
# *Regression Analysis: Real Estate Comparison Between Two Neighborhoods in Oregon and Texas*

Course Title: **Research Methods**
Course Number: **ETM565/665**
Instructor: **Dr.Anderson**
Term: **Winter**
Year: **2013**
Author(s): **Neil Brown**

ETM OFFICE USE ONLY
Report No.:
Type: Student Project
Note:


# Regression Analysis: Real Estate Comparison Between Two Neighborhoods in Oregon and Texas

ETM 565: Research Methods
Term: Winter 2013
Instructor: Tim Anderson
Student: Neil Brown

## 1.0 Introduction

When living and working in Texas on two occasions over the past 10-years, the author couldn't help notice the differences in residential real estate. Although only renting apartments and living in hotels during his time, the author found time and interest to look into house prices and talk with friends who lived there permanently about the subject of residential real estate. During these two occasions (2003 and 2011) the business environment was such that interesting and challenging construction projects were plentiful in Texas and not as much so in Oregon. So as a commercial/industrial construction manager who enjoys studying residential real estate, there were interesting observations to be made. However, no actual assessments were made, nor data collection or quantifiable research. So looking back on those experiences and with an opportunity to perform current research methods, the author decided on two major differences that were observed and could be studied:

1. Affordability of residential real estate in Texas versus in Oregon
2. Price volatility during national economic cycles seemed less prevalent in Texas versus in Oregon

Research Question

Regression analysis of residential real estate should validate (through quantifiable data) that Texas is more affordable than Oregon and that it has also performed with less volatility during national economic cycles. The analysis was performed in excel with best-fit polynomial regression trending.

## 2.0 Neighborhoods and Data Sources

Comparison of Neighborhoods

The author lives in the Cedar Mill neighborhood of NW Portland, Oregon 97229. And has previously lived in Houston and Beaumont, Texas but limited research data was available there, so he selected the Richardson neighborhood of Dallas, TX 75080. Past colleagues had lived nearby and the author was considering working for a company in North Dallas.

Data Source

Zillow has become a popular site for many home owners, buyers and sellers. The data is available for open analysis on the web in their "real estate research" site [1]. S&P/Case-Shiller is a well-known indicator of residential real estate and is studied and referenced by many. The author felt it would be best to analysis raw data such as what Zillow makes available, plus the highly integrated index from S&P/Case-Shiller as "Home Price Indices [that] are a consistent benchmark of housing prices in the United States" [2]. The indices measure changes in housing market prices given a constant level of quality. Changes in the types and sizes of houses or changes in the physical characteristics of houses are specifically excluded from the calculations to avoid incorrectly affecting the index value.

## 3.0 Data and Analysis

There are several criteria available to choose from on residential real estate, and practically every objective thing is tracked through title with local municipalities. The author selected two measurements that he believed to be informative and relatively normalized.

- Median Sold $/SF
- % of Homes Selling for a Loss

Median Sold $/SF Comparison: Cedar Mill, OR

The author initially trended a 4th level polynomial for best fit of the data, because the 3rd level polynomial did not trend the final years of the data very well. This trend line obtained an R^2 = 0.9086. See below Figure 1 for the scatter plot of data, the trend line and the equation fitted to the line plus the R^2 value.

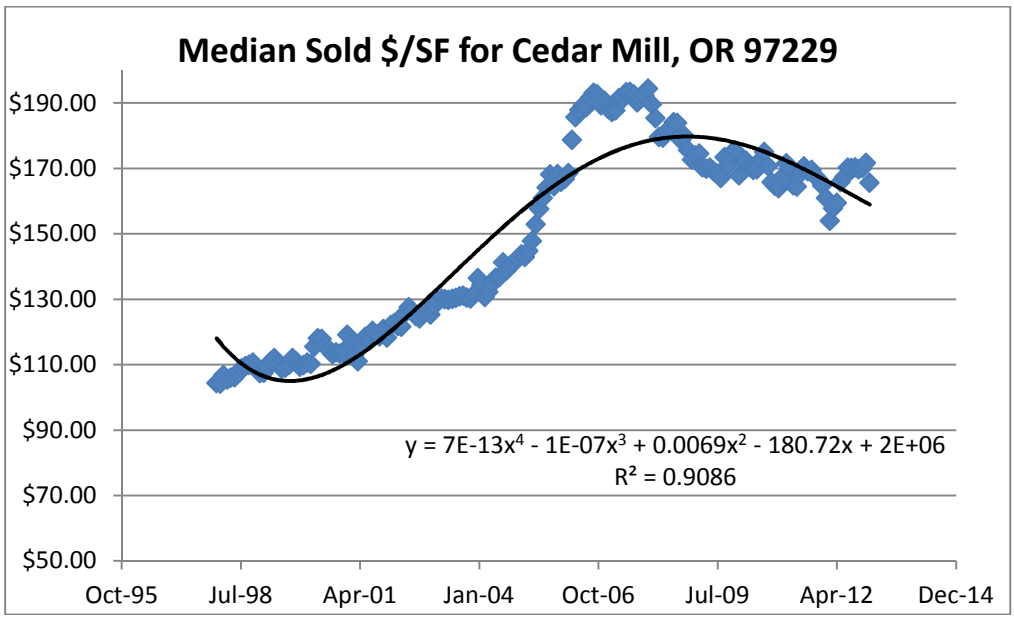


*Figure 1.*

But after presenting this data to the class of peers and his Professor, the author decided to revisit the complexity of data and determine if a lower level polynomial could be accepted. Taking the complexity of the trend from 4th level to 2nd level polynomial still left the data fitted to an R^2 value of 0.8098. So it is with this feedback from the earlier presentation that the author changes his proposed fitted trend line to the below chart and results in Figure 2.

