Data 515: Human-Centered Data Science
Juhi Choubey

# Project Part 2: Extension Plan

## 1. Problem Statement:

The objective of this project is to assess the impact of covid-19 cases on the real estate and housing market in the Multnomah County, OR. The duration of this analysis will include study between 2019-2021 to accommodate for the effect of time before and after covid-19 surge.

Real estate and housing market is subject to many ups and downs - and largely depends on the standing of customers (humans) either through politicial amendments, economy, spread of diseases (in this case covid-19), or many other human-driven factors. Not just that, the relative fluctuations in real estate is in turn a cause of stir among the general public - and not just people of one socio-economic class, but from every class. This problem is particularly a human-centered due to the direct impact on human society and the large population. Though, in this particular analysis, we will be narrowing the scope to Multnomah County, OR only, and for a fixed duration too (2019-2021).

Through this study, I aim to assess any connection between the rise/fall of confirmed covid-19 cases and the subsequent variation in real estate market (housing prices in particular). While a general trend is a base observation for housing prices, I hope to detect any specific/unusual fluctuations that can be connected to the pandemic in a certain way.

## 2. Research Questions:

The preliminary research questions I plan to assess from the dataset and answer are as follows:

- ➔ Was there a statistically significant difference between housing prices in 2020 and 2021 (during pandemic) as opposed to 2019 (prior pandemic) in Multnomah County, OR?
- ➔ Was there a statistically significant difference between housing prices of different house types (all residential, single family, condo, etc) between 2019 (prior pandemic) and 2020 (during pandemic) in Multnomah County, OR?
- ➔ Did the number of confirmed cases have a statistically significant impact on the housing prices between 2020-2021 in Multnomah County, OR?

I am aware that there are many considerations that I will have to look at, including political and economic effects, covid-19 severity (not just confirmed cases, but deaths and hospitalizations as well), etc. Due to lack in availability of such detailed data sources, I will make and highlight any relevant assumptions throughout the analysis.

3. **Data Sources:**

There are multiple datasets to be used during the analysis covering both covid-19 cases and real estate market in the specified duration (2019-2021).

Below lists the source of datasets and the link to download them. We are currently planning to use all the below datasets, but this is subject to change as the analysis progresses.

- ➔ The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data: https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv
- ➔ The RAW_us_deaths.csv file from the Kaggle repository of John Hopkins University COVID-19 data: https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_deaths.csv
- ➔ The weekly housing market data from Redfin: https://redfin-public-data.s3-us-west-2.amazonaws.com/redfin_covid19/weekly_housing_market_data_most_recent.tsv
- ➔ The monthly housing market data from Redfin: https://redfin-public-data.s3.us-west-2.amazonaws.com/redfin_market_tracker/county_market_tracker.tsv000.gz
- ➔ Interest Rate (fixed for 30 years): https://fred.stlouisfed.org/series/MORTGAGE30US

The redfin weekly housing market data has data for each county on a weekly basis. Furthermore, the data is categorized by property type (All Residential, Single Family, Condo, Multi-Family, Townhouse, etc.) Redfin has published a data dictionary to understand every column and interpret them. This data set is licenced under Redfin's Terms of Use. The guidelines for using the data states to cite the data source appropriately and provide a link to Redfin.

4. **Unknowns and dependencies:**

As specified in the research hypothesis section, there will definitely be other factors at work in explaining the variability within housing prices in Multnomah County, OR between 2019-2021, such as political and economic amendments, increased rental prices, lockdown restrictions leading to remote work, covid-19 severity (not just rise in confirmed cases, but deaths and hospitalizations as well), fluctuating interest rates etc. Due to limitations on both time and data availability, this project will narrow the scope to assessing impact of rising/falling covid-19 confirmed cases on housing rates, making necessary assumptions for other factors.

## 5. Methodology:

To perform the analysis, I have listed down the different tasks involved and have provided a description of the potential action items in each. This covers an overview of the steps under each task at a very preliminary level. The methodology is subject to change depending on the observations from the exploratory data analysis and the assumptions made.

| Task | Methodology |
|---|---|
| I. Data Cleaning | For all required data files:<br>a. Filtering the datasets for selected county: Multnomah, OR and Time Frame: 2019-2021<br>b. Treatment of null values (discarding/imputing)<br>c. Rollup #confirmed cases/deaths at weekly level<br>d. Merging the datasets to prepare 1 master file containing covid-19 metrics and housing prices.<br>e. *Optional:* Feature Extraction (high-level) - Infection rate, Moving Average of #confirmed cases |
| II. EDA (Visualizations) | Review following trends for Multnomah County, OR between 2019-2021:<br>a. Weekly trend on housing prices alone<br>b. Weekly progression of #confirmed cases and housing prices<br>i. Related view: #confirmed Moving average cases and housing prices<br>c. Weekly progression of #confirmed cases and #houses sold<br>i. Related view: #confirmed Moving average cases and #housing sold<br>d. Weekly progression of housing prices across different property types (all residential, single family, etc.)<br>Repeat the analysis for monthly level as well to align the results. Highlight if any discrepancy is observed. |
| III. Research Hypothesis | Hypothesis I: No difference b/w mean housing prices in 2019 vs 2020<br>Statistical Test: Two-sided Z-test of Population Means (appropriate for testing large sample, continuous numerical data)<br><br>Hypothesis II: No difference in mean housing prices b/w different house types in 2019 v/s 2020<br>Statistical Test: ANOVA (appropriate for testing large sample, continuous data across multiple groups)<br><br>Hypothesis III: No Impact of #confirmed covid cases on housing prices<br>Statistical test: Linear Regression (appropriate for testing correlation b/w two large sample continuous numerical data)<br>*Note: Further considerations on other metrics will be required* |

6. **Timeline to completion:**

The project is scheduled to be completed in a stipulated time frame of 6 weeks. The first and last weeks are set for project planning and report writing respectively. So, the actual analysis and presentation of results is to be completed in 4 weeks. Below is an initial draft of project timeline: