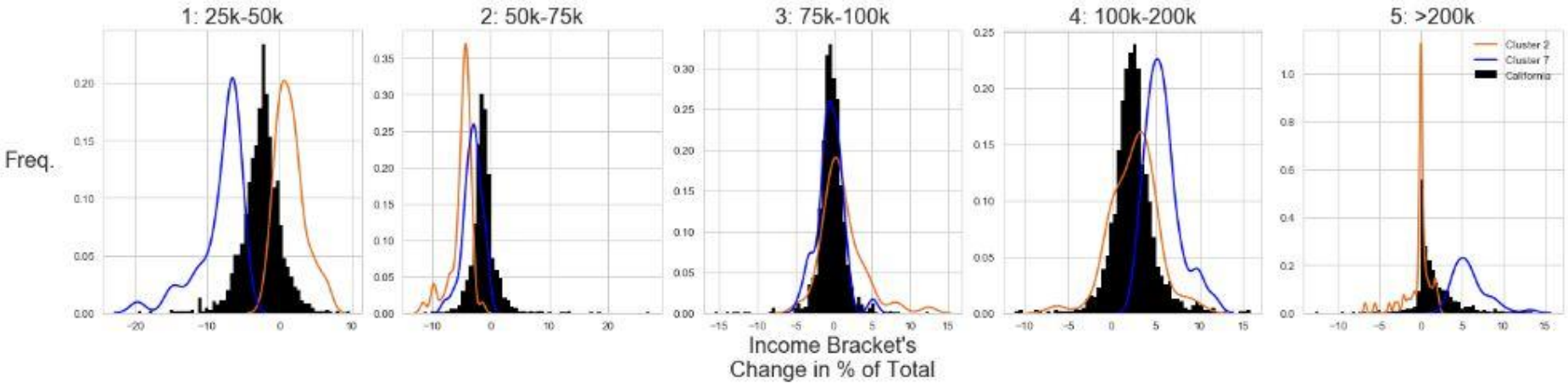
- Apply K-Means to **W** to group together regions which exhibit similar changes in their income distributions

- Different characteristics of values in **W** imply different underlying strengths of income growth

Analysis of Income:
Clustering Changes in Income Brackets

- Use diminished reduction in total sum of squares to centroids to determine cluster count
- Visual analysis suggests 9-15 clusters - use 12
- Use more robust criteria in practice, according to context

Analysis of Income:
Clustering Changes in Income Brackets



• Example distribution plots for two different clusters (kernel density estimations, in color) against distributions plots of the whole data set (normalized histograms, in black)

• Relative positioning of KDE plots to histograms suggest differences in strength of income growth over 2007-2014 period of respective clusters

Analysis of Income:
Income Growth Health Score - IGHS

• Want a more direct approach to comparing clusters, based on their implied strengths of income growth

• Use a real valued function taking summary statistics of each column of **W** as inputs, such that the function increases for sets of values in **W** implying stronger growth in income**,** *i.e.* greater positive values in higher brackets

• Simple approach: take linear combination of cluster bracket means, with greater weights for higher income brackets: *f*

• Transform output to standard normal, under Gaussian assumption

$$f(\vec{w}) = -2w_1 - w_2 + w_4 + 2w_5$$

Analysis of Income:
Income Growth Health Score - IGHS

- Right: Cluster centroid coordinates of **W**, sorted by their corresponding IGHS.

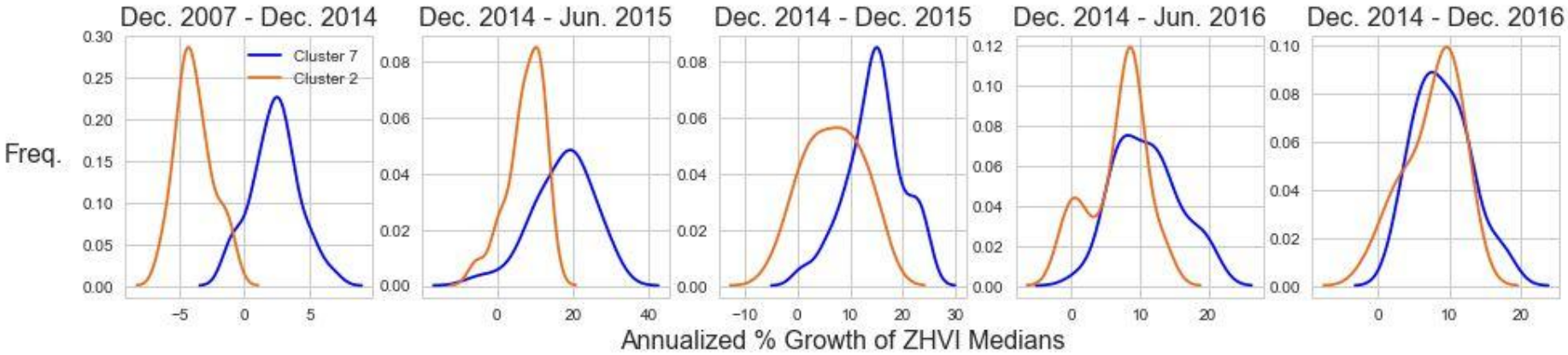
- Note that cluster #s are labels—the numerical values have no intended meaning.

- We use cluster 7 as a running illustration of a cluster with strong growth; cluster 2 for weak growth

	1: 25k-50k	2: 50k-75k	3: 75k-100k	4: 100k-200k	5: >200k	count	IGHS
cluster							
7	-8.187899	-3.084023	-0.465904	5.772686	5.965140	39	2.073688
9	-3.697212	-2.782451	-1.798880	-0.881259	9.159802	56	1.231902
5	-9.471188	1.983738	1.633733	5.960839	-0.107122	36	0.799000
0	-4.746756	-0.869686	0.442174	3.542077	1.632191	200	0.310942
1	-2.983695	-2.570535	-0.834074	4.943037	1.445268	120	0.240570
4	-2.193332	-1.881707	-0.984469	1.038351	4.021156	199	0.150422
6	-1.829289	-1.053150	-0.179817	2.320120	0.742136	377	-0.452020
8	-2.903538	-1.464789	-0.665759	9.699227	-4.665141	26	-0.529194
10	-3.614268	1.434829	1.373350	0.762242	0.043847	113	-0.617112
2	1.535650	-4.946841	1.098049	2.542466	-0.229325	59	-0.853779
11	-6.167152	13.573232	-10.259276	3.939187	-1.085991	11	-1.156288
3	1.038042	-0.942878	-0.962108	0.546975	0.319969	182	-1.198130

