|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **“Zestimate” Accuracy across Major Metros in California State**  **Introduction:**  The Zestimate's accuracy depends on location and detailed availability of real estate data in an area. Some counties have detailed information about the houses in its jurisdiction, such as number of bedrooms, bathrooms and square footage and others do not. The accuracy of Zestimates depends upon the availability of data, it is directly proportional to the data, and more the detail data available, the more accurate will be the Zestimate.  This report examines the accuracy of Actual Sold Prices of homes versus the “Zestimate” prices in three Major Metros, namely San Francisco, San Jose and Los Angeles, in the state of California that were sold in the month of February, 2015. The data for this report were extracted from [www.zillow.com](http://www.zillow.com) .  This report is to analyze and justifies the Zestimates STAR rating accuracy for the above mentioned metro areas The data for my analysis is extracted from the website “<http://www.zillow.com/zestimate/#acc>”.        According to this website, San Francisco Metro area are comes under fair Zestimate (2 star rating) between “Zestimates” VS “Actual Sold Price” compared to Metro area of Los Angeles Good Zestimate (4 star estimates.  **Analysis Description:**  **Data Set Name : Zillow.csv**  **GitHub Location :** [**https://github.com/UmaSuresh/Data-Science/blob/master/Data/zillow.csv**](https://github.com/UmaSuresh/Data-Science/blob/master/Data/zillow.csv)  **Source Code :** [**https://github.com/UmaSuresh/Data-Science/blob/master/Capstone%20Project/Zillow.R**](https://github.com/UmaSuresh/Data-Science/blob/master/Capstone%20Project/Zillow.R)  Simple linear regression results of the actual price sold vs the Zestimate shows a correlation coefficient of 0.93452 [cordat – source code variable]. My scatter plot [Figure 1:] below explains that there is a strong positive linear correlation between the Zestimates versus Actual sold price by looking visually, though it has some outliers. The coefficient of determination (R2) for the Zestimate price is 0.8733283, so 87.3% of the variability of selling price can be explained by the linear relation between the Zestimate price and the actual selling price which I was expected already by looking at the Figure 1.  Figure 1 is the scatter plot of Zestimate vs Actual Sold price, however, shows a positive linear correlation with apparent outliers with the **home prices goes up above 2000000 dollars**. Figure 1:   Figure 2 shows histogram plot for the same above data set. Figure 2   Calculated least-squares regression(R square), it is helpful in determining the percent of variability and also it is a number that indicates how well data fit a statistical model.  Below are some of the observations,  **Simple linear regression results for the overall Data of California(San Francisco, San Jose , Los Angeles):**  Dependent Variable: Actual Sold Price  Independent Variable: Zestimate  Actual Sold Price = 115450 + 0.7702 Zestimate  Sample size: 124  R (correlation coefficient) = 0.93452 R-square (Coefficient of determination) = 0.8733283 standard error: 133600 on 122 degrees of freedom  Since my examination involves the comparison study of accuracy between the three major Metro areas, I have  decided grouping the data set by each Metro area and do separate analysis.    **Simple linear regression results for the Metro Area - San Francisco:**  R (correlation coefficient) = 0.907830  R-square( Coefficient of determination ) = 0.8241558  82.4% of the variability of selling price can be explained by the linear relation between the Zestimate price and  the actual selling price.  Standard Error : 56062.48 **Figure 3:**   **Simple linear regression results for the Metro Area - Los Angeles:**  R (correlation coefficient) = 0.9845623 R-square ( Coefficient of determination ) = is 0.969362  96.9% of the variability of selling price can be explained by the linear relation between the Zestimate price and  the actual selling price.  Standard Error : 32283.12 **Figure 4:**   **Simple linear regression results for the Metro Area - San Jose:**  R (correlation coefficient) = 0.9469084 R-square( Coefficient of determination ) = 0.89663569  89.6% of the variability of selling price can be explained by the linear relation between the Zestimate price and the actual selling price.  Standard Error : 96712.13 **Figure 5:**   Compared to all of the above three Metro areas, San Jose standard error value shows more negative impact on  the STAR rating estimates.  According to the standard error values, San Jose shows poor Zestimates(1 star) vs Actual sold price,  San Francisco shows fair estimates (2 star) compared to other two and Los Angeles shows good estimate(4 star).  Further research on analyzing these **yellow highlighted data on the plot** might give more effective way of  Finding reasons affecting the results of Zestimates on this report.   **Figure 6:**  **Figure 7:**  **Figure 8:**   The boxplot with Zestimates and Actual price shows that there aren't many outliers and there are few extreme ones that would cause an alteration in the data. Those are explained in the below table.   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | Region Name | City | State | Metro | County | Zestimate | Actual Price | Residuals | | 95070 | Saratoga | CA | San Jose | Santa Clara | 2745000 | 1895500 | 849500.0000 | | | 94024 | Los Altos | CA | San Jose | Santa Clara | 2588000 | 2588461.5 | -461.5390 | | | 94010 | Burlingame | CA | San Francisco | San Mateo | 2695000 | 1462750.0 | 1232250.0000 | |   According to the above table Los Altos data point is not an outlier since the Residual (Green Highlighted) is  very low. Red Highlighted residuals (Saratoga , Burlingame ) are the two major outliers which disturbs the  entire linear model.  Below Density plots explains the Zestimates probability of distribution among the three Metro Areas.  Definitely justifies Bay Area (San Francisco +San Jose) has more probability of high value home prices than  Los Angeles , As I predicted earlier in this report , Since the bay area home prices are higher ,  Zestimate accuracy cannot be good compared to Los Angeles. **Figure 9:**   **Owner:** [**mworkm86**](http://www.statcrunch.com/profile.php?id=mworkm86) **Size: 7KB  Created: Mar 2, 2013**  **CONCLUSION:Owner:** [**mworkm86**](http://www.statcrunch.com/profile.php?id=mworkm86) **Size: 7KB  Created: Mar 2, 2013**  The data included in this report and resulting statistics derived from this data show just how greatly one or two outliers can affect the entire linear model.  My research identifies that the Zestimates are more accurate and good estimate for the lower value home prices.  Until 100,000 Zestimates seems to be very accurate. Between 100,000 and 200,000 Zestimates seems to be fair accurate , but when the home price increases above 200,000 dollars , Zestimates accuracy fluctuates a lot and may give poor estimates.  While the "Zestimate" seems to work fairly well in a linear model when no outliers are present, in today's market, I believe that in some cases Zillow.com is wonderful at predicting the prices of houses on the market .I would say that it is very useful tool right now when buying a house, however Zestimates should only be used as a tool for evaluating the current market, and **should not be relied upon too heavily particularly for higher value homes.** |