Group #67 Final Report

Executive Summary

Compared to other industries, the US real estate market has been slow to adopt data analytics as a key component of their business structure. In 2021, the US real estate market was valued at approximately $20 trillion yet most companies still use antiquated data pipelines and excel “number crunching” as their primary analytical tool.

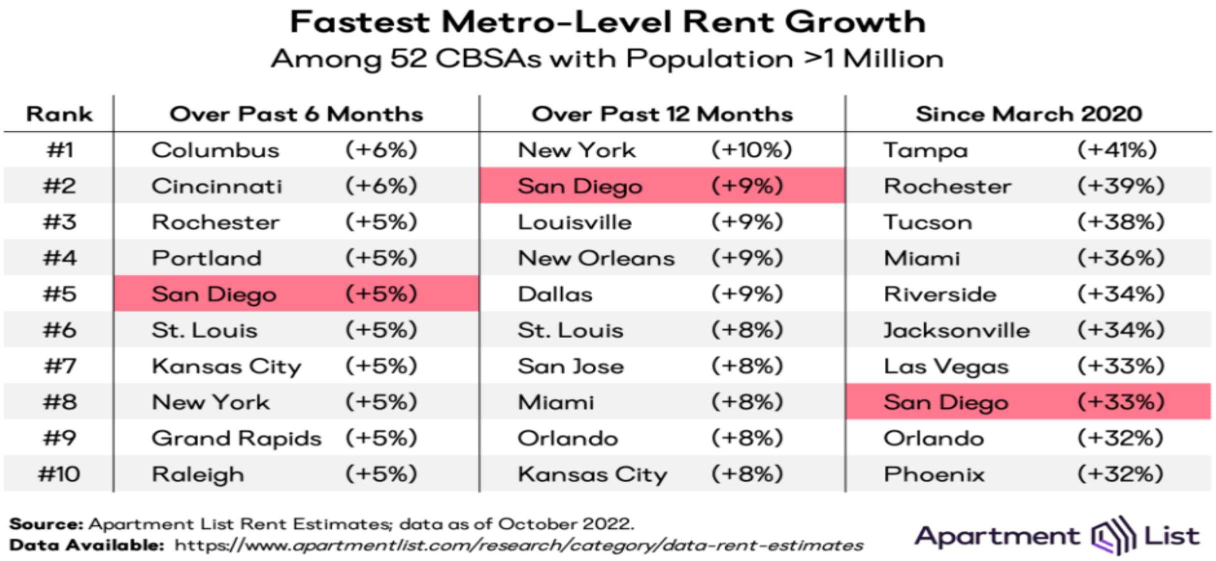
This report shows how analytics can be used in the real estate industry. More specifically, it focused on identifying the key factors that drive apartment rents. Using US census data, we have been able to find three key factors that drive rent: apartment size, population density, and occupancy rates. These factors can help real estate firms locate new projects and guide investment decisions.

Objective

The primary objective is to pinpoint the key demographic features within a population that have the biggest predictive power for rent prices in a given market.

Business Impact

The US apartment market, a subclass of the overall US real estate market, is estimated to be a $169.5 billion dollar industry based on revenue. Rent is the key economic driver of any apartment project. Market rents can vary greatly depending on location as shown in the chart below. The ability to identify profitable locations is critical to economic success of the project and to the investors.



Fastest Metro-Level Rent Growth
Among 52 CBSAs with Population >1 Million

Rank	Over Past 6 Months		Over Past 12 Months		Since March 2020	
#1	Columbus	(+6%)	New York	(+10%)	Tampa	(+41%)
#2	Cincinnati	(+6%)	San Diego	(+9%)	Rochester	(+39%)
#3	Rochester	(+5%)	Louisville	(+9%)	Tucson	(+38%)
#4	Portland	(+5%)	New Orleans	(+9%)	Miami	(+36%)
#5	San Diego	(+5%)	Dallas	(+9%)	Riverside	(+34%)
#6	St. Louis	(+5%)	St. Louis	(+8%)	Jacksonville	(+34%)
#7	Kansas City	(+5%)	San Jose	(+8%)	Las Vegas	(+33%)
#8	New York	(+5%)	Miami	(+8%)	San Diego	(+33%)
#9	Grand Rapids	(+5%)	Orlando	(+8%)	Orlando	(+32%)
#10	Raleigh	(+5%)	Kansas City	(+8%)	Phoenix	(+32%)

