Real Estate Market Analysis
A Data-Driven Approach

Pablo Guevara
September 2017

http://www.github.com/pandrewg/CA_housing_analysis

Abstract Traditional approaches to investment analysis in housing markets have been significantly dependent on human judgment, with data serving to guide the analyses. The development of increasingly rich data sets and the means to process the data is making it feasible to gain deeper insights as well as automate more of the research once performed by humans. We examine two components of this ongoing development: producing reliable novel metrics and producing systems which build upon established metrics. In particular, our applications will serve short-term real estate investors in discerning the growth prospects and current value of a property. The algorithms, when designed with the proper objectives and fueled by the data, will be able to interpret market conditions more effectively and efficiently than methods overly reliant on human intervention. The applications we develop are relatively nascent, and they serve as introductions to more sophisticated products.

Table of Contents

· INTRODUCTION

Real estate markets continuously evolve in response to changing economic climates. Taking the standpoint of researchers supporting investment decisions, we desire to produce intelligent analyses of different market elements. We have two primary goals with this project: (1) to illustrate the process of developing novel metrics which accurately encapsulate economic information—verified through data and (2) to illustrate the process of developing systems built upon these verified metrics—from processing the raw data to outputting the results.

We assume that the outputs of our systems will be read and used by a human client (a real estate investor developing higher level analyses); although, there is nothing restricting us from designing the systems to be integrated within a larger automated application. The project is far from all-encompassing with respect to investment analysis and serves only as an exploratory study for how a data scientist could develop more complex systems which draw from deeper data sources and are built upon more sophisticated models.

This report is organized according to the two major components we study: growth prospects (section 1) and valuation (section 2). Along the way, we develop code which enables our analyses and can be generalized to other data sets of the same nature.

For this report, we focus on only the key insights of our study. Workbooks containing detailed explanations on each section are available at the link on the cover page.

§ 1: GROWTH

1.1 : Overview

In this section, we study income data within zip codes in California and their relationship to real estate value data in those corresponding zip codes. Our goal is to develop an intuitive metric for measuring the relationship between growth in income and growth in real estate value over varying time periods. Sub-section (1.2) will be concerned with analyzing income data and exploring patterns that suggest growth in income strength. In sub-section (1.3), we study real estate value from a separate data set, but with the structure developed in sub-section (1.2) as an attempt to show connections with this structure. This is possible because data points in both sections are identified by zip codes.

In particular, after preliminary data processing, we cluster zip codes based on their net percentage changes in income distributions (from a derived data set we denote **W**) and describe a method for endowing clusters with a ranking based on their strength of income growth (our intuitive metric— what we denote 'IGHS'). We then apply the class labels (and ranking) to the data set on real estate value and study the characteristics of real estate value growth within each cluster using different measurements. Our goal is to find distinct patterns of real estate growth between distinct clusters which are consistent with our intuitive metric. Since the clusters are developed via analysis of income, this may be

an indication for a relationship between income growth (defined by our metric) and real estate value growth. Below, we have an overview of the analytical pipeline in our envisioned system.

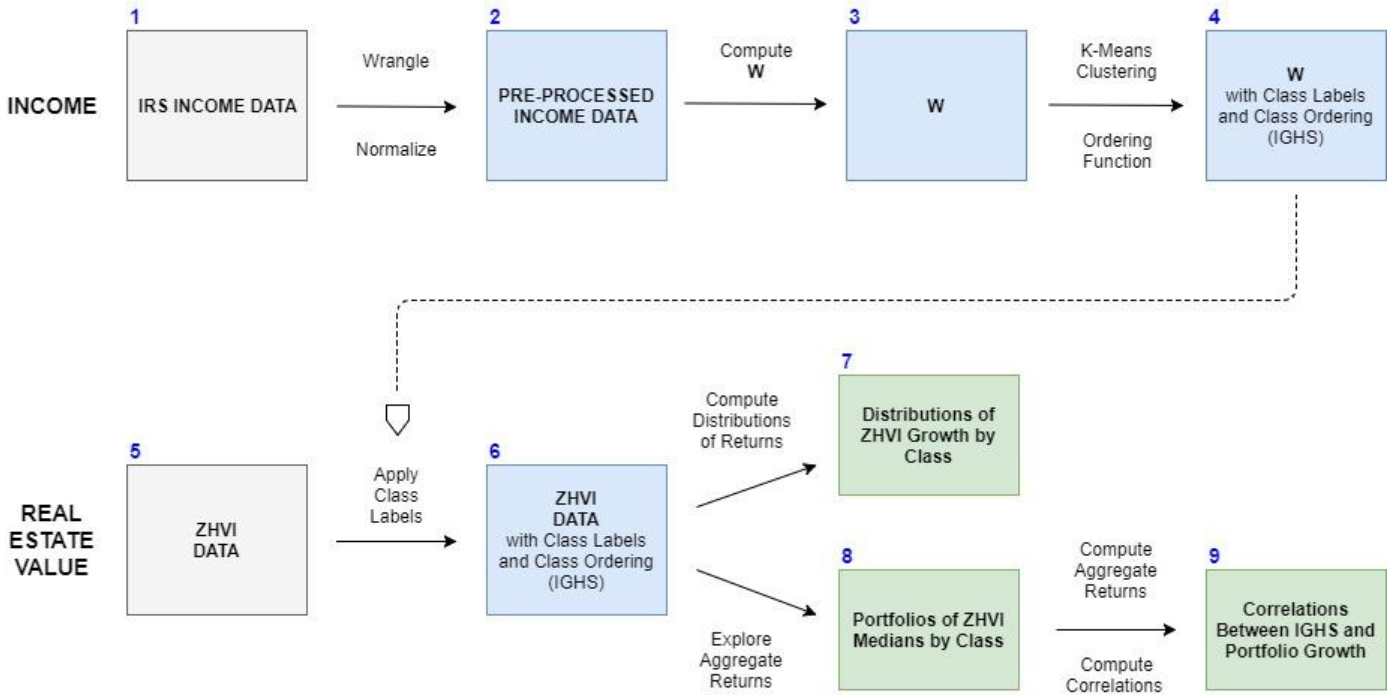

Figure 1.1 *GREY - raw data sets. BLUE - derived data sets. GREEN - computed metrics. SOLID ARROWS - algorithms and data manipulation. DOTTED ARROW - transfer of class labels and ranking.*

1.1.1: Background

The prices for properties in a region are dependent on demand for those properties. With an increase in demand in a region, suppliers will find buyers willing and able to purchase their properties at higher prices. Since pricing models in real estate markets depend heavily on comparable properties, price levels for properties in that region will thus tend to rise as a result of increased demand.

Demand in a market segment can be broken down into several components, but we will focus on measuring *purchasing power*. In particular, we study income levels—a fundamental determinant of total purchasing power in a region. Purchasing power is also dependent on population, and we account for this by normalizing income distributions and assuming that populations are non-decreasing. The population in California increased by about 2.4 million during the timeline of our data set, so it is a reasonable assumption that populations in most regions did not decrease. Furthermore, we study growth over collections of regions, giving us stronger confidence that populations did not decrease over these collections.

1.1.2: Data

In sub-section (1.2), we study data derived from a data set released by the U.S. Internal Revenue Service (IRS) [1] containing counts of income tax returns—separated by zip code and income bracket. In essence, the data set contains a six dimensional time series for each zip code (one dimension for each income bracket) over the 7 year period from 2007 to 2014. We reduce the data set to five dimensions as a more accurate representation of market participants and normalize each distribution to study percentages.

		%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

Figure 1.2 *Example subset of the pre-processed income data, representing the percentage makeup of each income bracket out of the income distribution.*

In sub-section (1.3), we study a data set released by Zillow Research [2] containing time series of real estate value estimates. In particular, we study a metric coined the "Zillow Home Value Index" (ZHVI), where each time series represents the median ZHVI for a zip code in California. ZHVI is one of many possible approaches to creating an index to approximate real estate value. The details of ZHVI will not be fully explored in this report and we will simply note that ZHVI is a reflection of real estate price levels, so analyzing the metric will give us an idea of how the 'true real estate value' of different regions changed. See Figure 1.3 for an example of the data.

