Emily Sun, Emily Zhou, Mariam Seshan, and Teresa Duong

DS 2002 Final Project

Data Science Final Project Reflection

Introduction

When deciding on a topic for our final project, we chose one that mirrored the challenges and complexities of the real world. Our project, which aimed to explore the relationship between housing market trends and COVID-19 case data in the United States, provided a valuable opportunity to navigate the entire data lifecycle, from selection and extraction to transformation, analysis, and cloud storage. While the process involved technical hurdles, unexpected results, and adjustments to our research question, it ultimately enhanced our understanding of data science workflows and strengthened our collaborative and analytical skills. This reflection highlights the obstacles we faced, the strategies we employed to overcome them, and the lessons we learned for future projects.

Challenges

Data Selection:

Some challenges we faced during data selection were finding datasets that were fit to be compared and could be analyzed with the methods we know of. We evaluated our datasets on recency, availability of features of interest, file format, and whether it was a representative sample for our purpose. Our original research question focused on comparing housing prices over time between two different U.S. states. However, it was challenging to find public datasets that were collected within the same time period, which contained our features of interest (a home value metric and time), and which had comparable populations (two U.S. states). Additionally,

we were hoping to have access to the data in either CSV or JSON file format to use techniques learned in class. This led us to revise our research question comparing more widely housing in the United States to COVID-19 cases from January 2020 to December 2021. This experience showed us the challenges of forming a research question and how it may be necessary to continuously adjust the scope of the question to the available data.

ETL setup and implementation:

One challenge we faced during ETL setup and implementation was in matching table formatting between the two datasets for comparison and visualization in the Analysis portion of the project. Both datasets are indexed by date, however, the Zillow Home Value Index (ZHVI) dataset reports the date once per month while the US Counties COVID-19 dataset reports the date daily. This reflected differences in the time scale of the collection of data of these two sets as many economic indicators like housing and employment are reported monthly, while data on an infectious disease might change significantly within a day. Additionally, the ZHVI dataset reports housing data per state, while the US Counties COVID-19 dataset reports per county within a state. To reconcile these differences, we had states as the rows which meant compiling the counties for the COVID dataset and we had the months for the years 2020-2022 as the columns which meant for the COVID dataset, we had to grab the cases from the last day of every month since it was a cumulative count. Through this process, we learned how domain knowledge can guide the ETL process and prevent errors that could compromise data integrity. Moreover, we gained practical skills in using Python libraries like pandas for reshaping and aligning datasets to enable meaningful comparisons.

Analysis:

Some challenges we faced during analysis came from the fact that our results were not as we expected. We originally predicted home value to decrease in housing markets as COVID-19 cases grew because of what we knew of the pandemic's effect of economic uncertainty during that time and how there might be less moving and home-purchasing activity. A line graph of the dataset however showed a seemingly normal increase in Zillow Home Value Index over the year 2020. At first, we thought the reason could be that pandemic impacts on homeowners required more time to be reflected in the housing market and Zillow's Home Value Index measuring system. However, expanding the timeline of our data an additional two years showed minimal changes in the indices outside of normal increases in housing prices over time and supported the observation that Zillow Home Value Indices of homes across the U.S. did not change significantly during the pandemic.

Cloud Storage:

The challenges we faced in setting up cloud storage for our data were in handling access control. We learned only some Google Cloud APIs could be authorized using API keys (e.g. Google Maps and Gmail), so we explored using a service account private key to access our data instead. This process taught us more about different modes of cloud storage access and the importance of secure access.

Lessons Learned:

This project demonstrated the complexities of working with real-world datasets. From reconciling differences in data granularity to managing cloud storage, each step required careful

planning and execution. The hands-on experience with ETL pipelines, data transformations, and visualizations provided us with valuable technical skills, including the ability to adapt and troubleshoot when things did not go as planned.

Effective team coordination proved essential for this multi-faceted project. While data science often allows for asynchronous work, we discovered that key decisions, such as research question refinement and dataset alignment, benefited from group discussions. Allocating tasks based on individual strengths helped us work more efficiently and meet deadlines.

Looking ahead, we could approach similar projects with greater emphasis on alternative metrics and data sources. For example, comparing COVID-19 case trends to the number of homes listed on the market might better capture the pandemic's effects on housing activity. Additionally, exploring other data providers or county-level housing data could provide more granular insights. We could also experiment with advanced tools for integrating datasets with differing granularities, such as using interpolation techniques for time-series data. These enhancements would enable us to tackle more complex research questions and extract richer insights.

Skills gained and areas for further development:

This project enhanced our proficiency in Python, particularly in data manipulation and visualization with pandas and matplotlib. We also developed skills in cloud storage management and ETL pipeline design. Areas for further growth include refining our ability to interpret unexpected results and incorporating more advanced statistical analyses into our workflows. By completing this project, we gained a deeper understanding of the challenges and rewards of working with real-world data, preparing us for future endeavors in data science.