Harvard
Undergraduate
Data Analytics

HDAG-Greystar Final

Director of Consulting: Mustafa Ansari
Case Team Lead: Yao Yu
Associate: Jay Pratap
Analysts: Eli Kirtley, Evan Jiang, Jason Tang, Lauryn Wu

Yao Yu

Jay Pratap

Eli Kirtley

Evan Jiang

Jason Tang

Lauryn Wu

Goals

1. Create a **correlation matrix** with Greystar's internal data to determine the potential predictability of variables
2. Identify variables and **derived variables** that offer true **predictive value**
3. Prescribe **optimal pricing** for each property based on identified variables with predictive power
4. Prescribe **enhancements** to specific properties and/or identify properties with highest potential for enhancements

Goals

1. Create a **correlation matrix** with Greystar's internal data to determine the potential predictability of variables
2. Identify variables and **derived variables** that offer true **predictive value**
3. Prescribe **optimal pricing** for each property based on identified variables with predictive power
4. Prescribe **enhancements** to specific properties and/or identify properties with highest potential for enhancements

Summary of Greystar Data

- **MSA:** Denver-Aurora, CO
- **Total Number of Properties:** 131 Properties
- **Total Number of Units:** 18,766 Units
- **Time Period:** Jan 1, 2015 – Mar 24, 2022
- **Format:** comma-separated values
- **Topics Covered:**
  - Overview of Properties
  - Lease Activity at the unit level
  - Daily status of units
  - Market unit price history
  - ILS Prospect Data
  - Kingsley Survey data from tenants

Executive Summary

- **Total units occupied** per property maintains a correlation with the **market rental price** after updating the correlation matrix with the full historical data (2015-Present)
- More interesting insights about **reasons for moving** in the Denver-Aurora-Lakewood region from American Housing Survey
- Time series analysis with the full historical data shows a strong **winter and summer seasonality boost** in market rental price
- **Lease likelihood** model from prospect activity data shows **87% accuracy** on test data
- Pricing model from internal variables, derived variables, and external variables shows strong potential as a foundation to build upon

Agenda

Midpoint Summary and Updates	Summary of midpoint findings and updates made since	▪ Updated correlation matrices ▪ Final variables pulled from American Housing Survey ▪ Time series with seasonality trends
Sentiment Analysis from Kingsley	Summary of findings from sentiment analysis of Kingsley survey data	▪ Natural Language Processing Sentiment Analysis ▪ Results in a derived variable for “sentiment” of tenants per property that is used in pricing model
Lease Likelihood and Pricing Model	Lease likelihood model (Logistic) and Pricing model (OLS)	▪ Lease likelihood model from prospect activity data ▪ Pricing model from internal variables, derived variables, and external variables

1. Midpoint Summary and Updates

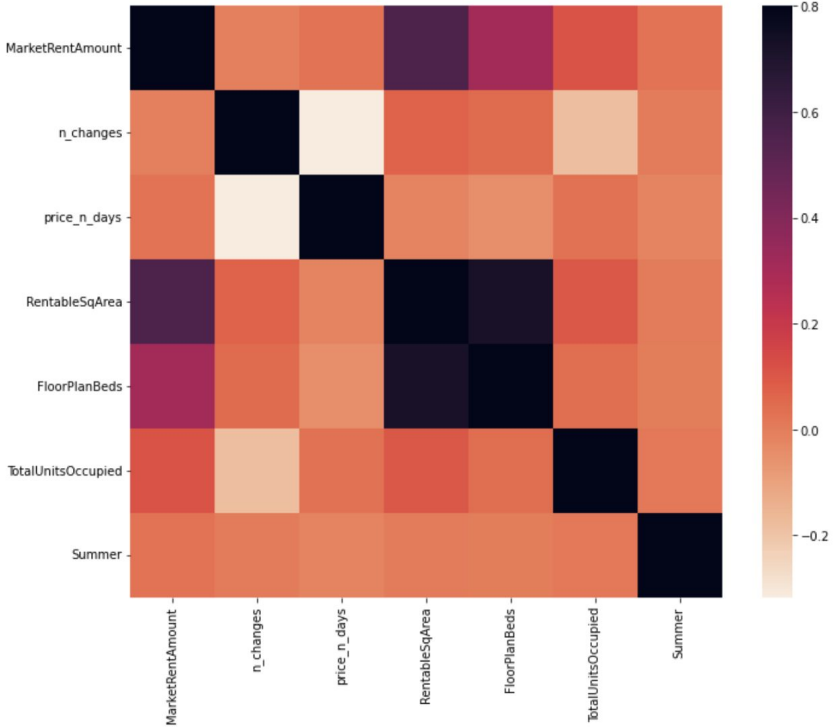
Outline:

1. **Updated correlation matrices**
2. **Final variables pulled from American Housing Survey (AHS)**
3. **Time series with seasonality trends**

Goals

1. Find interesting variables from all historical data (2015-Present)
2. Select final variables from AHS for the final pricing model
3. Find seasonality trends from all historical data (2015-Present)