	1996-04	1996-05	1996-06	1996-07	1996-08	1996-09	1996-10	1996-11	1996-12
zipcode									
94109	298200.0	295700.0	296400.0	298500.0	298800.0	298900.0	298700.0	297200.0	297800.0
90046	264500.0	265000.0	266600.0	269800.0	273600.0	276400.0	276900.0	277800.0	279700.0
94501	206900.0	207900.0	208100.0	208400.0	209500.0	211200.0	213100.0	215400.0	217600.0

Figure 1.3 *Example subset of the ZHVI medians data set.*

Pre-processing of the data will not be explored in this report, and details can be found in the tax EDA workbook. We will begin section (1.2) at step 2 in Figure 1.1.

1.2: Analysis of Income

1.2.1: Distribution Analysis of Changes in Income Brackets

We first create a table succinctly representing net percentage changes in income brackets. This enables us to directly study the underlying statistical characteristics, as well as process the data for clustering. To keep a running illustrative example, we create a table **W** representing changes in income distribution from 2007 to 2014. This is simply done by taking the difference between year 2014 and year 2007 from the table in Figure 1.2.

	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

Figure 1.4 *Subset of **W**: net percentage changes in each income bracket from 2007-2014.*

This allows us to study the distributions of each component of these vectors to gain insight on how income brackets all over California changed over the time period. As a caveat, the vectors in the table are derived only from years 2007 and 2014, and do not explicitly include information from years 2008-2013. The justification is that the end points are reflections of the entire time window, so measuring them gives a simplified representation of the entire period.

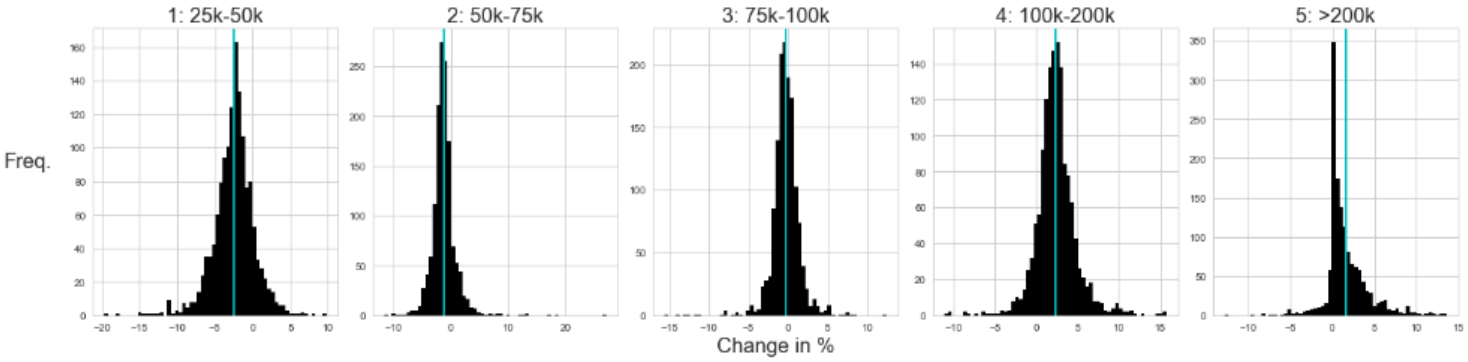

Figure 1.5 *Distributions of changes in each income bracket between years 2007 and 2014. The mean of each distribution are marked in cyan.*

We can see in Figure 1.5 that for most regions, the proportion of earners in the lower two brackets decreased. This is evident by the majority of the distributions being below 0%

(in particular, the means are under 0%). The third income bracket is more centered, suggesting no bias over the data set. Finally, we observe more frequent increases in the higher income brackets (the means are above 0%). This overall pattern suggests that income levels generally increased throughout California, with most regions having more residents move from lower brackets to higher brackets than vice versa.

1.2.2 Clustering Changes in Income Brackets

We apply K-Means clustering to classify each zip code into one of a set of clusters. By classifying zip codes, we are later able to able to measure changes in real estate value over samples rather than individual regions. We choose the number of clusters, 12, based on the error reduction in the total sum of squares when adding a cluster to a system (see figure 1.6). As we achieve diminished reduction in error, we stop the process to avoid overfitting.

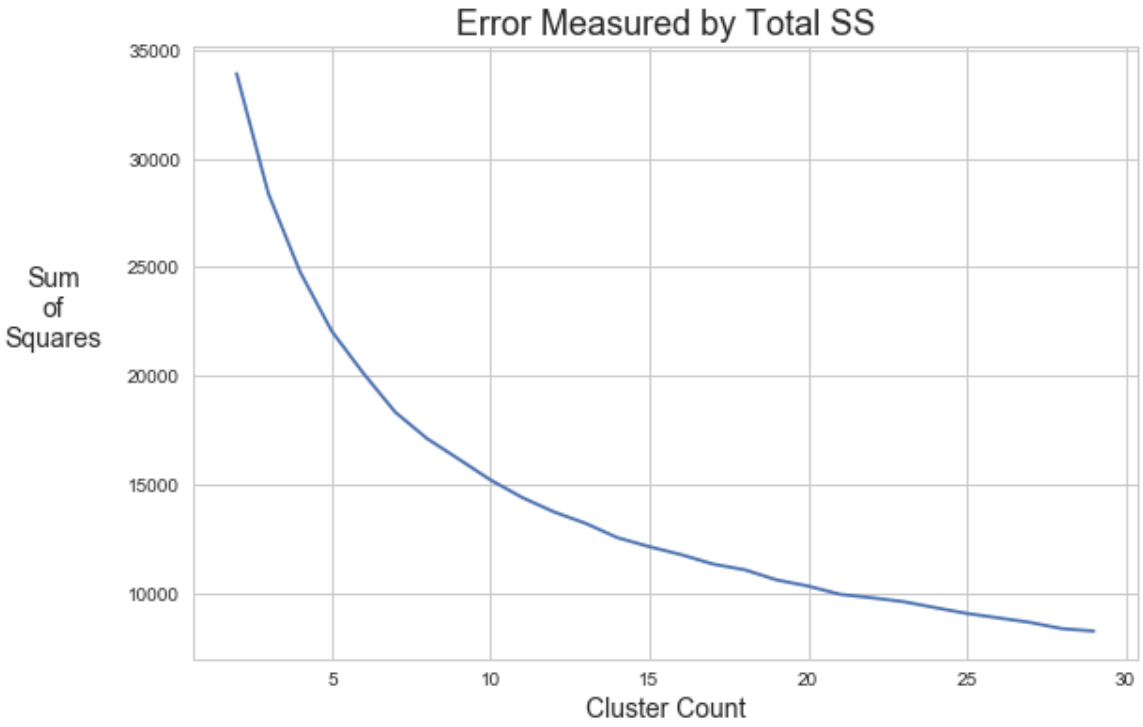

Figure 1.6 *Note the "elbow point" is not clearly distinguishable here. In practice, we would use more robust criteria in determining the number of clusters.*

Each cluster centroid can be interpreted as the general "behavior" of how income distributions changed—the behavior shared by zip codes belonging to that cluster. For example, a centroid with values close to 0 across all brackets can be interpreted as an indicator for stability and lack of change. On the contrary, centroids with large magnitudes in its components indicate greater change over the time period, with many residents moving from one income bracket to another.

