Community Data Platforms

SMARTER AND STRONGER COMMUNITIES

To: Jim Hemphill, Chloe Stuber, Christopher Morgan
From: Anna Tapp, Remy Beveridge
Re: *Charleston Housing Methods Summary*
Date: *February 4, 2021*

Contents

Introduction

Given the limitations of traditional sources in informing data-driven policy decisions, the city of Charleston asked Community Data Platforms (CDP) to generate insights about the city's housing stock. CDP has provided insights to Charleston by leveraging its extensive expertise in Machine Learning. Charleston has also asked CDP to apply its knowledge of statistics and data science to validate the rigorousness of the data provided. The following methods summary explains the process of data validation, with particular reference to the Affordable Housing Analysis.

There are many variables that can affect the housing market. Among these are macroeconomic trends, spatial differences, community-specific characteristics, and environmental features. Assets in real estate are diverse and influenced by hundreds of context-specific variables. Further, non-linear relationships between prices and variables, as well as interaction effects among variables, make housing analysis all the more complex. When prospective buyers, developers, tax assessors, and other stakeholders are investigating real estate assets, they often consult valuers, who compile recent sales evidence to generate current price estimates. The generated price estimates are usually based on a cost and sale price comparison,

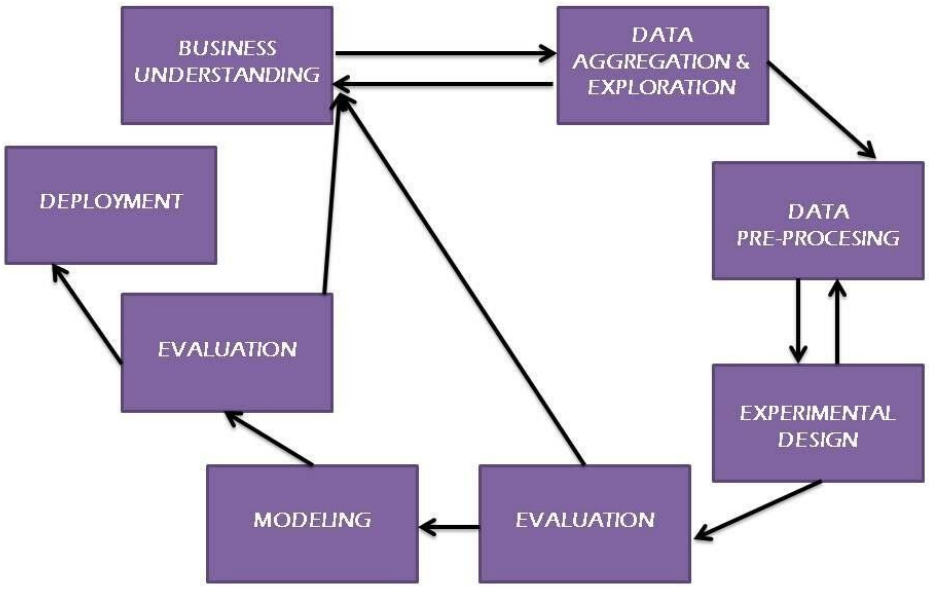
however, the process lacks a uniform and standard certification process. Thus, a house price prediction model proves very useful when making real estate decisions.

On a high level, CDP built three models, one to model the current valuations of single family residential homes and condos, one to model the rental value of all residential properties, and one to model the historical value of single family residential homes and condos. The simulations were iteratively built in order to create the most accurate and predictive models.

Process

In order to build precise and reliable models, CDP applies a sophisticated and rigorous Machine Learning process. The iterative process flow (Figure 1) allows for new discoveries and approaches to be incorporated at any point in the process.

Figure 1. *Iterative process flow*



Data ingestion

The purpose of the housing valuation model is to predict the current valuation of every housing unit in our dataset, that is, to identify the current market rates of Charleston's residential properties. Thus, housing units were our primary unit of analysis: the grist for our algorithms. Our predictive analysis was focused on single-family residential and condominium properties.

The City of Charleston provided data on the housing sale history of unique housing units belonging to three categories: city, county, municipalities.