Correlation matrix with historical data

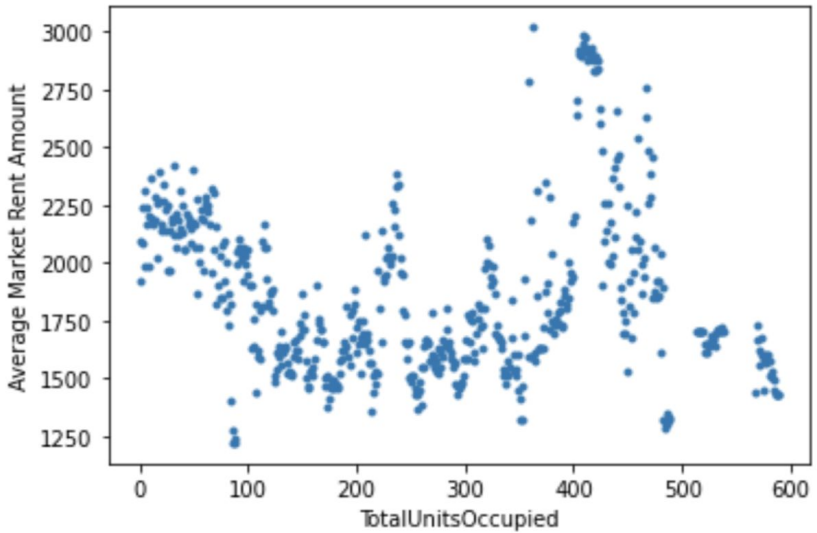


Observations

- Derived variables of interest include **price_n_days** (number of days a price was unchanged on the market), **TotalUnitsOccupied** (total number of other units occupied within a given unit's property), and **season** (season during which the price is being set).
- As expected, MarketRentAmount has **high correlation with static variables** like square footage and number of bedrooms.
- MarketRentAmount also has promising **moderate correlation with TotalUnitsOccupied and season**—in the following slides, we explore the relationship between these variables and market rent.

Relationship between total units leased and average unit market rent

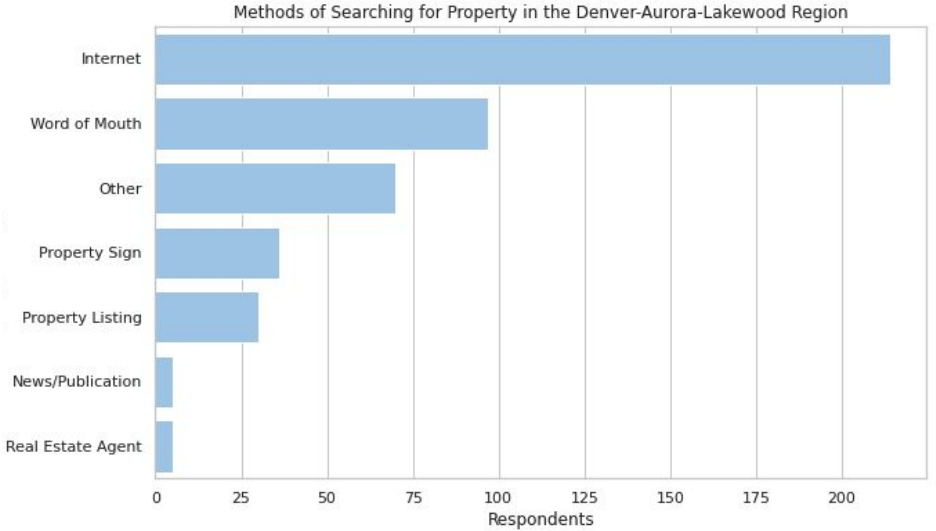
Avg. Unit Market Rent Amount vs. Total Units Occupied



Approach, Observations, and Takeaways:

- The scatter plot shows **relatively high rent when few units are occupied**, with a general downwards trend until ~200 units are occupied. One explanation is that as more units are already occupied, **demand decreases** and the **quality of remaining units decreases**—both drive rental price down.
- A return to higher rents around the ~400 unit mark may be due to rental prices of larger properties with **inherently higher quality and rental prices.**
- We also seek to relate market rent with **proportion of units occupied,** which may offer a more normalized metric on supply/demand for a given property. Our notebook contains the relevant code, but we ran into computing issues that prevented us from displaying the results here.

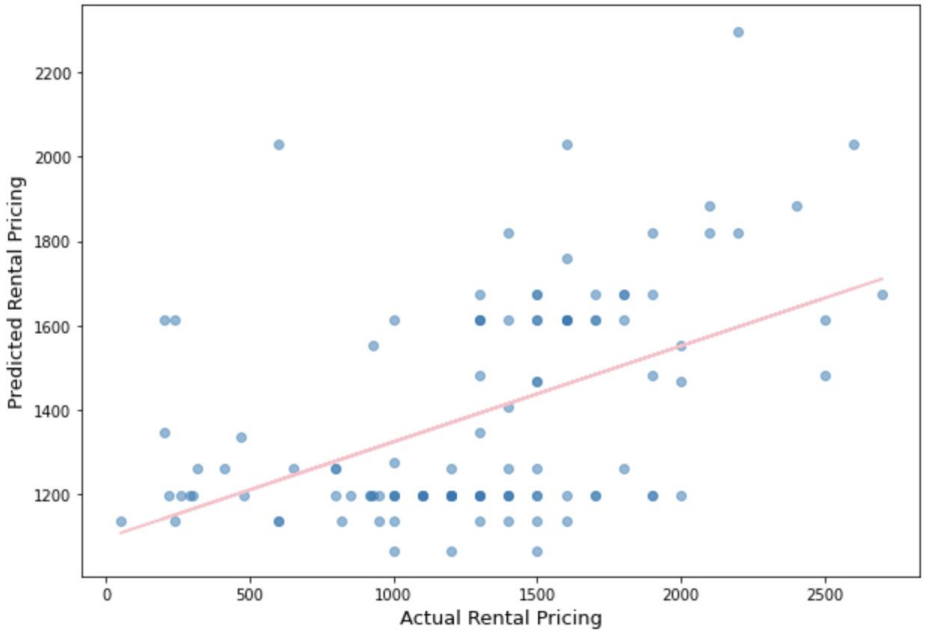
Reasons for previous residence exit and means of finding new apartment



