Capstone project 1 - Exploratory Data Analysis - Inferential Statistics

**Jupyter notebook in github:** <https://github.com/yulmee/springboard/blob/master/Cap_Stone_1/Unit_8_Inferential_Statistics_WC.ipynb>

**Methods used:**

1. Pearson Correlation Coefficient:
   1. Calculated Pearson Correlation coefficient to what variables are most correlated to Sales Price for each of the counties.
      1. Alameda: Unemployment Rate is -0.96 at strongest negative correlation. Mortgage rate has weakest correlation at -0.11. Only number of monthly listing in Zillow is over -0.5.
      2. Sacramento: Unemployment Rate is -0.95 at strongest negative correlation followed by Affordability indext at -0.89. Mortgage rate has weakest correlation at -0.09. Surprisingly, for this market, number of days on Zillow has correlation of -0.55 and number of monthly listing in Zillow is at 0.16. Just looking at Pearson coffecient, Sacramento market has different characteristics than Alameda.
      3. L.A.: Unemployment Rate is -0.93 at strongest negative correlation followed by Affordability indext and Monthly listing at -0.67. Mortgage rate has weakest correlation at -0.11. L.A. is yet again different from Sacramento and Alameda in that affordability index and number of monthly listing is over -0.5.
   2. Calculated Pearson Correlation coefficient for sale prices between counties and other counties in Bay area to understand how close are counties’ sales price and if Alameda is good representation of the prices in Bay area.   
      Alameda and Sacramento are more correlated to each other than between LA and Sacramento/Alameda.  
      Surprisingly, Pearson coefficient between Alameda and adjacent counties are lower than Alameda vs Sacramento or LA.   
      Therefore, Alameda is not the best representation of Bay area or not really a good one for other counties in CA state for that matter.
2. Linear Regression between sales price and variables that have Pearson Correlation Coefficient of more than 0.5 (or -0.5).
3. For Alameda County, to see if Mortgage rate in early 2013 that could affect the slow down in sale price growth starting a month later, performed linear regression and calculated Pearson Correlation Coefficient between Sales Price % change and Mortgage rate % change. Mortgage rate % change was lagged by 1 to 3 months to find out if there is any change in relationship between sales price and mortgage rate if the rate is lagged.   
   When Mortgage rate is lagged by 1 month, pearson R coefficient is slightly better than when not lagged. Lagging by more than 1 month actually make the R coefficient worse so lagging by 1 month seems to be sweet spot.