Team Leftovers - Progress Report 2
Project Title: “Geospatial Analysis of Housing Prices: A Crime and Amenities Perspective”

Project Description

Our project looks at how neighborhood amenities influence property values, with a special attention to the influence of crime statistics. This leads to the core research question: **How do the supply and the price of local amenities affect housing prices?** We are also still exploring: **How much of a difference do various types of amenities make at offsetting the adverse impact of crime on property values?**

Progress & Changes:

- We have successfully collected and cleaned data from multiple sources, including Redfin, Google Maps Places API, and Chicago Police API.
- Zhenning pointed out that the initial Redfin data included parking lots and vacant lots, which were subsequently removed to ensure data accuracy. Andrew updated the CSV file and moved it to the `redfin_data` folder.
- Marija, while conducting statistical analysis, shifted from using zip codes to latitude and longitude for more precise geospatial matching. This change has significantly improved the accuracy of our analysis, as it allows for more granular and accurate mapping of amenities and crime data to property locations.

Next Steps:

- Complete the integration of all datasets (amenities, crime, and property values) for a comprehensive analysis.
- Ensure all data is geocoded correctly for spatial analysis.

Data Sources

Progress & Changes:

- We have successfully scraped and integrated data from Google Maps Places API, Redfin, and Chicago Police API. We stop collection on other data sources as we have plenty of data we need.
- Andrew utilized md.dhr.wtf to return website pages as markdown files. This was a necessary step, as Redfin does not allow scraping of their search pages. From the search pages, ahrefs were pulled for each real estate listing. These ahrefs were then fed into scraperapi, which returned all relevant data for each real estate listing.

Next Steps:

- Cross-validate the data from Yelp and Google Maps to ensure accuracy and consistency.

Data Cleaning and Wrangling

Progress & Changes:

- We have standardized and combined property value data from Redfin.
- Zip code level amenity data, along with unique identifier pieces, have been aggregated.
- We have processed crime data by geospatially matching it with zip codes and cleaning missing and duplicate data.
- We have started the process of counting the number of amenities and crimes around each house using latitude and longitude coordinates, which will be joined with the redfin house data in “summary_redfin.csv”, in an effort to have a master csv file that contains everything for analysis.
- Marija has also implemented a different approach to count the number of amenities around each house using latitude and longitude, which has improved the accuracy of our geospatial matching.

Next Steps:

- Compare the relative costs of amenities and property prices to normalize costs.
- Verify that all datasets have been cleaned and prepared for the analysis; no missing or double entries. We will also pay particular attention to inferring values for square feet as we will go through some outliers.
- (Counts of each type of crime — theft, battery, etc.) could also be posted. The number of total crimes in just the count only at this stage.
- Finalize the geocoding process to ensure all data points are accurately mapped.
- Reorganize code into functions and restructure the repo.
- `price_per_sq_ft` data needs to be populated. Zhenning and Andrew will develop a method to average nearby sq. ft. pricing and use that to populate empty `price_per_sq_ft` cells.

Data Analysis and Visualization

Progress & Changes:

- We have also started some preliminary correlation analysis to test relationships between amenities, crime, and property values.
- Marija has been working on geocoding the datasets and conducting spatial analysis using latitude and longitude instead of zip codes, which has improved the precision of our analysis.
- Andrew has begun statistical analysis and regression modelling entries to see if any relationships can be found between variables and data sets.

Next Steps:

- Perform the correlation analysis and multiple regression modeling to quantify the effects of amenities and crime on property value.
- Generate heatmaps of amenity density and crime data, scatter plots of amenity levels vs. property values, and regression plots of the relationships between crime, amenities, and housing prices.
- Complete spatial analysis to demonstrate trends in Chicago zip code/ coordinates

Responsibilities

Progress & Changes:

- **Andrew**: Scraped and cleaned real estate pricing data from Redfin and Realtor.com. Hosted the repository and ensured data consistency.
- **Marija**: Conducted geocoding using latitude and longitude, and began correlation analysis, completing the correlation matrix analysis and scatter plot relationship analysis.
- **Zhenning**: Using Google API to scrape Amenities data, CPD API for crime data, and counting amenities and crime around each house.

Next Steps:

- **Andrew**: Continue cleaning Redfin data. Finalize the integration of all datasets and ensure the repository is up-to-date.
- **Maria**: Complete the geocoding and spatial analysis, and finalize the correlation analysis.
- **Zhenning**: Further data cleaning, graphical exploration and finalizing the statistical models.