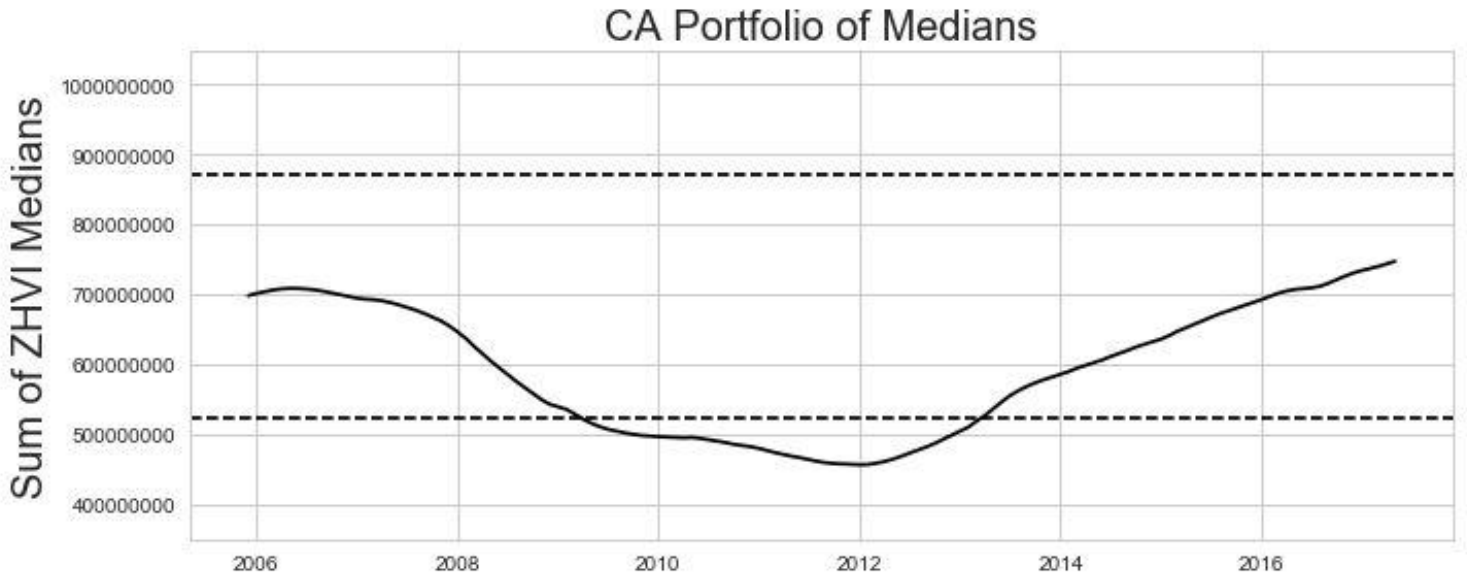
- ***We now move on to studying real estate value--according to the clusters we have developed***

Analysis of Real Estate Value: Distributions of ZHVI Growth



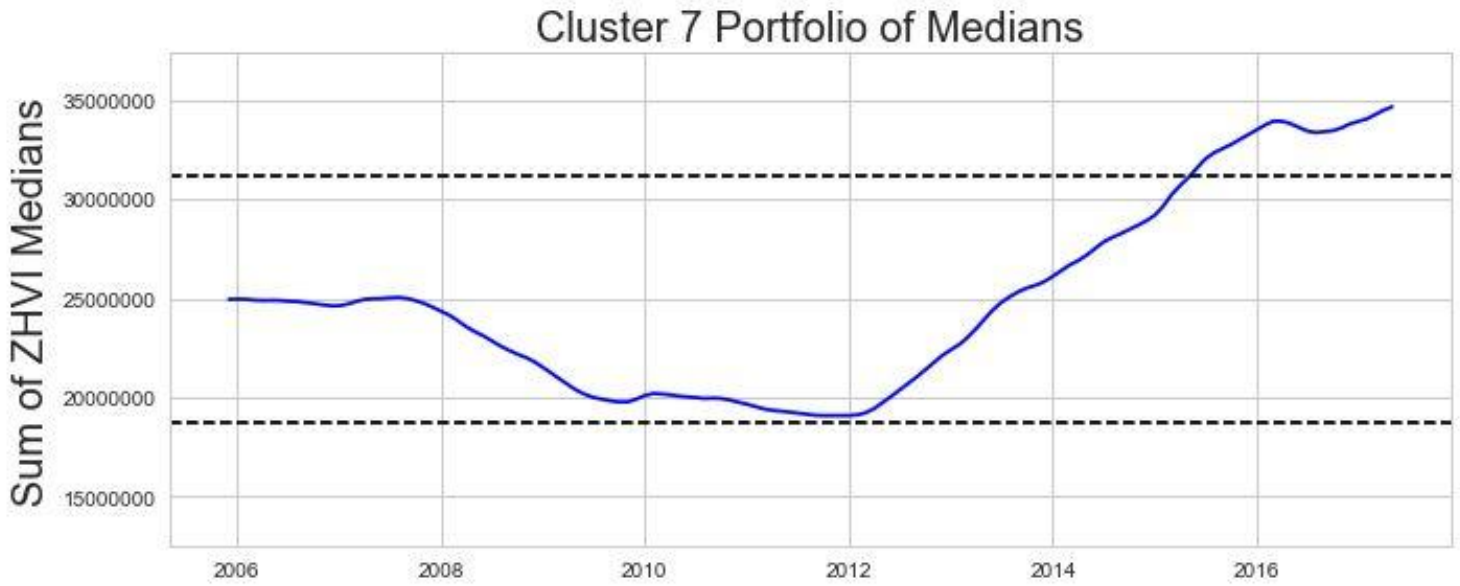
- Distribution estimations of real estate value growth are distinctly separated when measuring real estate data over same timeline as income data
- Distinction between clusters becomes less pronounced as we project into future

