Team REX Technical report

Preston Ching, Marcel Hedman, Nam Luu, Owen Schafer
Supervisor: Zona Kostic

March 2021

1 Introduction

The $9.6 trillion (as of 2019) global real estate market is one of the least technology disrupted markets today. Connections, transactions and relationships in residential real estate are often forged through traditional stakeholders and human interactions. Our partner organization, REX, leverages the power of machine learning and big data to match home owners and buyers, seeking to create a platform that significantly reduces transaction fees. Amongst the various approaches to reduce information asymmetry, REX intends to establish a more robust and fine-grained real estate index for realtors and property investors. Current solutions include price indices from the National Association of Realtors (NAR) and automated valuation machines (AVMs) implemented in Zillow or Google. However, these price indices are usually only based on state-level or city-level metrics, and do not provide clarity into more fine-grained submarkets. Moreover, AVMs ascertaining the spot price of residential real estate are noisy metrics mired with varying bid-ask spreads and mortgage approvals. Prices from sale transactions are also considered second-order effects since they are reflecting the equilibrium of supply and demand dynamics. Therefore, this project aims to create a real estate index for predicting demand via home listings and sale transactions. Due to data availability, modelling challenges and ease of comparison with the analysis conducted by the IACS Team in Fall 2020, the focus will be on predicting demand for homes in the Denver market. Moreover, a model for supply prediction is not in the scope of this project due to influence from off-market homes amongst other factors.

Instead of census-tract based submarkets (census tracts are more granular than zipcodes and are strictly defined geographical demarcation), this project will explore a Bayesian hierarchical approach to more accurately and effectively identify submarkets which can be a hybrid of geographical (proximity to schools, business districts), environmental (crime rates, demographics) and physical (size, apartment view) factors. To better inform realtors and property investors, one of the key goals is for our model(s) to be interpretable and allow insights into the various submarket classifications and properties. Nevertheless, there are several challenges in implementing a hierarchical model for a real estate index. Firstly, there is no clear definition of the demand target variable as a function of home listings and sale transactions as proxy variables. Apart from using the number of listed properties and sale transactions within a particular discretized time period, we have also explored using the number of days on market as a better proxy for demand. Secondly, the Bayesian hierarchical model involves fixing k number of submarkets which implicitly requires potentially faulty assumptions on submarket classification types. In terms of defining success metrics, we intend to use traditional machine learning metrics of accuracy and ROC-AUC on test datasets as well as the ease of identifying submarkets based on similar attributes within a particular model-generated submarket. Moreover, we will benchmark the efficacy of the Bayesian hierarchical model against non-Bayesian baseline models, ranging from models without submarkets and those incorporating submarkets.

2 Literature Review

Firstly, most real estate indices are focused on prices, ranging from stock market performance of real estate companies and appraisal-based indices analyzing returns on real estate institutional investor portfolios to transaction-based prices indices. The more reliable price indices are based on actual transaction prices, such as in the Paris Housing Market by Maurer et. al. (2004) [1] and Gourieroux et. al. (2006) [2]. Instead of naive price indices based on mean or median prices, these transaction-based indices are constructed from hedonic price models with data availability and consistency in the French markets. The hedonic assumption here refers to home prices being predicted fully from the individual homes' attributes, shown to be plausibly fallible based on our preliminary developed models. Another hedonic price index by Jiang et. al. (2014) [3] in the Singapore residential real estate market uses data on single-sale and repeat-sale properties, outperforming the SP Case-Shiller price index for single-sale properties. Nevertheless, this is predicting on a city or national level rather than providing information on different submarkets.

There are two key pieces of literature that motivate aspects of our model. Previous work towards building a more granular real estate index was developed by the IACS REX team from the Fall 2020 semester [4]. The model described in their work predicts a number of listings and sales over a specified time-frame at a market level, and then predicts these same quantities within predefined geographical submarkets (census tracts) by scaling their market prediction by the mean ratio of sale/listing rates within a particular submarket to those of the market as a whole. This approach can be viewed as "top-down", in that predictions made at a more granular submarket-level are generated in part from the model's previously determined market-level prediction. This can also potentially be construed as being similar to a fixed-effects regression model where census-tracts demarcation provide valuable and statistically significant information in predicting demand and supply. As mentioned earlier, we intend to test our models using the same predictive domain (the Denver metro area) with the goal of predicting identical proxies for demand (number of listings/sales) for ease of comparison.

Liu et. al. [5] introduce a Bayesian framework for modeling real estate valuation using hierarchical clustering to capture distinct trends between learned submarkets. This approach can be viewed as a "bottom-up" approach, aiming to first predict at a submarket level and then generate an aggregate prediction of demand at the market level. One major assumption in this Bayesian hierarchical framework is also the hedonic assumption where prices can be fully predicted based on the individual homes' attributes. Moreover, this model seeks to predict noisy price signals instead of demand.

Therefore, our proposed developed model leverages the strengths from these two key literature by using a Bayesian hierarchical approach to forecast demand (through proxies described in subsequent sections) in place of a property's monetary value. Our model has the capability of simplifying the structure presented in Liu et. al. (e.g. allowing for a linear hedonic demand/supply model instead of a Gradient Boosting Tree hedonic price model) in order to allow a reasonable future adaptation for time series data.

3 Data

For milestone 3, REX provides us with a new dataset that contains fewer observation points (257,057 units compared to the previous 439,427 units). The dataset contains 45 features that provide information about housing profile for each unit. Compared to the previous dataset, this dataset has fewer NaN values, and all duplicate listings are removed.

3.1 Data Preparation and Exploratory Data Analysis

In this section, we outline the steps that we did to prepare and analyze this new dataset. Our work for the orginal datasets that are used in the first two milestones can be found in Appendix A.

We first merged the new dataset from REX with the condensed dataset that we got from previous milestone, together with the census track datasets. Unlike previous milestones, in this iteration, we wanted to use more census features. Therefore, instead of handpicking a few census features, we kept all features whose variances are higher than a certain threshold. We also remove multicollinearity by plotting covariance matrix and manually removing features that are highly correlated. The covariance matrix is shown in figure 3.