Method and Insights

- Migration to Denver-Aurora-Lakewood for a **home quality upgrade and neighborhood are predominant factors**
  - Reasons for migration **correspond with the trends of rising socio-economic standard** in the region and its **rising housing prices**
- **Internet and word-of-mouth** are overtaking the traditional methods of sourcing for new apartments
- Over **80% of respondents** report that their new home is the same/better than their old home and over **82% of respondents** report their new neighborhood is the same/better than before

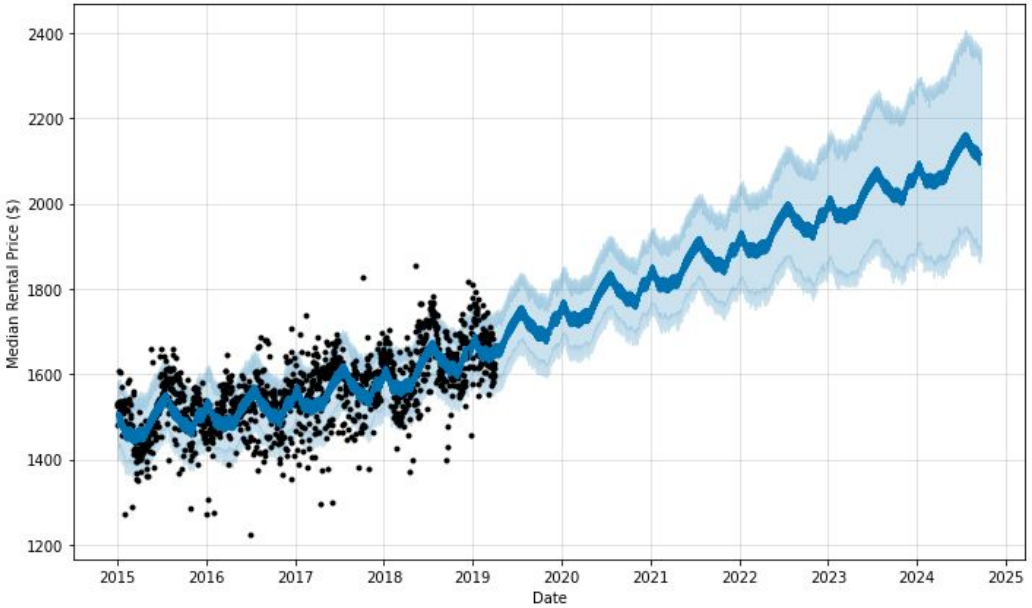
Preliminary OLS Pricing Model using data from the American Housing Survey



Method and Insights

- Filtered AHS for the following factors:
  - Numeric variables include **rent, total_rooms, bedrooms, and bathrooms**
  - Categorical variables include **good_schools, good_public_transportation, petty_crime, and serious_crime**
- Dropped outliers and total_rooms, bedrooms, and good_schools to avoid multicollinearity
- With **R^2 of 0.856** and **332 data points**, the coefficients from OLS include
  - **Bathrooms: 394.8**
  - **Good_public_transportation: 597.0**
  - **Petty_crime: 28.3**
  - **Serious_crime: -35.7**

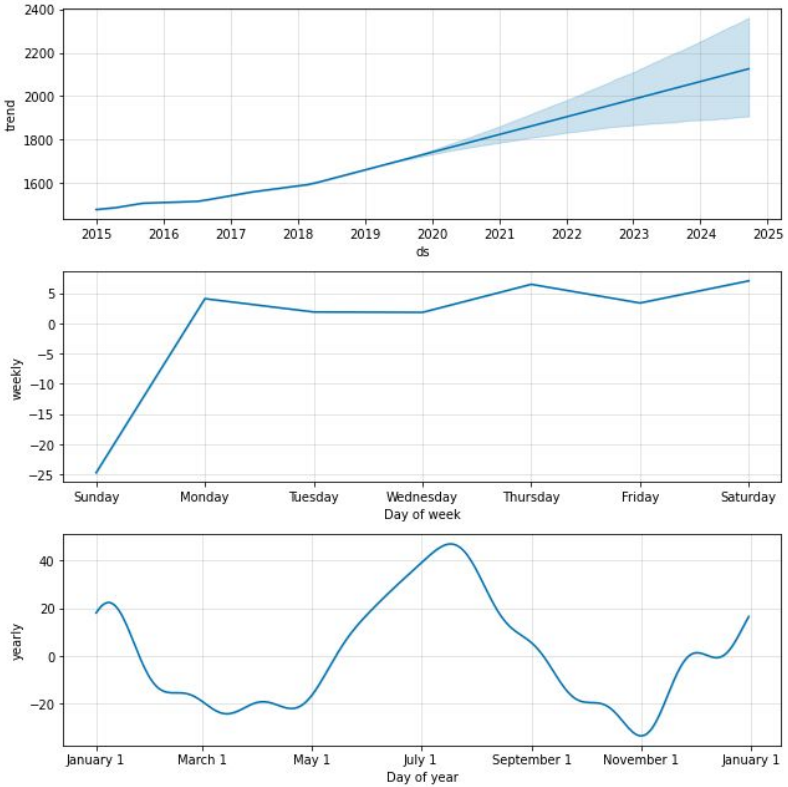
Median Rental Prices Time Series Model



Historical Trends and Forecast

- Predictions from the median of rental prices from 2015 to mid-2019
- Removed data from Covid-19 to obtain a long term forecast
- Overall upward trend
- Would recommend a more dynamic model to account for Covid
- Used Prophet procedure