Analysis of Real Estate Value: Portfolios of ZHVI Medians



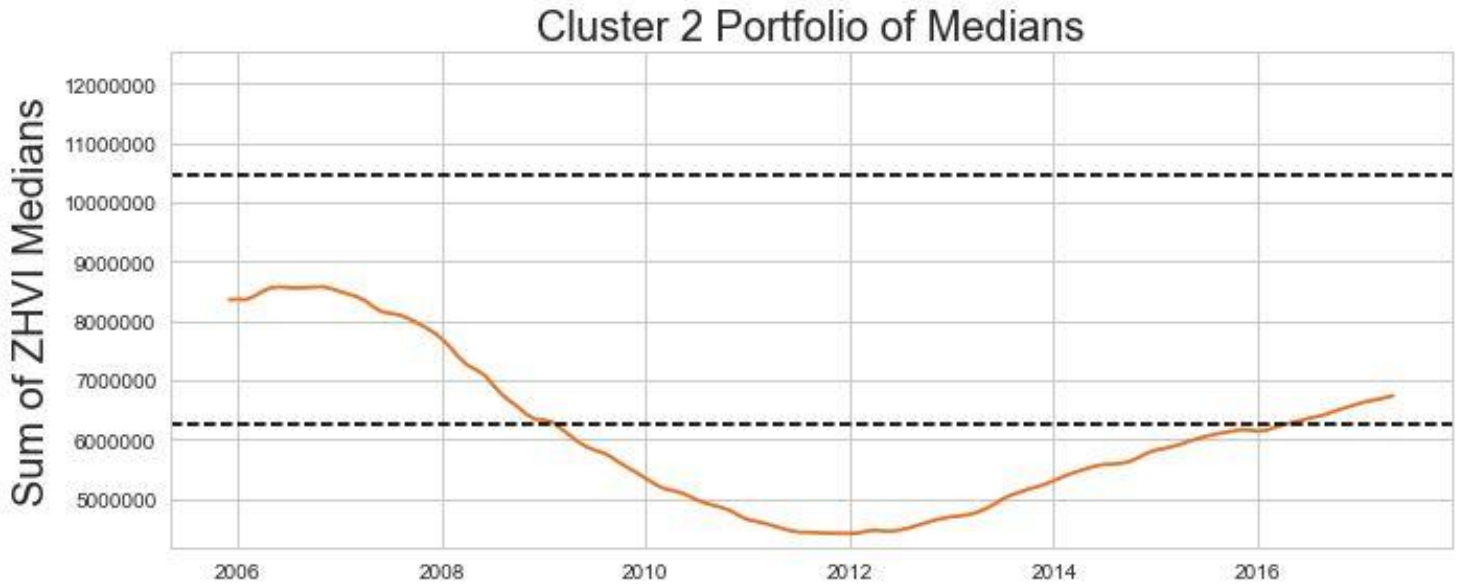
• Growth of California's housing market. Axis limits at +/- 50% of earliest point plotted; dashed lines at +/- 25% of earliest point plotted

Analysis of Real Estate Value: Portfolios of ZHVI Medians



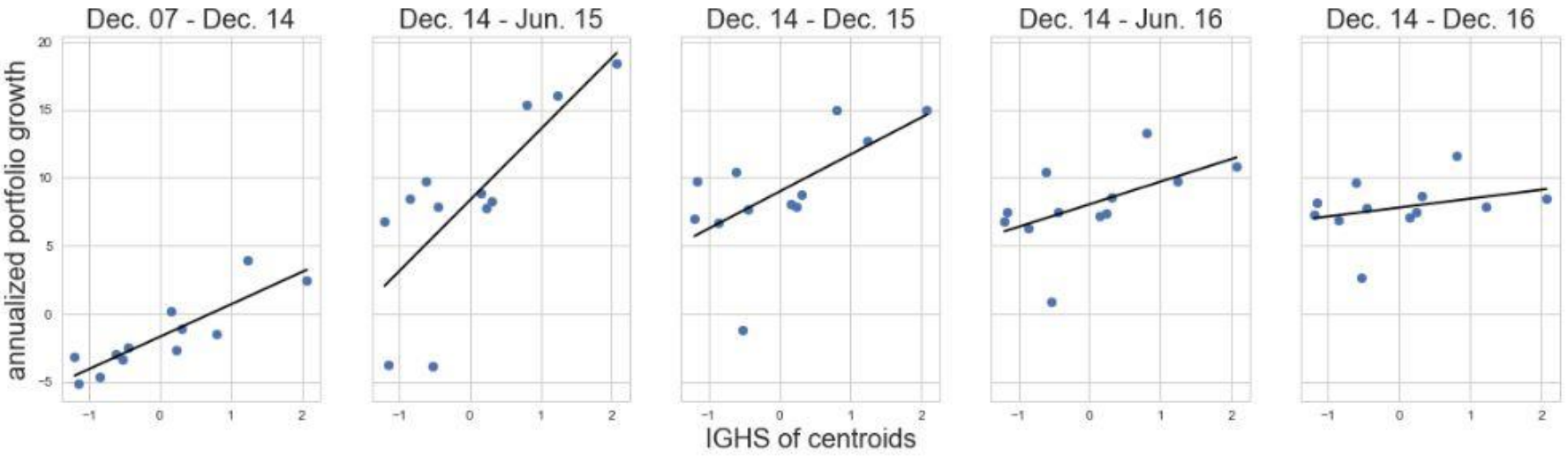
- Sample sum of medians from cluster 7. Note the better overall growth compared to CA

Analysis of Real Estate Value: Portfolios of ZHVI Medians



- Sample sum of medians from cluster 2. Note the worse overall growth compared to CA

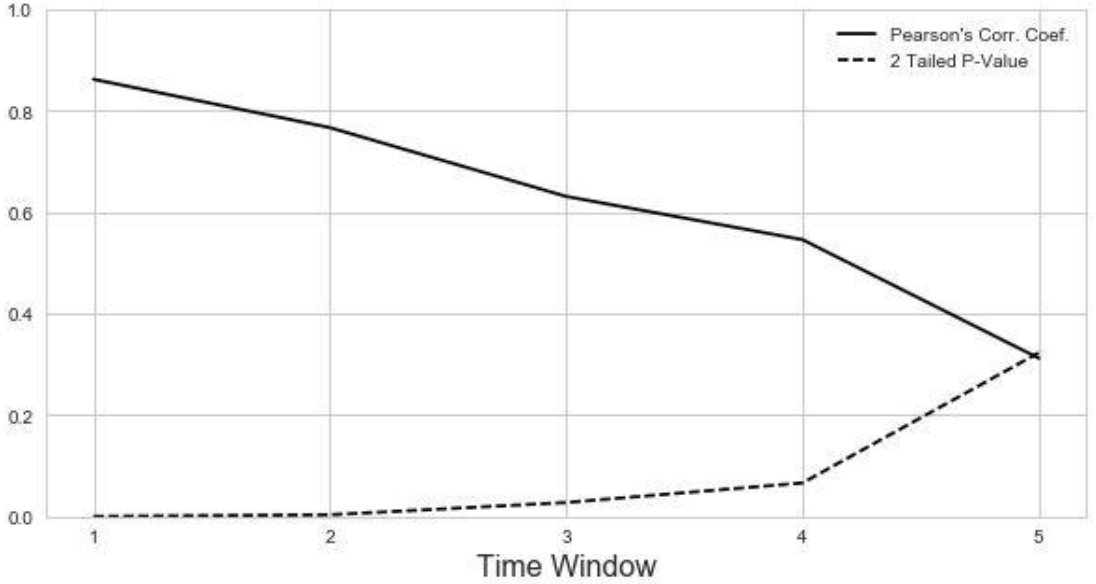
Analysis of Real Estate Value: IGHS-ZHVI Growth Portfolio Correlations