This process leaves us with a total of 62 features. However, in the first attempt to run our pymc3 and EM models on this dataset, we ran into memory issues, even with AWS instances with very high memory. This calls for the need to trim down the number of features even further. To do so, we use XGBoost to select the top features with highest prediction gain. We used the submarket classification from previous milestones, and used XGBoost on each submarket to get the top features for each submarket, and keep the features that perform consistently well for all submarkets. Figure 2 shows the feature importance plot for 4 out of 5 submarkets. The remaining submarket has very few units so we omit from this process. Using feature importances, we reduced our feature space to 16 features. We also found out that there was conflict between *totalrooms* and *bedrooms* and *bathfulls*. There are many units whose *totalrooms* is smaller than the other two combined. Therefore, we also removed *totalrooms* and we used 15 features to build the model. These 15 features are:

- *18-59*: the percentage of residents in the neighborhood whose age is in the range 18-59
- *mean_household_income*: mean household income of the households in the census
- *built 1995 or later*: whether the unit is built in 1995 or later
- *MULTI_FAMILY, CONDO, OTHER*: whether the unit is a multi family, condo, or not multi family, condo nor single house. These features are the one-hot version of the property type feature
- *mobile_home_pct*: percentage of properties that are mobile
- *annual_births_per_residence*: the ratio of births per residence in each census
- *farm score, luxury communities score*

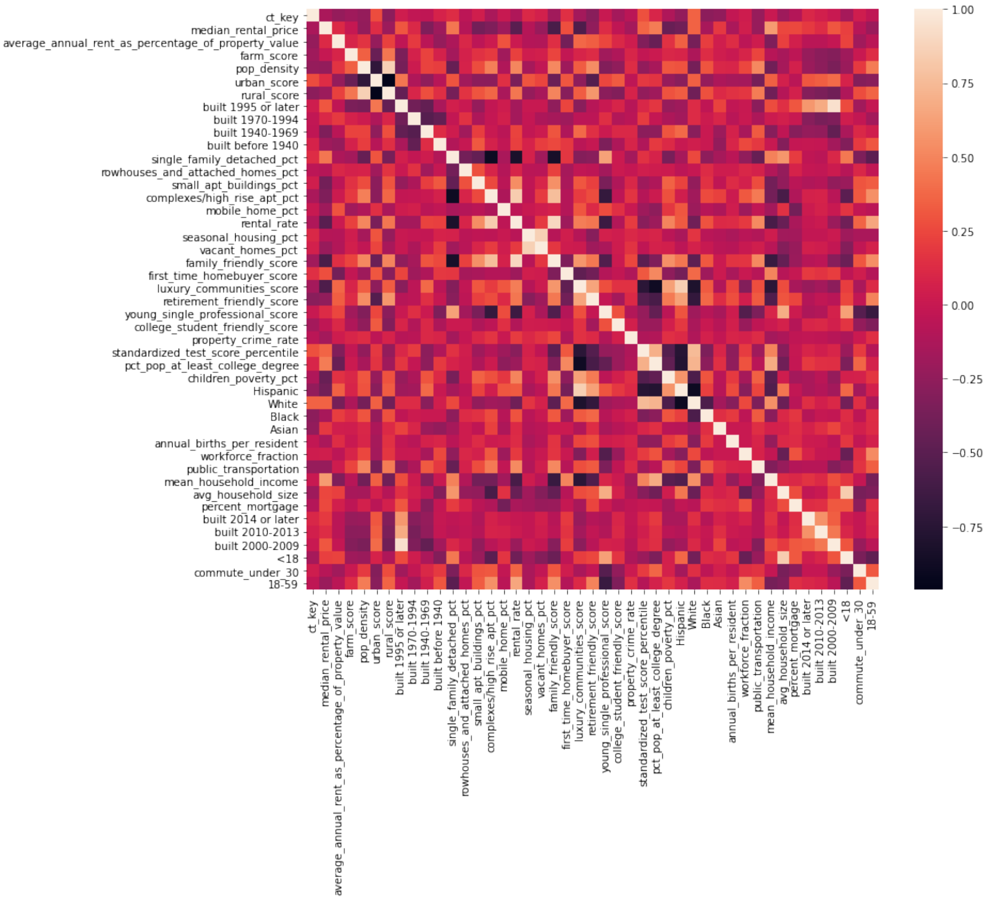

Figure 1: Correlation matrix

- *property crime rate*: crime rate of properties in the same census
- *bedrooms*: number of bedrooms
- *bathfulls*: number of bathrooms
- *small_apt_building_pct*: the percentage of small units in the same census
- *standardized_test_score_percentile*: standardized test percentile of residents in the same census.

4 Baseline model

Our baseline approach falls into two categories, a submarket and non-submarket approach. The non-submarket baseline acts as a baseline for performance to ensure that our allocations of subamrkets perform better than if we had considered Denver as just one market. The submarket baseline is to ensure that the submarkets chosen in the advanced model are better than naively assigning submarkets.

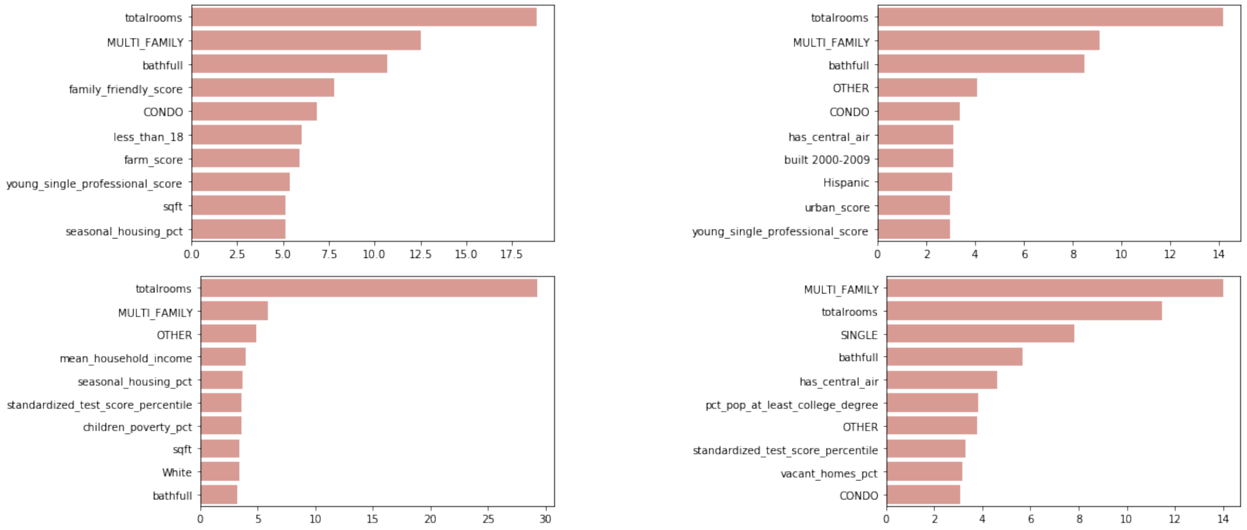

Figure 2: Feature importances

4.1 Non-submarket baseline

4.1.1 Results

Methods	Testing Accuracy	ROC-AUC	$\frac{\text{Predicted number of sales}}{\text{Actual number of sales}}$
Logistic Regression	0.63	0.54	2.07

Table 1: Non-submarket baseline models

4.1.2 Key Insights

For non-submarket baseline, in previous milestones we built a logistic regression model, a random forest classifier and a neural network model. In this milestone, given the performances of the models, we focused on the logistic regression and XGboost models.