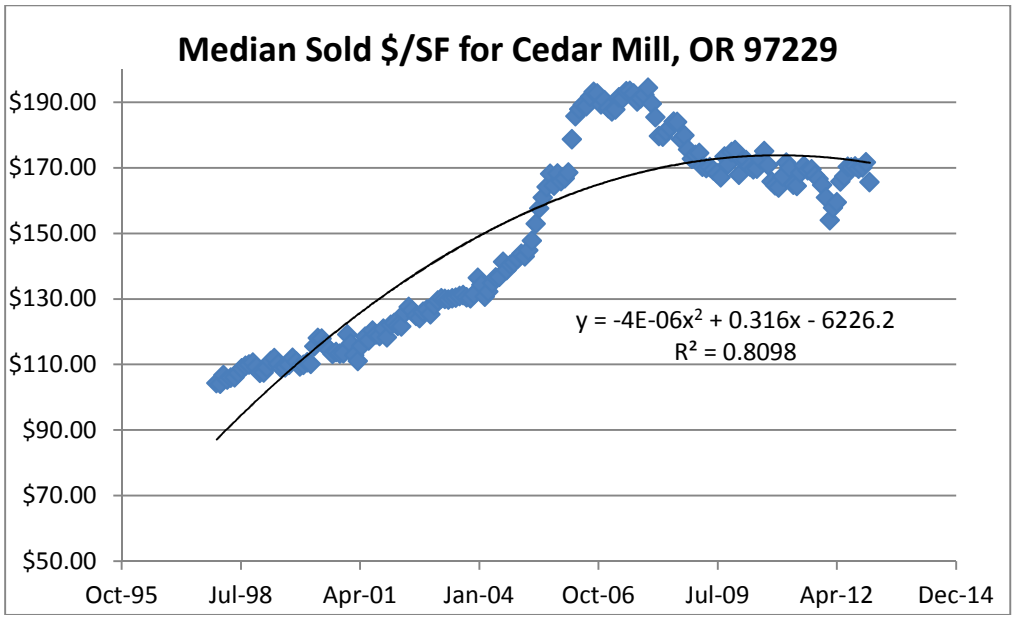


*Figure 2.*

Median Sold $/SF Comparison: Richardson, TX

A 2nd level polynomial for Richardson produced a higher R^2 value than the 4th level in Cedar Mill, R^2 = 0.9323. This is shown in the below figure with a common axis to what we observed for the Cedar Mill data. It is a very high R^2 value so we can say with certainty that the scatter plot and trend line suggests a less volatile group of data. It also is well behaved in the lower portion of the chart and therefore quantifies the better affordability of residential real estate in the selected neighborhood for Texas versus Oregon. See Figure 3 below.

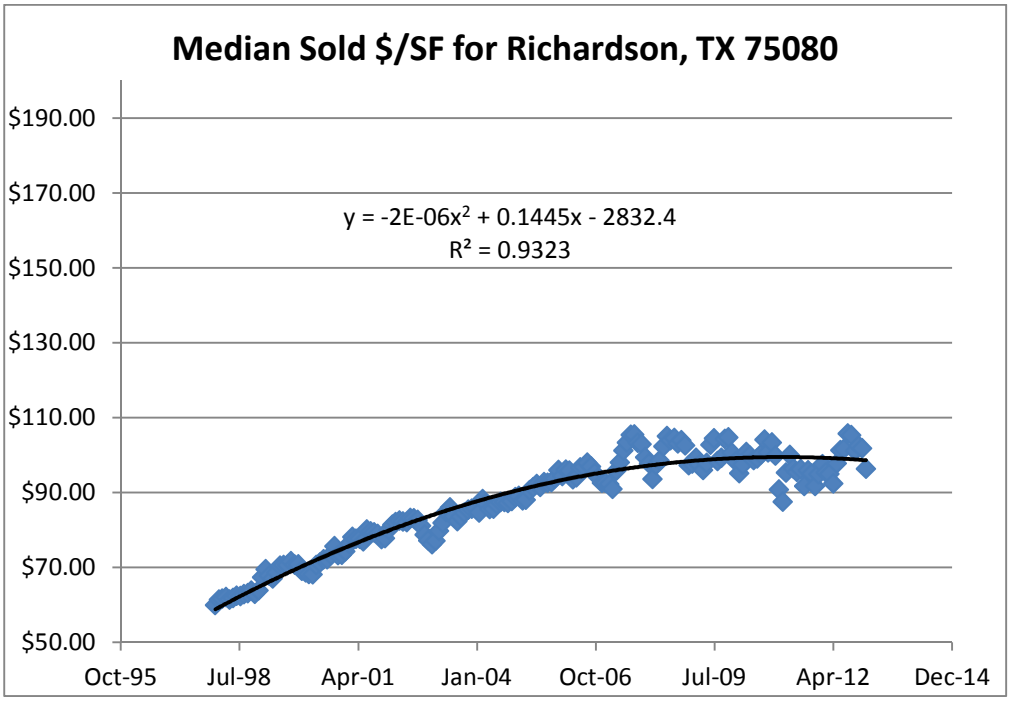

*Figure 3.*

Again, after peer review during the class presentation of this data, the author wanted to see if standard linear regression could fit the data group within a reasonable level of confidence. Below figure 4 shows that linear regression provided a decent R^2 = 0.8314. Therefore, this opportunity to have a good R^2 value, plus the most simple trend line available, means this first level linear basic should be acceptable for the data group. See Figure 4 below.