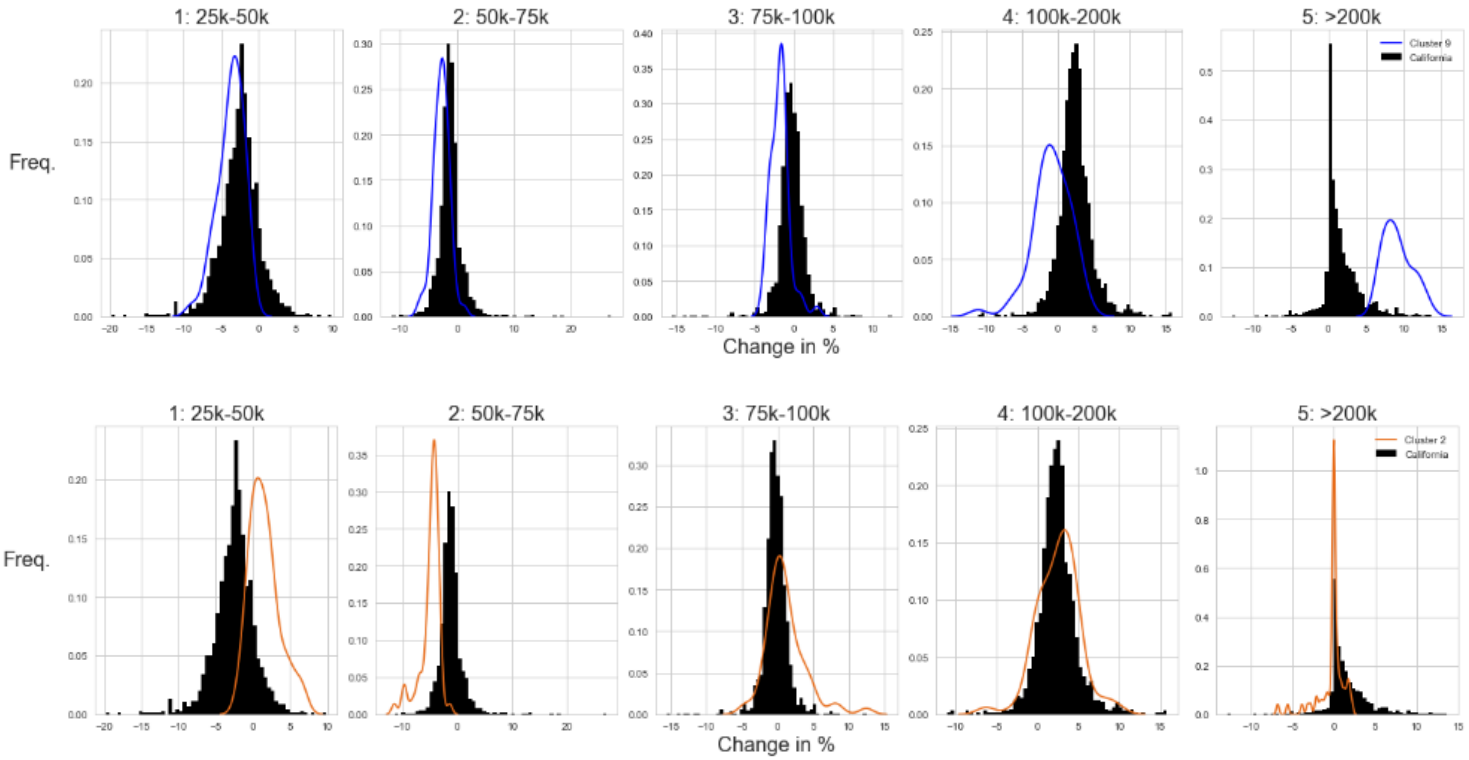

Figure 1.7 *Example distribution plots for two clusters (kernel density estimations, in color) against distributions plots of the whole data set (normalized histograms, in black).*

In Figure 1.7, the distributions on top (with blue KDE plots) represent a cluster that we interpret as having relatively good income growth over the time period. In contrast, the distributions on bottom represent a cluster that we interpret as having relatively bad income growth. We arrive at these interpretations by studying the positions of the KDE plots relative to the positions of the histograms representing the entire data set. Cluster KDE plots centered leftward on the lower income brackets and rightward on the higher income brackets suggest relatively stronger income growth, because this suggests relatively more of the income distributions shifted *away from the lower brackets* and relatively more shifted *towards the higher brackets*.

We will use these clusters—labeled 9 and 2, respectively—in section (1.3) to illustrate metrics of real estate growth in what we interpret as clusters with strong income growth and weak income growth, respectively.

1.2.3: Ranking Clusters: Income Growth Health Score (IGHS)

The plots in Figure 1.7 enable us to visually analyze the characteristics of different clusters, but we wish to develop a method to compare clusters more directly. Thus, our next goal is to associate with each cluster a value computed from statistics of the distributions, such that the values reflect strength of income growth. In turn, these values will endow our clusters with a ranking.

We take a weighted sum of the means of each distribution (i.e. the centroids we have established for **W**) such that the output increases for larger values in the higher brackets. Ultimately, we will test the viability of this function with the ZHVI data set—under the assumption that increasing purchasing power indeed causes real estate prices to increase. In particular, the weights -2, -1, 0, 1, 2 are used (in order of lowest income bracket to highest). That is, the IGHS function is:

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

In Figure 1.8, we can see that centroids higher to the top (greater IGHS) tend to have higher values in the higher brackets and lower values in the lower brackets than centroids lower to the bottom (lower IGHS). Still, we see different sorts of combinations of values near the top—suggesting that income strength can grow in different ways when measured with our IGHS function.

	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

Figure 1.8 *Cluster centroid coordinates of* ***W****, sorted by their corresponding IGHS. Note that cluster #s are labels—the numerical values have no intended meaning.*

1.3: Analysis of Real Estate Value

Having developed a structure on our income data, we now move on to studying real estate value data—with the goal of finding a parallel between the income data and the real estate value data.

1.3.1: Distributions of ZHVI Growth

Our first approach to measuring growth in real estate value will be to compute the annualized percentage return in ZHVI medians for each zip code and analyze the distributions when we separate the data by the clusters we have established while studying income. We do this for each zip code for which we have ZHVI data and for five different windows of time:

- **Dec. 2007 - Dec. 2014** (7 years overlapping with income data)
- **Dec. 2014 - Jun. 2015** (6 months following income data)
- **Dec. 2014 - Dec. 2015** (12 months following income data)
- **Dec. 2014 - Jun. 2016** (18 months following income data)
- **Dec. 2014 - Dec. 2016** (24 months following income data)

The first 7-year period overlaps with the time window of our income data, whereas the last four windows are 6, 12, 18, and 24-month periods following the period represented by our income data. Thus, the latter four windows of time are used to test the viability of our models in predicting future changes in real estate value.

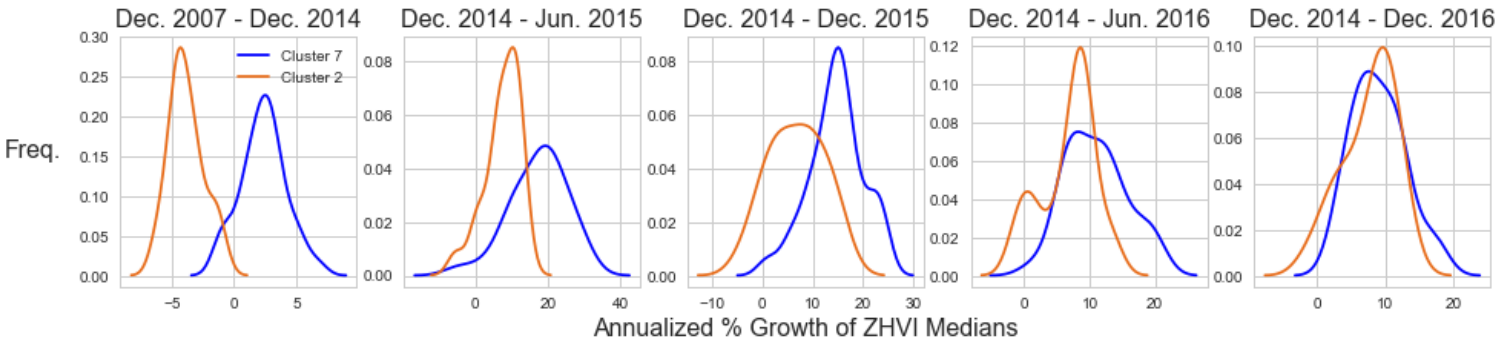

Figure 1.9 *KDE plots of ZHVI medians' annualized percentage growth for clusters 9 and 2.*

We can see the KDE plots in the leftmost graph are more distinctly separated than the right four graphs. Real estate from regions in cluster 9, our example cluster of strong income growth, tended to grow at a greater rate than those in cluster 2 (as measured by their medians—this is a limitation of our ZHVI data set). As we measure the annualized growth further into the future, the distributions of returns overlap more. In the right two plots, there appears to be no significant distinction between the distributions.