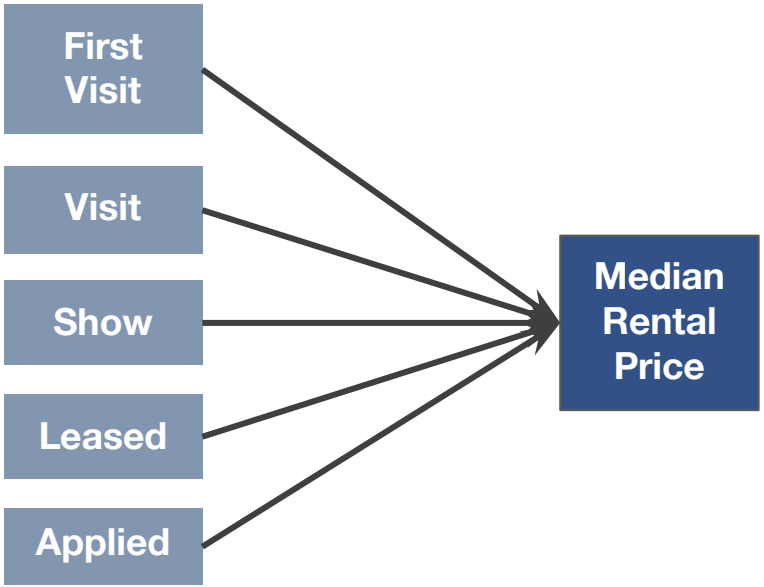
Median Rental Prices Time Series Model Decomposition



Forecast Components

- Broken down into trend, weekly seasonality, and yearly seasonality
- Overall trend of future market price
- Consistent weekly pricing with drop on Sunday
- Peak in late summer and second peak in January

Prospect Activity and Rental Prices Granger Causality Testing



Causality Test Results

- Analysis of relationship between aggregated prospect activity and median rental prices
- Numerous statistically significant cause-effect relationships
- First visit, visit, show, leased, and applied factors “Granger Cause” the rental prices
- All contact types “Granger Cause” a lease, except for first contact text, first contact self-guided tour, and first contact Internet

2. Sentiment Analysis from Kingsley

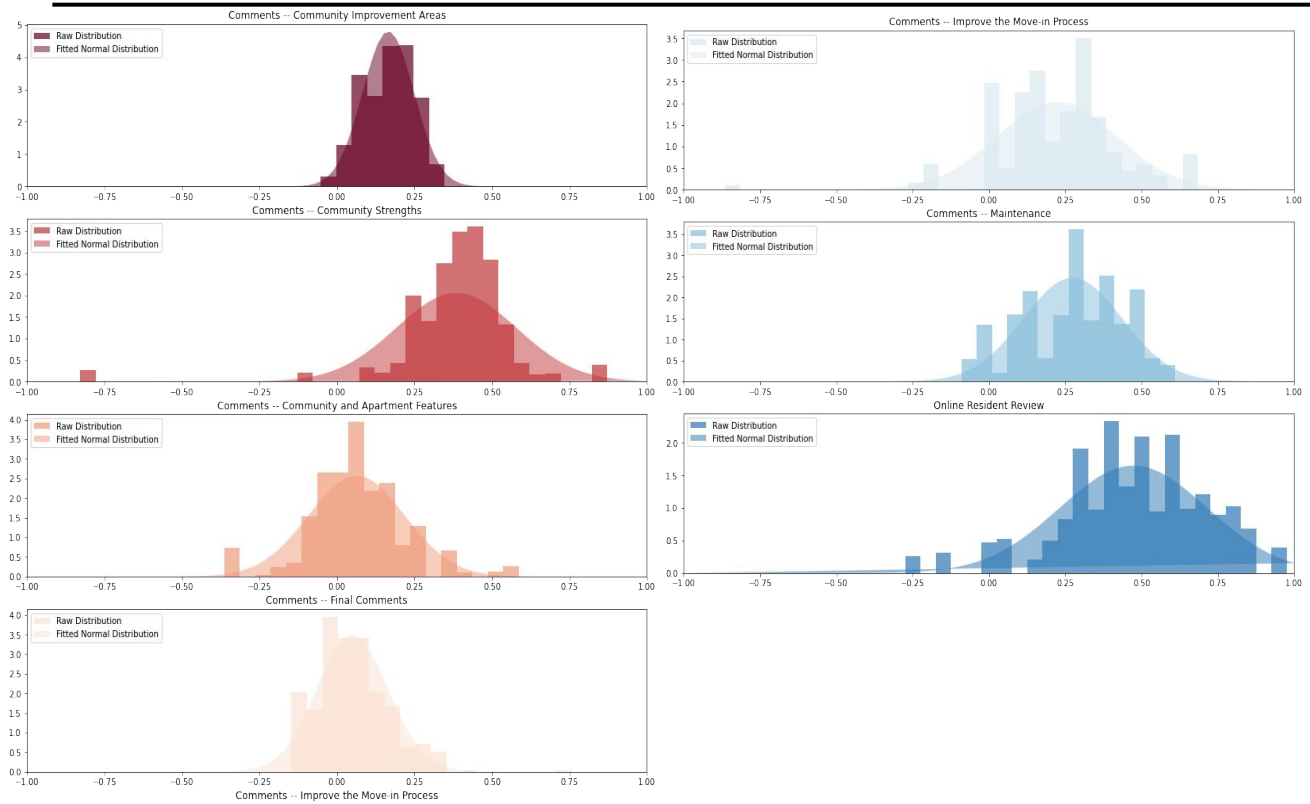
Outline:

1. **Sentiment Analysis from Kingsley survey data**
2. **Distribution of sentiment scores**
3. **Principal component analysis for relevant survey questions to predict market price**

Goals

1. Quantify tenant feedback for statistical analysis
2. Create derived variable from tenant “sentiment” for pricing model
3. Understand tenant feedback at the property level

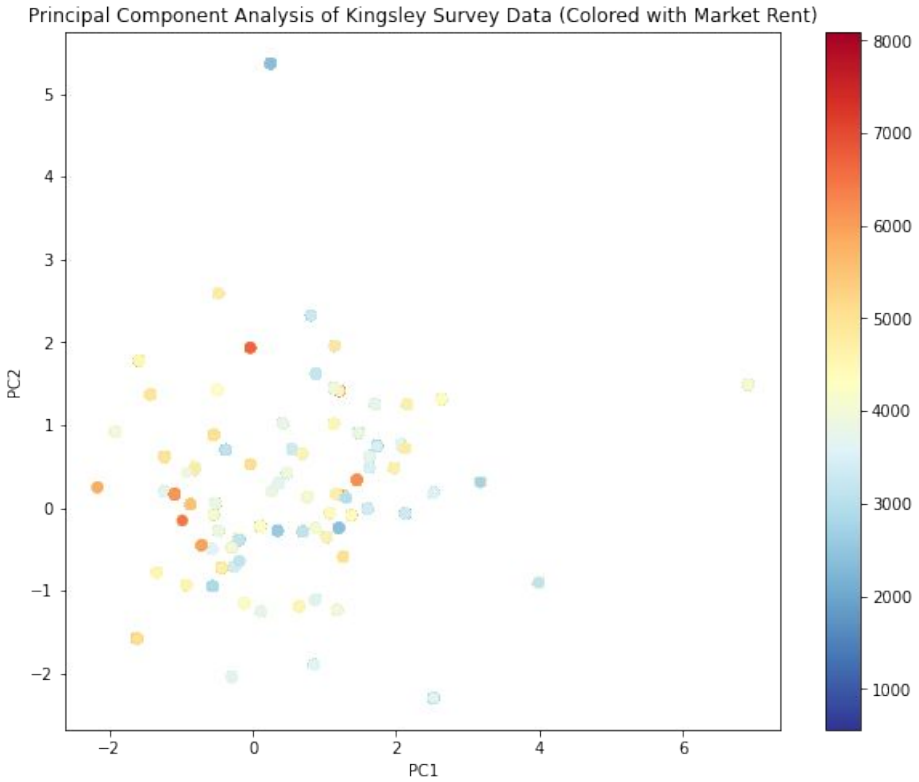
Sentiment Analysis of Kingsley Survey Data



Sentiment Score Distributions

- Performed sentiment analysis (positive, negative, neutral) on Kingsley survey responses
- Scores between -1 and 1 generated for each survey response
- Find top candidates to explain variance in market rent
- The density distributions of the sentiment scores are shown left, for various questions within the Kingsley survey data

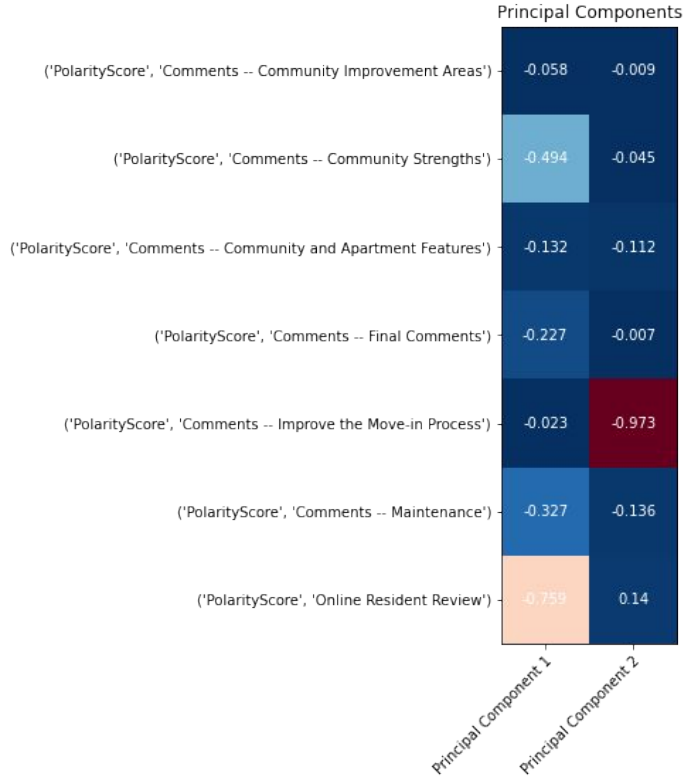
Principal Component Analysis of Kingsley Survey Data



Trends in Principal Component Space

- Perform dimensionality reduction with the sentiment score data from the survey questions
    - Extract two axes (principal components) explaining the variance of the sentiment score data
- We observe a **potential negative trend between PC1 and market rent** – what scales with PC1?