- Floods FEMA Floodplain
- Hurricane Surge Data
- Observed Number of Road Closures

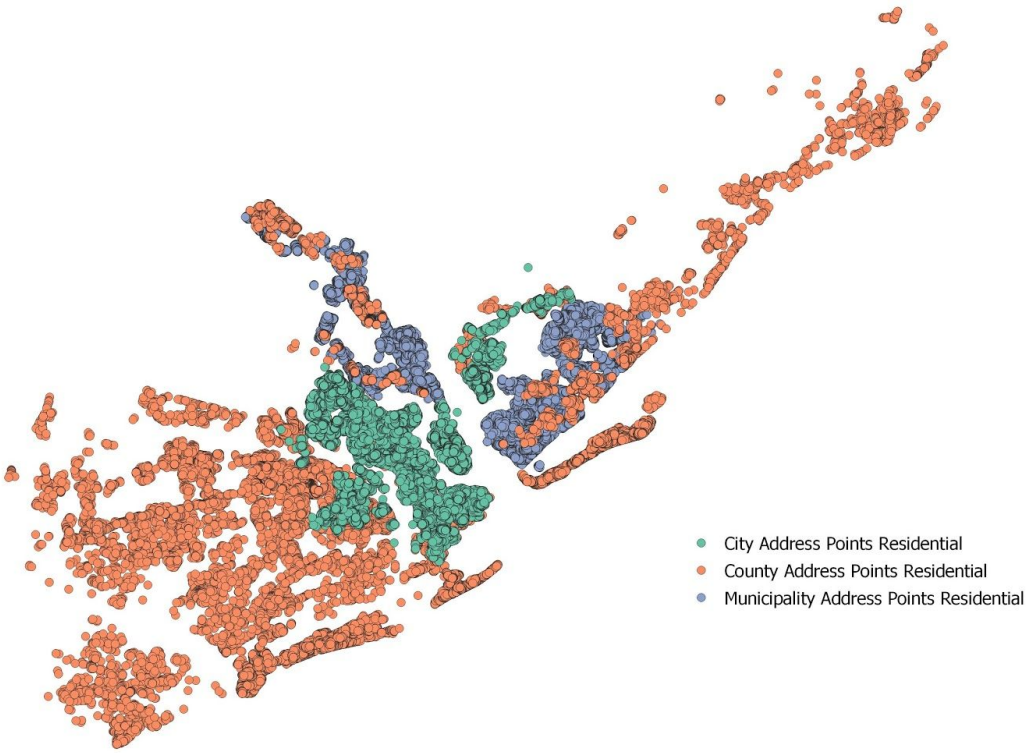
Based on research and numerous conversations with stakeholders, CDP ingested data related to the following areas.

- Public policy and regulations
- Geographic characteristics and location
- Vulnerability to storm surge and flooding
- Property attributes and improvements
- Distance to salient locations
- Advertised Real Estate Data

In order to understand the data and how the different variables interact, our team grouped housing units based on location: within the city; within the county; and within municipalities. The City of Charleston provided housing sale history data on unique housing units in each of the three categories (city, county, municipalities). Their geographic distribution is visualized below in Figure 2.

Visual diagnostics

Figure 2. *Geographic Distribution of Housing Units*



- City: 52,926 properties
    - 9,232 condos
    - 91 mixed
    - 2,674 multi-family residential
    - 40,929 single-family residential

- County: 48,188 properties
    - 3,919 condos
    - 728 multi-family residential
    - 43,541 single-family residential
- Municipalities: 63,227 properties
    - 6,014 condos
    - 5,025 multi-family residential

- 52,188 single-family residential

Non-residential properties and units within subsidised housing complexes were excluded from the dataset. This left 155,624 unique properties to be assessed.

Data diagnostics

As part of the Machine Learning process, our data science team used statistical tests and visual diagnostics to assess the quality of the data gathered. We identified missing, inconsistent, and duplicate values in the dataset and used industry-standard methodologies to resolve data quality issues. When necessary for modeling purposes or data cleaning, variables were transformed based on their type (e.g., numeric, factor, etc.) and/or standardized.

Ground truth

In the realm of Machine Learning, the “ground truth” is the value measured in the training, validation, and test datasets. Typically, the output is a label, or classification, but given that the housing valuation is a regression problem, it outputs a number as opposed to a label. The ground truth in this case is the value of the house.

Train, validation, and test

CDP acquired sales data from the Charleston and Berkeley County assessors, which included sales and improvement history. However, given that most houses do not sell every year, CDP used the most recent sales information available for each unit as the indicator for the valuation. The valuation was done via sophisticated Machine Learning techniques based on principles of statistics and data science. As is standard, the data include some noise, in this case originating from the outdated sales information and potentially low sales prices from assessors, which are sometimes below market value. However, the target is to make the model ‘learn’

the underlying patterns in the data in order to produce the best result. No model gives 100% accuracy, but the goal is to be as close as possible.

In a typical valuation, the relevant data is divided into two sets: training and test. However, given the complexity of the project at hand, CDP divided the data into three sets: train, validation (evaluation) and test. The train dataset is used to train the model candidate. Before testing the model on the test dataset, a set built from a random selection of the data points with known sales prices, the model is tested on a validation dataset. The added validation step ensures an unbiased evaluation of the model candidate when applied to the training dataset. The validation step also makes the tuning of the model's hyperparameters possible to ensure that the most optimal combination of hyperparameters is chosen. Once the model is completely trained and validated, it is tested based on the test dataset to provide the final evaluation of the model candidate. Finally, the outputs for the test and train sets are compared to evaluate how well the model has learned. The model has learned when it is stable and properly fits the data. The process can be reiterated to achieve the lowest possible error rate and the highest possible accuracy (Figure 3).

Figure 3. *Train, validation, and test datasets*

Feature engineering

Feature engineering involves using domain knowledge to extract features from data via data mining techniques. Feature engineering can be used to combine variables when their predictive ability depends on their being evaluated simultaneously (for instance, longitude and latitude) or, conversely, when a single component of a variable should be isolated because it has greater predictive value on its own (for instance, time of day in a date-time variable). Feature engineering is a crucial part of the Machine Learning process because choosing the optimal features ensures model stability and improves performance of the Machine Learning models. For this valuation, CDP used feature engineering to ensure that the variables used in the model were relevant, specific and highly predictive of home valuations.

CDP used two categories of features:

- Brainstorm features: based on the business problem, e.g. applying external knowledge of short-term rentals to create variables that were specific to Charleston's short-term rental market
- Devise features: automatic feature extraction, manual feature construction and mixtures of the two (e.g. mean, counts, etc.)

Feature selection

Next, we used a process called feature selection, whereby features not relevant to the model are eliminated, either because of deficiencies in the information provided by the variable, or because of the variable's covarying relationship. Using feature selection, we eliminated features not predictive of home valuations.

Dimensionality reduction

We then conducted a dimensionality reduction, or a transformation of the data from a high-dimensional space to a low-dimensional space, which ensures that the low-dimensional space retains the meaningful properties of the raw data. Two

distinct methods of dimensionality reduction were employed in order to assess the significance of the selected variables:

- Filter methods
- Wrapper methods

We relied upon two methods of dimensionality reduction so as to take full advantage of the distinct information provided by each. Filter methods assess a particular feature's relevance by testing its correlation to the dependent variable. Wrapper methods, on the other hand, assess the relevance of *a subset* of features by training a model using every possible permutation of features and selecting the subset that has the best results, judged by standard model evaluation techniques.

Further, while filtering approaches use statistical methods, wrapper methods use a technique called cross-validation. Cross-validation ensures that the model does not overfit to the training data. When a model is overfitted, it is highly susceptible to changes in the underlying data set and thus cannot generalize to other data, making it an ineffective predictive tool. Using wrapping in addition to filtering allowed us to generate a housing valuation model that can generalize to other data and thus be used as a predictive tool.

We used forward selection, which starts with a null model and fits n to simple linear regression models, each having one predictor and one intersect (n equals the number of predictors). The model having the lowest residual sum of squares is chosen, then the search continues using the remaining n-1 predictors to find the model to be added next. The process continues until there is no more improvement in the model.

Model evaluation and selection

Three different types of Machine Learning models were chosen for the home valuations analysis:

- Linear regression
- Random Forest
- XGBoost

For each model (housing valuations, rental valuations, and historical valuations), a subset of models was developed for the three housing unit datasets (City, County, and Municipality), for both dimensionality reduction techniques outlined above (Filter and Wrapper), and for each minimum and maximum threshold for County and Municipal datasets.

First, hyperparameter tuning was used to find the optimal tradeoff between bias and variance for each model (Linear Regression, Random Forest, XGBoost). Then, each model was assessed using five model evaluation metrics:

- Mean absolute error (MAE)
- Mean squared error (MSE)
- Root mean squared error (RMSE)
- R^2
- Adjusted R^2

Using the models described above (Linear Regression, Random Forest, XGBoost) and the Wrapper Method for each dimensionality reduction, we found the most predictive *subset* of housing valuations and rental valuations features for the properties located in each of the three datasets. The key distinction between a predictive subset and a predictive feature is that in a predictive *subset*, because of the statistical relationships between the subset's components, the variables are only predictive when evaluated in tandem. Therefore, analysis of the individual features of a subset is not useful, given that the features can only predict home valuations when analyzed in tandem.