Source: Apartment List Rent Estimates; data as of October 2022.
Data Available: https://www.apartmentlist.com/research/category/data-rent-estimates

In this analysis, we’ve developed a model to help identify areas with favorable rent demand and potentially good areas for future apartment development.

Problem Statement

Real estate investment and development companies, such as Equity Residential REIT or AvalonBay Communities REIT, need to understand how the populations and demographics of a given area affect rental rates. Rent is the main driver of profitability for any rental property and apartments are especially sensitive to rent fluctuations. We have analyzed key census data information, namely population density and income, and how it affects apartment rental rates.

Understanding the Data

For our analysis, we are leveraging three main datasets:

- **2015 and 2020 Rent Price Data:** This data provides median rent prices at the census tract geographic level based on the Census. An overall median rent price value is provided, along with separate estimates for units with one through five or more bedrooms.
- **2020 Demographic Data:** This dataset contains age and population data that we can use to establish population composition. These will each be features in our model to predict average rent price within each market.
- **2020 Housing Characteristics:** This dataset contains information about the housing market including: total housing units, vacant housing units, and homeowner vacancy rate.
- **Additional Income and Population Density Variables:** We developed population density and average income variables because we believed these to be strong potential predictors of rent pricing. Income and total population data by census tract was available directly from US census data, while the geographic data of the census tracts was loaded into GIS software. Using GIS analysis, we calculated the land area of each census tract and joined the population, income and land area variables to estimate population density.

Data Cleaning and Transforming

Housing Price Datasets

The 2015 and 2020 Rent Price datasets provide median rent for rental units classified by number of bedrooms and grouped by geographical location (Census Tract). There are multiple types of non-numeric data in the estimated rent and margin-of-error columns, but the distinctions are irrelevant for our initial analysis, and we are simply removing those points from our dataset. Preprocessing of these datasets also included the decoding of the original cumbersome column names, and the relabeling of all data to under new, concise headings.

Estimated rent was initially provided under separate variables for each number of bedrooms. To use number of bedrooms as a variable in our regression we had to collect all the individual median rent figures for an area under a single column and create a series of dummy variables to represent the number of bedrooms. With 0 bedrooms as the initial base case, we created dummy variables br_1 through br_5. Initial data exploration revealed that many geographical areas do not have estimated rents for all six-bedroom categories. Most common were rent estimates for 1, 2, and 3 bedrooms, with roughly 47% of surveyed geographical areas having rent estimates for all three of these bedroom categories. We are limiting our initial analysis to this subset of our data.

It is worth noting that rental prices by no means follow a normal distribution, as can be seen in the Q-Q plot of the Estimated Median Rent. Even after log transformation there is still significant skew in the rent prices. We will need to be aware of this skew and closely watch the distribution of residuals when modeling.

Housing Characteristics Data

The raw housing characteristics data contained a total of 1,147 columns, and similar to the housing price data, the variable names were long and cumbersome. Many of the columns contained annotations or other information about the housing characteristic estimates that will not be used in the model. In addition, the raw data contained a large number of columns with percentage values, which were calculated based on other columns in the dataset. The annotation type columns and the percentage data type columns were removed from the data.

Once these columns were removed, there were 143 variables left in the dataset. Analysis of the columns in the data shows that most columns have no missing data, however 10 columns do have more than 500 missing data points:

These ten columns contain information related to:

- Vacancy rates
- Median number of rooms
- Average household size
- Telephone service availability
- Mortgage information
- Rent price information

The data in some of these columns is better represented by other columns in the dataset. For example, there are several columns with counts of the number of properties with numbers of rooms between 0 and 9, which provides more detail than the single column with the median number of rooms. For other columns, such as the rent price, we have information from other sources that is more valuable for our analysis. As a result, these ten columns were removed from the housing characteristics dataset.