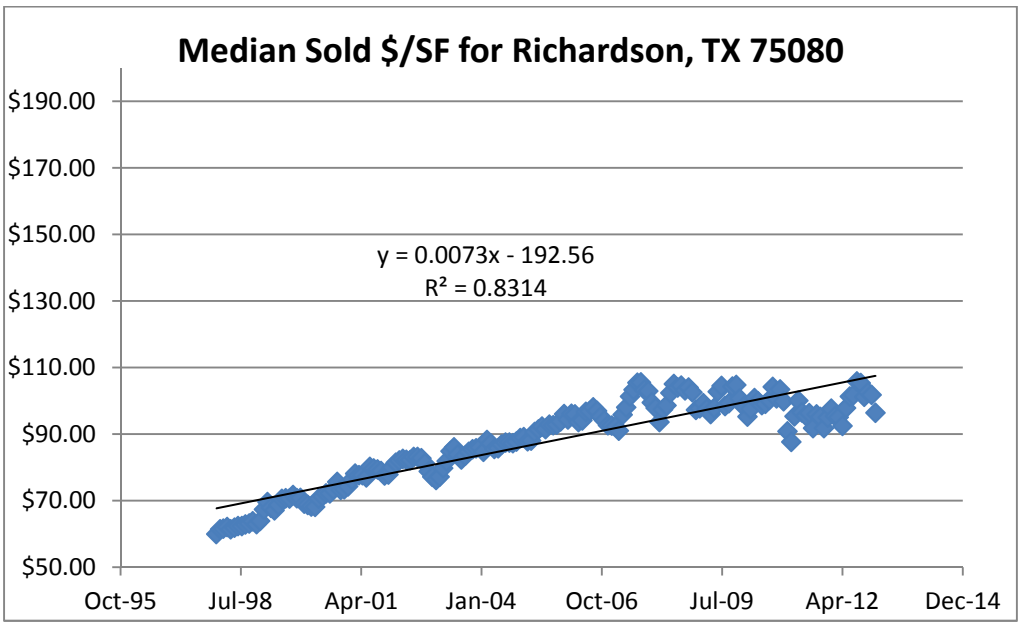


*Figure 4.*

For the next section of data analysis, the author quantifies even more interesting findings when studying the percentage of homes selling at a loss (% Homes Selling for Loss Comparison).

<u>% Homes Selling for Loss Comparison: Cedar Mill, OR</u>

The author fitted a trend line at a 3rd level polynomial for best fit of this explosive data and received an R^2 = 0.9285 as shown in Figure 5 below. If we scale back the ambition and fit a 2nd level then we receive an R^2 = 0.8855, so still quite high. See Figure 6 for 3rd level. During the presentation of Figure 5 to the class, we spent time looking at the data around the 2008 timeframe. It really begins to tell a story about the residential real estate market that things are "shifting" unfortunately, for the worse. We discussed the "what ifs" that if someone was tracking this data with an eye for investment how they could short the market and bet again the residential real estate in our local region based on regression analysis. By mid 2009, it becomes quite clear that the trend is not just continuing but even increasing in slope. The percentage of homes selling for a loss in Cedar Mill, OR goes up and up, even hitting more than 35% at one point in time. With the story that this trend line tells us (most importantly along the X-axis of time), and how well it fits overall at 0.9285, the author would in this case select the slightly more complex 3rd level polynomial. Figures 5 (preferred) and 6 are below.