These models were trained on the updated dataset for milestone 3 using the feature selection approach described in the data preparation section. We can see that the logistic regression gives poor performance and predicts a figure for the number of homes sold which is over double the actual figure.

4.2 Submarket baseline

The submarket baseline approach was developed to allow comparison with the developed model and to ensure that the developed model performed better than naive classification of homes into submarkets. In this milestone we iterated over the number of submarkets in the range 3-20 to find the optimal number of submarkets (as defined by the MSE score on the difference between expected number of houses sold and actual number sold). The optimal number was found to be 11 for the logistic regression and this is what we then used when finding our final results. Note, submarkets were still created using k-means clustering on the numerical features. For each value of K submarkets, we trained K distinct models on each of the discovered submarkets in line with the hedonic assumption and found the weighted average scores to arrive at final results for test accuracy.

Compared to the previous milestone, we have also implemented a new evaluation metric which is the predicted number of homes sold at both submarket and market levels. This metric is valuable as the models capture meaningful information about the submarket on these scales which are larger than the per home basis.

4.2.1 Results

Methods	Optimal submarkets	Testing Accuracy	ROC-AUC	$\frac{\text{Predicted number of sales}}{\text{Actual number of sales}}$
Logistic Regression	11	0.63	0.55	0.9964
XGBoost	7	0.65	0.69	1.0004

Table 2: Submarket baseline models, marketwide accumulative results

Submarket number	Number of homes	Predicted number of sale	Actual number of sale
Submarket 0	933	301	302
Submarket 1	1978	827	834
Submarket 2	71	16	18
Submarket 3	1603	567	613
Submarket 4	64	24	20
Submarket 5	1255	431	438
Submarket 6	1332	496	488
Submarket 7	289	87	72
Submarket 8	1973	774	791
Submarket 9	1353	520	475
Submarket 10	410	131	138

Table 3: Number of homes in each of the eleven submarkets from K-means for logistic regression

Submarket number	Number of homes	Predicted number of sale	Actual number of sale
Submarket 0	1741	568	551
Submarket 1	2724	1031	1028
Submarket 2	115	26	35
Submarket 3	2165	767	773
Submarket 4	1647	613	602
Submarket 5	382	112	116
Submarket 6	2485	1032	1042

Table 4: Number of homes in each of the seven submarkets from XGBoost

4.2.2 Key Insights

The major takeaway from the submarket baseline approach is that classifying homes to submarkets based on the similarity of their numerical features is not enough to drive performance boosts from our models on a per home basis and this is the reason for the low accuracy and AUC scores. However, the final metric which compares the predicted number of homes sold to the actual number shows good performance. We will therefore look to the developed model to improve the per home performance while maintaining the level of performance in predicting the number of total sales in the market.

5 Developed model

Our developed model attempts to improve upon our baseline models in two ways. Firstly, we aim to have our model better learn to define optimal submarket classifications, rather than being limited by a predefined clustering of the data for each home. Secondly, we aim to be able to interpret the hedonic demand model used within each submarket in order to understand the relative importance and effect of each feature of a given home on demand and define rough geographic regions corresponding to each submarket.

In order to accomplish these goals, we introduce a Bayesian hierarchical model where the submarket classification of a home into one of K submarkets is a latent variable, and each predicted quantity for that home is distributed according to a (potentially stochastic) hedonic demand function of its features as determined by its submarket classification. Below, we discuss two subtly different realizations and implementations of this modelling framework.

5.1 Hedonic Demand Function #1: Logistic Regression

5.1.1 Model Specification

For this model, we assume that our target variables for each home are functions of a deterministically-weighted linear combination of the home's features (akin to logistic regression), and that these weights are determined by the home's submarket membership. That is, for a particular draw $n \in \{1, \ldots, N\}$ from our hierarchical structure (corresponding to a single home), our model assumes the following structure and dependencies amongst variables (displayed graphically in Figure 3):

1. $z_n \sim \text{Cat}(\pi)$: This variable represents the submarket classification of a home. The parameter $\pi \in \mathbb{R}^K$ represents the prior probabilities of a home belonging to each of the K submarkets, where $\pi_k = 1/K$ for $k \in \{1 \ldots K\}$.

2. $\mathbf{h}_n \sim P_h(\Theta_{z_n})$: These random vectors represent the features of each home. These features are drawn from one of K total distinct distributions (for each submarket), and each distribution is parameterized by Θ_{z_n}, a set of parameters characterizing the multivariate-normal, Poisson and Bernoulli distributions of continuous, discrete and boolean home features (respectively). Note that we no longer assume a multivariate Gaussian distribution of *all* home features within this model, nor do we assume that the covariance matrix of continuous home features is shared across submarkets.

3. $\mathbf{l}_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$: This variable corresponds to a home's latitude and longitude. This is treated separately from $\mathbf{h}_n$ because we do not wish to include location features as inputs to a linear hedonic demand function (this contributes to model over-fitting). However, we choose to include it in this model hierarchy with the goal of obtaining more geographically-distinct submarkets for better interpretability.

4. $y_n = f_{z_n}(\mathbf{h}_n)$: This variable represents the probability that a home will be sold during this specified time interval, given that it is listed. The function f_{z_n} is a hedonic demand function of $\mathbf{h}_n$ determined by the home's submarket membership. Note that f_{z_n} corresponds to a logistic regression of $\mathbf{h}_n$ in this implementation, although we explore another functional form in a later section.