- We are measuring *changes* in income against *changes* in real estate value
- Relationship and goodness of fit weaken as we project further into future

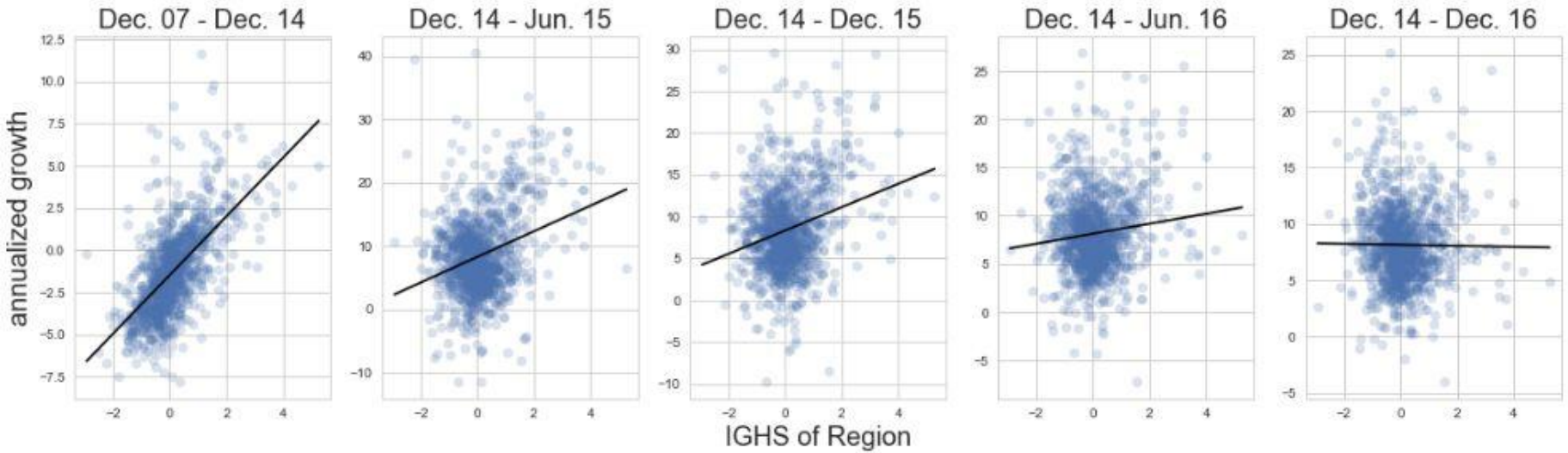
Analysis of Real Estate Value:
IGHS-ZHVI Growth Portfolio Correlations

- 1 - 12/2007 - 12/2014
- 2 - 12/2014 - 06/2015
- 3 - 12/2014 - 12/2015
- 4 - 12/2014 - 06/2016
- 5 - 12/2014 - 12/2016



- Time windows corresponds to the periods from previous slide
- Gradual weakening of both correlation and confidence in correlation

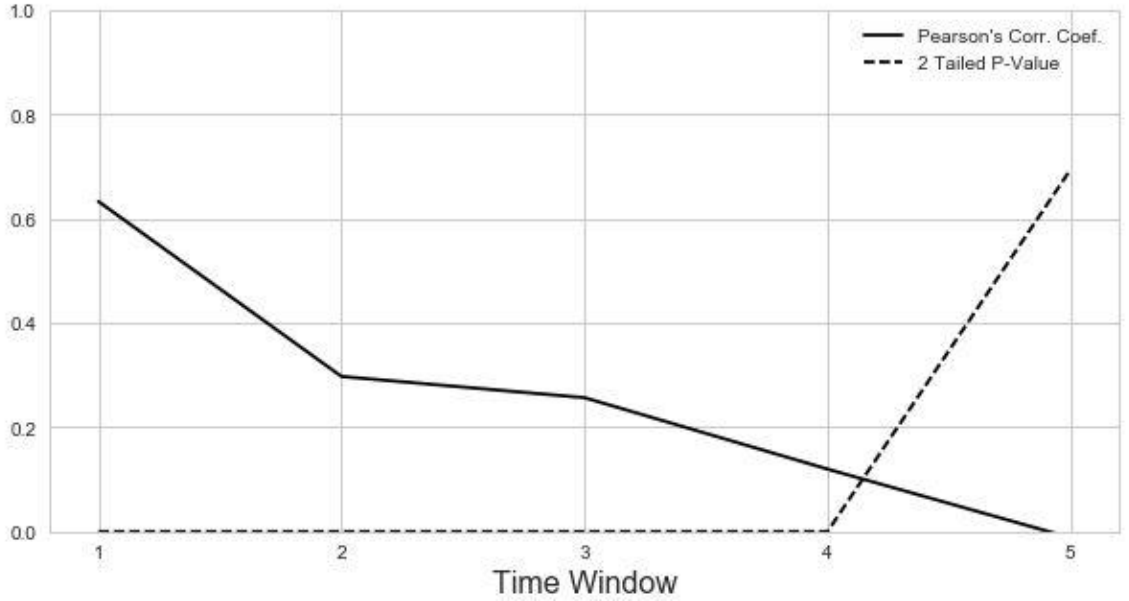
Analysis of Real Estate Value: IGHS-ZHVI Growth Individual Correlations



- Similar analysis with IGHS applied to individual regions rather than cluster centroids

Analysis of Real Estate Value: IGHS-ZHVI Growth Individual Correlations

- 1 - 12/2007 - 12/2014
- 2 - 12/2014 - 06/2015
- 3 - 12/2014 - 12/2015
- 4 - 12/2014 - 06/2016
- 5 - 12/2014 - 12/2016



- P-values are very small in the earlier/shorter time windows due to large sample size
- Despite large sample, there is a large decrease in confidence in the measured relationship in last window of time: model reliability deteriorates

§1: Key Takeaways

• Changes in income bear a relationship to changes in real estate value over short term

• To reliably measure new changes in real estate value, we need new income data

• With valid transformations, we can perform deeper analysis on even simple data sets