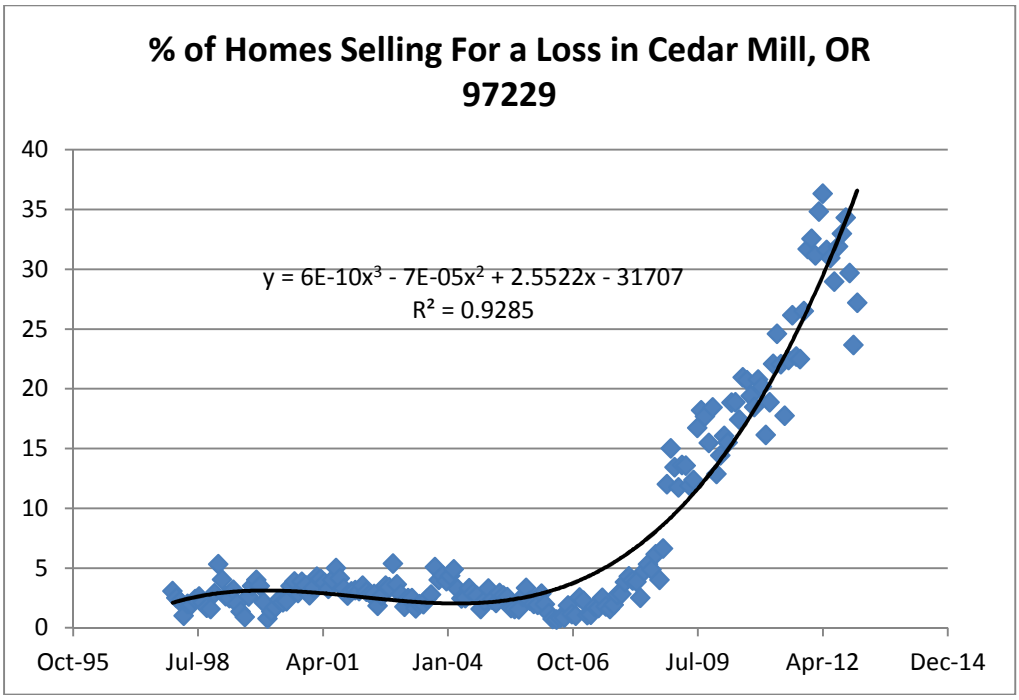


*Figure 5.*

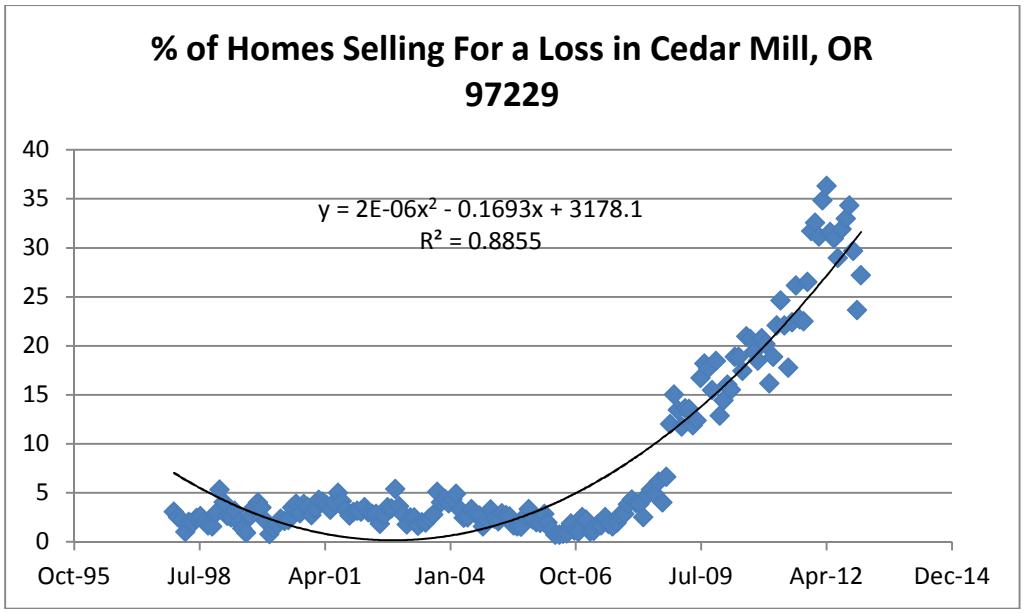


*Figure 6.*

<u>% Homes Selling for Loss Comparison: Richardson, TX</u>

Interestingly, a 4$^{th}$ level polynomial was needed for Richardson's percentages of homes selling at loss to obtain a good fit. There is a steady 10% rolling spread in the cluster of data over the past 10-years, through moderate ups/downs. The author only obtained a value for R^2 = 0.8002 for this trend. So while the percentages of homes selling for a loss doesn't get as high as Cedar Mill, OR, nor the trend line as steep, it retains some complexity. See Figure 7 for data and results.

