Capstone Report

Arnold Wang

TABLE OF CONTENTS

MAIN POINTS COVERED

PROBLEM STATEMENT

Wether you're looking for a place to settle down, investing in assets for the future or even just looking for a place to live, navigating the real estate market has never been easy to maneuver. That last statement gets amplified with Toronto. With prices soaring and talks of a bubble burst, how can we keep ourselves financially safe in making wise decisions with where we live?

VALUE PROVIDED

As someone who has delt with real estate in the past and currently, finding someone to appraise property can be not only troublesome but expensive. Through applying the innerworkings of datascience, this issue can be worked almost instantaneously through the click of a button. The optimization of a prediction model would moreover immensely benefit companies that cycle through real estate as this would act as a cornerstone to their business model.

SOLUTION

This project aims to provide insight into the innerworkings of the realestate market in Toronto providing users with a respectable number in guiding them towards a proper assessment of wether or not a unit is fairly priced or not. Ideally, this model will be further improved upon to utilize live market data for time series analysis and provide immediete

PREVIOUS USE CASES

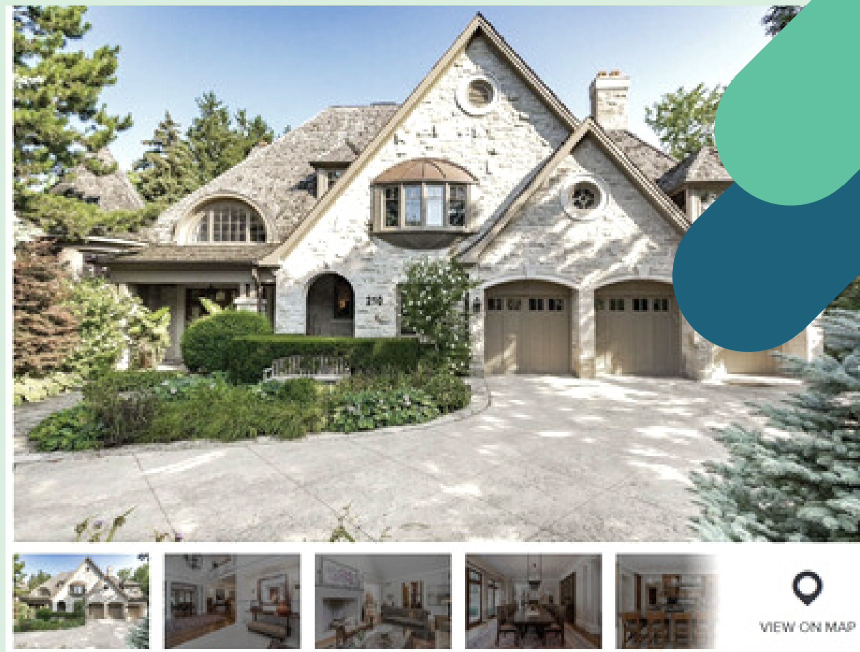
When researching the applications of a prediction model, the 'property appraiser' role came up frequently as this could act as a pseudo replacement prior to recieving a professional quote without any of the hassle. Businesses and consumers alike rely on these appraisers to assess property both industrial and residential on the daily to which machine learning can quickly aid/take over this position

DETAILS ON DATASET

Data was provided via Slava Spirin a former BS student who utilized webscraping on Zoocasa.

