

Predicting Prices of Second-Hand Items

Price Predictor: nlei, mali37, imeza4

Hypothesis

Sites online such as Craigslist have increasingly become the main way in which consumers buy second-hand items. We sought to analyze the relationship that item prices have with their attributes to give people a better idea of what the usual price of an item may be. Our study aimed to investigate potential variations in pricing based on location and categories the items were listed as. Additionally, we sought to explore the impact of mileage on vehicle listing prices.

Data

We obtained our data by scraping this website:

<https://providence.craigslist.org/search/sss?purveyor=owner#search=1~gallery~0~0>. We focused on the Providence section where items were listed for sale by owners. Our data was formatted into a JSON file where each listing contained a title, price, location, metadata, link, images, number of images, category, metadata length, date, and mileage. Mileage was only for the vehicle listings where it is applicable. We cleaned our data to ensure that the item's attributes were consistent within their category.

Findings

Claim #1: Across different locations, there is a significant difference in the mean price of items.

Support for Claim #1: Using the Kruskal-Wallis, we used price as our base variable and location as our grouping variable to determine if there was a difference in the mean price of items across locations. Our result was a statistic of 914.337 and a p-value of 0.00. Because our p-value was below our alpha level of 0.05, we were able to reject the null hypothesis. This indicated to us that there was a significant difference in the mean price of items across different locations.

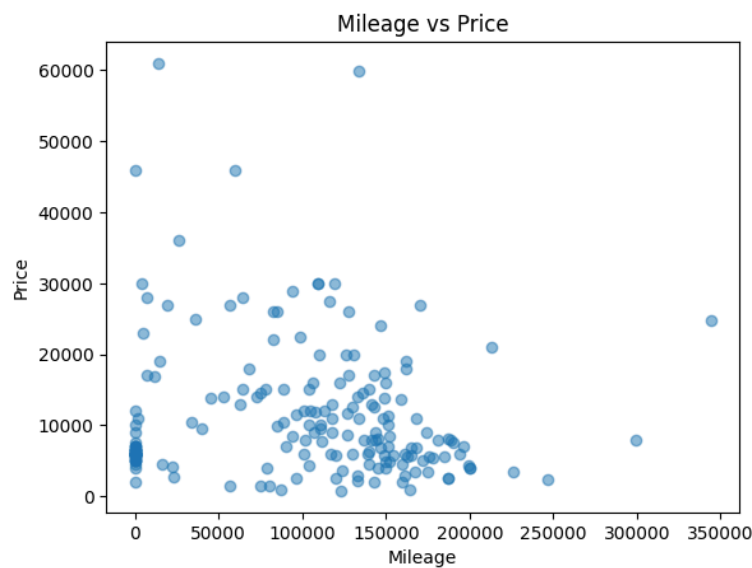
Claim #2: The mean price of items will be different depending on the category.

Support for Claim #2: We attained a statistic of 1163.701 and a p-value of 0.00 using the Kruskal-Wallis test. We used price as the base variable and category as the grouping variable. We compared the p-value to our alpha level (0.05), and with a p-value close to zero, we rejected the null hypothesis, indicating that there was a significant difference in the mean price of items across all categories.

Claim #3: Vehicles with lower mileage will have a higher price.

Support for Claim #3: We conducted a simple linear regression analysis to examine the relationship between mileage and price in vehicle listings. Our test revealed a correlation coefficient (r^2) of -0.08, indicating a negative linear relationship between mileage and price. While the r^2 value suggests that the relationship is not strong, upon reviewing the produced graph (placed below), we can observe some correlation between the two variables. However, it's important to note that the correlation is not substantial. Despite the weak correlation, the evidence from the analysis suggests that we can reject the null hypothesis and conclude that there may be some degree of linear relationship between mileage and price, although it is not substantial. Further analysis may be needed to understand other factors influencing vehicle prices.

See graph below:



Claim #4: We could not predict the price of second-hand items using the information present (including image) on the Craigslist listing page using Linear Regression.

Support for Claim #4: We attempted to utilize Linear Regression to predict item prices. The analysis revealed a median Root Mean Squared Error (RMSE) of \$262.70 and a slightly positive R^2 of 0.08, indicating a weak correlation between predictors and prices. Although the model's accuracy may be limited, it still provides valuable price suggestions for sellers.

Produced visualization below:

