Project Proposal:

Team Members: Tonnar Castellano, Connor Mignone, Ty Painter

The housing market is a prime target for statistical analysis and machine learning. Zillow, a major player in the housing space, attempted to create machine learning models to predict housing prices and find arbitrage. Zillow eventually scrapped the project saying that due to an immense amount of data in an ever-changing space, it was currently impossible to accurately predict prices. The company currently does have a value called ‘zestimate’ which predicts the assets value. However, due to fluctuations in the economy; housing price predictions are often very inaccurate.

Our group overall disagrees with Zillow. We believe that through leveraging the power of big data we can account for fluctuations in the market. We are using Kaggle housing price data sampled from the Pacific Northwest of The United States. The data needs some, but minimal cleaning. The columns seem easy to engineer, with plenty of opportunity for additional feature creation. Given the geographically localized data, our main objective will be to use a regression model to predict housing prices. This entails us minimizing the loss function of mean squared error. This process will be completed using PySpark which will allow us to handle vast amounts of data in a distributed environment. Additionally, modeling will be done using PySpark’s MLlib package.