This suggests that the classification we established by studying income data may be a good indicator for real estate value growth *only when measuring over the same time period.* When we try to predict future changes in real estate value, our classification becomes less reliable as we see a greater overlap in the distributions of growth.

Our next approach to accessing the practicality of our income-growth health scores will be to measure the aggregate returns in samples of ZHVI medians when we separate the regions by the classification established in studying income.

1.3.2: Portfolios of ZHVI Medians

In the last sub-section we computed the percentage returns of individual zip codes and subsequently grouped together the zip codes to establish distributions. Here, we reverse the approach: we first group together the zip codes into their respective groups and subsequently measure the percentage return of the sum of the samples. That is, we create hypothetical portfolios of real estate properties and study their performance.

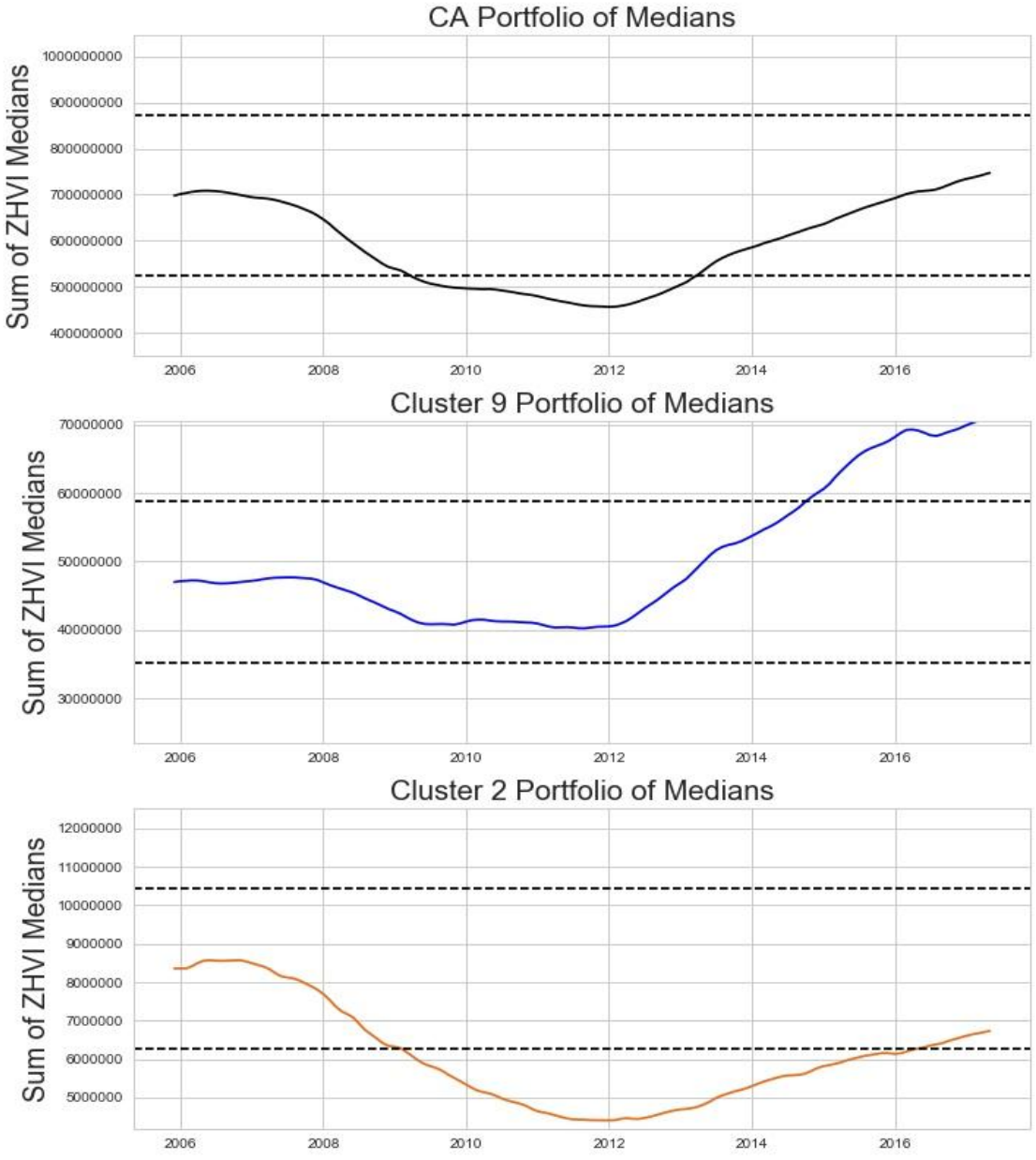

Figure 1.10 *Growth of hypothetical portfolios. Axis limits are scaled to +/-50% of the earliest point plotted. Dashed lines represent +/-25%.*

We can see in Figure 1.10 that in comparison to California as whole, regions from cluster 9 were much more resilient against the loss that occurred from 2007 to 2014. On the contrary, cluster 2 lost even more value during that time period.

1.3.3: IGHS and ZHVI Growth Correlations

Now, we step away from studying clusters 9 and 2 in isolation and instead look at the behavior of all of the clusters. We revisit the income-growth health score and measure its correlation to annualized growth in real estate value in regions corresponding to distinct clusters. This is similar to the approach we took in (2.2) in that we first aggregate the regions and subsequently measure their growth, but we return to studying the five windows of time established in (2.1) so that we can plot an aggregate percentage return for each cluster portfolio in a more succinct manner.

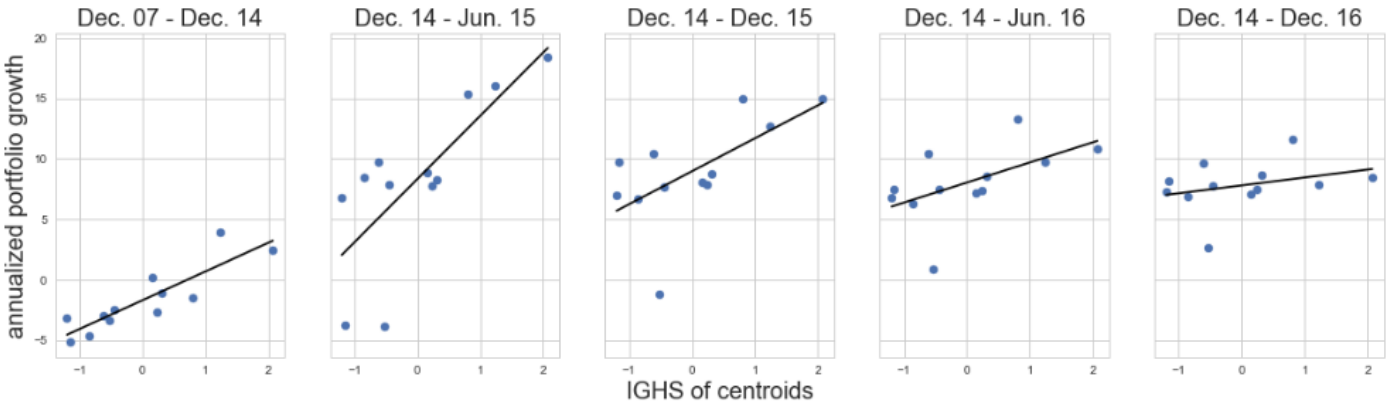

Figure 1.11 *Scatter plots of IGHS against annualized portfolio growth of each cluster.*

Figure 1.11 illustrates perhaps the most important lesson of this study: income data bears a strong relationship to real estate value data when measured over the same time period, but the relationship greatly weakens when projecting into the future. It seems that to reliably measure new changes in real estate value, we require new income data.

