Milestone 1
David Ziman and Rebekah Doochin

Datasets:

For the final tutorial we will be looking at various data sets to determine the impact that the COVID-19 pandemic has had on the housing market in the United States. The first dataset contains data on the sale prices of houses in various cities from the past twelve years. Cities included in the study were the largest cities in the US in 2008.

Although the cleaning and sorting the data initially seemed challenging, it was fairly straightforward to tidy the data after familiarizing ourselves more with the proper techniques. In the initial CSV file, all of the dates in which the data was stored were column titles rather than values in a column titled “Dates.” To remedy the problem, we created “Dates” and “Price” columns to hold the information that was previously stored in multiple columns. After doing this and dropping the rows with NaN values, our dataframe appeared much cleaner and easier to look at.

The second dataset we looked at contains the median sale prices of houses in the US going back to 2005. This data was already fairly tidy, however this may be attributable to the fact that it lacked detailed data on a city and state wide level. Although this dataset was smaller than the first, it can be used in conjunction with other data. We would like to merge this dataset and compare median sale prices in cities to overall median sale prices for that year.

While we have not yet broadened the scope on the types of data we are looking at, we think that it will be important to look at factors that affect the housing market in non-pandemic years as well. Possible datasets that we will look at for this might include data regarding economic growth or stock market prices.

Ultimately, the pandemic has affected different real estate asset classes very in various ways, thereby changing the landscape of our cities and suburbs. It will be interesting to analyze how the industry has changed in the past, and build data driven models to predict where it will be going in the future.

Collaboration Plan:

To collaborate on our final tutorial we plan on utilizing a private GitHub repository so that we both have access to current versions of our code and Jupyter notebooks. As we worked through milestone 1 we found that it was easiest to communicate over FaceTime for less technical questions or planning. However, for harder questions it was easiest to talk via Zoom so that we could use the screen-sharing feature. Because we are also working on our Capstone project together and communicate daily, we found that it was easier to talk as needed rather than have set meeting times.

To work on code simultaneously we will utilize Visual Studio Code’s Live Share feature. This will allow both of us to be hands on in the coding portion of the final tutorial and make communication regarding specific lines of code easier. We had a few technical difficulties getting VS Code working the way we liked for milestone 1, however the issues have been resolved and it will be a useful tool for the remainder of the project.

Links to Datasets:

https://www.kaggle.com/paultimothymooney/zillow-house-price-data?select=Sale_Prices_City.csv
https://fred.stlouisfed.org/series/MSPUS