Housing valuation

Findings

Table 1. *Evaluation results for each model (city, county, municipality)*

Table 1a. *City evaluation results*

Dimensionality reduction	Filtered				Wrapper					
Model	LR		RF		LR		RF		XGBoost	
Data set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R2	0.5476611	0.5577250	0.7937433	0.7436395	0.5379838	0.5527573	0.7724164	0.7155416	0.7868762	0.7363583
Adj R2	0.5470568	0.5523478	0.7934764	0.7406211	0.5376754	0.5500549	0.7722739	0.7139308	0.7865737	0.7329500
MAE	0.2934803	0.2943086	0.1898926	0.2153220	0.2958807	0.2944002	0.1966401	0.2239994	0.1893645	0.2141036
MSE	0.1910869	0.1853691	0.0871315	0.1074475	0.1951750	0.1874513	0.0961409	0.1192241	0.0900324	0.1104992
RMSE	0.4371349	0.4305452	0.2951804	0.3277918	0.4417862	0.4329564	0.3100659	0.3452884	0.3000540	0.3324142

Table 1b. *County evaluation results*

Cut off	75K+	
Dimensionality reduction	Filtered	Wrapper

Model	LR		RF		LR		RF		XGBoost	
Data set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R2	0.5011574	0.4768877	0.8242873	0.6921354	0.5116177	0.4928527	0.8095060	0.6900261	0.7980982	0.7047548
Adj R2	0.5003973	0.4696222	0.8240307	0.6880401	0.5112179	0.4881785	0.8093246	0.6873493	0.7977263	0.6997855
MAE	0.4348645	0.4436536	0.2425310	0.3193802	0.4325131	0.4430735	0.2528259	0.3236783	0.2603250	0.3155738
MSE	0.3410074	0.3660735	0.1201167	0.2154433	0.3338568	0.3549012	0.1302212	0.2169194	0.1380195	0.2066123
RMSE	0.5839580	0.6050401	0.3465785	0.4641587	0.5778034	0.5957359	0.3608618	0.4657461	0.3715098	0.4545463

Table 1c. *Municipality evaluation results*

Cut off	75K+									
Dimensionality reduction	Filtered				Wrapper					
Model	LR		RF		LR		RF		XGBoost	
Data set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R2	0.5736926	0.5862723	0.7847822	0.7262182	0.5780363	0.5897566	0.7825567	0.7237768	0.7410598	0.7208587
Adj R2	0.5733668	0.5834092	0.7846259	0.7244189	0.5777784	0.5874884	0.7824321	0.7223456	0.7408322	0.7186349
MAE	0.2956570	0.2964255	0.1904818	0.2197979	0.2929779	0.2948691	0.1922830	0.2192729	0.2122296	0.2264598

MSE	0.18915 27	0.18500 28	0.09549 22	0.122424 5	0.1872 254	0.18344 48	0.096479 7	0.1235162	0.114891 8	0.1248211
RMSE	0.43 491 00	0.43011 90	0.30901 81	0.349892 1	0.4326 955	0.42830 46	0.310611 7	0.3514487	0.338957 0	0.3533002

Based on results from the five model evaluation metrics, we ultimately selected the following models for each dataset

- City: XG Boost using a Wrapper method with threshold values of 100,000 - 10,000,000
- County: XG Boost using a Wrapper method with threshold values of 75,000 - 10,000,000
- Municipality: XG Boost using a Wrapper method with threshold values of 75,000 - 10,000,000

For each of the three datasets, we have indicated below the subset of features that is predictive of home valuations (Table 2).

Table 2. *Housing unit dataset for each model (city, county, municipality)*

Table 2a. *City housing unit dataset*

Variable	Description
addruse_desc	Land use categories
county_name	County name where address point belongs
slosh_cat2_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat3_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat4_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category

slosh_cat5_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
old_city_district	It shows name of old city district if an address point belongs to it
number_of_road_closures	How many times the closest road was closed
fld_zone_2004	Categorical variable for flooding zone type in 2004. It shows flood zone code where point location belongs
fld_zone_2016	Categorical variable for flooding zone type in 2016. It shows flood zone code where point location belongs
building_flooded	It shows if any part of the building (building footprint that corresponds to the address point) overlaps a flood zone (i.e. not X). This includes flood zones from both years 2004 and 2016
legal_acreage	Legal acreage
city_limits_dist	Distance to the city limits. Distance in feet
grade	Assessor's rating of the condition of the structure
type_of_foundation	Type of foundation
type_of_roof	Type of roof
heating	Type of heating unit
exterior_wall_materials	The type of wall materials used for the exterior of the house
number_of_half_bathrooms	Number of half bathrooms
number_of_full_bathrooms	Number of full bathrooms