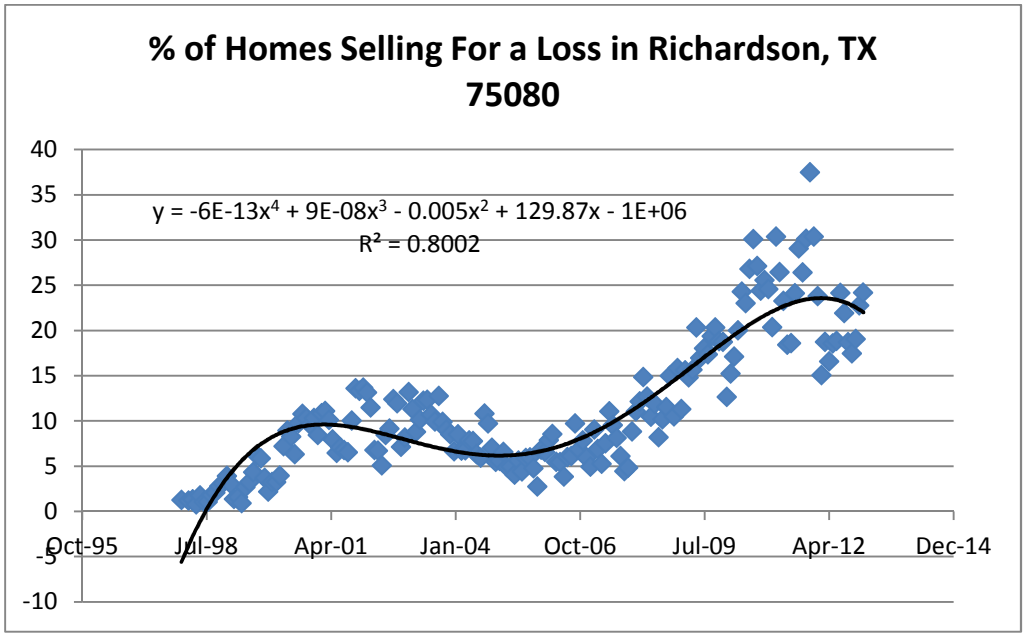


*Figure 7.*

By reducing the complexity of the polynomial, the trend line becomes less accurate (as anticipated). Our visual analysis shows that the standard deviation is high, much as it was in the previous figure, and that the data just doesn't seem to fit as well as the previous data for Median $/SF at both locations and the % selling for loss in Cedar Mill, OR. Our Professor often reminds us that a great statistician once said, "All model are wrong, some are useful" (generally attributed to George E. P. Box). Well, the below model isn't useful. At least it's wrong, so it has that going for it.

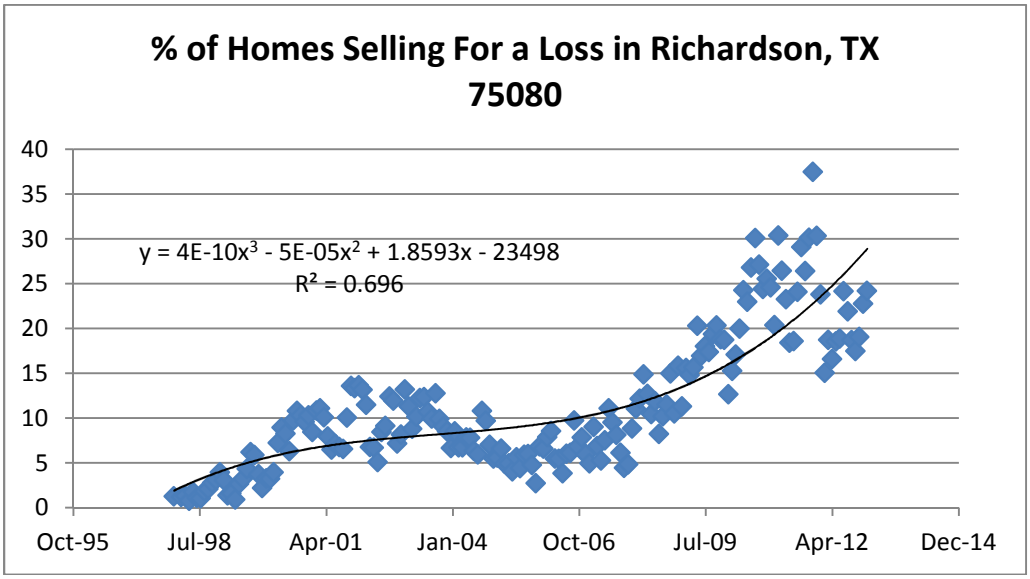


*Figure 8.*

Here is an even more basic trend line for this challenging data group. The 2nd level gets an R^2 = 0.6637. In review of each of these 3-options for a fitted trend line, the author selects this one because it has less complexity and the data doesn't support much in the way of a better match.

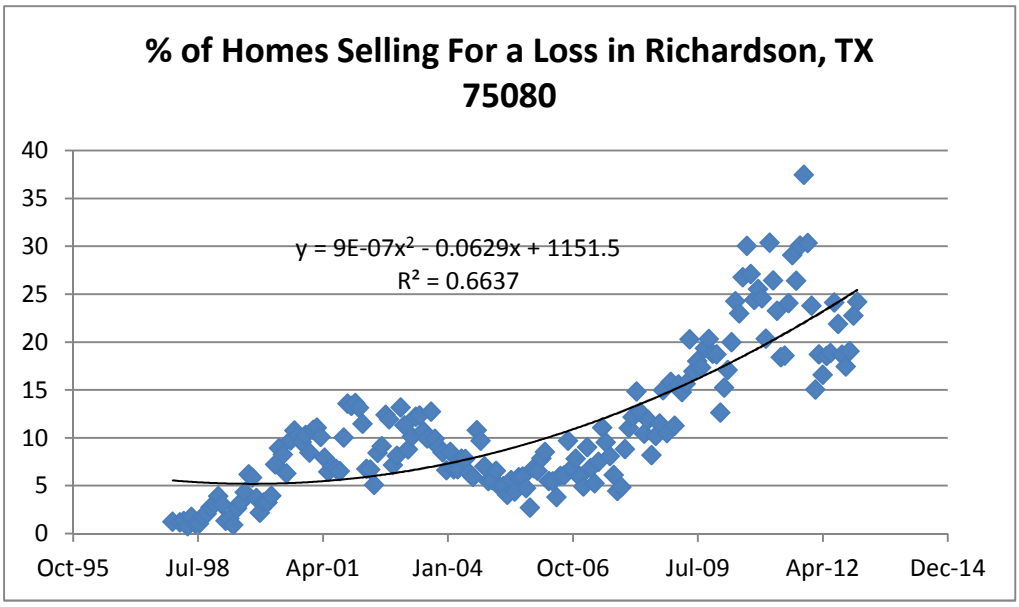


*Figure 9.*

Case-Shiller Home Price Index

Case-Shiller makes corrections for inflation and takes into account rolling averages of home prices. And the fact that it is an index, its units are easily transferable. These points, plus its popularity and use in the financial and other markets made it a must-have in this regression analysis exercise on residential real estate. Case-Shiller only tracks 20-metropolitan areas so the author had to concede on the localized neighborhood study from before and roll up to the city level. This was not anticipated to alter the results for this comparison analysis given its relatively general nature for quantifiable research, and the less technical excel charting functions versus other, more powerful analysis tools.