Note: not only do the regression coefficients approach 0 as we project further into the future (suggesting a weakening relationship), the model's goodness of fit greatly deteriorates when measuring even just 6 months into the future. That is, the points on the leftmost plot are much more tightly spread over the best-fit line compared to the right four plots. These results suggest our model can estimate relative growth prospects for only a short period into the future, given several years of income data (although, to a limited degree of accuracy).

Below, we show a related visualization of corresponding statistical metrics between the IGHS and annualized portfolio return for each time period. The 2 tailed p-value represents the probability of an uncorrelated data set achieving more extreme levels of correlation than the ones calculated.

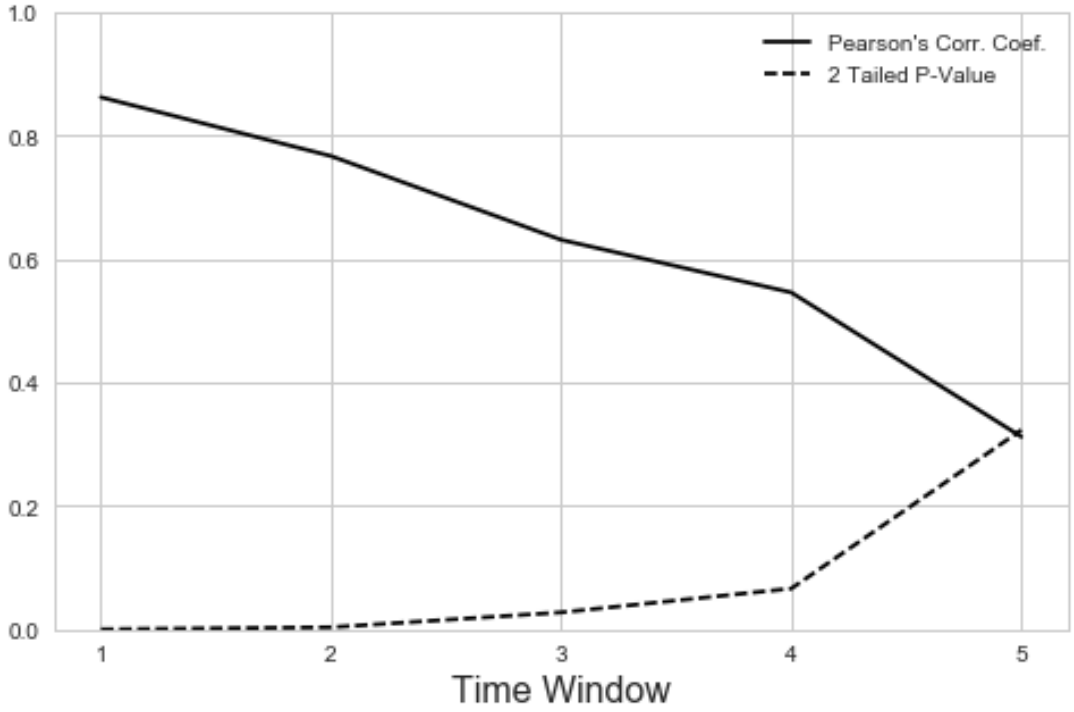

Figure 1.12 *Time windows 1 – 5 correspond to the windows in Figure 1.11, respectively.*

It seems our system would be best suited for the short-term real estate investor, since the decreasing correlation coefficients and increasing p-values suggest a weakening relationship as we measure changes further into the future. For short periods of time following the income data, however, we have been able to measure differences in real estate value growth—differences which can aid in choosing better investments.

Lastly, we re-run the analysis of this sub-section but apply all metrics to individual regions rather than entire clusters. Below are the corresponding visualizations for similar interpretation.

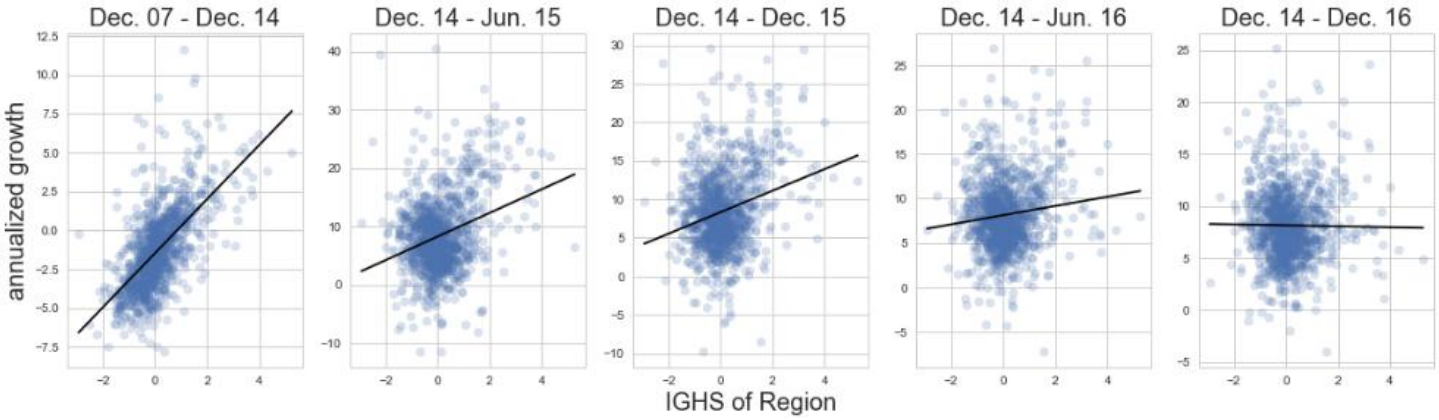

Figure 1.13 *Scatter plots of IGHS against annualized median growth of each region.*

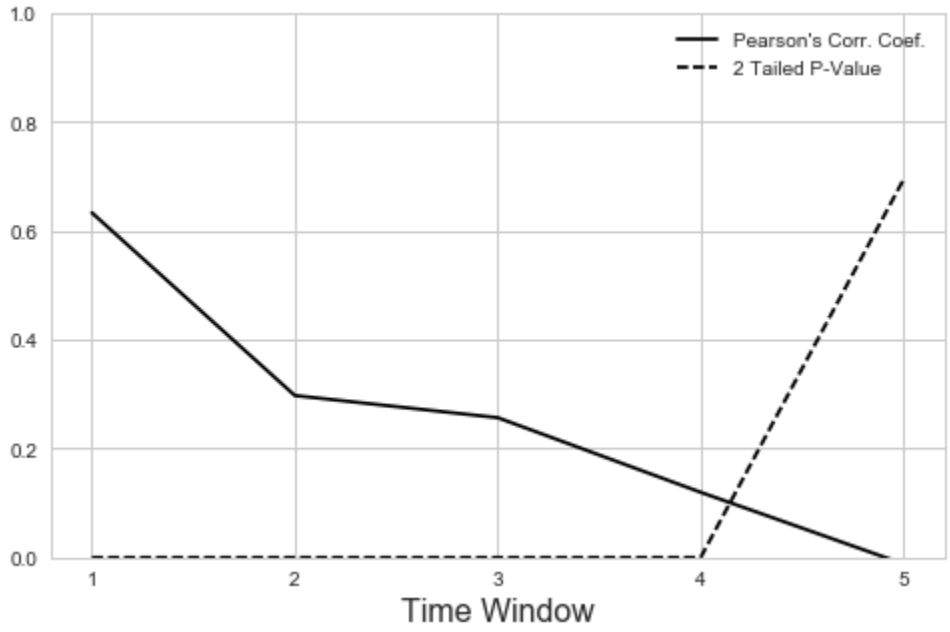

Figure 1.14 *Notice the more erratic behavior in our statistical metrics when we measure over individual regions rather than clusters.*

1.4 Summary

Ultimately, the output of this system—the income growth health score (IGHS)—is almost trivial to calculate after some preliminary data processing. The focus of this section, however, was the analysis used to develop the metric. That is, the goal of this section was to show how to develop a novel metric, given some objective. In this case, the objective was to show a relationship between income data and real estate value data over varying time periods. In retrospect, the data set behind this system is very simple. We use algorithmic manipulations to show a relationship between a time series of income distributions and a time series of real estate value, made possible because of shared keys.