number_of_living_units	Number of living units
type_Commercial	Number of commercial permits at a subdivision
type_Residential	Number of residential permits at a subdivision
LNB_STR_max	Max number of nights booked at a zipcode since Sep 2019 until Aug 2020
cafe_dist_nan	Binary indicator to show no data was available for cafe_dist and was imputed with the median
school_dist_nan	Binary indicator to show no data was available for school_dist and was imputed with the median
year_annexed_nan	Binary indicator to show no data was available for year_annexed and was imputed with the median
elevation_nan	Binary indicator to show no data was available for elevation and was imputed with the median
building_area_nan	Binary indicator to show no data was available for building_area and was imputed with the median
heated_space_nan	Binary indicator to show no data was available for heated_space and was imputed with the median
eff_year_built_nan	Binary indicator to show no data was available for eff_year_built and was imputed with the median
imp_DWELL_nan	Binary indicator to show no data was available for imp_DWELL and was imputed with 0
PR_STR_range_nan	Binary indicator to show no data was available for PR_STR_range and was imputed with 0

type_Total_per_nan	Binary indicator to show no data was available for type_Total_per and was imputed with 0

Table 2b. *County housing unit dataset*

Variable	Description
city_limits	Binary indicator if the property is away from the city limits
slosh_cat2_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat3_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat4_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat5_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
base_zoning_code	Base zoning code where address point belongs
rivers_dist	Distance to closest river. Simple euclidean distance in feet.
number_of_road_closures	How many times the closest road was closed
building_flooded	It shows if any part of the building (building footprint that corresponds to the address point) overlaps a flood zone (i.e. not X). This includes flood zones from both years 2004 and 2016

parcel_area	Area of a parcel where the address point belongs
condition_x	Code for the assessor's rating of the condition of the structure fetched from the building table
grade	Assessor's rating of the condition of the structure
type_of_foundation	Type of foundation
type_of_roof	Type of roof
cooling	Presence and type of AC unit
exterior_wall_materials	The type of wall materials used for the exterior of the house
number_of_half_bathrooms	Number of half bathrooms
number_of_living_units	Number of living units
condition_y	Code for the assessor's rating of the condition of the structure fetched from the building table
type_Total_per	Proportion of the total of Bed_breakfast, Commercial and Residential permits at a Subdivision
PR_STR_max	Max number of private room advertisements at a Zipcode since Sep 2019 until Aug 2020
school_dist_nan	Binary indicator to show no data was available for school_dist and was imputed with the median
university_dist_nan	Binary indicator to show no data was available for university_dist and was imputed with the

	median
elevation_nan	Binary indicator to show no data was available for elevation and was imputed with the median
building_area_nan	Binary indicator to show no data was available for building_area and was imputed with the median
eff_year_built_nan	Binary indicator to show no data was available for eff_year_built and was imputed with the median
number_of_living_units_nan	Binary indicator to show no data was available for number_of_living_units and was imputed with the mode
imp_DWELL_nan	Binary indicator to show no data was available for imp_DWELL and was imputed with 0
imp_POOL_nan	

Table 2c. *Municipality housing unit dataset*

Variable	Description
slosh_cat2_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat3_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat4_2011	Storm Surge by Hurricane Category. It shows

	true if an address point is within a category.
slosh_cat5_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category.
number_of_road_closures	How many times the closest road was closed.
fld_zone_2004	Categorical variable for flooding zone type in 2004. It shows flood zone code where point location belongs.
parcel_area	Area of a parcel where the address point belongs.
legal_acreage	Legal acreage
municipality	Municipality the address point belongs
grade	Assessor's rating of the condition of the structure
type_of_foundation	Type of foundation
type_of_roof	Type of roof
exterior_wall_materials	The type of wall materials used for the exterior of the house
number_of_half_bathrooms	Number of half bathrooms
number_of_living_units	Number of living units
PR_STR_range	Range of the private room advertisements at a Zipcode since Sep 2019 until Aug 2020
university_dist_nan	Binary indicator to show no data was available for university_dist and was imputed with the

	median
elevation_nan	Binary indicator to show no data was available for elevation and was imputed with the median
building_area_nan	Binary indicator to show no data was available for building_area and was imputed with the median
eff_year_built_nan	Binary indicator to show no data was available for eff_year_built and was imputed with the median
number_of_living_units_nan	Binary indicator to show no data was available for number_of_living_units and was imputed with the mode
imp_ATTGAR_nan	Binary indicator to show no data was available for imp_ATTGAR and was imputed with 0
imp_DWELL_nan	Binary indicator to show no data was available for imp_DWELL and was imputed with 0