Principal Component Analysis of Kingsley Survey Data, cont.



Top survey questions for market rent prediction

- As seen in the previous slide, principal component 1 (PC1) scales negatively with market rent
- All entries of PC1 are negative, indicating that **positive sentiment in all survey questions is associated with a higher market rent**
- The highest magnitude is given to **online resident review**, followed by **community strengths**
    - These are potentially the most influential factors on market rent with respect to prospect/resident sentiment

3. Lease Likelihood and Pricing Model

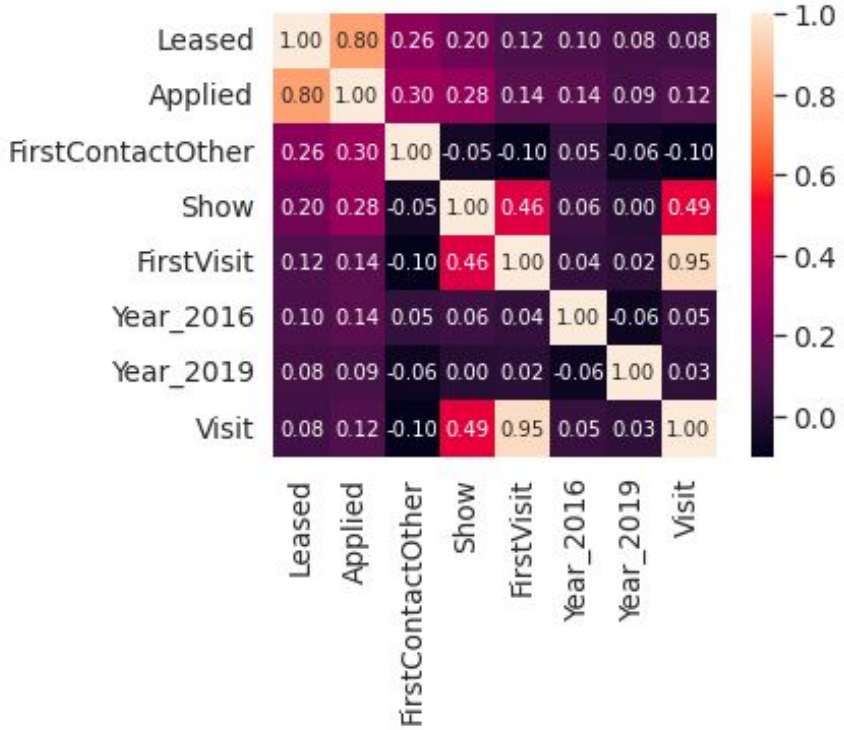
Outline:

1. **Lease likelihood model from prospect activity data**
2. **Pricing model from internal variables, derived variables, and external variables**

Goals

1. Predict likelihood to lease from prospect data
2. Predict future pricing of generalized unit characteristics
3. Build a foundation for deeper analysis in the future

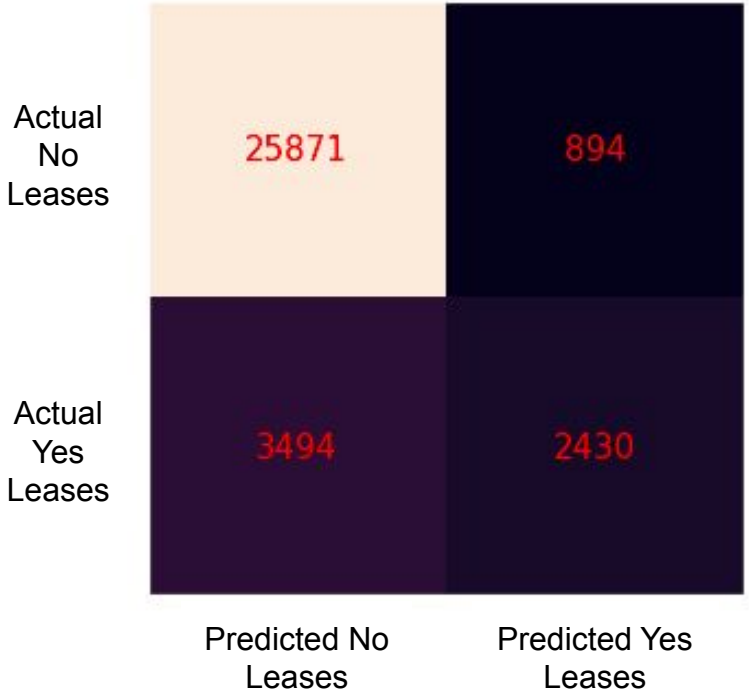
Preliminary Analysis of Lease Likelihood



Correlation Matrix of Top Factors

- Aggregated prospect activity data by the prospect
- Correlation matrix of the top 8 features contributing to a lease
- Applied, first contact other, and show are most highly correlated with leasing