We focus on measuring changes over time periods, but we could also consider static information from points in time. For example, a region with a high proportion of earners in the lowest bracket and a big increase in the second lowest bracket could indicate good growth. With our current system, increases in the second lowest bracket are not interpreted as good indicators of growth.

A big limitation of this system is in our data. Our income data is separated into five brackets, but a hypothetical data set with many more income brackets (or even a "perfect" data set with incomes listed for every individual) would allow us to model income growth much more accurately. The analysis we performed in developing the IGHS could also be greatly improved with bigger data sets to test over. Additionally, our ZHVI data only lists median values. A more complete data set would allow deeper statistical analysis and better measurements.

§ 2: VALUATION

2.1: Overview

Our section on growth was focused on exploratory analysis to develop a novel metric, with outputs of the system (e.g. income growth health scores: IGHS) resulting as products of our analysis. In this section, we focus not on developing novel methods through analysis of data, but rather on customizing implementations of established methods. In particular, we develop a system to build regression valuation models that are specific to each input property.

By specifying a set of characteristics, we filter through the main dataset to construct a customized training set—a '*comparable'* set. From this subset of the main dataset, we build and optimize valuation models. Below, we have an overview of our envisioned system.

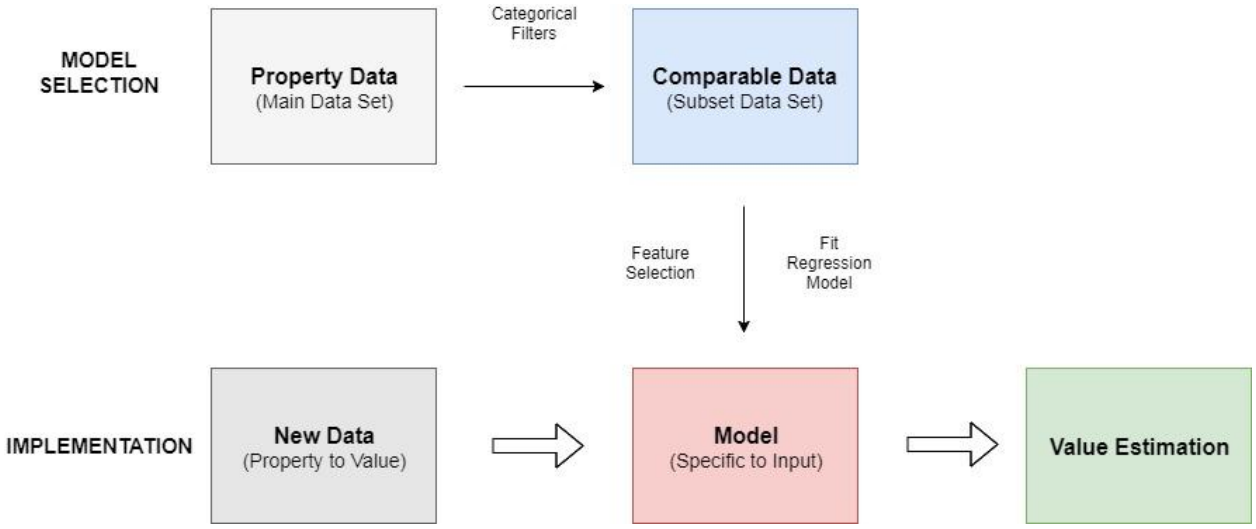

Figure 2.1 *Overview of the system.*

2.1.1: Background

One of the fundamental components of investment analysis is the study of value and market price. Put simply, we want to search for opportunities in the market which maximize value for a minimal price.

In this section, we develop a system to value real estate properties using the *comparables ('comps')* approach and test the valuation against the market price. The *comps* approach seeks to value properties by using established valuations for properties very similar to the one in question. That is, for a given property with categorical characteristics, we seek properties with characteristics as similar as possible.

2.1.2: Data

We use a data set representing about 3 million properties in southern California released by Zillow[3], with a non-uniform distribution of property types. Thus, as we shall see, the strength and complexity of our models will depend on each input property and the corresponding availability of relevant training data.

Features of each data point can be grouped into two subsets: categorical information and numerical information. Categorical information will be first used to filter our main data set into an input-specific training set, and numerical information will be subsequently used to fit the regression models.

2.2: Model Selection

The first component of our valuation system will reduce our main data set to a subset of data that is categorically similar to the property we are valuing. That is, we automate the development of the *comparables* appraisal framework ("comps") by re-building our models from data that is relevant to each individual property. Although we wish to build models using data as similar as possible to the property in question, we also want to build the models off of sufficient data points. With too few comps, there is a high probability of building a skewed and inaccurate model.

Following the selection of comparable properties, we subsequently select a subset of numerically-valued features to build regression models from. In different regions and with different types of properties, certain features may affect the value of real estate where they otherwise would not. We automate the process of discerning these features so that our models depend on the most relevant variables.

2.2.1: Categorical Features

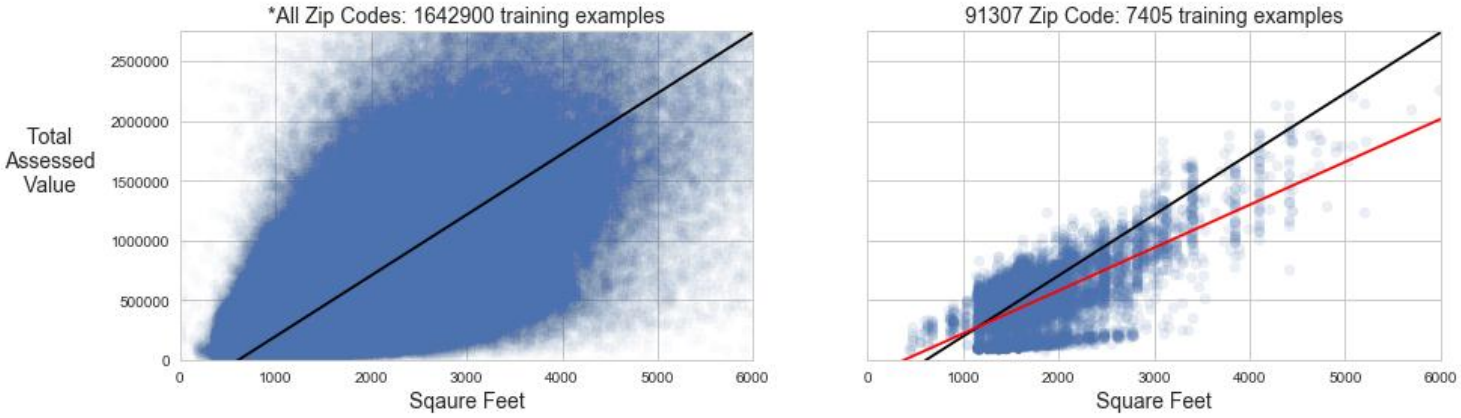

Figure 2.1 *Comparison of models: the left plot uses the entire data set; the right plot uses only those with the zip code 91307. The filter causes the best fit line to adjust from black to red.*

As an example, we show in Figure 2.1 the much higher variance observed in the relationship between square footage of properties and their predicted assessed value, when we do not specify a zip code. A linear assumption almost seems inappropriate in the general case (at the very least, it is unreliable) However, in the location-specific plot, linear regression appears to be a reasonable model (at least from a visual standpoint).