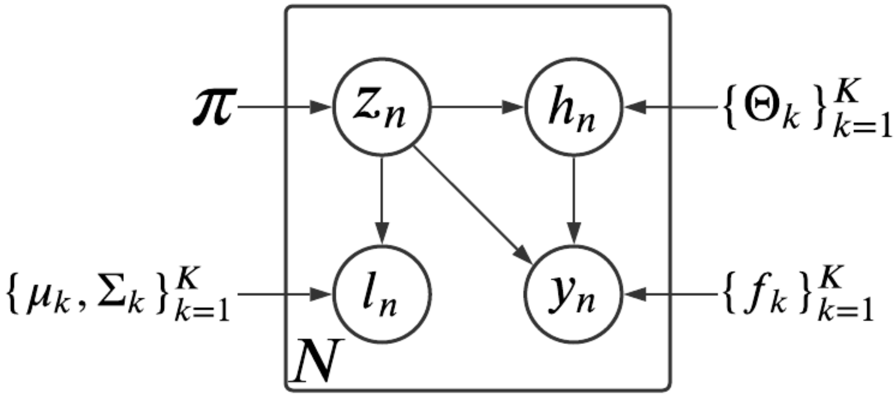

Figure 3: Proposed Hierarchical Model #2

5.1.2 Estimation with Expectation Maximization (EM)

Setting the number of submarkets parameter K to be 11 (based on the cross-validation discussed earlier), the Bayesian hierarchical model with a frequentist logistic regression hedonic demand function (estimated by 1000 iterations of EM) resulted in 9 submarkets identified. The training and testing accuracy, ROC-AUC, predicted home sales and actual home sales for the individual submarkets is shown in Table 5, as well as cumulative metrics weighted according to the various submarket sizes, MSE and R^2.

It is observed that the weighted average testing accuracy and ROC-AUC of this model indicate poor performance on predicting an individual home's sale, but the low MSE and high R^2 indicate that the model does learn properties of the home sale distribution within each submarket. This implies that our hierarchical structure is able to capture distributional properties of the home data used in training, but the logistic regression hedonic demand function may not be a good fit. The geographical representation of the various submarkets are depicted in Figure 4.

Submarket	# of Homes	Train Acc.	Test Acc.	ROC-AUC	Predicted	Actual	MSE	R^2
1	20,979	63.0%	64.0%	0.500	2,319.5	2,264	-	-
2	1,729	63.9%	63.6%	0.500	187.5	189	-	-
3	4,205	62.7%	63.0%	0.500	469.2	467	-	-
4	413	60.2%	57.3%	0.500	49.3	53	-	-
5	429	68.3%	73.6%	0.500	40.8	34	-	-
6	55	52.6%	41.2%	0.500	8.9	7	-	-
7	5,195	61.3%	59.3%	0.500	603.2	635	-	-
8	1,037	63.0%	65.4%	0.500	115.3	108	-	-
9	0	-	-	-	-	-	-	-
10	0	-	-	-	-	-	-	-
11	65	68.9%	60.0%	0.500	6.2	8	-	-
Total	34,107	62.8%	63.2%	0.500	3,800.0	3,765	383.0	0.999

Table 5: Submarket Metrics using EM

5.2 Hedonic Demand Function #2: XGBoost

5.2.1 Model Specification

This model has identical structure to that described in section 5.1, with a change in hedonic demand function. That is, our model assumes the hierarchical structure displayed graphically in Figure 3, where f_k is a gradient boosting tree (XGBoost).

5.2.2 Estimation with Expectation Maximization (EM)

Setting the number of submarkets parameter K to be 7 (based on the cross-validation discussed earlier), the Bayesian hierarchical model with XGBoost as a hedonic demand function (estimated by 1000 iterations of EM) resulted in 7 submarkets identified. The training and testing accuracy, ROC-AUC, predicted home sales and actual home sales for the individual submarkets is shown in Table 6, as well as cumulative metrics weighted according to the various submarket sizes, MSE and R^2.

Submarket	# of Homes	Train Acc.	Test Acc.	ROC-AUC	Predicted	Actual	MSE	R^2
1	22,615	78.4%	66.4%	0.614	2,488.8	2,503	-	-
2	132	98.9%	90.0%	0.750	5.8	8	-	-
3	81	100.0%	80.0%	0.798	10.3	12	-	-
4	514	99.4%	65.8%	0.651	65.9	64	-	-
5	5,078	88.8%	66.0%	0.629	573.4	607	-	-
6	4,447	90.2%	65.4%	0.616	472.4	519	-	-
7	1,240	97.1%	69.1%	0.631	117.4	126	-	-
Total	34,107	82.6%	66.4%	0.619	3,734.1	3,839	511.8	0.999

Table 6: Submarket Metrics using EM

It is observed that the weighted average testing accuracy and ROC-AUC of this model indicate reasonable performance on predicting an individual home's sale, and this fit can possibly be improved through additional

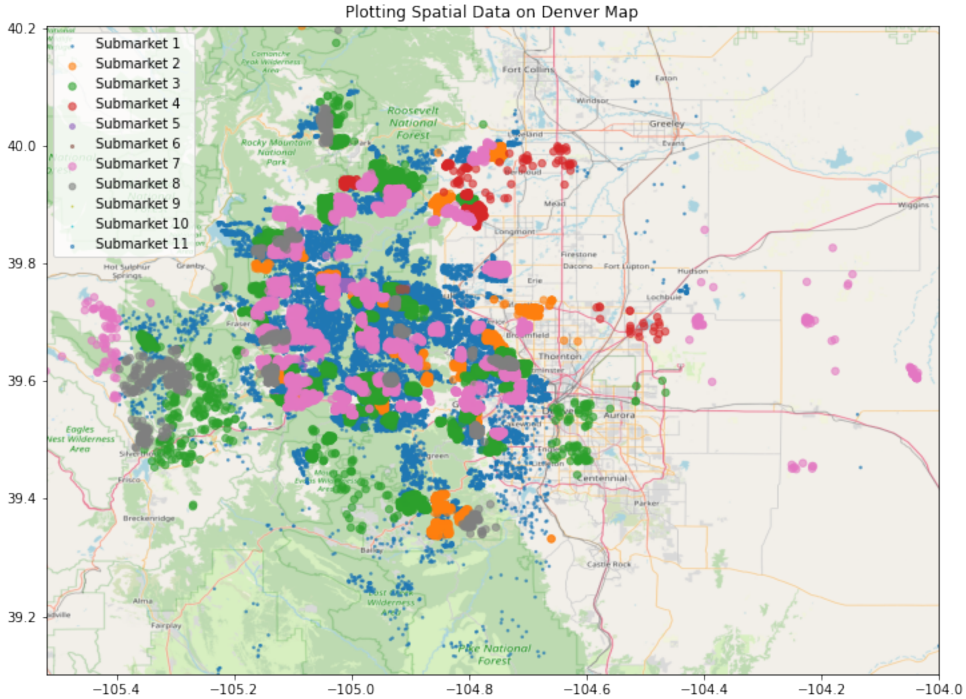

Figure 4: Locations of 9 Submarkets Generated by 1000 EM Iterations

data or EM iterations (as evidenced by the difference between training and test accuracies at the market level and within each submarket). The model's low MSE and high R^2 indicate that it learns properties of the home sale distribution within each submarket. The geographical representation of the various submarkets are depicted in Figure 5. Additionally, the importances of the three most important features within each submarket are shown in Table 7. While there are some features that are important across many submarkets, no two sets of feature importances are identical, indicating that this model is learning unique hedonic functions within each submarket.