At the complex 5th level polynomial regression trend line, Portland gets its up and down (read: volatile) data fitted to an R^2 = 0.9547 in the below Figure 10. This is ambitious, no doubt. So the author re-ran the experiment at a 3rd level and still received an acceptable value, R^2 = 0.838. After further review, the interest in the 5th level trend is because of the tightness of fit in the last 1-year of data. By simply being (generally) aware of the residential real estate market in our area, it seems the market is in a slight recovery. The 5th level captured that point very well and so it was presented in class that way. The author acknowledges the ambition may have gotten in the way of the best, most balanced trend line for the data group. See Figure 10 below, with a special focus on the last year or two.

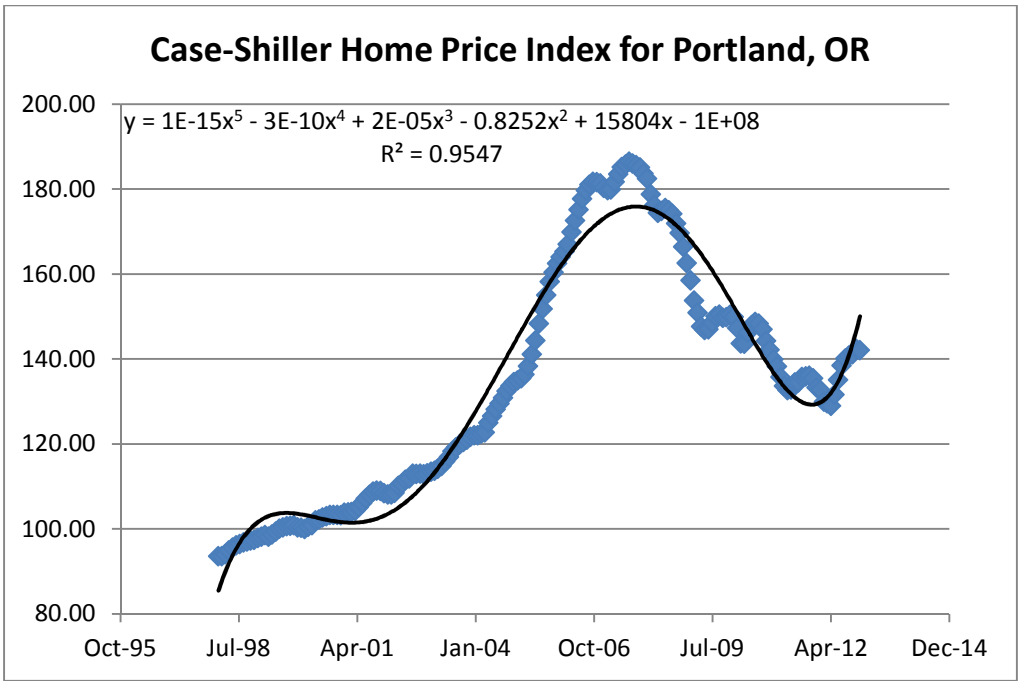


*Figure 10.*

Below is the lower, 3rd level polynomial regression for Portland.