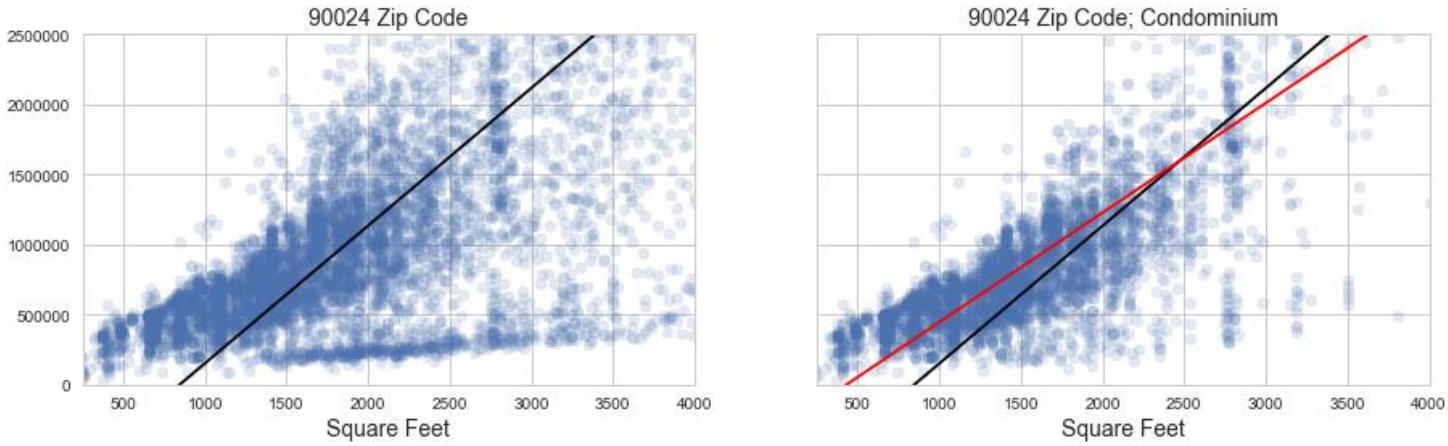

Figure 2.2 *Illustration of model adjustment due to a second categorical filter.*

When we add additional filtering criteria, our fitted line is again adjusted from the black line to the red line. (In particular, notice that the cluster of data points near the bottom of the left-hand graph disappears from the right-hand graph. It seems that this was a group of properties of a different class which were skewing the valuation model for condominiums.) Since the linear model on the right is built from training examples which have more similar characteristics to our hypothetical input, it is reasonable to assume that it will do a better job in valuing the property.