Submarket	Most Important	Importance	2nd Most Important	Importance	3rd Most Important	Importance
1	MULTI_FAMILY	0.333	luxury_communities_score	0.068	bathfull	0.066
2	MULTI_FAMILY	0.368	OTHER	0.337	18-59	0.190
3	built 1995 or later	0.839	bathfull	0.060	bedrooms	0.039
4	property_crime_rate	0.256	bathfull	0.135	mobile_home_pct	0.132
5	MULTI_FAMILY	0.163	bathfull	0.085	small_apt_buildings_pct	0.078
6	MULTI_FAMILY	0.160	bathfull	0.102	CONDO	0.089
7	luxury_communities_score	0.155	bathfull	0.141	small_apt_buildings_pct	0.097

Table 7: Submarket Feature Importances

5.3 Key Insights

We find that none of our models perform particularly well when predicting whether or not an individual home will be sold (although the model using a XGBoost hedonic demand function does substantially outperform that using a logistic regression hedonic demand function in this area). However, our models

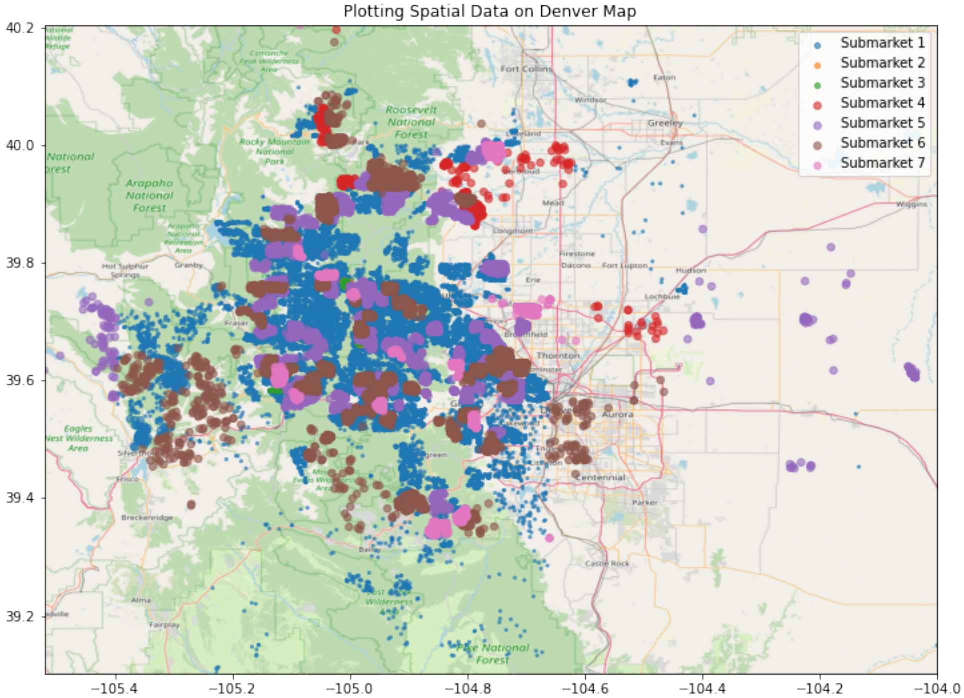

Figure 5: Locations of 7 Submarkets Generated by 1000 EM Iterations

are able to provide a reasonable estimate of the expected number of home sales within each submarket. We note the following from our observations:

1. **Hedonic Demand Function:** The use of a logistic regression hedonic demand function produces generally worse results than XGBoost, so we prefer the latter. However, there may be other hedonic demand functions that are interpretable (we can understand the importance/weight of each home feature).

2. **Hedonic Assumption:** A key assumption of our current framework is that a listed home's probability of being sold in a specified time period (as well as an unlisted home's probability of being sold) is given by a (possibly submarket-dependent) function of the home's features and other macro factors (currently at the level of a home's census tract). The set of 15 features selected via the process described above contains only 5 features that are specific to a particular home.

3. **Generality:** The submarkets identified by our models have been evaluated using a single time interval. While they exhibit reasonable performance, it is unclear whether or not these submarket classifications provide comparable performance when evaluated on data from other time periods.

6 Further work

Our future work ahead of the next project milestone will be focused on addressing the observations cited in the ”Key Insights” of our developed model discussion. Some key goals that we will work towards include:

- We plan to explore other hedonic demand functions beyond logistic regression and XGBoost. In this exploration, we will prioritize a balance between interpretability and model performance.

- We plan to test the generality of our submarket classifications by evaluating the performance of our trained model on data from other time periods of equal length.

- As a stretch goal, we hope to begin our adaptation of the model for time series data. This process will begin once once we are satisfied with our model's structure and performance within the context of a fixed time frame.

Appendix

A Data

In this section we provide the EDA of the original dataset that we used for milestones 1 and 2.

A.1 Data Profile

For the first two milestones, we used three sets of data collected from multiple listing services (MLS), REX and the Census Bureau.

The first MLS dataset, compiled from numerous cooperating real estate brokers, provides information on home listings and sale transaction data for our target variables. This dataset contains 439,427 observation points, which have the transaction data of 399,883 unique properties in the greater Denver area from 03/2016 to 11/2020. For each transaction, we have the list date, sale date, withdrawn date, expired date, status (whether the property is sold, expired or withdrawn), the property's zip code and sale price.

The second datasource is REX, which provides information about the profile of each property. This dataset provides an additional 70 features, both numeric and categorical, such as the square footage area, the number of rooms and presence of garage amongst others. This dataset has information of 200,000 listing properties and 1,200,000 non-listing properties.