	title	final_price	list_price	bedrooms	bathrooms	sqft	parking	description	mls	type	...	full_address	lat	long	city_district	mea
	1303 - 38 Grenville St, Toronto (C4461599) | Z...	855000	870000	2 + 1 beds	2	850.000	1	Luxurious And Spacious Murano Tower. 2+1, 2 Ba...	C4461599	Condo Apt	...	38 Grenville St, Toronto , Ontario, Canada	43.662	-79.386	Bay Street Corridor	
	2 Cabot Crt, Toronto (W4502992) | Zoocasa	885000	898000	3 beds	2	NaN	6	Fantastic Opportunity To Live Within The Histo...	W4502992	Semi-Detached	...	2 Cabot Crt, Toronto , Ontario, Canada	43.647	-79.530	Islington-City Centre West	
	1504 - 30 Roehampton Ave, Toronto (C4511330) |...	550000	549900	1 beds	1	550.000	0	Bright Sunfilled Spacious 1 Bdr Unit; Floor To...	C4511330	Condo Apt	...	30 Roehampton Ave, Toronto , Ontario, Canada	43.708	-79.397	Mount Pleasant West	
	514 - 65 East Liberty St, Toronto (C4515763) |...	665000	600000	1 + 1 beds	1	650.000	1	Rare Loft-Like Condo In Liberty Village W/ 18'...	C4515763	Condo Apt	...	65 East Liberty St, Toronto , Ontario, Canada	43.638	-79.414	Niagara	
	61 Twelfth St, Toronto (W4519375) | Zoocasa	825513	839000	2 beds	2	NaN	1	Location! Location! Location. Your Cottage In ...	W4519375	Detached	...	61 Twelfth St, Toronto , Ontario, Canada	43.597	-79.510	New Toronto	

× 21 columns



Sold

[redacted]

$10,200,000
List price ~~$12,400,000~~
Sold 4 months ago

beds | 6 baths | N/A sq. ft. | 6 parking
irtual Tour / Photos

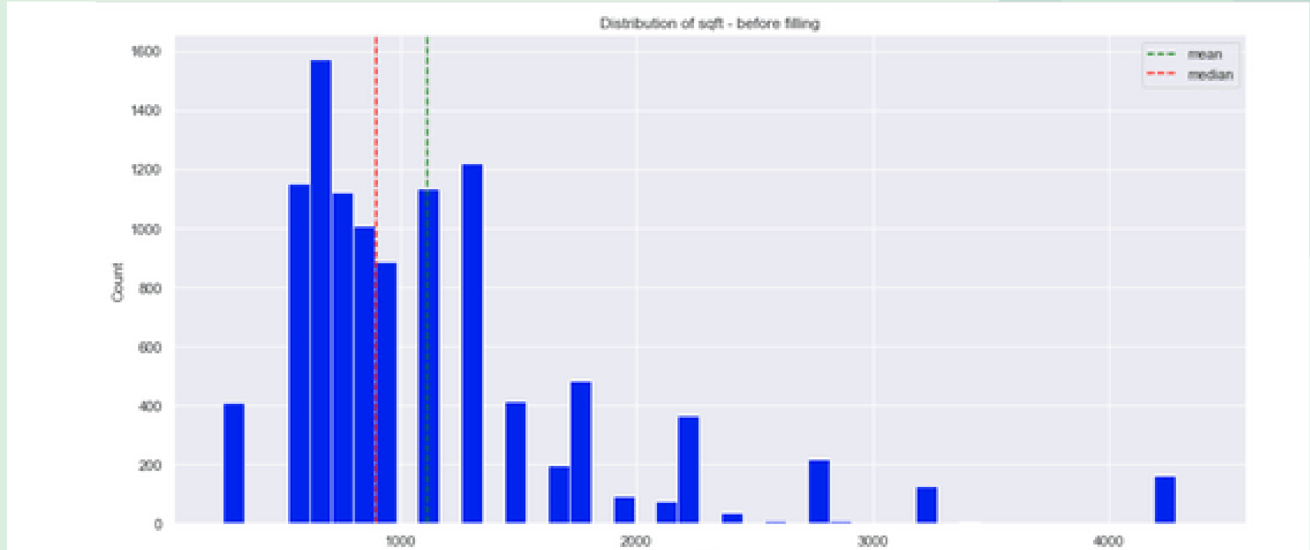
acking Onto A Mature Ravine. A Peaceful Sanctuary To Retreat, Yet Minutes To The 401 ighway And In The Heart Of Bustling Toronto. Every Aspect Of This Magnificent Stone esidence Has Been Carefully Considered. Quality Abounds Throughout The 12,365 Sq Ft f Comfortable Luxurious Living Space (Including 4100 Sq Ft In The Lower Level). Indiana mestone Exterior.Cedar Shake Shingles.Soaring Ceilings.Massive Windows And Doors. 72 Acre Park Like Setting.

