Test of Normality of Distribution of Listings With Bedroom Attribute of 1:

Statistical significance pointing at the non-normality of the distribution of one-bedroom listings suggests that the distribution may be bi-modal, in which case one of the groupings may be listings for studio apartments, and the other, listings for true 1 bedroom apartments. This suggests that, in order to better estimate affordability - using the fair market rent as an estimate -, there may be a need to reclassify some of the listings listed as one bedroom as studio apartments or listings for an actual bedroom in a multi-bedroom unit. If listings that are truly Studio apartments are being listed as 1-bedroom apartments, this would overestimate affordability of zip codes. While there is no fair market rate for a single bedroom in a multi-bedroom unit, perhaps this could be estimated by taking the average of the 2 bedroom rate divided by two, the 3 bedroom rate divided by three and the 4 bedroom rate divided by 4. This value would be equal to a rate of $504 dollars for the price of renting average bedroom in a mult-bedroom unit in the Twin Cities. However, this is all under the assumption that there is a significant amount of people who are using the 1 bedroom attribute allowed by craigslist to advertise a physical bedroom space in a mult-bedroom unit. While further investigation is needed for deciding whether or not to include this third listing type as a classification category, there is statistical evidence to suggest that building a machine learning model to attempt to classify studio and one bedroom listings may lead to a more accurate dataset.

Two Sample Proportions Z-test To Test Whether St. Paul Is More Affordable Than Minneapolis

A two sample proportions z-test at the .05 level of significance between the proportion of listings in the City of Minneapolis at least one standard deviation over the FMR, and those in St. Paul shows that those in Minneapolis are significantly higher. This is indicative of a difference in price congruent with the location of a listing. In the context of the problem of predicting prices for craigslist housing rental ads, this test suggests location - at least at the city level - as an important predictor variable to include in a machine learning algorithm.

Two Sample T-test of Means to Test If There Is A Difference in Affordability Between Zipcodes Below And Above Median Wealth Rating:

A two sample t-test was performed on two groupings of zipcodes - those above, and those below the median wealth rating for all zipcodes. The t-test was performed using the means of the proportions of listings over the fair market rent per zipcode. The test failed to reject the null hypothesis with a p-value of .96, suggesting no significant difference between the means of the two groups. This suggests that income levels by zipcode may not be an important predictor of price for craigslist rental listings in a machine learning algorithm. A mann-whitney test, which does not require the assumption of normal distributions of the input samples was was also performed on the sample populations. While the p-value was much lower - p = .40 - , the test also failed to reject the null hypothesis.

Two Sample Z-test to Analyze Affordability Between St. Paul/Minneapolis and Their Surrounding Suburbs

A two sample proportions z-test at the .05 level of significance between the proportion of listings in Minneapolis/St. Paul at least one standard deviation over the FMR, and those in the suburbs reveals that the proportion of Minneapolis/St. Paul is significantly larger, suggesting suburb/main city as a potentially important attribute to include in a machine learning algorithm in predicting price and/or affordability.

Pearson Correlation Coefficient Between Price and Square Feet

As is indicated by the results Pearson Correlation Coefficient between Price and Square feet, at the .05 level of significance we can accept the alternative hypothesis that there is a relationship between the Price and Square Feet variables such that 47% of the variance of the Price residuals can be explained by the variance of the residuals of the Square Feet residuals. This correlation suggests that Square Feet may be an important variable to include in a linear regression model that attempts to predict the price of a listing.

Autocorrelation and Partial Autocorrelation Tests

Performing autocorrelation and partial autocorrelation tests on the Price and Time variables of the dataset produced some statistically significant results. An autocorrelation plot with 8 lag windows to represent approximately 2-week periods in the dataset shows a statistically significant correlation between time and price variables for all 8 lags. The highest correlation coefficient around 60%, and the lowest slightly above 40%. A partial autocorrelation plot with the same number of lag windows also shows a statistically significant correlation between time and price variables albeit, smaller and decreasing over time. A partial autocorrelation may be better representative of the true correlation between price and time variables, as it removes any variables besides price and time alone that may contribute to the correlation coefficient. In this sense, the time variable may not be an important price predictor in a machine learning model.