

## Project Part 2 Writeup

1. **Motivation/problem statement:** Why are you planning to do this analysis? Why is it potentially interesting and useful, from a scientific or practical perspective? What do you hope to learn? Your motivation statement should explain how/why the problem is human centered. That most of the COVID-19 data reflect people as data points does not make a problem immediately human centered.

In this analysis, I would be investigating the impact of the pandemic on the housing prices in Polk, Florida. During the pandemic, the housing prices surged in the United States and presently it is still showing the sign of further uprise. This would suggested a lot of human centered implications, in terms of the economy, the society and political relationships. For homeowners, this could help them gain insights on their personal wealth and better manage their cost of livings. For investors, the relationship between the pandemic and house prices could be a significant factor when making investment decisions. Moreover, if there does exist a relationship between pandemic and house prices, it could say a lot of things about the economy as the house prices both reflect the ongoing economic/political policies, and signal potential future policies.

2. **Research questions and/or hypotheses:** These describe what you hope to discover or determine. Keep in mind this should be related to the Common Analysis that you conducted in Part 1. There are lots of possible questions related to the pandemic, and the Part 1 data set has day-by-day data for your county. For example you might ask:
  - Example research question: How was the local economy influenced by the pandemic?
  - Example hypothesis: Every 10% change in confirmed COVID cases resulted in a 5% decrease in employment for service sector workers in <CITY> within the county.

### **Research Question:**

Does the pandemic have an impact on the housing prices of Polk, Florida? If so, how was the local housing prices influenced by the pandemic?

3. **Data to be used:** What additional data do you plan to use, and why? Summarize what is represented in the dataset, **provide a link to the dataset**, and specify license/terms of use. Highlight any possible ethical considerations to using this dataset. Briefly explain how this data will allow you to answer your question and how it expands on the Common Analysis.

For this analysis, I would be also using the US Housing Market Dataset from Zillow: <https://www.zillow.com/research/data/>. The home type of “ZHVI all homes time series, smoothed, seasonally adjusted” is selected and geography location is specific to county. According to the description on the webpage, ZHVI is “a smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type”, and “reflects the typical value for homes in the 35th to 65th percentile range”.

**License:** none found on the Zillow website

**Data Description:**

This data contains the housing prices in the United States specific to the county level, including Polk, Florida, the target country that I am investigating. It also contains all other counties in the United States as well as their housing prices. The housing prices are a time series data taken as an average from all houses in the specific region at the end of each month. Using this data, I would be able to investigate the relationship between the covid confirmed cases and the house prices in Polk, as well as in other regions or US as a whole as well.

4. **Unknowns and dependencies:** Are there any factors outside of your control that might impact your ability to address the supplementary questions? The purpose of this section is to get you thinking, in a practical sense, about your ability to answer the supplementary research question in the time allotted.

There might be factors out of my control that influences the house prices. In fact, housing prices could be affect by a huge amount of factors: inflation, unemployment rate, infrastructure development, location, currency rates – so many that I couldn't possibly list them here. More importantly, the houses in US is a market, and market has its own parties: suppliers, consumers, agencies, and each one of them play an important role in the prices of houses. The complexity of this market makes it extremely difficult to predict the exact prices of the houses, even though leading companies such as Zillow and Redfin are already doing a decent job in estimating the prices.

This analysis considers Covid as a macro-level activity that could possibly have a general impact on the housing prices. It is believed that the role of Covid – if there is any – would not be significant enough to become the primary cause.

5. **Methodology:** Describe how you plan to investigate this phenomenon. Don't just describe what your analytical methods are (e.g. "ordinary least squares", "student's t-test", "heatmap visualization", or "recurrent neural network"), it's critical to justify why these are appropriate methods for gathering and analyzing your data, or presenting your findings. You are expected to be thorough here: please describe to the best of your ability the entire series of gathering, analysis, and presentation methods you plan to use.

In this analysis, I would be using cross-correlation methods to analyze the impact of covid confirm cases. The time series of the covid confirm cases and the time series of the housing prices would be used for the corelation analysis. The two time series would be normalized and cross-correlation would be conducted with a set of lags. Moreover, the overall US home prices and confirmed cases would also be analyzed as a comparison to

the local level. To present the results, a cross-correlation graph would be created showing the correlation level with different lags.

Cross correlation does a great job in presenting the correlation between different time series. They show the correlation level at different time lags, and we could use a specific confidence level for the correlation indicator to determine if there is any statistically significant correlations. However, even if there does exist correlation, we could not determine that the pandemic does have an impact on the house prices as it doesn't imply causation. Theoretically, it would be more convincing if we could isolate out the factor such as simulating the housing prices without the pandemic playing a role, but that would be unrealistic and many other factors would also change.

6. **Timeline to completion:** You should list each of the milestones and significant tasks you will need to complete the extended analysis. You might need to: collect data, build a model, test or analyze the model, visualize the results, and document your process. Think of the tasks that you need to complete. You should then provide a timeline that would allow you to complete the analysis on deadline. Keep in mind that you have a presentation (Course Project, Part 3) and final report (Course Project, Part 4) to complete as well. You'll need some time for those activities.

Collect data: Done

Data Processing: Nov. 17

Run Cross-Correlation tests: Nov. 25

Data Visualization: Dec.1

Presentation: Dec.5

Report: Dec.12