SUMMARY OF CLEANING

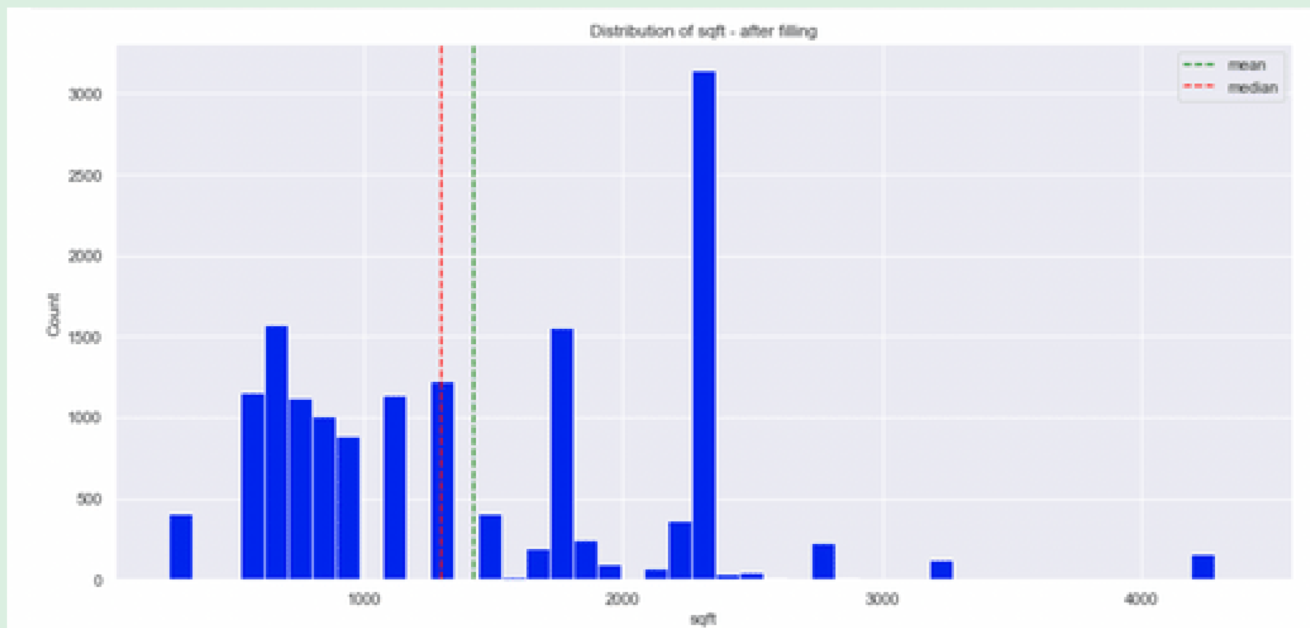
DATA CLEANING

Data Was mostly cleaned and processed missing only a few values in the sqft column which were filled based on a mean value of the type of home. These rows could've also been removed as it represented a small fraction of the total dataset. At first, the null values were all filled with a median however upon realizing it would be more ideal to factor in the type of home into filling these 2 methods were combined.