Variables present in all subsets

- "slosh_cat2_2011"
- "slosh_cat3_2011"
- "slosh_cat4_2011"
- "slosh_cat5_2011"
- "number_of_road_closures"
- "grade"
- "type_of_foundation"
- "type_of_roof"

- "exterior_wall_materials"
- "number_of_half_bathrooms"
- "number_of_living_units"
- "elevation_nan"
- "building_area_nan"
- "eff_year_built_nan"
- "imp_DWELL_nan"

County Only

- "city_limits"
- "base_zoning_code"
- "rivers_dist"
- "condition_x"
- "cooling"
- "condition_y"
- "type_Total_per"
- "PR_STR_max"
- "imp_POOL_nan"

City Only

- "addruse_desc"
- "county_name"
- "old_city_district"
- "fld_zone_2016"
- "city_limits_dist"
- "heating"
- "number_of_full_bathrooms"
- "type_Commercial"
- "type_Residential"
- "LNB_STR_max"
- "cafe_dist_nan"
- "year_annexed_nan"

- "heated_space_nan"
- "PR_STR_range_nan"
- "type_Total_per_nan"

Municipality Only

- "municipality"
- "PR_STR_range"
- "imp_ATTGAR_nan"

Common variables in city and county

- "slosh_cat2_2011"
- "slosh_cat3_2011"
- "slosh_cat4_2011"
- "slosh_cat5_2011"
- "number_of_road_closures"
- "building_flooded"
- "grade"
- "type_of_foundation"
- "type_of_roof"
- "exterior_wall_materials"
- "number_of_half_bathrooms"
- "number_of_living_units"
- "school_dist_nan"
- "elevation_nan"
- "building_area_nan"
- "eff_year_built_nan"
- "imp_DWELL_nan"

Unique variables in city and county (do not intersect)

- "addruse_desc"
- "base_zoning_code"
- "cafe_dist_nan"

- "city_limits"
- "city_limits_dist"
- "condition_x"
- "condition_y"
- "cooling"
- "county_name"
- "fld_zone_2004"
- "fld_zone_2016"
- "heated_space_nan"
- "heating"
- "imp_POOL_nan"
- "legal_acreage"
- "LNB_STR_max"
- "number_of_full_bathrooms"
- "number_of_living_units_nan"
- "old_city_district"
- "parcel_area"
- "PR_STR_max"
- "PR_STR_range_nan"
- "rivers_dist"
- "type_Commercial"
- "type_Residential"
- "type_Total_per"
- "type_Total_per_nan"
- "university_dist_nan"
- "year_annexed_nan"

Common variables in city and municipality

- "slosh_cat2_2011"
- "slosh_cat3_2011"
- "slosh_cat4_2011"
- "slosh_cat5_2011"

- "number_of_road_closures"
- "fld_zone_2004"
- "legal_acreage"
- "grade"
- "type_of_foundation"
- "type_of_roof"
- "exterior_wall_materials"
- "number_of_half_bathrooms"
- "number_of_living_units"
- "elevation_nan"
- "building_area_nan"
- "eff_year_built_nan"
- "imp_DWELL_nan"

Unique variables in city and municipality (do not intersect)

- "addruse_desc"
- "building_flooded"
- "cafe_dist_nan"
- "city_limits_dist"
- "county_name"
- "fld_zone_2016"
- "heated_space_nan"
- "heating"
- "imp_ATTGAR_nan"
- "LNB_STR_max"
- "municipality"
- "number_of_full_bathrooms"
- "number_of_living_units_nan"
- "old_city_district"
- "parcel_area"
- "PR_STR_range"
- "PR_STR_range_nan"

- "school_dist_nan"
- "type_Commercial"
- "type_Residential"
- "type_Total_per_nan"
- "university_dist_nan"
- "year_annexed_nan"

Common variables in county and municipality

- "slosh_cat2_2011"
- "slosh_cat3_2011"
- "slosh_cat4_2011"
- "slosh_cat5_2011"
- "number_of_road_closures"
- "parcel_area"
- "grade"
- "type_of_foundation"
- "type_of_roof"
- "exterior_wall_materials"
- "number_of_half_bathrooms"
- "number_of_living_units"
- "university_dist_nan"
- "elevation_nan"
- "building_area_nan"
- "eff_year_built_nan"
- "number_of_living_units_nan"
- "imp_DWELL_nan"

Unique variables in county and municipality (do not intersect)

- "base_zoning_code"
- "building_flooded"
- "city_limits"
- "condition_x"

- "condition_y"
- "cooling"
- "fld_zone_2004"
- "imp_ATTGAR_nan"
- "imp_POOL_nan"
- "legal_acreage"
- "municipality"
- "PR_STR_max"
- "PR_STR_range"
- "rivers_dist"
- "school_dist_nan"
- "type_Total_per"

Long-term rentals valuation

Process

Two distinct processes were used to model the rental valuations for the properties specified above: single-family, condos and mixed properties were modeled according to one process; multi-residential properties were modeled according to a second process.

