Jeremiah, Fayol, Colin, Sunny, Sean

Beagley

DATA151

07 May 2023

Final Project Report

Executive Summary

After the COVID-19 global pandemic, the United States economy was heavily impacted.

One of the biggest impacts that seemed to happen was a huge growth in the price of

housing in the United States. This is what lead to the original purpose of this project is to find out if there were any discrepancies between the housing market now and before covid. After further research, the group realized that there was not enough data set to determine a true difference in the price of housing by only using data from 2020 to 2022 so the group decided to use data going all the way back to 1970. After understanding this new guideline for the project the group began working on finding and cleaning data that was relevant to our project. The group found data that was based on the average price of housing every year for the purpose of comparison of the average price of housing over the years. The group also found data based on the price of housing in different regions to see how the location of a house impacts the price of housing as well. Lastly but most importantly the group found data on the amount of houses sold every year as well as other economic factors that were happening each year as well. At this point in the project, it was time to make the model which the group did. After further breaking down the data in the model and finding correlations amongst all variables the group came to the realization that the housing price had the strongest correlation to the year that it was. By this logic the price of housing would have gone up since covid and confirmed our alternative hypothesis however this hypothesis would have been true for every year prior to covid as well. However, the next biggest correlation to the price of housing though was the GDP of the country which had also dropped tremendously during the pandemic which had affected the price of housing. After finally getting an output from the model the group came to the conclusion that GDP and Year had the biggest impact on the price of housing in the United States and that the price of housing had gone up every year so our alternative hypothesis was true about the rise in the price of housing after the global pandemic.

Background Page

Link to gitub: <https://github.com/jjohn120/DATA151_group_project>

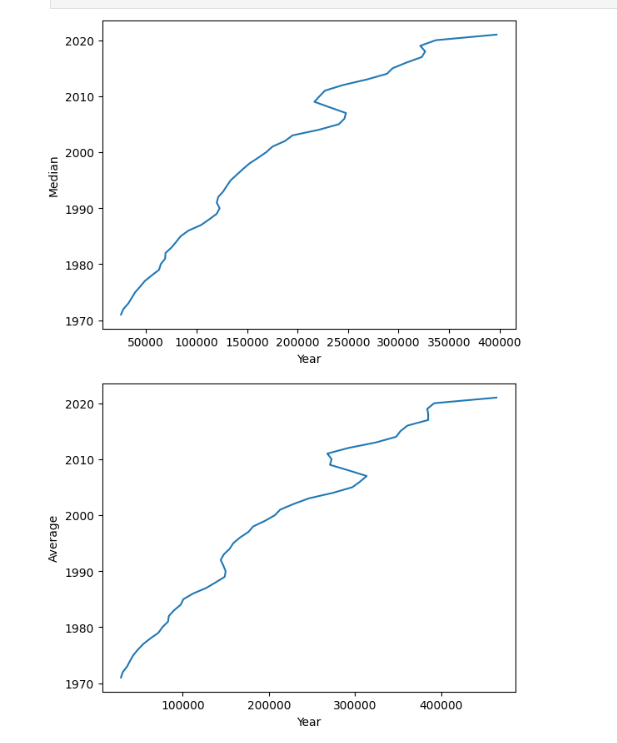
The link above contains access to all code, write-ups, and presentations to the project as well as the model. This project’s purpose is to analyze data, find correlations and help answer our research question of if Covid-19 made the price of housing go up in the United States. This was done by gathering data that is also in the repository of the link above. By using specific features of the Python coding language correlations were able to be found within the data and that helped us answer our research question. In order to understand the method and process of the creation and usage of the model it’d be best to look at the code written for it which can be found in the link above.

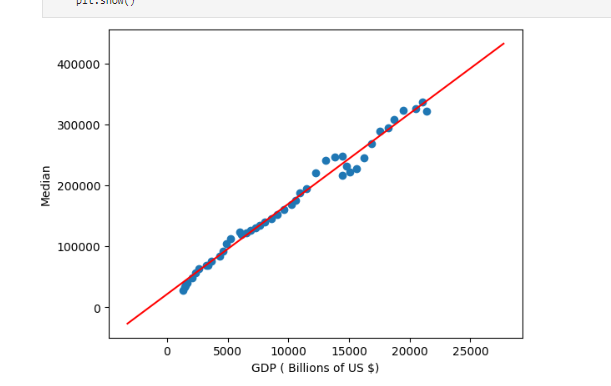
Methods page

The methods we used in order to come to the final conclusions were first finding correlations between each variable within our data, understanding the type of correlation each variable had, and how much those variables impacted one another. The next method was obtaining a better visual and understanding of this information by creating charts and graphs in order to see how big certain correlations were as well as just the overall message of the data. After doing this a linear regression on every heavily correlated variable was done in order to make a prediction based on the model. After this the RMSE of the linear regression was measured however it didn’t give the greatest result so then the data was put through a ridge regression. After obtaining the prediction from the ridge regression the group got a much more accurate result because of the minimization of overfitting. The result was within 5000$ of precision. The group also tried a lasso regression however there was still not as much accuracy as the ridge regression.