• Our application is best suited for the short term real estate investor

§1: Looking Ahead

• Bigger data sets would allow for more robust verification of the metric

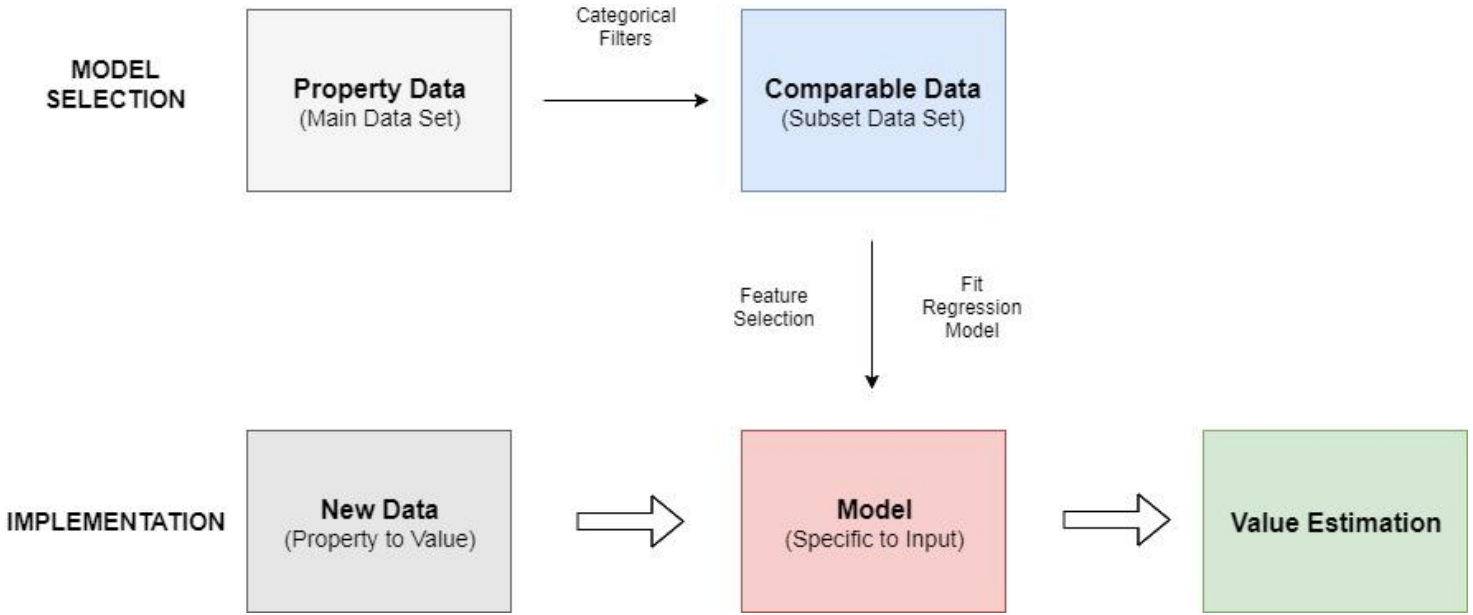
• More specific data sets would allow for more accurate measurements

• IGHS function can be improved to take inputs other than changes in income distributions

§2: Valuation

Development of regression models for valuation based on the *comparables* appraisal method

Overview: Valuation

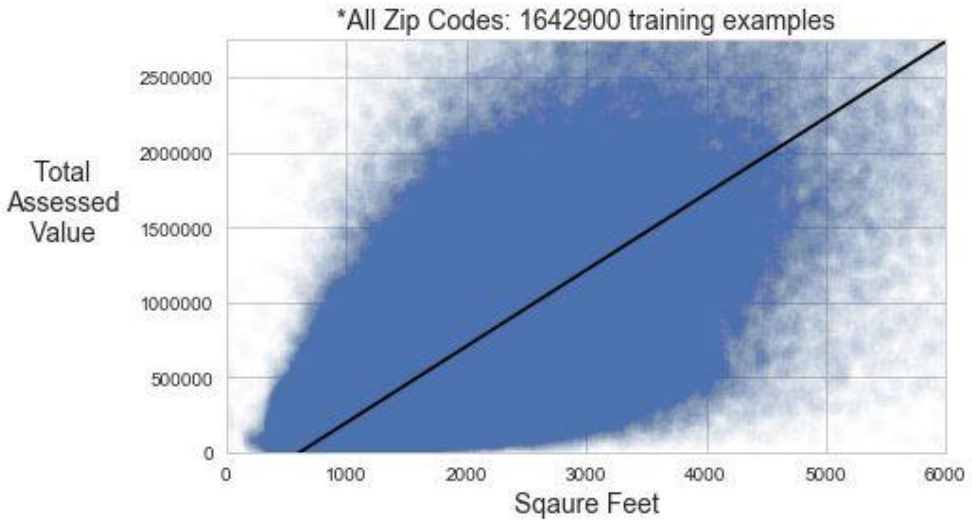


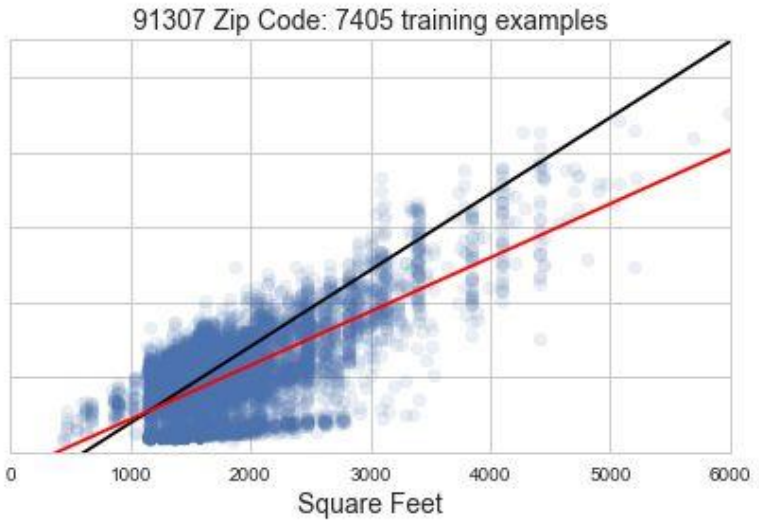
Data:
Zillow Research - Los Angeles Properties