BEFORE FILLING



AFTER FILLING



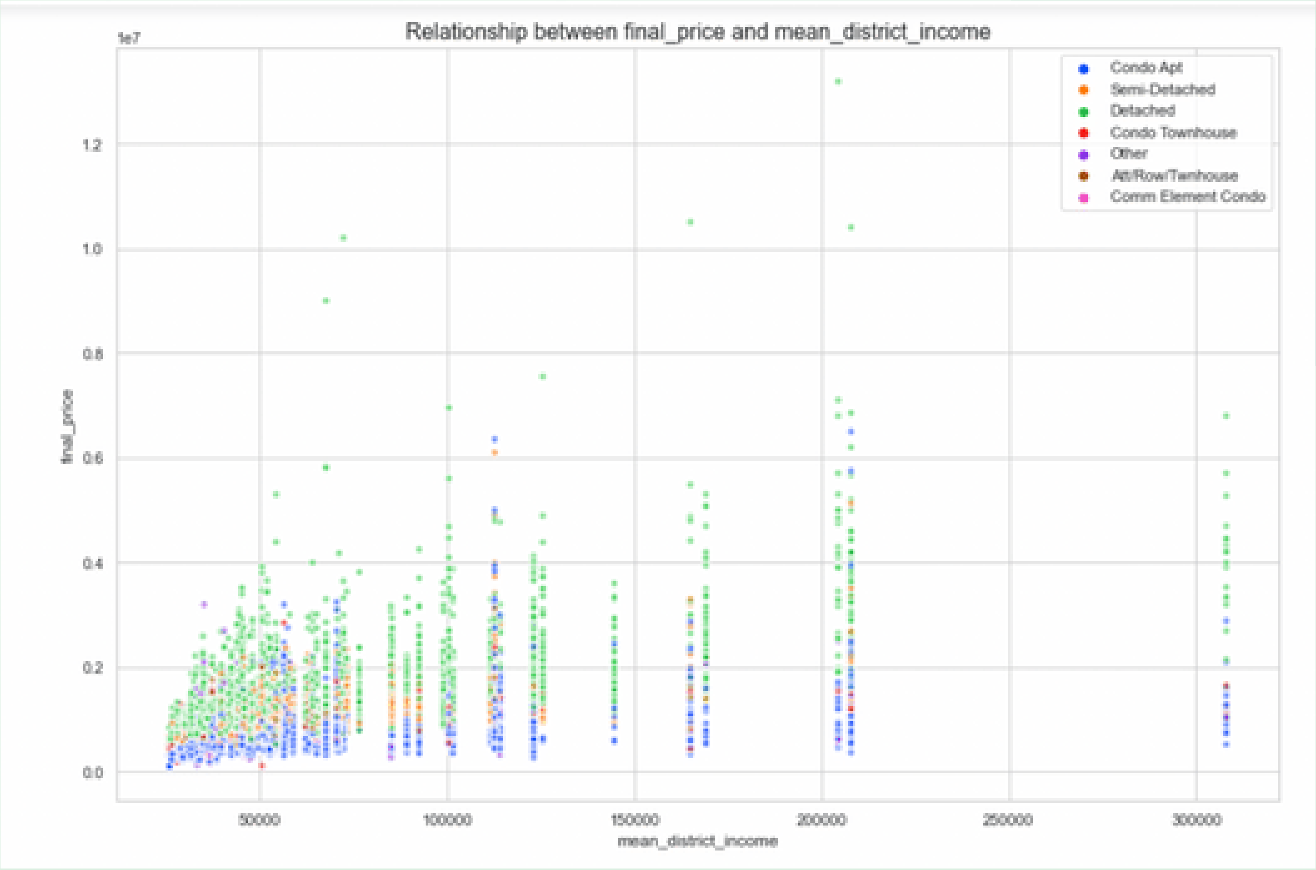
SUMMARY OF DATA ANALYSIS

BASIC EDA

Through the magics of pandas profiling, this program was able to generate a full summary of the distribution of each column in the dataset. Through this, the data was incredibly easy to interpret and see how the distribution of values including any skews, mean/median, extreme values, etc.

ADVANCED EDA

The goal with advanced Data Analysis was the explore the relationships of data to a greater depth in comparing columns to eachother. Obviously there are too many permutations to consider every combination so the focus was on the target variable (final_price) and its relationship to sqft, district_income and district.



From the data outlining final price vs average disctrict income it can be concluded that while a majority of less expensive homes (under 400,000 (0.4 x 10^7)) are districts with lower average income, as the average income increases, the range of pricing also increases with higher priced houses generally being more expensive than houses.

DATA PREPROCESSING

Processing the data is the process of converting all categorical columns into numerical values that can be interpreted by computers and categorized by importance when it comes to modelling. While a number of columns were dropped, they were not useful for the model to interpret its importance such as link, description and MLS. The type of homes were converted just into a 0 and 1 value of wether or not the row represented a house. Districts were rallied up and converted into dummy variables where each district has its own column.