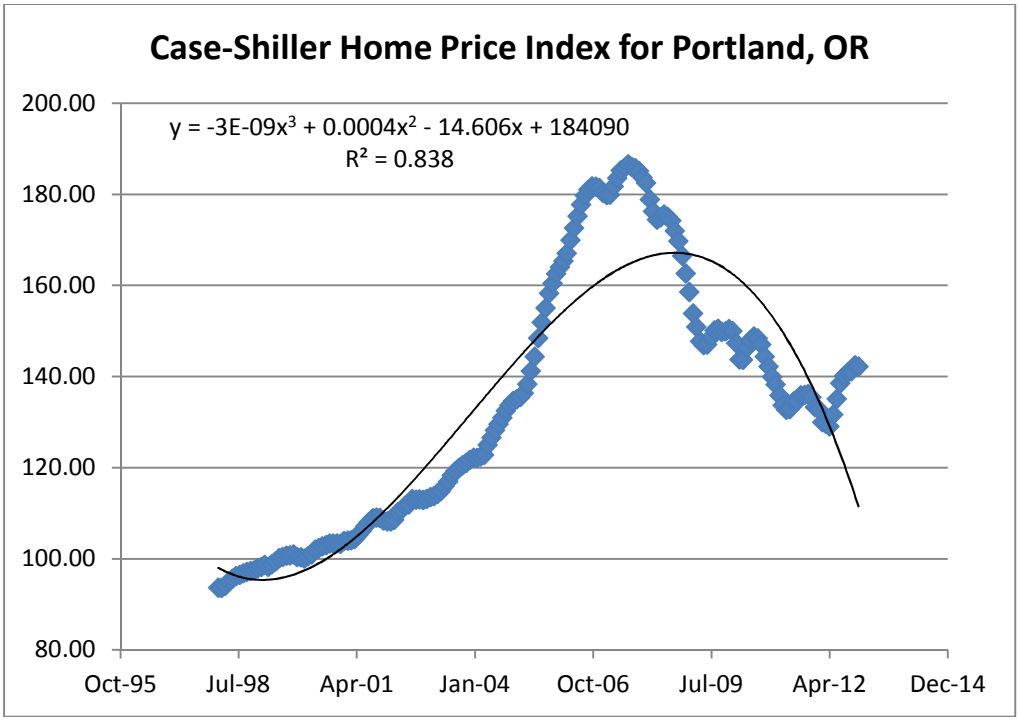


*Figure 11.*

When studying the Case-Shiller home price index for Dallas, the author was careful to use a common axis as was used for Portland, plus fit a similar level polynomial to compare R^2 values. But upon observing the simplistic data group in the scatter plot, it's hardly reasonable to assume we need a 5th level polynomial. Nonetheless, the R^2 was a good, acceptable 0.8498, see Figure 12 below.

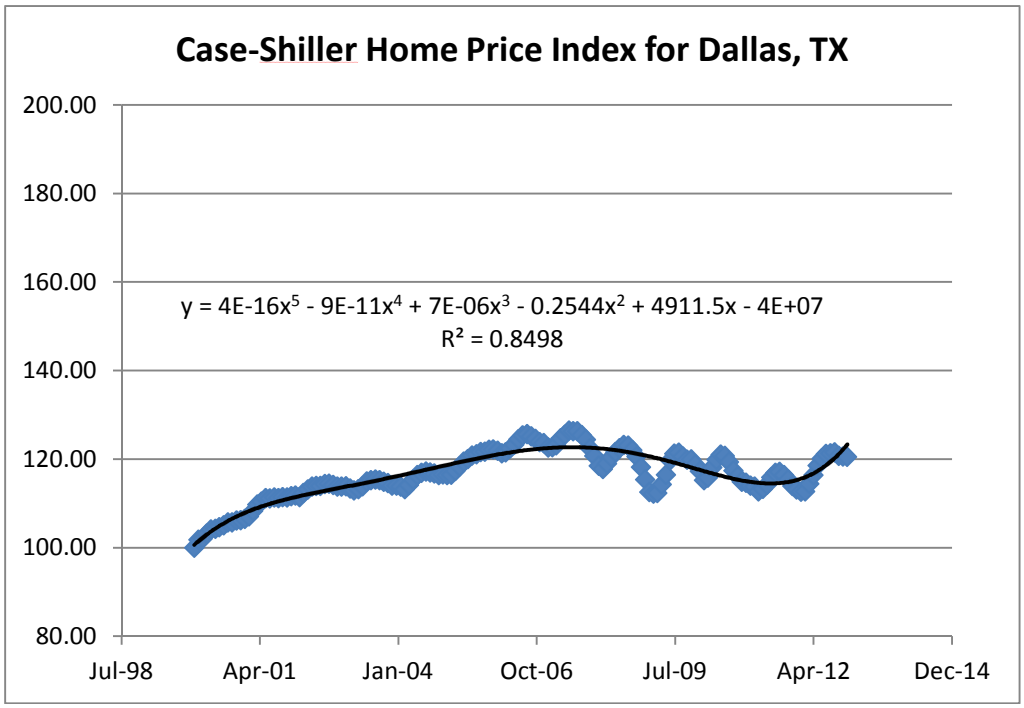


*Figure 12.*

In hindsight, the below 2nd level trend still may be too much complexity, but it does capture the slight up and slight down of the trend well enough to be justified. And with an R^2 = 0.7279, its just fine to be an acceptable fit for the data group. See Figure 13 below.