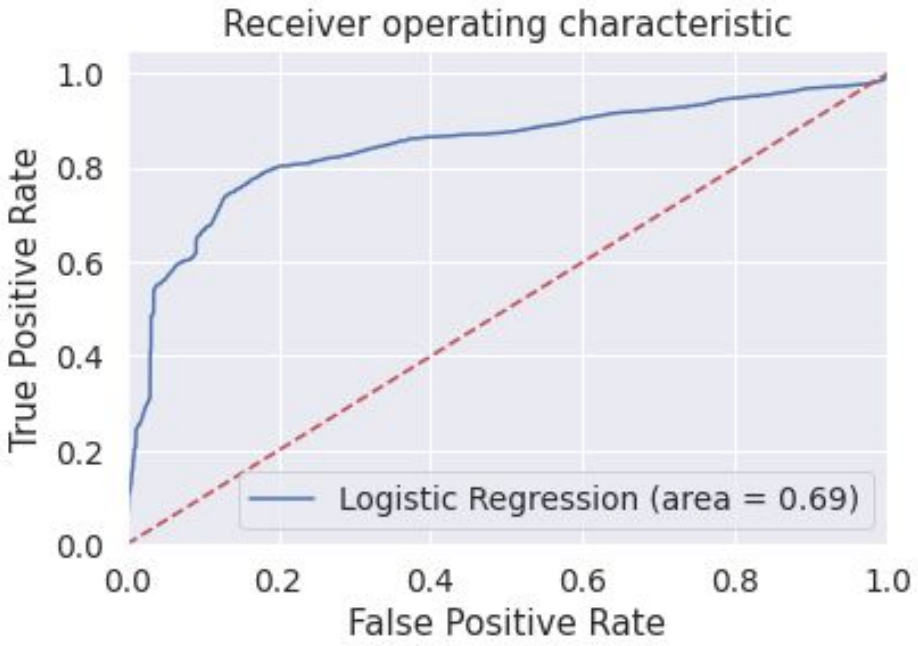
Logistic Regression for Probability of Application to Lease



Model Fit and Evaluation

- Binary logistic regression model to predict the probability of whether a prospect applies for lease
- Predictor variables from prospect activity data, aggregated by prospect
- 87% accuracy on test data
- Confusion matrix of the model

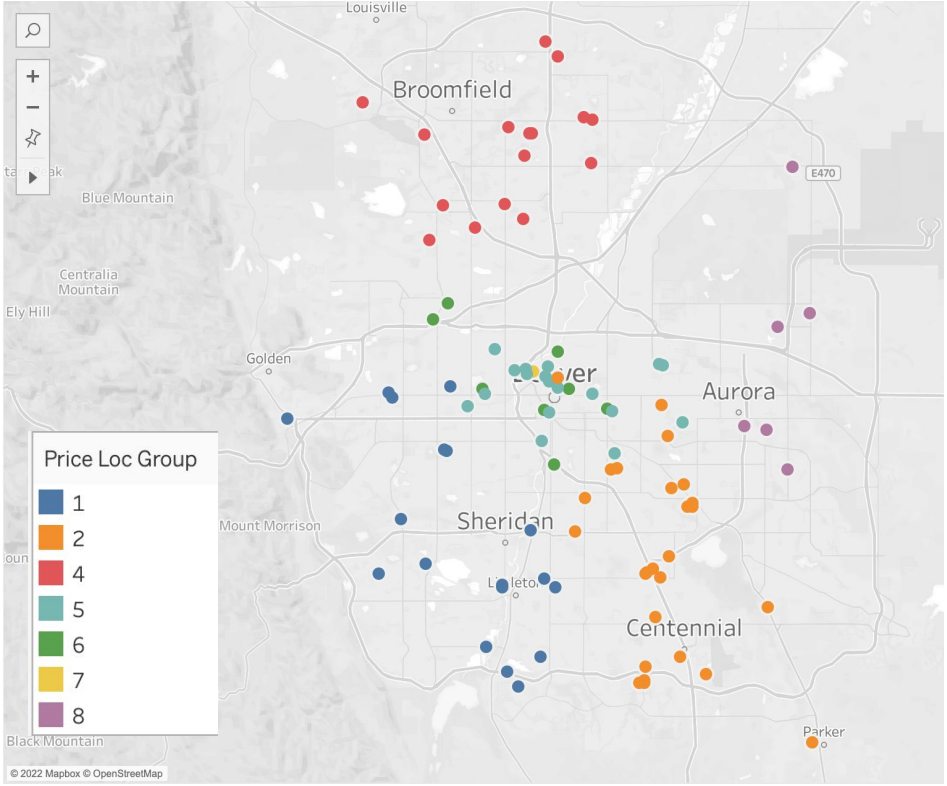
Logistic Regression for Probability of Application to Lease



Model Evaluation

- ROC curve shows the predictive power of the model
- Blue line shows the tradeoff between trying to maximize the true positive rate and minimizing the false positive rate
- Area under the curve is 0.69
- Future improvements: add prospect's income, whether a discount is being offered, and market competition variables