Other than datasets for transactions and house profiles, we believe that the socio-economical environments of the residential community where the houses are located are also important indicators for supply and demand, thereby augmenting our analysis with the census dataset. This dataset is a combination of data from NeighborhoodScout, a database of US neighborhood analytics, and other public census records, generating an additional 122 features containing information on city-level demographics as well as census-level metrics on friendliness to college students, professionals and transportation score amongst others.

A.2 Exploratory Data Analysis

A.2.1 Data preparation

Since we have a dataset that is rich in features, we need to trim down the number of attributes. We first handpicked the features that we think are good indicators for our model. Of the remaining features, we removed those that have too many missing values, or have very low variance. We also modified the *sale price* attribute to *sale price per sqft*, to remove the correlation between the area of a property to its price. This way, we shortened our list of features to the following, which could be classified into 2 categories:

- House profile features: *zipcode*, *area*, *sale price per sqft*, *property type*, *has central air*, *pool*
- Residential community features:*farm score*, *median rental price*, *population density*, *home buyer score*, *retirement friendly score*, *young single professional score*, *college student friendly score*, *violent crime rate*, *walk score*, *transportation score*, *carpool score*, *densely urban score*, *suburban score*, *rural score*, *luxury communities score*, *farm score*

These features all have numerical or boolean/categorical values except for property type. There are 6 property types (single, multi-family, condominium, land, townhouse and other) and we map each type to a number from 0 to 5.

When we plot the number of listings per week each year (figure 6), we observed that the number of listings in the area follow almost the same pattern every year: the number of listings gradually increased and peak in the summer before slowing down towards the end of the year. Therefore, to account for the seasonal effect in the real estate market, we created an additional temporal feature, that records the average days on market of other properties in the same communities in the past 30 days. We called this dependent variable *mean_dom*. The intuition behind this feature is that houses that are temporally sold near each other may have similar DOM values, since they share similar market conditions.

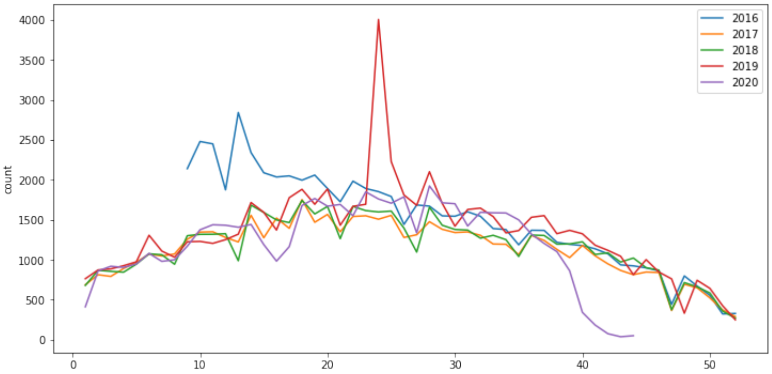

Figure 6: Number of listing each year

Figure 11 show the correlation matrix of these features. After examining the correlation matrix, we further removed the following features: *transportation score*, *carpool score*, *densely urban score*, *suburban score*, *rural score*, *luxury communities score*, *farm score*, *bedrooms*, *bathrooms*, because they strongly correlated with other existing attributes in the dataset.

To deal with outliers, we first plot the geographical distribution of all transactions (figure 8). We observe that the dataset we received not only contain data about the Denver area, but also from other nearby cities as well. Since we only focus on the supply and demand of the Denver metropolitan area, we only consider properties that have latitude inside the range of $[39.5, 40]$ and longitude in $[-105.25, -105.55]$

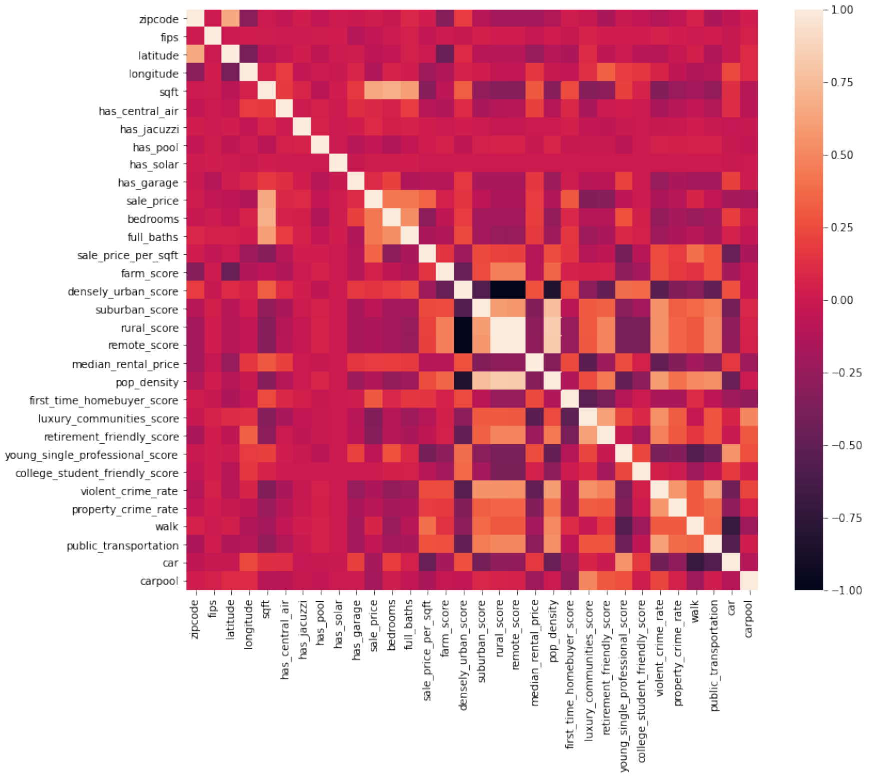

Figure 7

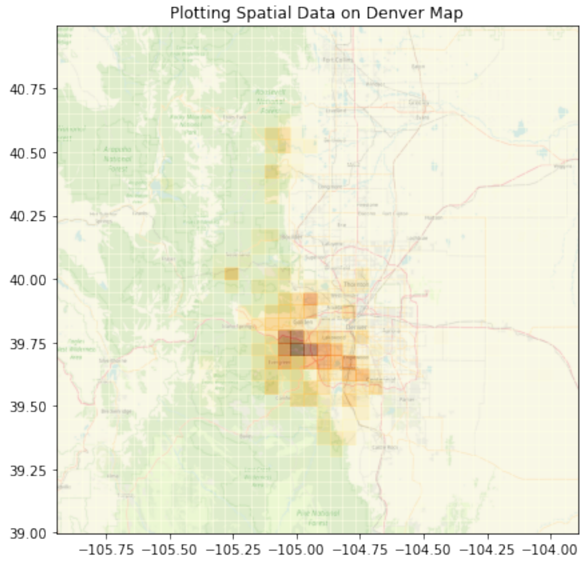

Figure 8: Geographical distribution of data

We also detected outliers by plotting the distribution of each features. In figure 9, for example, when plotting the distribution of days on market, we decided to remove the data point that have days-on-market equal to 0, or list date and sale date are the same, because we speculate that there are some agreements being made before the houses were listed, and the numbers do not truly reflect the days-on-market of the properties. The visualization of other features' distributions can be found in our GitHub repository.