CDP conducted an exhaustive analysis of the impact of increased short-term rental properties on long-term rental prices in order to evaluate current industry best practices. Property owners who can make more money through short-term rentals than long-term rentals will generally choose the former. However, there is some debate about the impact that short-term rentals have on the long-term rental property market. Intuitively, it would seem that if a large number of properties transition from the long-term rental market to the short-term rental market, the decreased supply would drive up the prices of long-term rentals (barring the simultaneous construction of long-term rental properties in the area contributing to real estate stock).

However, if the short-term rental market is small compared to long-term rental market, changes in the long-term market might have a tiny effect or no effect at all on the prices of long-term rentals. Other research demonstrates a positive correlation between the presence of short-term rentals and long-term rental prices. According to the 2019 study *The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb,* "Airbnb has a positive impact on house prices and rents. This effect is stronger in zip codes with a lower share of owner-occupiers, consistent with non-owner-occupiers being more likely to reallocate their homes from the long- to the short-term rental market. At the median owner-occupancy rate zip code, we find that a 1% increase in Airbnb listings leads to a 0.018% increase in rents and a 0.026% increase in house prices." Conversely, other studies have found that the presence of short-term rentals makes the price binding stronger on lower price properties, preventing some people from getting housing.

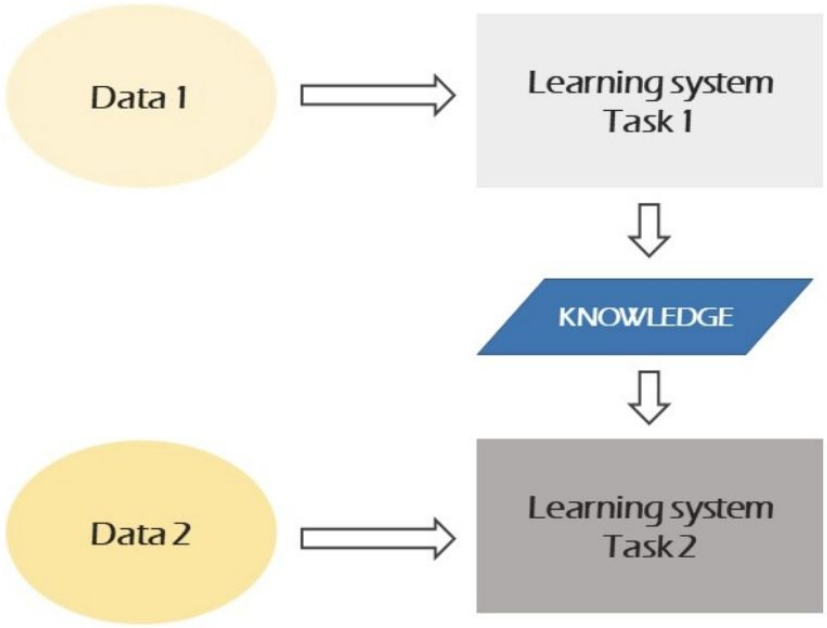
To test these effects as well as to see whether short-term rentals are significant to the model, CDP incorporated data on short-term listings, grouped by zip code, ranging from September 2019 to August 2020.

- Entire home advertisements
- Private room advertisements
- Listing nights booked
- Number of booked properties
- Permit at the zip code level.

Model evaluation and selection

The data only included rental records for 393 properties (approximately 0.2532% of all properties in Charleston). While this presented a challenge, CDP's Data Science Team was able to overcome this hurdle through a process called transfer learning (Figure 4).

Figure 4. *Transfer learning*



While traditional learning is rather isolated and no knowledge is retained, in transfer learning the knowledge can be leveraged and transferred from one model to another. Transfer learning proves especially useful in situations like this one, where there is limited data available. Using transfer learning, CDP created a synthetic dataset to fill in the missing rental prices (Table 3). The synthetic dataset is constructed using tiers of rental prices as percentages of the values of the properties.