Approach & Methodology

After cleaning and transforming the initial data sets, we hypothesized that population and income would be key rent drivers. In general, we thought that large populations may increase rent due to supply and demand dynamics and that household income would also play a role.

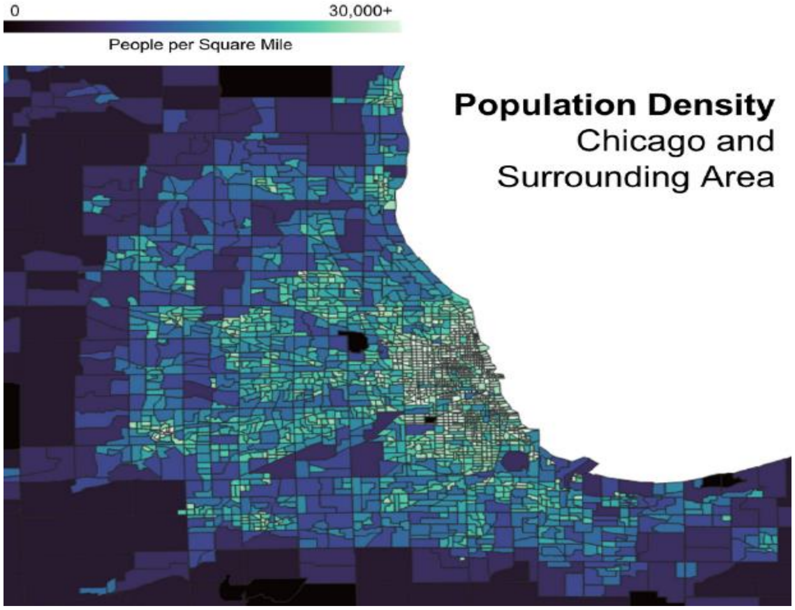
Initial Modeling

The first model we tried was a linear-log regression model with all available variables which had an adjusted R-squared value of 65%. This is a decent model, over performing random guesses, but we continued to iterate over it to make it more robust. Our approach was to start off with all available data and then do a feature selection process to identify our most important variables. We also looked for additional data that could help us explain more of the rent price variance.

Adding Population

To make our model more robust, we created population density and average income variables because we believed these to be strong potential predictors of rent pricing. Income and total population data by census tract was available directly from US census data, while the geographic data of the census tracts was loaded into GIS software.

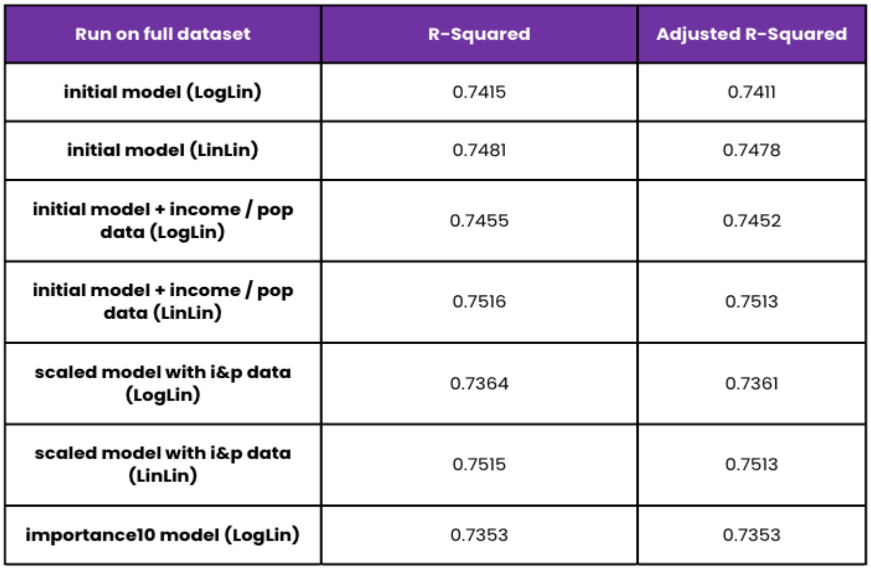
Using GIS analysis, we calculated the land area of each census tract and the population to extrapolate the population density of each tract. Population density was calculated by dividing the population by the land area to get the number of people per square mile. A heat map of a Chicago area is shown below as a sample.