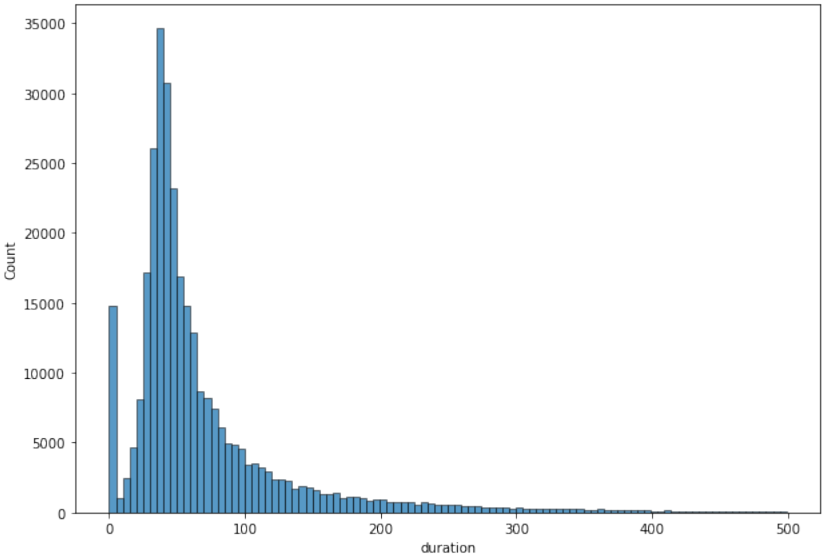

Figure 9: Distribution of days-on-market

A.2.2 Target variables

Initially, in our milestone 1, we created four target variables, two to predict the probability that a property is sold and listed, and two target variables to predict the days-on-market of that properties. However, from milestone 2 onward, we decided to focus on one target variable: the probability of a property being sold, given that it is listed, to keep it consistent between the baseline and developed model. We will later expand our models to predict supply if we are able to successfully address key challenges to our demand-side modelling approach.

In addition, we selected the second quarter of 2019 as the time-interval for all of our models, meaning we will predict given that a property is listed before or during the second quarter of 2019, then will it be sold during this timeframe at all. The reason that we chose 90 days because when we looked at the DOM distribution, we found out that most properties are sold within 90 days.

One challenge arises after we select this target variable, is that the number of houses that are not sold in each time frame is always much bigger than the portion of house that sold. Figure 10 shows this imbalance, where 0 in each graph indicates the number of houses that are not listed and sold, and 1 represents the portion of houses that are listed and sold, in the second quarter of 2019. We will discuss how this imbalance affects our models in the next section.

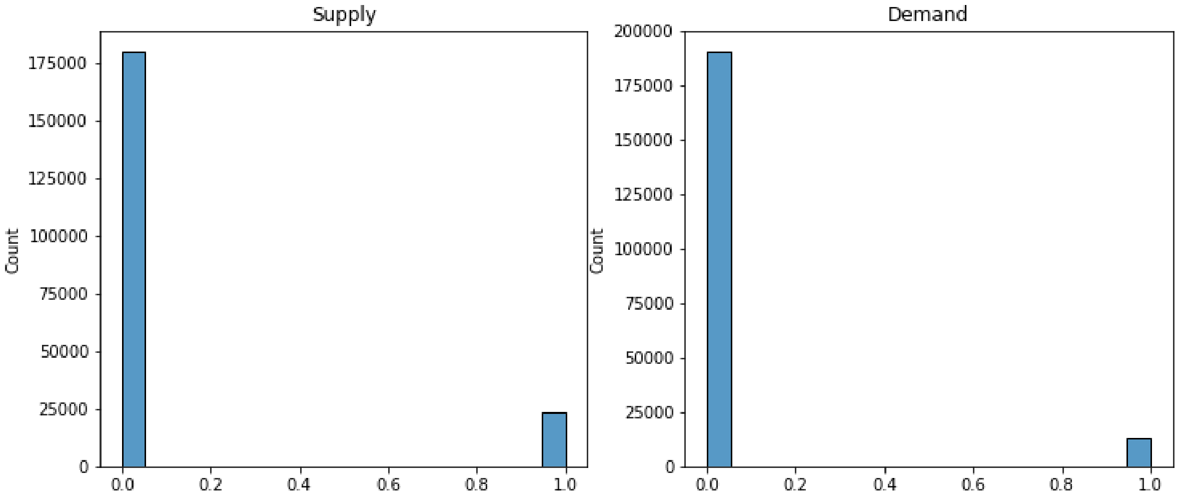

Figure 10: Distribution of supply and demand in Q2 2019

A.3 Hedonic Demand Function: Bayesian Logistic Regression

The Baysian logistic regression method estimated using PyMC3 was trained successfully on the old dataset. However, there were several challenges in achieving convergence due to a significant increase in computational time and memory complexity on the larger new dataset. Nevertheless, the specifications and results from previous milestones are elaborated below.

A.3.1 Model Specification

For this model, we assume that our target variables for each home are functions of a stochastically-weighted linear combination of the home's features (akin to Bayesian logistic regression), and that the distribution of these weights is determined by the home's submarket membership. That is, for a particular draw $n \in \{1, ..., N\}$ from our hierarchical structure (corresponding to a single home), our model assumes the following structure and dependencies amongst variables (displayed graphically in Figure 11):

1. $z_n \sim \text{Cat}(\pi)$: This variable represents the submarket classification of a home. The parameter $\pi \in \mathbb{R}^K$ represents the prior probabilities of a home belonging to each of the K submarkets, where $\pi_k = 1/K$ for $k \in \{1 \ ... \ K\}$.