INSIGHTS & MODELING

WHERE THE MACHINE LEARNING HAPPENS

LINEAR REGRESSION

THE FIRST MODEL WAS A BASIC LINEAR REGRESSION WHERE THE RESULTS WEREN'T IDEAL AS EACH FEATURE HAD A STRONG AMOUNT OF MULTICOLINEARITY AMONGST THE OTHERS SO THIS METHOD HAD TO BE SCRAPPED IN FAVOR OF MORE ADVANCED MACHINE LEARNING TECHNIQUES

ADVANCED MODELLING

THE MODELS SELECTED FOR THIS PROCESS WERE XGBREGRESSOR AND RANDOM FOREST MODELS. THESE MODELS WERE PICKED APART MATICULOUSLY AS EACH PARAMETER WAS EXAMINED AND REFINED TO CREATE A VERSION THAT COULD PREDICT PRICES WITH THE GREATEST ACCURACY

CHOOSING THE RIGHT ADVANCED MODELS

	MLA Name	MLA Parameters	Total Time MAE	MLA Train MAE Mean	MLA Test MAE Mean	MLA Test MAE 3*STD	Total Time RMSE	MLA Train RMSE Mean	MLA Test RMSE Mean	MLA Test RMSE 3*STD	Total_Time
0	BaggingRegressor	{'base_estimator': None, 'bootstrap': True, 'b...	2.110	17735.642	44670.547	4127.787	2.069	52294.969	111559.351	162391.613	4.179
1	GradientBoostingRegressor	{'alpha': 0.9, 'ccp_alpha': 0.0, 'criterion': ...	6.196	40166.111	43831.624	5158.430	6.255	88462.802	115133.071	186110.218	12.451
2	XGBRegressor	{'objective': 'reg:squarederror', 'base_score'...	4.941	25325.848	43140.058	4515.136	4.691	47070.189	106295.826	155444.522	9.631
3	AdaBoostRegressor	{'base_estimator': None, 'learning_rate': 1.0,...	2.975	82635.645	83057.312	8471.818	2.877	191564.485	193849.669	186048.439	5.852
4	LGBMRegressor	{'boosting_type': 'gbdt', 'class_weight': None...	0.588	37639.649	44192.146	6725.443	0.514	129546.898	134729.068	216389.618	1.102
5	KNeighborsRegressor	{'algorithm': 'auto', 'leaf_size': 30, 'metric...	0.020	286439.155	346659.989	20853.322	0.018	570422.912	638351.248	118930.525	0.038
6	NuSVR	{'C': 1.0, 'cache_size': 200, 'coef0': 0.0, 'd...	17.113	352394.166	352417.779	18158.293	17.037	660293.927	659260.501	119841.487	34.150
7	SVR	{'C': 1.0, 'cache_size': 200, 'coef0': 0.0, 'd...	22.223	352297.051	352423.911	18142.894	22.323	659861.726	658855.516	118741.143	44.546
8	DecisionTreeRegressor	{'ccp_alpha': 0.0, 'criterion': 'mse', 'max_de...	0.333	-0.000	57541.211	5178.571	0.332	0.000	167844.971	198247.738	0.665
9	ExtraTreesRegressor	{'bootstrap': False, 'ccp_alpha': 0.0, 'criter...	16.398	0.000	43226.261	4499.080	16.549	0.000	110397.859	158903.041	32.946
10	RandomForestRegressor	{'bootstrap': True, 'ccp_alpha': 0.0, 'criteri...	20.289	16185.637	43114.737	4273.823	20.508	47701.591	109197.002	159060.240	40.796

WIth so many models as possible candidates the best 2 were picked via experimenting with all the base model regressors and determining the lucky 2 based on the mean squared error and mean absolute error. These factors are key when assessing the quality of regressive models.

OVERALL FINDINGS

For XGBoost, Random Search was implemented however GridSearch was used for RandomForest. Despite this, results tended to be similar as trained model appears somewhat be able to interpret the testing set with an absolute error rate of 40,000 dollars.

