# Project-1-EDA-Git-Collab

Project 1 of the NU Data Analytics

Proposal: To use the Cook County Assessor's Office's historical data to track potential effect of policy changes on rents or home values, comparing the county assessed values with Zillow's dataset's rental prices. We think finding a sample of zip codes might help us focus on particular patterns.

After looking at the assessor's office, we think we might need another dataset, so we will start with the American Community Survey for property values.

topic and why you chose it: This topic combines public policy and real estate transactions the datasets seem available, and it is a good narrative. Our hypothesis is that the new policy at the assessor's office had a material impact on stabilizing property rents.

initial questions for data (at least 3)

1. Are the assessor's office dataset and the Zillow datasets compatible?
2. Is using zip codes as a basis for comparison feasible?
3. Can our computers handle the volume of data, its 41million rows.
4. What will we use to match the values of property? Property identification vs. address?
5. Since the real estate market has been so hot in recent years, will this hypothesis be noticeable?

initial data sources (at least 2):

<https://www.census.gov/programs-surveys/acs/data/experimental-data/1-year.html>

<https://data.nasdaq.com/databases/ZILLOW/data>

<https://datacatalog.cookcountyil.gov/Property-Taxation/Assessor-Historic-Assessed-Values/uzyt-m557>

GROUP MEMBER-

|  |  |
| --- | --- |
| **Henry** | **Andrew** |
| Pandas Guru | GitHub Wagonmaster |
| Matplotlib Expert | Master of Ceremonies |
| API Enthusiast | Project Manager Carny |
| Statistician Extraordinarie | Presentation Barker |

Data acquisition and cleaning

Process of finding data

1. Are the assessor's office dataset and the Zillow datasets compatible?
   1. Not very. The assessor’s office had over 41million rows of data, it covered all of Cook County. It was raw numbers of individual parcels, so it contained a lot of information. Too much for our needs, frankly. Meanwhile, the Zillow Observable Rent Index, (ZORI) was a pre-weighted dataset based on zip codes of the top 100 metropolitan areas across the country. ZORI goes back to 2014. According to Zillow, they observe rent over time on the same rental unit then aggregate properties repeatedly listed for rent. Our two datasets were very different in scale and scope.
2. Is using zip codes as a basis for comparison feasible?
   1. We chose using zip code as a common denominator to focus our project’s scope, even then we had issues narrowing our focus. We agreed to limit the timespan of the project to 2016 forward. The API of the Cook County Assessor’s Office (Assessor’s) still returned hundreds of thousands of parcels for a specific zip code and took a substantial time to query. Ultimately, we set a cut off at 20,000 parcels per zip code and picked the total assessed value for the land and buildings as our key metric. But that left us with 120,000 total parcels in that dataset. Meanwhile, in the Zori dataset 2418 rows for the whole country and 105 columns. After selecting the zip codes and dates, we only had about 6 rows and about 80 columns of data.

Process of acquiring data

Process of cleaning data

1. Can our computers handle the volume of data, its 41million rows.
   1. **HENRY CAN COMMENT**
   2. The API queries were very time consuming so ultimately, we exported the API results to csv files and used that.
2. What will we use to match the values of property? Property identification vs. address?
   1. Neither worked, because the ZORI data was based on zip codes. Metaphorically, we had to find a way to walk the two datasets towards each other to find a scope that we could use. Even in the final analysis we had to find a way to compare about 6 rows and 80 columns with thousands of individual parcels. **TO BE CONTINUED**
3. Since the real estate market has been so hot in recent years, will this hypothesis be noticeable?
   1. There are signs of inflation in the ZORI database as it is more dynamic and subject to short term flux, like inflation. Graphically, we can see similar changes.

Link to final version of data cleaning jupyter notebook in main branch on repo

**Analysis and Conclusions**

Within our timeframe and selected zip codes, rent over time increased. There is a sin wave-esque aspect to the individual rent trend, this is likely attributable to the natural rent cycle in the city, the lower quality and cheaper rent apartments are available in the colder months, with the market peaking in May and June of each. But year over year the peak rents increased. There is a dip that is pronounced during the pandemic, 02/2020 -06/2020. The charts below are organized from lowest range of rent to the highest range of rent. The lower rent zip codes had a fair number of fluctuations over a tighter range, around $250 dollars. While the more expensive areas had a higher amount of fluctuation within a larger range, about $400. The steepest decline happened in the pricier rent zip codes, but it also had the fastest recovery and steepest return to rent. In the overall zip code comparison, the lowest rental zip code is relatively static and the priciest is much more dynamic in terms of peaks and valleys.

A picture containing chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

The ANOVA analysis below demonstrates the variance of rent among the observed zip codes. There are two interesting points here. First, zip codes 60618 and 60647 are neighboring zip codes. It makes sense that the high range of the 60618 zip code would be equal to the mean rent of the more developed, more commercially rich neighborhood to the east, 60647. Second, 60624 and 60614 are two very different neighborhoods with two very different stocks of rentals. Further investigation would warranted as to the square footage of units available.

Chart, box and whisker chart

Description automatically generated

Shape

Description automatically generated with medium confidence

text that explains the story of the visualization

analytical answers to proposal questions

statistical backup

link to analysis jupyter notebook in main branch on repo

Links and Resources

final data sources

link to presentation