2. $\beta_n \sim \mathcal{N}(\mu_{\beta,z_n}, \Sigma_\beta)$: This random vector represents the weights of each home feature used in the hedonic models demand. These weights are drawn from two of K total distinct distributions (for each submarket), and each distribution is parameterized by μ_{β,z_n} and Σ_β. Note that for these distributions and each of the following distributions in this hierarchy, we are currently making the simplifying assumption that prior (co)variance parameters are identical across all submarkets.

3. $\mathbf{h}_n \sim \mathcal{N}(\mu_{h,z_n}, \Sigma_h)$: These random vectors represent the features of each home. These features are drawn from one of K total distinct distributions (for each submarket), and each distribution is parameterized by μ_{h,z_n} and Σ_h. Note that within this model, we currently assume a multivariate Gaussian distribution of home features, some of which may be discrete or boolean.

4. $y_n = \sigma(x_n),\ x_n \sim \mathcal{N}(\beta_n^T \mathbf{h}_n, \Sigma_y)$: The variable y_n represents the probability that a home will be sold during this specified time interval, given that it is listed. This is a function of another random variable x_n with variance Σ_y and mean given by our linear hedonic demand model $\beta_n^T \mathbf{h}_n$. The sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ is used to ensure that the predicted quantities y_n can be interpreted as probabilities.

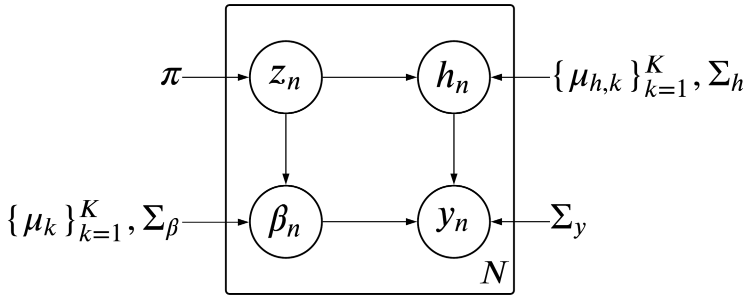
π
z_n
h_n
$\{\mu_{h,k}\}_{k=1}^K, \Sigma_h$
$\{\mu_k\}_{k=1}^K, \Sigma_\beta$
β_n
y_n
Σ_y
N

Figure 11: Proposed Hierarchical Model #1

A.3.2 Estimation with PyMC3

This developed model using PyMC3 focuses on predicting demand, or the probability of a listed home being sold in a given discretized period. Setting the number of submarkets parameter K to be 5, the Bayesian hierarchical model with a Bayesian logistic regression hedonic demand function (estimated by PyMC3) resulted in 3 submarkets identified (note that we select K arbitrarily here, as convergence of our sampler requires an increasing number of iterations with K, providing a barrier for cross-validation). The training and testing accuracy, along with the ROC-AUC for the individual submarkets is shown in Table 8, as well as the weighted average according to the various submarket sizes.

Submarket	Number of Homes	Training Accuracy	Testing Accuracy	ROC-AUC
1	124	69.8%	43.2%	0.404
2	7402	72.6%	72.1%	0.538
3	41	62.5%	85.7%	0.788
Weighted Average	-	72.5%	71.7%	0.537

Table 8: Submarket Metrics using PyMC3

It is observed that the weighted average testing accuracy and ROC-AUC of this model have similar performance compared to the submarket baseline models using logistic regression. This could be due to the PyMC3 model not fully converging or an inappropriate submarket initialization. The geographical representation of the various submarkets are depicted in Figure 12, where there are significant overlaps and no clear demarcation between the submarkets.

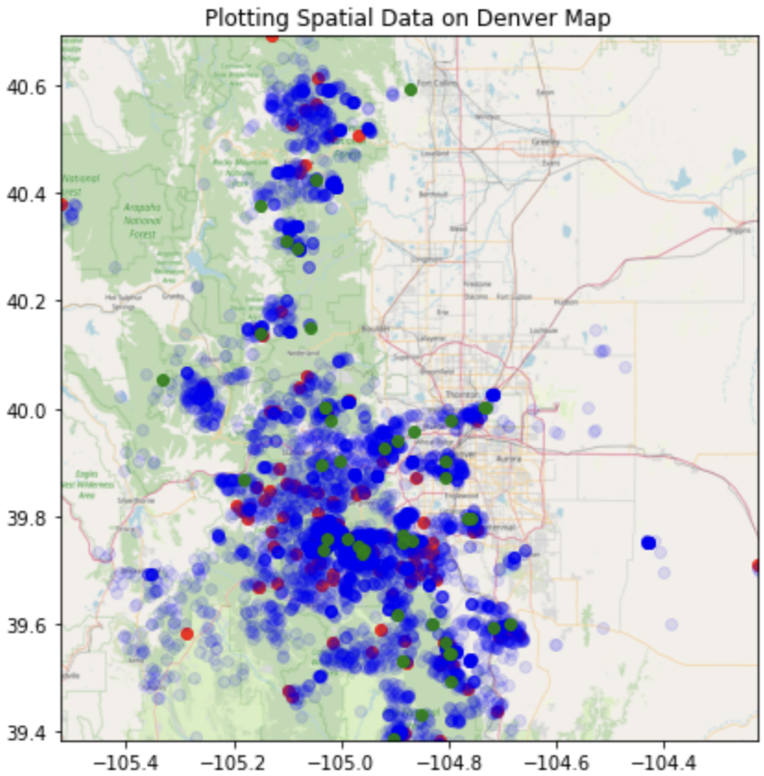

Figure 12: Locations of 3 Submarkets Generated by PyMC3

References

[1] Maurer, R., Pitzer, M., Sebastian, S. (2004). Hedonic price indices for the Paris housing market. Allgemeines Statistisches Archiv, 88(3), 303-326.

[2] Gouriéroux, C., Laferrère, A. (2009). Managing hedonic housing price indexes: The French experience. Journal of Housing Economics, 18(3), 206-213.

[3] Jiang, L., Phillips, P. C., Yu, J. (2014). A new hedonic regression for real estate prices applied to the Singapore residential market.

[4] Will Fried, Jessica Wijaya, Shucheng Yan and Yixuan Di, ”Towards a Revamped Real Estate Index Towards a Revamped Real Estate Index”, Accessed 2021, https://towardsdatascience.com/towards-a-revamped-real-estate-index-c48ae27b33c5

[5] Z. Liu, J. Cao, R. Xie, J. Yang and Q. Wang, ”Modeling Submarket Effect for Real Estate Hedonic Valuation: A Probabilistic Approach,” in IEEE Transactions on Knowledge and Data Engineering, Accessed 2021, doi: 10.1109/TKDE.2020.3010548.