Clustering 94 Greystar-managed properties in the Denver-Aurora MSA



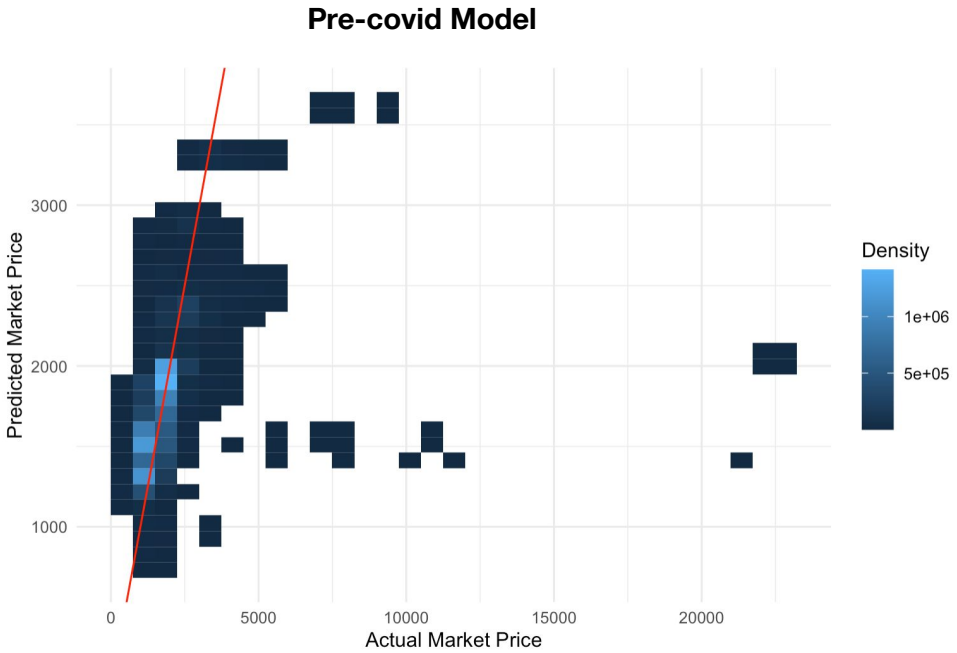
We cluster similarly properties to compare fluctuations in price among them with the EM algorithm

- Greater range of prices among properties as the number of bedrooms increases
- Most **expensive** properties located in the **downtown Denver** region
- Other potential clusters centered around Broomfield, Aurora, Sherida, and Centennial
- Final groups are properties that **share similar location and price per number of bedrooms**

Final Pricing Model Parameters

- **Type:** Ordinary least squares (OLS) linear regression model
- **Outcome:** Market Rent Price
- **Predictors:**
    - Number of Bedrooms and Bathrooms (unit component)
    - Total number changes made to the price
    - Total number of days that the price has been active
    - Year in which the property was built
    - Location cluster group (location component)
    - Number of Days since Jan 1, 2015 (time component)
    - During covid (Indicator variable for time period of COVID-19 - June 1, 2019 to June 1, 2021)
    - Season (Indicator variable representing seasonality trends)

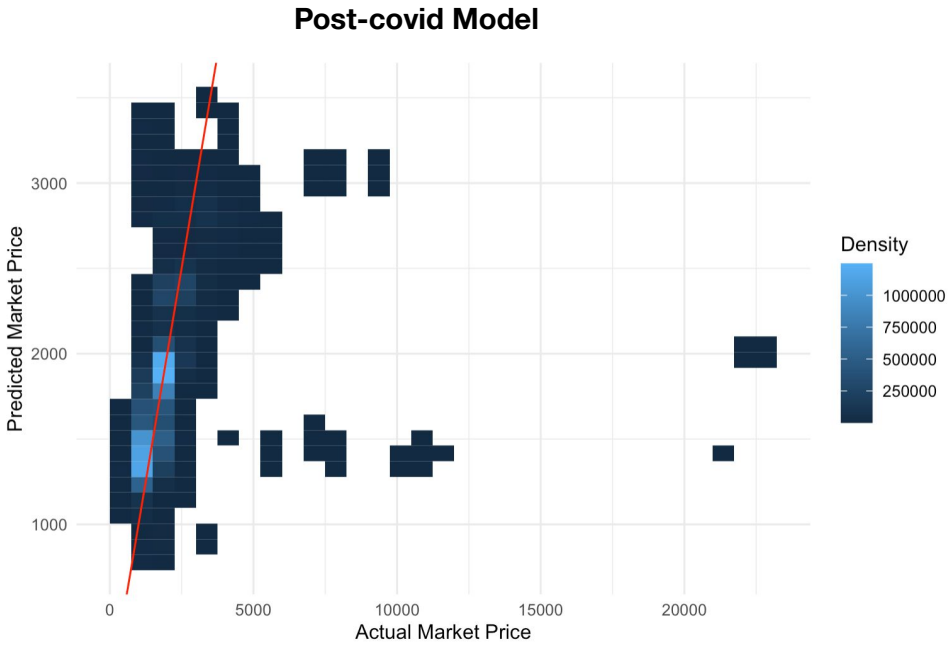
Pricing model from internal variables, derived variables, and external variables



OLS Model Details: Pre-covid

- Train the model on market data from 2015-2018 and make predictions for market prices in 2018
- Results show an average price difference between actual and predicted market of $563
- Outliers: Outliers are luxury properties that have higher actual market prices
- Conclusions: Not perfect, the predictions are too aggregated and not specific enough for each individual unit

Pricing model from internal variables, derived variables, and external variables



OLS Model Details: Post-covid

- Train the model on market data from 2018-2020 and make predictions for market prices in 2020-Present
- Results show an average price difference between actual and predicted market of $572
- Outliers: Outliers are luxury properties that have higher actual market prices
- Conclusions: Very similar to the pre-covid model, still too aggregated

Conclusions

	Summary	Next Steps
Midpoint Summary and Updates	• Total units occupied • Reasons for moving - AHS • Winter and Summer seasonality	• Examine total units occupied more thoroughly - is there an underlying cause to this correlation? • So much more interesting data from the American Housing Survey to examine • Continue investigating the seasonality trends
Sentiment Analysis from Kingsley	• Positive sentiment correlated with higher market prices on average	• Examine potential use of survey data for a model or showcase in leasing brochures • Test out different types of sentiment models and compare them
Lease Likelihood and Pricing Model	• Likelihood to lease model • Foundation Pricing Model	• Continue building and expanding these models over the summer and potentially next fall

Questions?