2.2.2: 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
```

Figure 2.3 Correlations with total assessed value: all data (left) vs. 91304 region (right).

As an example, notice how correlations with assessed value change when we specify location. Not only are many of the correlations more significant, certain variables now play a role in valuation due to location-specific behavior. In particular, notice in the

91304 correlations vector that longitude has a significant correlation with assessed value. In the 91304 region, a more negative longitude (farther eastward) corresponds to a property higher up in the hills. Resultantly, the valuation tends to be higher.

For our application, we use a minimum value of 0.2 of the magnitude in the correlation to total assessed value of a feature as a cutoff for whether it is included in the regression model. This is to avoid unnecessary model complexity and to remove features which may otherwise create noise in predicting the value.

2.3: Total Assessed Value

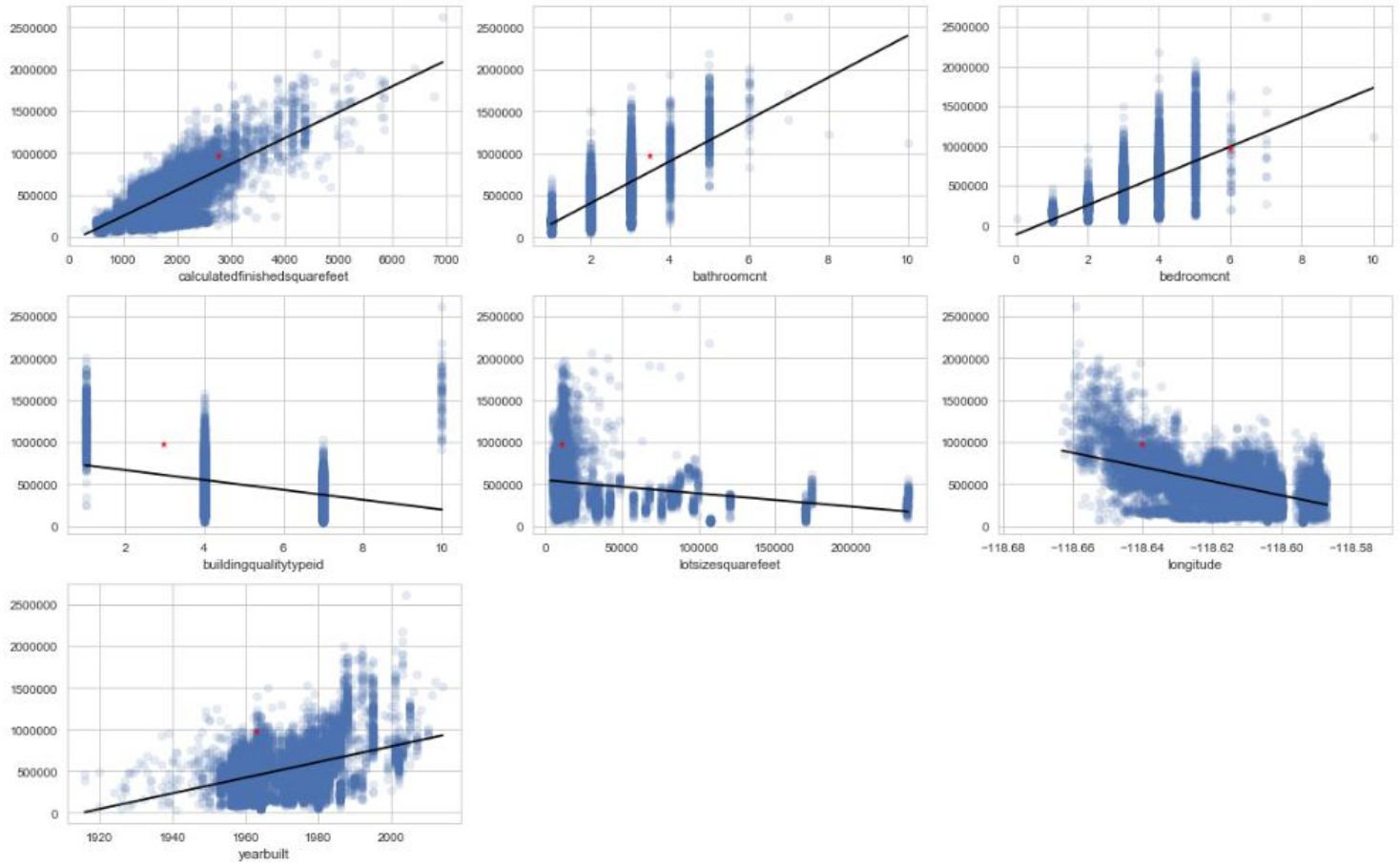

Figure 2.3 *Hypothetical property valuation (in red) shown relative to the values of properties used in its model—for each of the features used.*

We can plot the listed price against the predicted value and see how much of a bargain we are getting with respect to each characteristic of the house. For example, this plot shows that the list price of 969000 is fairly high with respect the averages for most of the features it has. However, for a house with 6 bedrooms, it seems the list price is just about average.

2.4: Summary

Real estate valuation is not straight forward and requires a significant amount of human deliberation, but we have seen that it is possible to automate much of the grunt work necessary to arrive at an estimate. With sufficiently rich data sets and reliable models, it is conceivable to develop a system which automates much more of the process and arrives at an estimate to a greater level of accuracy.

§ 3: EVALUATION

We have developed a test set representing 20 properties listed on Zillow's website in May 2017. For each property, we extracted a collection of categorical information and features which can be input to our investment applications. This includes each property's list price, address, square footage and bed/bath count—among others. We have manually transformed the data from Zillow's webpage to a format suited for our applications (e.g. latitude/longitude, unit conversions, etc.) but this step would not be difficult to automate.

There exist practical limitations in the data used to develop our applications. First, the timelines of the growth section and valuation section do not directly align. Thus, we observe how our applications could *hypothetically* be used in combination--assuming that the growth section was derived from income data during the time period May 2010 - May 2017.

Additionally, the growth data outputs an assessed value for December 2015. We adjust each property's output value to a May 2017 value by the % growth of the median ZHVI in their respective regions over the time period. Lastly, we adjust the assessed value to a list price to account for brokerage fees, legal fees, etc. by taking ratios of list price medians to assessed value medians in respective regions. This is a major weak point in our model that overly simplifies the pricing process and would not be suggested in practice.

There are several approaches we can take to evaluate our applications, but we illustrate just one approach that can be generalized to other test sets. We skip a growth evaluation of our test set, since that is equivalent to picking out a subset of the tests we already performed in section 1.3. We study the valuation section by developing confidence intervals for each suggested price and testing them relative to the actual list prices. Of our 20 test list prices, 14 fell into their respective 68% confidence intervals—suggesting that our model is indeed performing to the level of accuracy implied by the statistical parameters.

We note that there is a property in our test set with a list price 28.8 standard deviations from its suggested price. This shows that although our models work reasonably well for most properties, there may be outliers for which our applications do not perform well. When we do not consider this property, the test set has an average magnitude of standard deviations from the suggested price of 0.72.

Out of the 20 properties, only 2 exist in regions which we measure as having relatively strong growth prospects (IGHS greater than 0) on top of having bargain list prices (list price below our model's suggested price). Thus, from the perspective of a client filtering through potential properties to invest in, we have achieved a 90% reduction in our test set size when we use these criteria for supporting investment decisions.

More details on the test set's evaluation can be found in the presentation and project notebooks.

§ 4: CONCLUSION

Not only does investment research require analysis of many different economic elements, there also exist many valid ways to approach the analyses. We have explored in this report just two systems for analyzing two different economic elements in a reproducible and generalizable way. Specifically, our applications have been built to serve the short term real estate investor—for instance, investors working in the 'fix and flip' industry, who purchase debilitated properties, rehabilitate them, and sell them for a profit.

For simplicity of illustration of how an investor would utilize our applications, we think of the outputs of each system as being binary: good growth vs. bad growth and underpriced vs. overpriced. (In reality, the outputs lie on a spectrum—IGHS of a property as standard deviations from the mean and list price of a property as standard deviation from the estimated value.) Then, there exist four possible outputs of the combined system. An underpriced property in a region with good growth prospects is an attractive investment. On the contrary, an overpriced property in a region with bad growth prospects is an unattractive investment. The other two possible combinations are ambiguous—the investment decision will depend on other factors. As mentioned, in practice we would want a system which analyzes many different economic elements to support the client's decision making.

The key points to creating reliable systems are that the systems use verified metrics for the different economic elements we want to measure and that the systems are built upon sufficient data pipelines. The systems we have developed in this project are built upon many assumptions that have not been rigorously tested and must be further refined and tested to be used in practice. With the proper resources and guidance, however, a researcher could develop much more sophisticated systems.

Lastly, we note that the applications we have developed in this report have automated lower level analyses for a human client to use in developing higher level analyses, but with sufficient data and models it is feasible to automate the higher level analyses as well. That is, the 'market analysis' component in Figure 3.1 is currently assumed to be undertaken by a human analyst, but is potentially replaceable with data and algorithms. Organizations operate with many intertwined components like in Figure 3.1, and the ongoing data revolution will cause an overtake of lower level human analysts with automated applications. This overtake, however, will be possible only through proper data engineering and intelligent systems development.

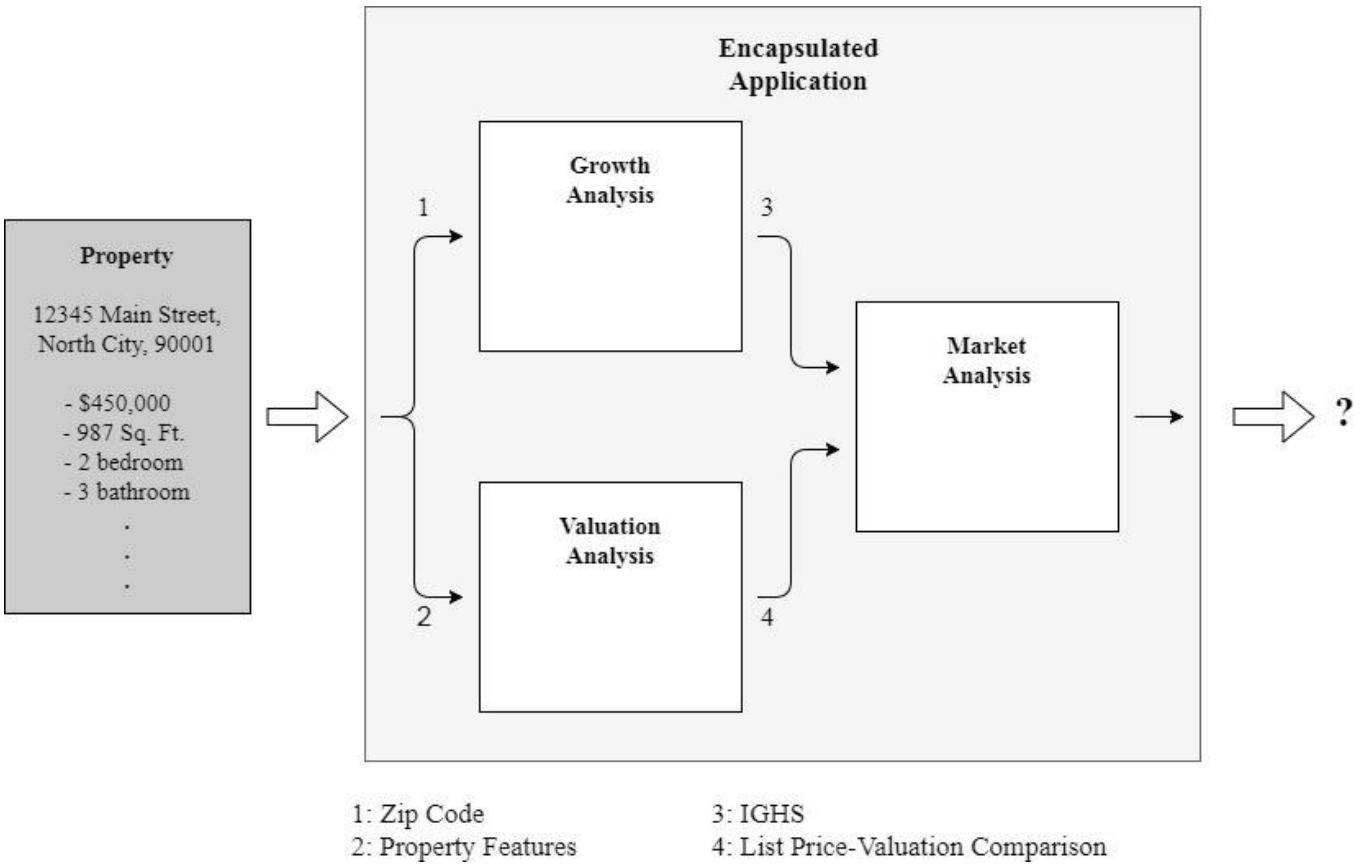

Figure 3.1 *A simplified application to automate more levels of analysis. There is room for creativity and flexibility in the output, according to the needs of the client.*

[1] https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi

[2] https://www.zillow.com/research/data/

[3] https://www.kaggle.com/c/zillow-prize-1/data