	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	latitude	longitude	lotsizesqu	poolcnt	poolsizesi	pooltypei	pooltypei	pooltypei	propertyc	propertyl	propertyz	rawcensu	regionidci	regionidc	regionidn	regionidzi
2	34144442	-118654084	85768						010D	269		60378002	37688	3101		96337
3	34140430	-118625364	4083						109	261	LCA11*	60378001	37688	3101		96337
4	33989359	-118394633	63085						1200	47	LAC2	60377030	51617	3101		96095
5	34148863	-118437206	7521						1200	47	LAC2	60371412	12447	3101	27080	96424
6	34194168	-118385816	8512						1210	31	LAM1	60371232	12447	3101	46795	96450
7	34171873	-118380906	2500						1210	31	LAC4	60371252	12447	3101	46795	96446
8	34131929	-118351474							010V	260	LAC2	60371437	12447	3101	274049	96049
9	34171345	-118314900	5333						1210	31	BUC4YY	60373108	396054	3101		96434
10	34218210	-118331311	145865						010D	269	BUR1*	60373101	396054	3101		96436
11	34289776	-118432085	7494						1210	31	SFC2*	60373202	47547	3101		96366
12	34265214	-118520217	3423						1200	47	LAC2	60371112	12447	3101	31817	96370
13	34447747	-118565056	81293						010D	269	SCUR3	60379201		3101		96377
14	34465048	-118568166	6286						010D	269	LCA25*	60379201		3101		96377

- Data set is sparse but substantial (3 million + properties), subset illustrated above
- Remove empty columns and rows to work with full data set - reduced to 1.6 million

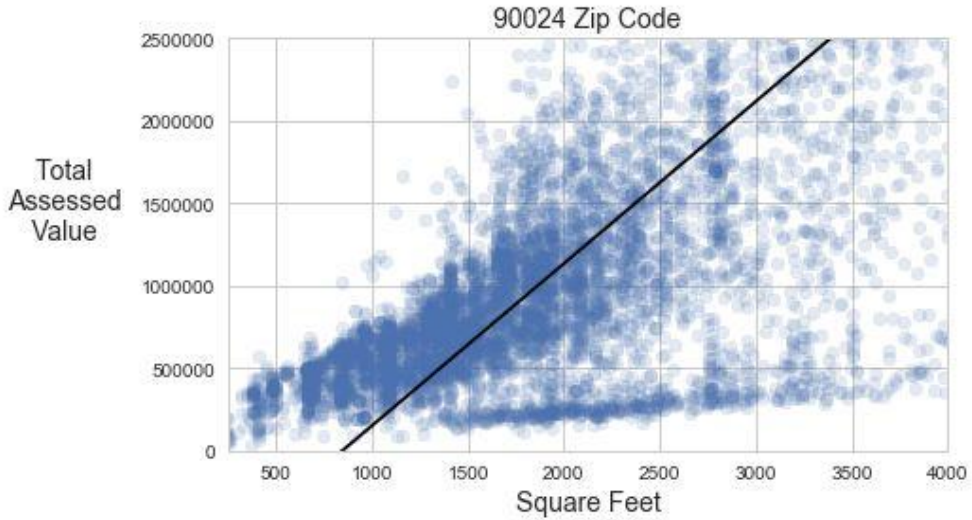
Model Selection: Categorical Filters

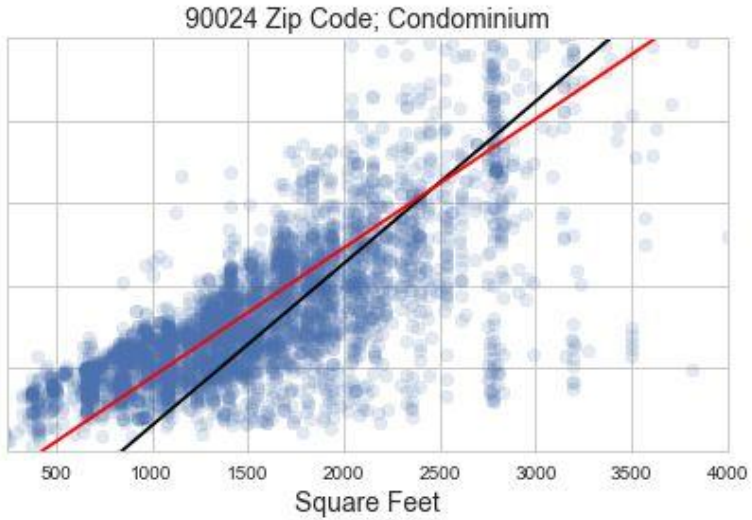




- The first and most important filter specifies the region of our property (defined by zip code)
- Note the substantial reduction in data used to develop the linear model

Model Selection: Categorical Filters





- Adding filters continues to adjust our linear model to be based on more relevant properties
- Important that our models use sufficient data to minimize probability of a skewed model

Model Selection: Feature Selection

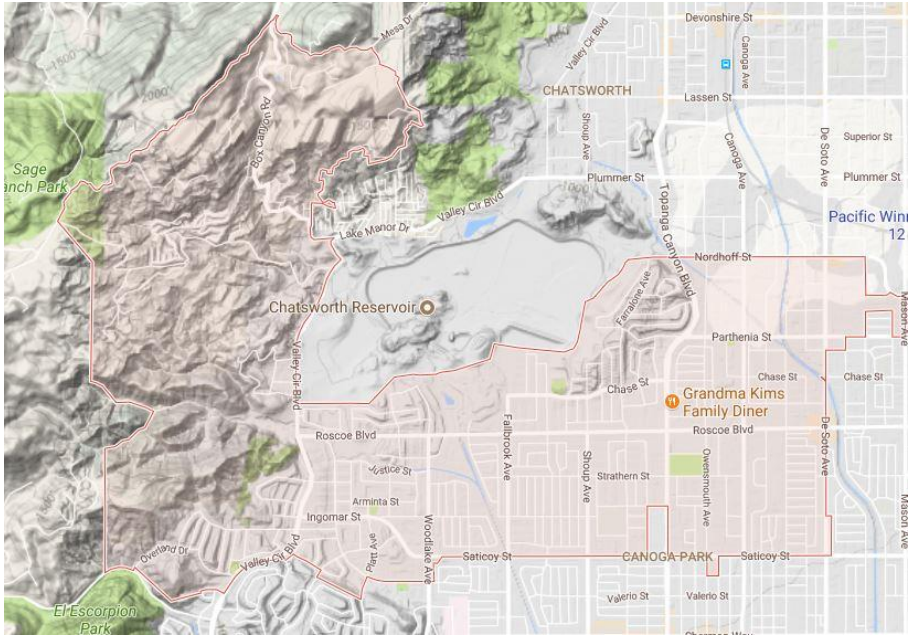
```
totalassessedvalue              1.000000
calculatedfinishedsquarefeet    0.647142
bathroomcnt                     0.536924
bedroomcnt                      0.303769
yearbuilt                       0.144349
lotsizesquarefeet               0.003141
roomcnt                         0.001399
latitude                       -0.017477
buildingqualitytypeid          -0.090570
longitude                      -0.169651
Name: totalassessedvalue, dtype: float64
```

```
totalassessedvalue              1.000000
calculatedfinishedsquarefeet    0.756331
bathroomcnt                     0.637797
bedroomcnt                      0.554183
yearbuilt                       0.410167
latitude                       -0.198977
lotsizesquarefeet              -0.233884
buildingqualitytypeid          -0.313282
longitude                      -0.527611
roomcnt                              NaN
Name: totalassessedvalue, dtype: float64
```