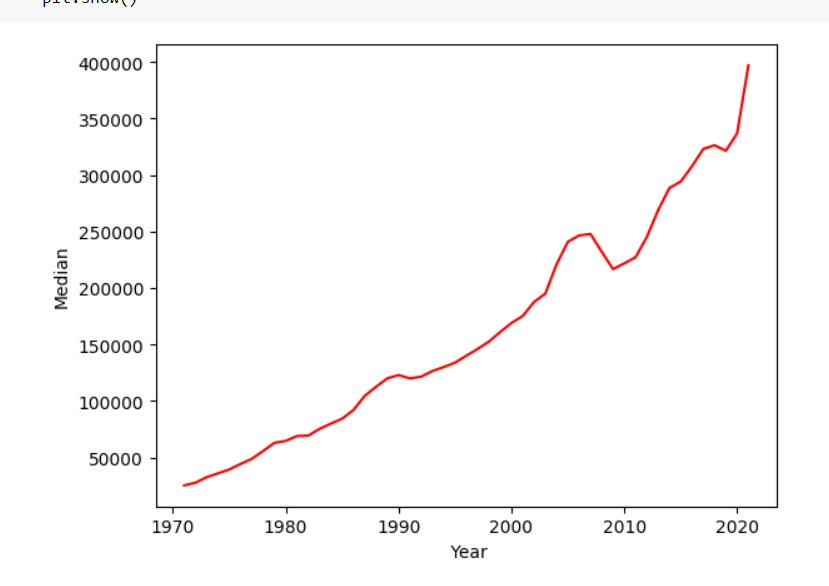
Results Page

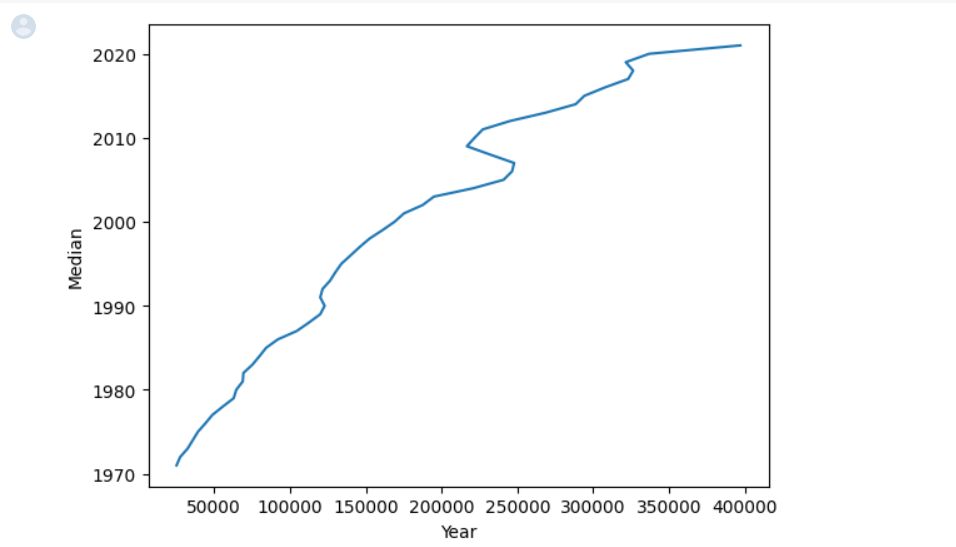
To conclude ultimately the answer to our research question was our alternative hypothesis that ended up being true that the price of housing had gone up since the global pandemic. However, evidence showed that the price of housing had gone up every year and that COVID-19 may have not been the biggest factor in this.