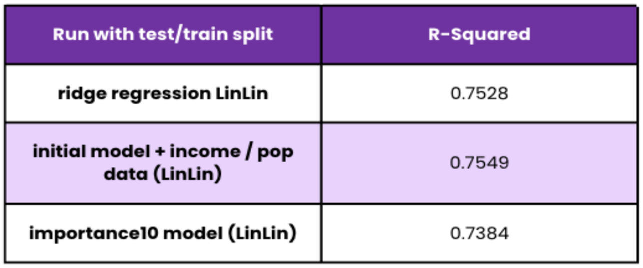
Testing Various Models

Our initial assumption was that we would need to log-transform the rent price, as it did not follow a normal distribution. However, during our feature analysis we tested each model with both linear-linear and log-linear methodology, and the linear-linear model routinely had a higher R-squared value. In light of this, we went with a linear-linear regression for our final model.

In addition to standard linear regression models, we trained and tested a ridge-regression model. We were concerned about the potential for high collinearity in our data, and thought that a ridge-regression model would reduce the likelihood of overfitting our model. However, the ridge-regression model performed slightly lower than our best straight linear regression model, and our final model is a standard linear regression. If the project were to continue, our next step would be to run a thorough k-fold validation of both models to be absolutely certain which model performed better.



Run on full dataset	R-Squared	Adjusted R-Squared
initial model (LogLin)	0.7415	0.7411
initial model (LinLin)	0.7481	0.7478
initial model + income / pop data (LogLin)	0.7455	0.7452
initial model + income / pop data (LinLin)	0.7516	0.7513
scaled model with i&p data (LogLin)	0.7364	0.7361
scaled model with i&p data (LinLin)	0.7515	0.7513
importance10 model (LogLin)	0.7353	0.7353



Run with test/train split	R-Squared
ridge regression LinLin	0.7528
initial model + income / pop data (LinLin)	0.7549
importance10 model (LinLin)	0.7384

Improved Model

The model’s R-squared improved to 75.49% after adding the population density. This means that our model accounts for more than three quarters of the variability in our response variables.

Conclusion and Key Takeaways

Our initial objective was to show that analytics can be used to pinpoint key demographic features that have the biggest predictive power for rent prices. We have been able to determine the following relationships:

1. **Apartment Size:** Knowing whether an apartment has 1-3 bedrooms will help predict rent.
2. **Population Density:** Densely populated areas have higher demand which increases rent pricing.
3. **Occupancy Rates:** When more units are occupied, the scarcity of available units can drive rents higher.

Based on these factors, we should suggest that real estate companies specializing in apartments should 1) focus on identifying areas with high population density and good population growth 2) construct apartment units with primarily 2-3 bedroom as they tend to have a higher average rental price and 3) avoid areas with low occupancy rates and an over-supply of apartment units and this will drive down rental rates.

References

- Bureau, US Census. “Census Datasets.” Census.gov, 12 Sept. 2022, https://www.census.gov/data/datasets.html.
- Team, Apartment List Research. “Apartment List National Rent Report.” Apartment List - More than 5 Million Apartments for Rent, Apartment List, 31 Oct. 2022, https://www.apartmentlist.com/research/national-rent-data.
- Zumper. “Zumper National Rent Report.” The Zumper Blog, 25 Oct. 2022, https://www.zumper.com/blog/rental-price-data/.