Executive Summary

Key Findings:

- Higher Airbnb density in the north, but competition is a problem.
- Homicide density is a good predictor of prices.
- Random Forest model outperformed others in price prediction.

Future Improvements:

- Meta-Learning and Automated Hyperparameter Optimization.
- Incorporate points of interest and amenity data.

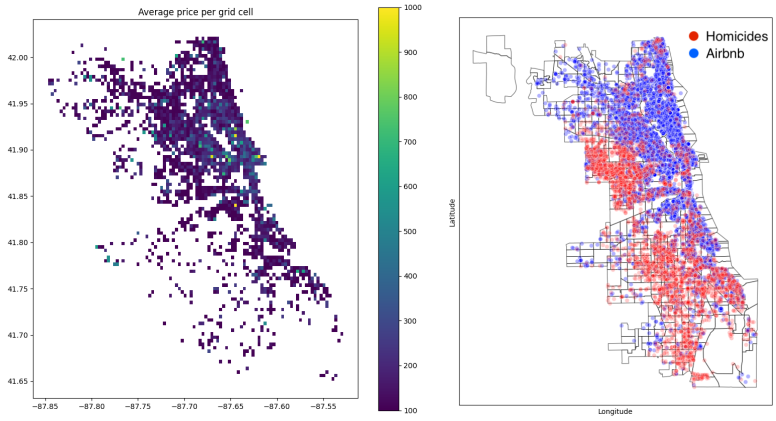
Introduction, Objective & Scope:

Intuitively one assumes that housing, hotel and Airbnb prices are all dependent on the “quality” of the neighborhood. However, as initiation might mislead someone, we tried to actually quantify the “quality” of a neighborhood by deriving them from crime rates and population data.
Therefore, this project preprocesses, analyzes and combines three datasets(Airbnb data, crime data & population data) in order to obtain insights and to additionally answer the following questions: (a) Can we analyze and predict the price of Airbnbs within Chicago? (b) What are the influencing factors behind Airbnb prices and can we model this meaningfully?

Key Performance Indicators

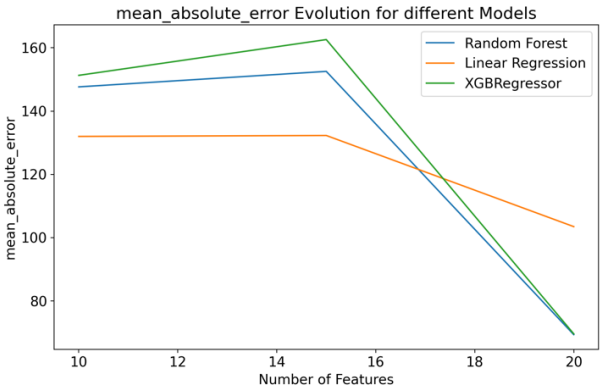
Our main KPIs where the **correlation of features** (to get a better understanding of the importance for single features for the model) and the **Root-Mean-Squared-Error** (indicates the error between prediction and actual values)

(a) Can we analyze and predict the price of Airbnbs within Chicago?



While we were able to analyze the average prices per grid cell (left plot) we discovered a high density of Airbnbs in the north of the map. However, this observation aisre only partially useful as an explanatory variable as a higher density in Airbnbs also results in a higher competition and thus a self regulation in the prices. Conversely, we observed that the density of homicides is much higher in the south-west of Chicago. This however, is one of the top 10 most relevant features as we discovered in our combined model.

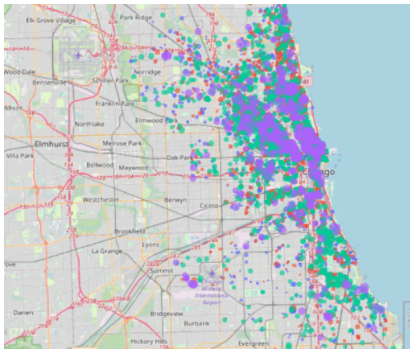
(b) What are the influencing factors behind Airbnb prices and can we model this meaningfully?



The analysis aims to predict Airbnb prices using different combinations of features. Three sets of features (10, 15, and 20) were evaluated with the Random Forest, Linear Regression, and XGBRegressor models. The results reveal an improvement in performance with an increasing

number of features, especially with 20 features. The Random Forest model outperforms the others, showing a net with a positive r2_score, suggesting a better fit to the data.
Potential reasons for these findings could include the inherent complexity of relationships between Airbnb location features and price. Adding more features may have allowed the models, particularly Random Forest, to better capture the variability in the data. However, this can also lead to overfitting to the training data, hence the need for validation on independent test datasets. Alternatively, the models' moderate performance may be explained by the essentially complex and dynamic nature of the Airbnb rental market, where many factors can influence prices and more likely models are needed to capture these nuances. Parameter adjustments, deeper exploration of features, and use of advanced techniques could also help improve Airbnb price prediction.

Additional Insights - Clustering



cluster	accommodates	bedrooms	beds	availability_60	availability_90	availability_365	number_of_reviews	price
0	3.178176	1.416635	1.641358	6.774132	14.731782	113.337657	45.734834	139.520031
1	2.782878	1.223599	1.417066	44.522388	73.160800	268.811884	40.435089	142.023655
2	6.790881	2.731132	3.513627	40.481656	67.002096	250.603774	62.592243	233.807128
3	12.595982	4.642857	6.816964	36.087054	59.794643	226.265625	32.328125	853.475446

Our clustering analysis aimed to discern patterns within different Airbnb groups. Facilities from cluster 0 are rather small, however, their prices are also the lowest. In comparison, cluster 1 also settles on the lower tail of the price distribution with the big difference of higher availability across the year. Hence, facilities of cluster 0 seem to be either booked out more frequently, or simply not put on the market by their host. Showing a degree of similarity to cluster 1, cluster 2 on average accommodates more people, which in the end reflects in the higher price. Lastly, cluster 3 represents the largest facilities which, on average, accommodate 12 people; hence the highest average price of around 850$ per night. On the map, we can see that there are more of smaller AirBnBs (cluster 0 & 1) than larger ones (cluster 2 & 3).

Future Outlook & improvements:

Combining Meta-Learning (for finding the best suited algorithm) and Automated Hyperparameter Optimization (HPO) might allow for further improvements in the model, through which we might be able to make better predictions. Additionally, paywalls hindered us from accessing datasets on points of interest which might also be a relevant feature for our model. Then, we want to include amenities without disrupting the existing model in order to emphasize our analysis.

Conclusion:

The optimal place is difficult to find (at least for Chicago)

- High density in the north, suggesting greater activity (but also more competition!).

The price and the localisation of an airbnb in Chicago depends on many factors

- Crime count is one of the 10 relevant features selected to predict price of a Airbnb.
- Whether it's the type of accommodation, the number of beds, the number of rooms or the featured amenities - many variables jointly impact the price.