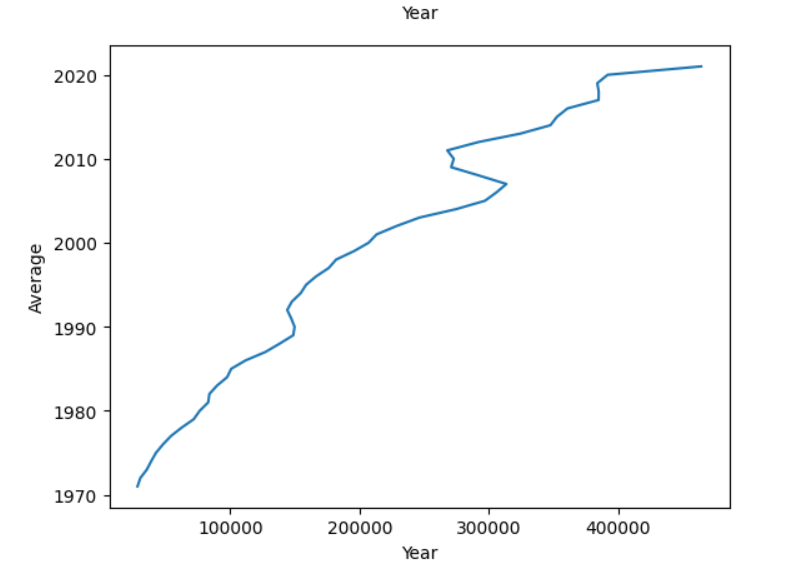
The next conclusion found was how the year had the strongest correlation with the price of housing however this correlation wasn’t necessarily valid because the year variable can only rise. This is what lead to the final and biggest conclusion found was that though GDP alone could give results with an RMSE value of 15000 dollars, combining many features does not always gives the best results in modeling. Furthermore, depending on the model, different features are given more weight. This sheds some light on WMD by highlighting how even for simple linear regression models, our algorithms can pick strange features to be important. With respect to our housing problem, we can confidently say that we can predict the price of houses within 20,000 dollars in accuracy. It may seem big, but that's within 7% of house prices nowadays.

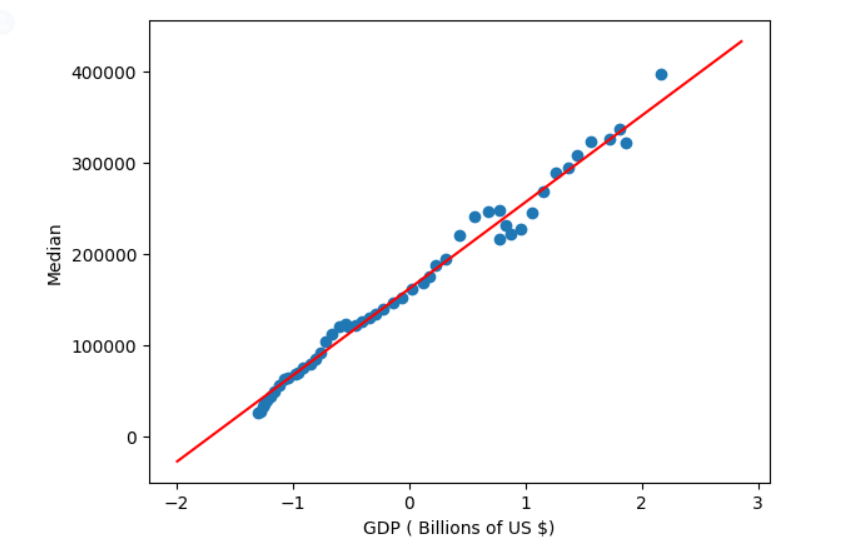


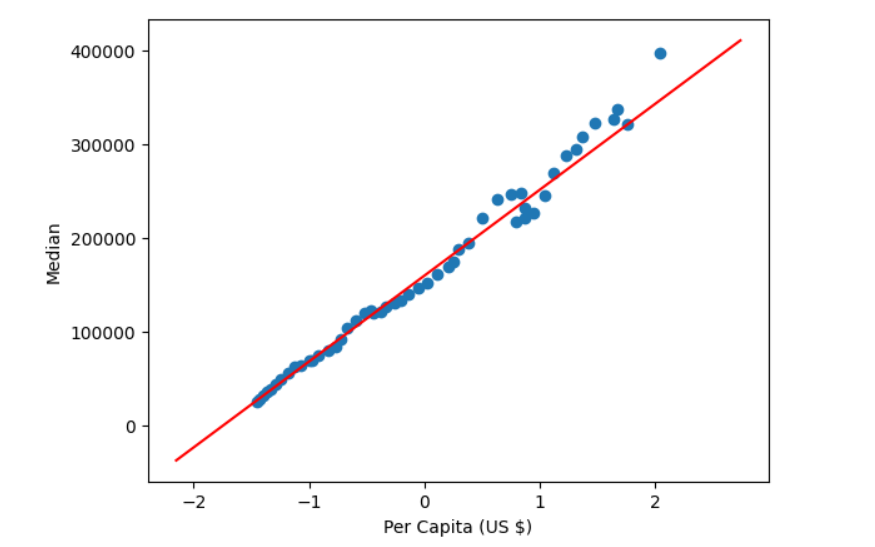
Here are all the graphs within the model in order to completely understand the data.











Along with a final link to the [notebook](https://colab.research.google.com/github/jjohn120/DATA151_group_project/blob/main/Project_part_4.ipynb)