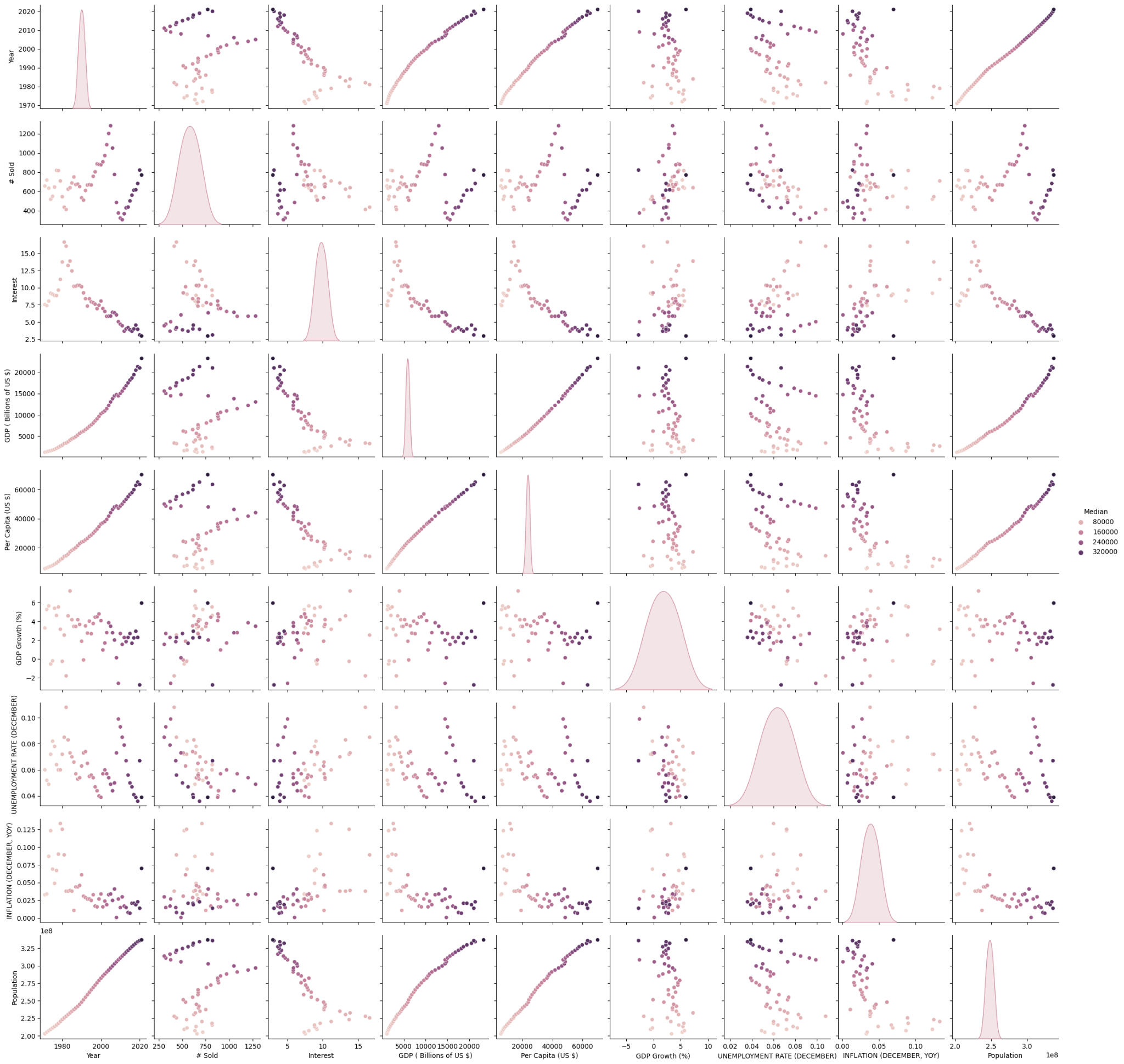
DATA 151 Research Paper

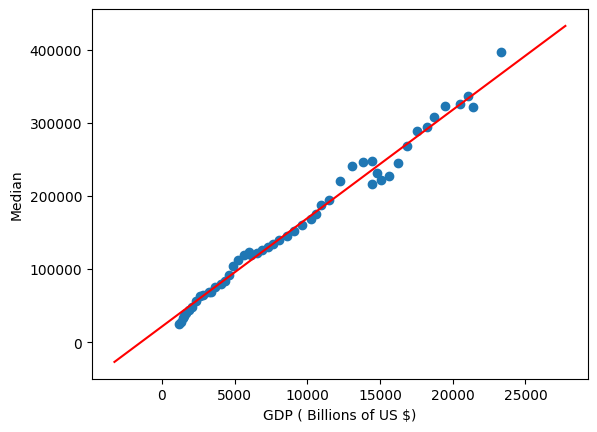
In recent years, the housing market has undergone a significant increase in prices. Understanding the underlying factors contributing to this trend and identifying the most important variables in predicting house prices is of utmost importance. There could be multiple reasons for the inflation in housing prices, including the construction of larger and more expensive houses, a reduction in the availability of affordable housing, or the impact of inflation. Given that housing is a basic need and represents a significant investment for most people, it is crucial to make informed decisions based on a thorough analysis of the market. However, the high prices in the market may seem inexplicable and require further investigation.

We analyzed median and average housing prices, per capita income, population size, GDP, inflation rate, average interest rates on houses, the number of houses sold, and unemployment rate from 1971 to 2021. We utilized Python libraries such as Pandas, Seaborn, and Scikit-learn for data analysis.

We used single and multivariable regression to determine the most influential factors on house prices. Our analysis revealed that GDP had the highest correlation and accuracy in predicting median house prices. The root mean square error of our models on the test data was used to measure the success of our analysis.

****

The pairplot revealed that population, GDP, and per capita income were closely related. Eliminating one of these variables may seem like it would have little impact on the results, but our analysis showed that this was not the case. Other variables did not exhibit clear covariance, except for the interest and inflation rate, which negatively correlated with GDP, population, and per capita income.



This graph shows how GDP may be a very good indicator of house prices.

**Our linear regression analysis of the top five variables most correlated with median house prices showed the following RMSE values on the test data:**

**GDP (in Billions of US $): $15,000**

**Per Capita Income (in US $): $18,000**

**Year: $34,000**

**Population: $37,000**

**Interest: $91,000**

Although GDP, per capita income, population, and year had almost perfect correlation with median house price, their RMSE values differed significantly. This indicates that correlation alone is not sufficient to determine a variable's predictive power. Interestingly, the inflation rate did not appear to be a significant factor in predicting house prices

.

**Our analysis using all variables combined resulted in an RMSE value of $6,000, which was a 60% improvement over using GDP alone. This indicates that using all variables together significantly improves the accuracy of predicting median house prices. Furthermore, removing population size from the data resulted in a drop in accuracy from 60% to only 14%. Attempting to predict house prices using population size alone led to imprecise results. This observation underscores the importance of using a combination of weak models to generate a strong model, as in the concept of boosting.**

We also utilized Ridge regression, which uses L2 normalization to minimize overfitting and improve test accuracy. Our analysis using Ridge regression resulted in an RMSE value of $5,000, indicating that we can accurately predict median house prices with a precision of $5,000.

In conclusion, our analysis revealed that GDP was the most influential factor in predicting median house prices, but using all variables together significantly improved the accuracy of our predictions. The use of Ridge regression further enhanced our analysis, highlighting the importance of minimizing overfitting when analyzing complex data.