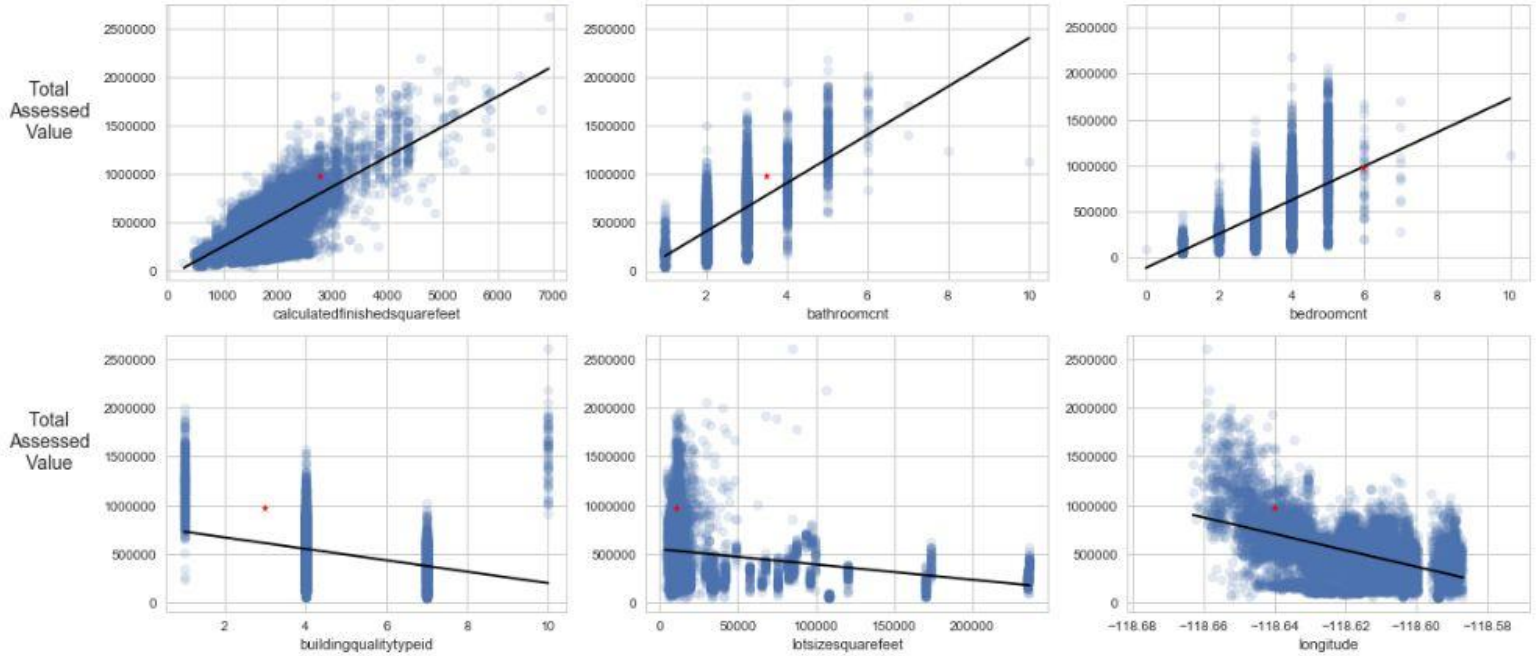
- With different inputs, features will affect value to different extents, *e.g.* longitude in figure above - All properties vs. 91304 properties
- Use minimum magnitude of 0.2 in correlation to total assessed value as cutoff, to mitigate unnecessary model complexity and noise

Model Selection:
Feature Selection Example

- More negative longitude implies farther eastward
- In the 91304 region, elevation climbs with eastward direction
- Properties in the hills tend to be valued higher
- Thus, we observe a large negative magnitude in longitude's correlation to value



Total Assessed Value: Breakdown by Feature



- RED: listed price for input property. BLUE: *comps* used to model BLACK: linear fit

§2: Key Takeaways

• Our modeling approach seeks to replicate the *comps* appraisal method

• Want to strike an optimal balance between relevance and sufficiency of training data

• Restrict features in regression to variables bearing relatively strong relationship to value

§2: Looking Ahead

• Can improve valuation with non-linear models

• Data used as regression features can be used to create a better *comps* training set, e.g.:

- Use latitude/longitude to specify closer location

- Use bedroom/bathroom counts to specify exact property type

• Can develop valuation model not based on *comps*

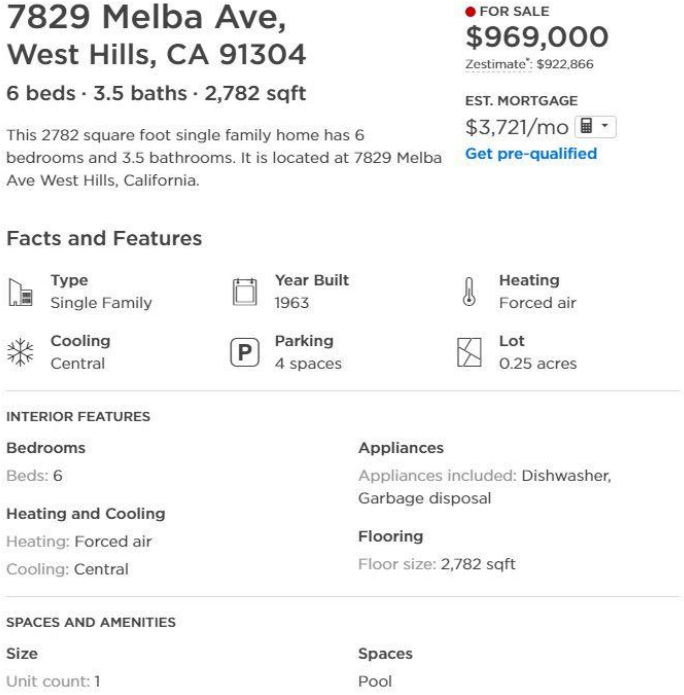
Test Set

• Extract new input data for 20 properties from listings on Zillow, e.g. right figure

• Make appropriate conversions and calculations (e.g. acres, latitude, longitude) - trivial to automate

• Use “sale prices conversion” to adjust model’s output (total assessed value) to account for brokerage fees, legal fees, etc. (Not recommended in practice)

• **Disclaimer**: the results of our analysis are hypothetical, since the timelines of the data sets do not match.



