ETL Project

August 4, 2020

TEAMMATES

Verina Hanien, Kenny Wu, Shivani Chopra – honorary member: Joshua Steier

Dedication: In memory of our former teammate, Eline VE

OVERVIEW

This project requires our team to demonstrate our learnings around data extraction, transformation and load to create a more dynamic and robust dataset for our analytical needs.

GOALS

1. Utilize 2 or more comparable datasets to be able to run meaningful analysis off of
2. Submit a project write up, outlining sources, methodology, challenges (if any) and key findings

SOURCES USED

1. Home values from Zillow: https://www.zillow.com/research/data/
2. NYC Covid 19 Data: https://github.com/nychealth/coronavirus-data/blob/master/data-by-modzcta.csv

Source Rationale: These two sources were utilized to understand the relationship between the effects of COVID-19 and the drastic effect it potentially had on the NYC housing market.

HYPOTHESIS

The amount of COVID cases within a zip code will have a direct correlation to the housing prices in that area.

METHODOLOGY

1. **Data Extraction:** download data from relevant sources in .csv format. API was not used due to time constraint.
2. **Data Cleanup & Transformation:** data cleanup was primarily done on the data from Zillow.
    a. Zip codes that did not fall under NYC were dropped (i.e. where the column, City <> New York City, was dropped)

b. Data from 2020 was retained (January – June 2020)
c. Average of data from January – June was created
d. RegionName column was renamed to “zip_code”
e. Specifically on the NYC Covid 19 data file, the column, ZIP_CODE was renamed to zip_code for consistency

3. **Data Join:** A left join was done on SQL with the Zillow housing data against the NYC Covid 19 Data, using the zip code field
4. **Data Filtering:** The NYC Covid 19 data had 177 zip codes vs. Zillow’s 175. This discrepancy was likely due to the lack of housing availability in the other 2 zip codes. The 2 records where there was no match for the zip code, and were NULL, were dropped.
5. **Data Aggregation:** All relevant fields across both datasets were aggregated to come up with one master database, including the average of housing prices
6. **Data Analysis:** Though we saw a slight downward trend in overall housing prices MoM, while reviewing the aggregated Dataset, we realized we would need to bring in additional data to make a direct correlation to COVID 19 cases. Since most of NYC residents are renters, the decline in housing prices isn’t as steep at the moment, and there is currently no direct correlation between the housing prices to the amount of COVID cases in the zip code.

Next Steps

Our recommendation is to revisit both the datasets and append with FY 2019 - 2021 data to see if there is indeed a direct correlation between the amount of COVID rates per zip code and the housing prices in that area.