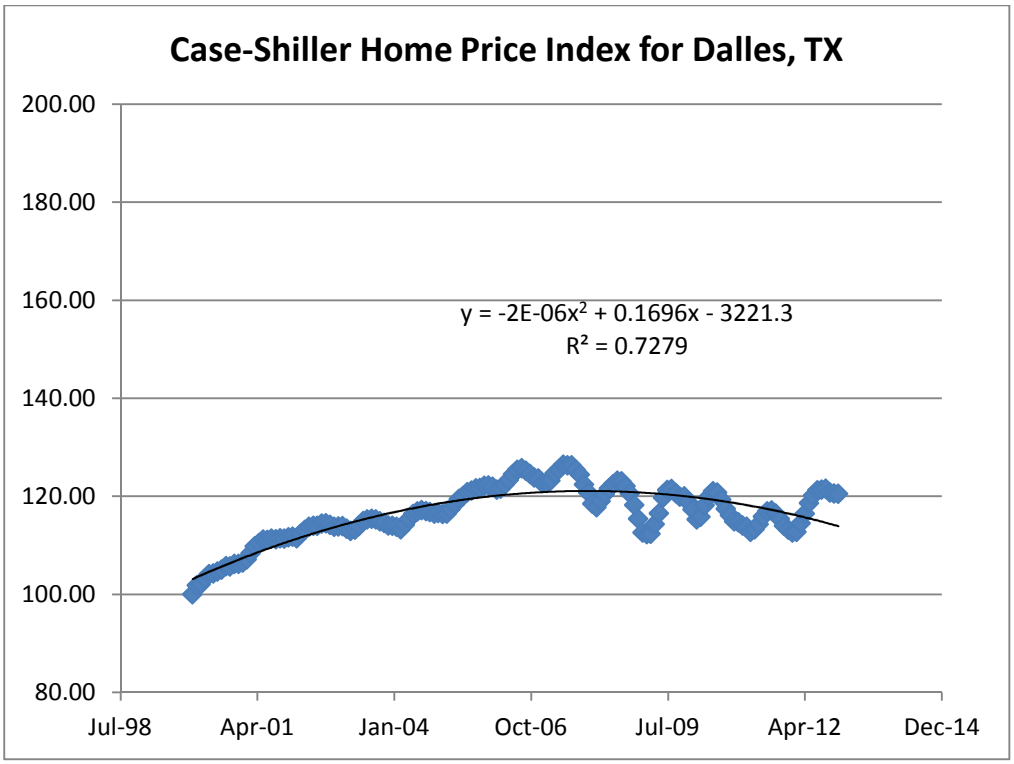


*Figure 13.*

## 4.0 Conclusion and Recommendations

Summary of Findings

All of the above figures and their relevant data are summarized below for review alongside each other. Initially, the author was trying to achieve a high value for R^2 but the complexity of the fitted trend line grew and grew. But with the review and new analysis at lower level polynomials, and associated R^2 values seem to be just fine. The below highlighted trends show the recommended polynomial level and their associated R^2 value for each of the data groups that was analyzed.

| Data Group | Trend Line | R^2 Value |
|---|---|---|
| Median Sold $/SF – 97229 | 4th Level | 0.9086 |
| **Median Sold $/SF – 97229** | **2nd Level** | **0.8098** |
| Median Sold $/SF – 75080 | 2nd Level | 0.9323 |
| **Median Sold $/SF – 75080** | **1st Level** | **0.8314** |
| **% Selling for Loss – 97229** | **3rd Level** | **0.9285** |
| % Selling for Loss – 97229 | 2nd Level | 0.8855 |
| % Selling for Loss – 75080 | 4th Level | 0.8002 |
| % Selling for Loss – 75080 | 3rd Level | 0.696 |
| **% Selling for Loss – 75080** | **2nd Level** | **0.6637** |
| Case-Shiller Index, Portland, OR | 5th Level | 0.9547 |
| **Case-Shiller Index, Portland, OR** | **3rd Level** | **0.838** |
| Case-Shiller Index, Dallas, TX | 5th Level | 0.8498 |
| **Case-Shiller Index, Dallas, TX** | **2nd Level** | **0.7279** |

*Table 1.*

The data in the figures helps us quantify the cost of living as less in Richardson by a factor of approximately 1.7, and it has been for the life of the data. Plus, the Richardson data was generally less volatile than the Cedar Mill data. The author defines less volatility as being set in tighter clusters and having less overall directional shifts. Both of these two findings from the data analysis help validate the research question, as repeated below:

> *Research Question*
> *Regression analysis of residential real estate should validate (through quantifiable data) that Texas is more affordable than Oregon and that it has also performed with less volatility during national economic cycles. The analysis was performed in excel with best-fit polynomial regression trending.*

The Case-Shiller data suggests that with a similar 5th level polynomial regression fitted to each neighborhood, the subtleties of the Richardson data shifts could not be as accurately described as the large, sweeping changes in Cedar Mill. Taking the complexity of the fitted trend line down, we see that Portland is still better defined by the curve. The Portland Case-Shiller index spanned between 95 to 185 points. Literally, it more than doubled its value from the beginning of the measurements. Plus, then it trended back down to hover around the midpoint of 140 points. Dallas only ranges from a value between 100 and approximately 125 during the life of the home price index. Clearly, the Case-Shiller index has documented that residential real estate is less volatile in Dallas, TX then in Portland, OR. The data suggests exactly that.

Future Research

This research could be extrapolated from residential real estate into other economic indicators. While the author enjoys the basics of studying residential real estate and believes it is a contributing factor to the over health of our economy, it certainly is not the only indicator. As such, it might be interesting to know if (and how) other indicators compare in a city to city analysis. So for this example, a researcher could continue the greater Portland, OR versus Dallas, TX study. And extrapolate the residential real estate data into other sectors, then perform a

compare and contrast review. And as such, it may show broader, more defendable differences in the overall business environments at each location.

Researchers with a deeper understanding of statistical analysis and computer programming could run more complex studies. The author spent lots of time reading and testing the functions within R, but ultimately was not able to produce publishable results from the software to easily quantify and discuss the data. If a researcher was more proficient in R or another statistical analysis package, they could no doubt producer even richer findings.

**Bibliography**

[1] “Zillow Website.” Yahoo!-Zillow Real Estate Network, 2006-2013 Zillow.com, all rights reserved, 2013.
[2] “S&P/Case-Shiller Website.” Copyright © 2013 Standard & Poor’s Financial Services LLC, a subsidiary of The McGraw-Hill Companies, Inc. All rights reserved., 2013.
[3] A. P. Field, *Discovering statistics using R*. London ; Thousand Oaks, Calif: Sage, 2012.
[4] J. Adler, *R in a nutshell*, 1st ed. Sebastopol, CA: O’Reilly, 2010.