7829 Melba Ave,
West Hills, CA 91304
6 beds · 3.5 baths · 2,782 sqft

This 2782 square foot single family home has 6 bedrooms and 3.5 bathrooms. It is located at 7829 Melba Ave West Hills, California.

FOR SALE
$969,000
Zestimate®: $922,866

EST. MORTGAGE
$3,721/mo
Get pre-qualified

Facts and Features

Type Single Family	**Year Built** 1963	**Heating** Forced air
Cooling Central	**Parking** 4 spaces	**Lot** 0.25 acres

INTERIOR FEATURES

Bedrooms
Beds: 6

Heating and Cooling
Heating: Forced air
Cooling: Central

Appliances
Appliances included: Dishwasher, Garbage disposal

Flooring
Floor size: 2,782 sqft

SPACES AND AMENITIES

Size
Unit count: 1

Spaces
Pool

Growth - Test Set Results

- Filter results to regions with above average income growth, *i.e.* IGHS greater than 0
- Sort regions by IGHS
- In effect, this ranks the regions by their implied strength of income growth

	1: 25k-50k	2: 50k-75k	3: 75k-100k	4: 100k-200k	5: >200k	cluster	IGHS
zipcode							
90292	-4.146367	-3.174431	-1.608475	4.630890	4.298383	1	1.354536
90045	-5.646175	-1.945737	0.644221	2.629204	4.318486	0	1.334581
90046	-5.273282	-0.311768	0.882458	2.738580	1.964013	0	0.603720
91307	-1.983989	-1.857438	-2.145410	3.105439	2.881398	4	0.307233
91406	-3.550964	-0.920699	0.315629	3.221998	0.934036	0	0.141679
91403	-3.980963	-0.008550	0.106258	2.639672	1.243583	0	0.140068
91344	-2.120445	-1.983828	-0.634051	3.336529	1.401795	6	0.063364
90049	-2.418760	-0.812035	-0.136197	0.378598	2.988394	4	0.025677

Note the zip codes marked by the colored arrows - they reappear when we apply a similar filter to find properties with good value

Valuation - Test Set Results

- Filter results to underpriced properties, *i.e.* suggested price > offered list price

- Order by list price's number of standard deviations from the suggested price

- Properties are indexed for anonymity, but represent actual listings

	Zip Code	Suggested Price	List Price	+/- SDs	id
3	91344	$1,809,985.56	$1,529,000.00	-1.023408	3
12	91304	$680,158.51	$439,000.00	-0.980891	12
6	91343	$948,010.60	$749,950.00	-0.748384	6
11	91401	$789,285.50	$549,000.00	-0.714486	11
5	90057	$559,544.14	$390,000.00	-0.331496	5
16	91406	$805,109.92	$745,000.00	-0.194159	16
4	91342	$583,321.25	$550,000.00	-0.117995	4
14	91352	$506,970.81	$469,964.00	-0.105863	14

** Valuation is actually for May 2017, but we hypothetically suppose it is for Jan. 2015--directly following the period of our income data--for the sake of illustration*

Market Analysis - Test Set Results

• Only two properties appear in both the filtered tables, *i.e.* have above average growth prospects and are underpriced: ID 3 and 16, marked with arrows

• We have reduced our test set of 20 properties to 2 properties which represent the best investment opportunities according to our analysis:

- The properties are in regions with good growth prospects for the short term future
- The properties are underpriced: their list price is below our estimates for their “true value”
- One has better growth prospects, the other offers better value

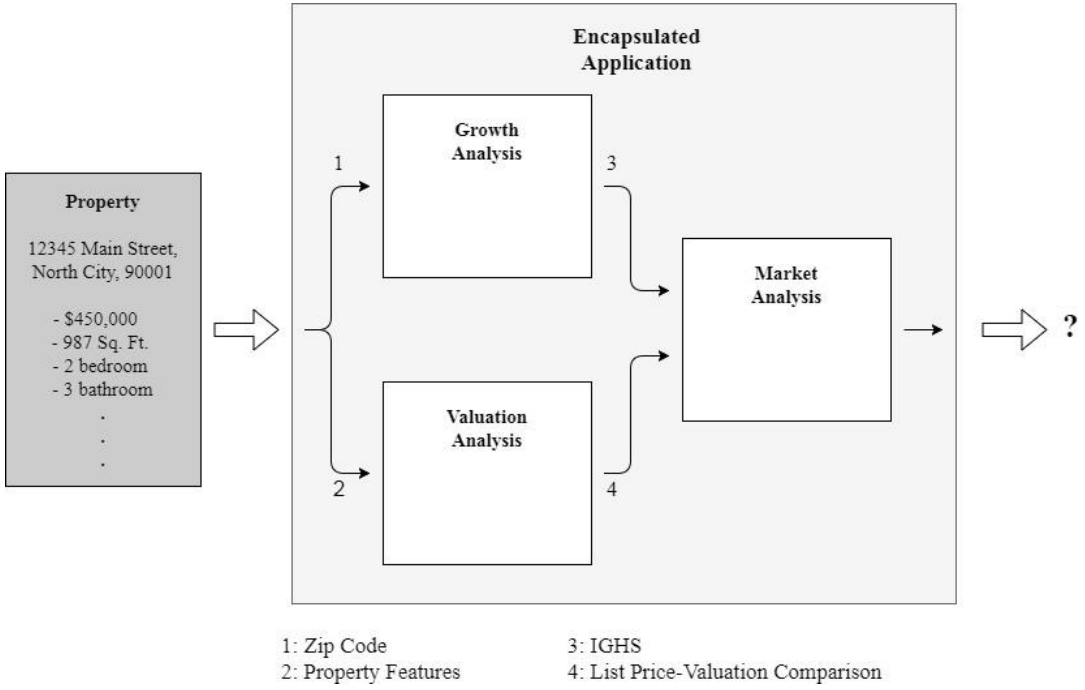
• It is feasible to take the statistics from each analysis and further determine which of the two is the better investment with deeper analysis.

Automated System - Overview

- The system thus far automates lower level analysis, growth and valuation, while leaving higher level decision to a human analyst

- Assuming valid lower level analysis, we can develop systems to automate the higher level analysis, i.e. the ‘Market Analysis’ component in the right figure

- Choosing the output depends on the objective of the system



Summary

• Growth and valuation are two major economic elements of investment analysis that we've included in our system

• We've measured just two elements with limited data (limited both in volume and complexity). Our systems can be improved with more elements and better data sets

• Automated applications can be improved with more sophisticated algorithms and encapsulated system architecture

• ***The principles of this study can be generalized to other assets--we simply need the means to measure and model variables of interest***