ggplot3: Jordan Chiantelli-Mosebach, John Martin, James Winkeler

We sought to understand the behavior of rent prices and their correlations within the data set. The rent price data is provided as the average per market per quarter, with separate values for the high-tier properties and the non-high-tier properties. This is such that there is exactly one averaged rent price in the data set for the high-tier properties sold in Chicago in Q1 of 2018, one for those in Q2 of 2018, etc. As a result, when exploring its relationship with other variables, we assured that they were also averaged like this. Furthermore, for the sake of comparison across markets, rents were normalized according to the formula below, where $rent_q$ is a given average rent and $rent_0$ is the rent for that property type in that market in Q1 of 2018. This provides a percentage change from the baseline value for each quarter.

$$100 * \frac{rent_q - rent_0}{rent_0}$$

This new normalized variable is approximately normally distributed, skewed right, with a range of -12.2% to 62.9%, a median of 9.3%, and a mean of 11.2%. To determine its relationship to other variables, we employed a Random Forest model using the function randomForest() from the package of the same name. The function tuneRF() was used to show that the out-of-bag error is minimized when mtry = 4, which was then used in the creation of a Random Forest model. Our input variables were the Market (factor), the Time (numeric, by quarter), the Availability Proportion (numeric), and the Internal Class (factor, whether the lease was high- or non-high-quality). This model was trained on a subset of 70% of the data. Using it to predict the rent prices for the remaining 30% of data showed a high level of success, with the model explaining 87.3% of variance in predicted residuals. It also showed that the Market variable was the most important variable in the model. It's removal would result in a 135.4% increase the model's MSE, effectively tied with the Time variable, and it is has the most impact on node purity. This suggests that investigations into individual market patterns may be particularly fruitful.

To this end, we performed time series analyses, which showed that in most markets, the "O" market showed much stronger seasonal patterns in rent than the "A" market. San Francisco, Baltimore, and Manhattan did not display these patterns, as in general, their patterns were very distinct from the other cities. In general, a standard ARMA model of some sort was appropriate for the premium market. For example, for Los Angeles, an ARIMA (3, 1, 0) model was used. However, in the non-premium market, seasonal terms were pretty much always necessary. In the LA market, a SARIMA (1,1,1) x $(0,0,1)_4$ was needed. This pattern of utilizing a seasonal term for the "O" market and not for the "A" market repeated itself across the majority of markets.

Furthermore, we observe trends for premium rent subtracted by non-premium rent by market across time. This analysis was done in Stata, mostly incorporating collapse and two-way functions. We highlight San Francisco and Los Angeles, two cities with similar conditions and the same state, and observe little relationship. Total leases by region, however, had clear correlation across time, with observed shock response and shock recovery to the COVID-19 pandemic.