CONCLUSIONS

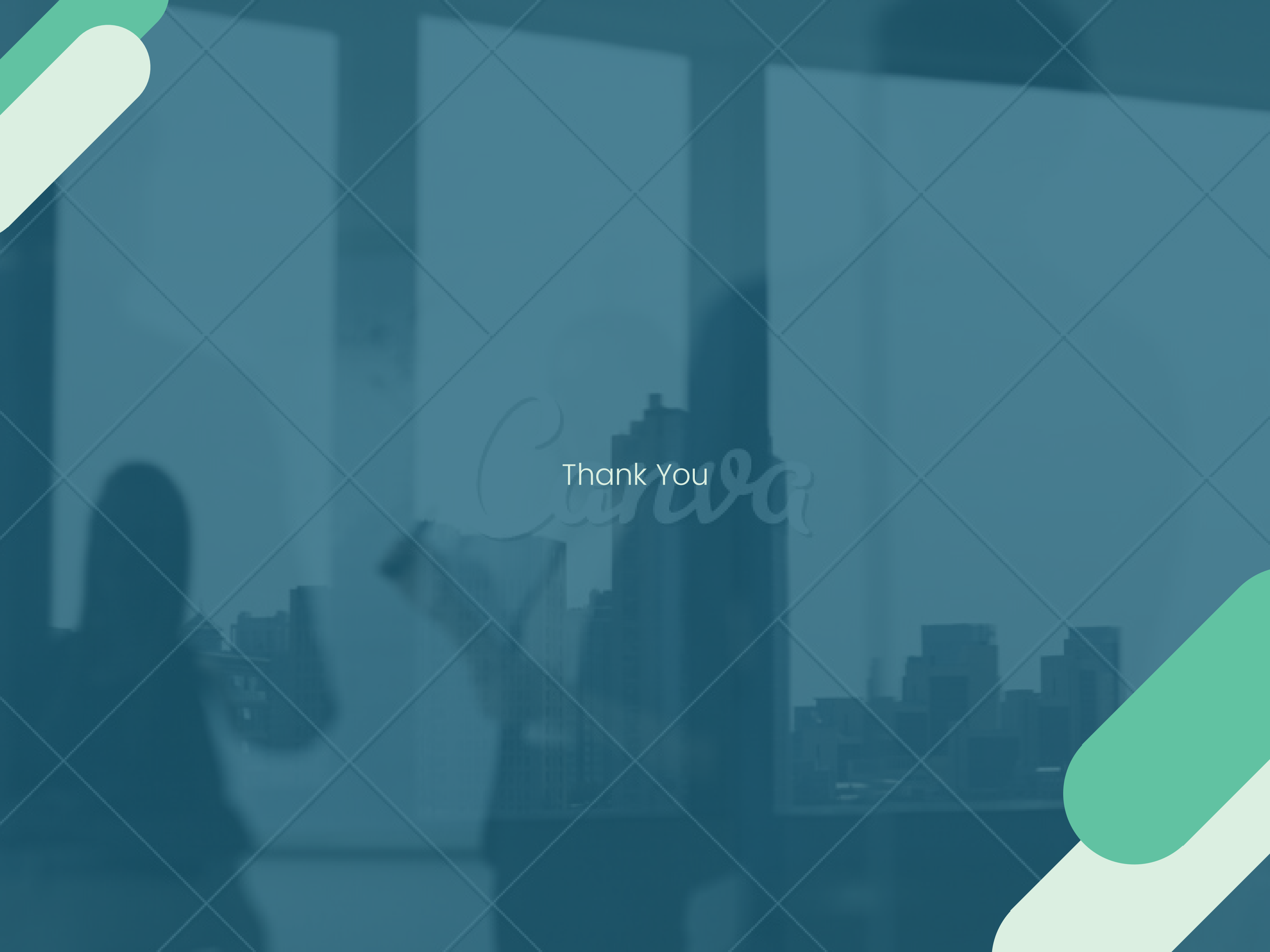
Conclusions: This may seem like a lot of money however when considering the mean/median price of a home is 450,000 and more, this means there is only about a 5-10% error rate which ideally would be considered a success given the variability of other outside factors when determining a true price of real estate. It may be worth attempting to forego the use of transformations in favor of scaling or removing both entirely future iterations. Overall consider these beginning models to be a success

NEXT STEPS

THINGS TO IMRPOVE

In revisiting the capstone for refinement (in preperation for demo day and personal use) these are the improvements to touch upon:

- Revising the dataset to include listings from previous and recent dates via better research and webscraping
- Include sold dates for time series analysis and seasonality feature importance
- When measuring accuracy, focus on not only mae and rmse but other factors as well
- Better hyperoptimization:
    - Use hyperopt and bayesian search methods
    - Use more models


Thank You