

Term Project Summary

DSC530

Eliyahu Forta

08/14/2021

After going through the Exploratory Data Analysis, we concluded that a high covariance exists between the price of a home and how many bathrooms it contains.

Since the correlation was low, we assume that there is a non-linear relationship between these two variables.

There was certainly more data that could have been used to improve the efficacy of our analysis. The way cost exists in our dataset does not paint the full picture of what that house should be worth. Thus, variables which almost certainly affect the value of a given home were not readily evident.

The assumption I did not agree with was that zip codes within similar geographic areas are good indicators of home value. While this is almost certainly correct in most cases, the varying geographies of our dataset are remarkably like one another in price and size. Therefore, I question why the column exists to begin with. Variables should be checked for usefulness before being plugged into a dataset.

The challenges I faced were mostly code related. I do not come from a technical background, and so, I can never be sure how to approach a problem in the most efficient way possible. While I used the Think Stats book for guidance, it was frustrating to have my creativity quashed by the breadth of my own knowledge.

Additional struggles came from using multiple packages while not fully understanding the scope of their functionality. Pandas and Thinkstats have been incredibly useful, but when encountering an error, it is less clear how to resolve the issue.

In conclusion, there is likely a non-linear relationship between our variables, the dataset was lacking in information, and the struggles encountered in this analysis were primarily due to my limited coding experience.