Table 3. *Rental prices as percentage of property value*

Bin	Average
250K	1.10%
500K	0.62%
1M	0.48%
1M+	0.22%

Further, specific tasks were applied to this dataset and then transferred to the dataset which was used for further modeling. Rents for single-family properties as well as mixed properties and condos were modeled and evaluated using three different models:

- Linear regression
- Random Forest
- XGBoost

Findings

Table 4. *Evaluation results for each model (city, county, municipality)*

Table 4a. *City evaluation results*

Dimensionality reduction	Filtered				Wrapper					
Model	LR		RF		LR		RF		XGBoost	
Data set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R2	0.4105005	0.3766152	0.7167895	0.6422874	0.4316571	0.4113744	0.7297877	0.6614534	0.7606722	0.6782514
Adj R2	0.4099157	0.3703729	0.7165213	0.6388698	0.4312471	0.4070993	0.7296050	0.6591494	0.7603807	0.6742882
MAE	0.1949737	0.1996756	0.1318163	0.1451437	0.1924320	0.1961421	0.1297270	0.1433512	0.1225903	0.1405936
MSE	0.0684118	0.0740198	0.0328668	0.0424743	0.0659566	0.0698925	0.0313583	0.0401985	0.0277741	0.0382040
RMSE	0.2615565	0.2720658	0.1812919	0.2060929	0.2568201	0.2643720	0.1770828	0.2004957	0.1666558	0.1954584

Table 4b. *County evaluation results*

Dimensionality reduction	Filtered				Wrapper					
Model	LR		RF		LR		RF		XGBoost	
Data set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R2	0.3948455	0.4102898	0.5753302	0.5409705	0.3984639	0.4151075	0.5864643	0.5540644	0.6019027	0.5567765
Adj R2	0.3943770	0.4056917	0.5750220	0.5376167	0.3979983	0.4105469	0.5861642	0.5508062	0.6015174	0.5524481
MAE	0.2760681	0.2770808	0.2236806	0.2347321	0.2750872	0.2759448	0.2188894	0.2297214	0.2148721	0.2296408
MSE	0.1350859	0.1323830	0.0947971	0.1030467	0.1342782	0.1313015	0.0923117	0.1001073	0.0888654	0.0994985
RMSE	0.3675404	0.3638448	0.3078914	0.3210089	0.3664399	0.3623555	0.3038284	0.3163970	0.2981032	0.3154338

Table 4c. *Municipality evaluation results*

Dimensionality reduction	Filtered				Wrapper					
Model	LR		RF		LR		RF		XGBoost	
Data set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
R2	0.4322187	0.4332427	0.7333723	0.7027272	0.4239765	0.4228477	0.7202325	0.7004324	0.7361392	0.7133728
Adj R2	0.4311839	0.4294360	0.7332048	0.7008487	0.4236146	0.4192007	0.7300736	0.6986584	0.7359526	0.7113335

	6									
MAE	0.233516 4	0.2405429	0.1591455	0.1702172	0.2346244	0.2413774	0.1588051	0.1689268	0.1570757	0.1658048
MSE	0.090872 1	0.0951158	0.0426731	0.0498897	0.0921912	0.0968603	0.0431757	0.0502748	0.0422303	0.0481031
RMSE	0.301449 9	0.3084084	0.2065748	0.2233599	0.3036301	0.3112238	0.2077875	0.2242204	0.2055001	0.2193241

Based on this evaluation, we ultimately selected the following models for each dataset:

- City: XG Boost using a wrapper filtering method
- County: XG Boost using a wrapper filtering method
- Municipalities: XG Boost using a wrapper filtering method

Since prices tend to fluctuate over time, CDP outputted the range of rental prices based on 90% confidence intervals. The confidence level represents the proportion of confidence intervals that contain the true value of the unknown parameter (population). With a 90% confidence level, the range contains the true value of the parameter.

Multi-residential properties

Multi-residential properties contain multiple housing units within one complex, arranged either adjacent to one another or stacked, as in an apartment building. This creates certain restrictions and complications for the modeling process, given that without hard data onhand, it is impossible to know the layout of a particular multi-residential property.

Further, each multi-residential property has different characteristics which all drive price (including varying quality, floor plan, location etc.). To overcome these restrictions, CDP collected as much data as possible on multi-residential properties

in the city and county with real characteristics and prices for every type of apartment (size, floor plan, number of bedrooms and bathrooms, number of units within the building etc.) and ran feature engineering to generate additional information on distances to the places of interest (supermarkets, schools, cinemas), maturity of the building, neighbourhood reputation etc.

CDP ran unsupervised clustering via k-means clustering to partition the data into k clusters, grouping similar data points into clusters. Statistical tests established that the optimal number of clusters are 3 for the city and 2 for the county. Thus, all of the multi-residential properties in the city were divided into 3 clusters and all of the multi-residential properties in the county were divided into 2 clusters.

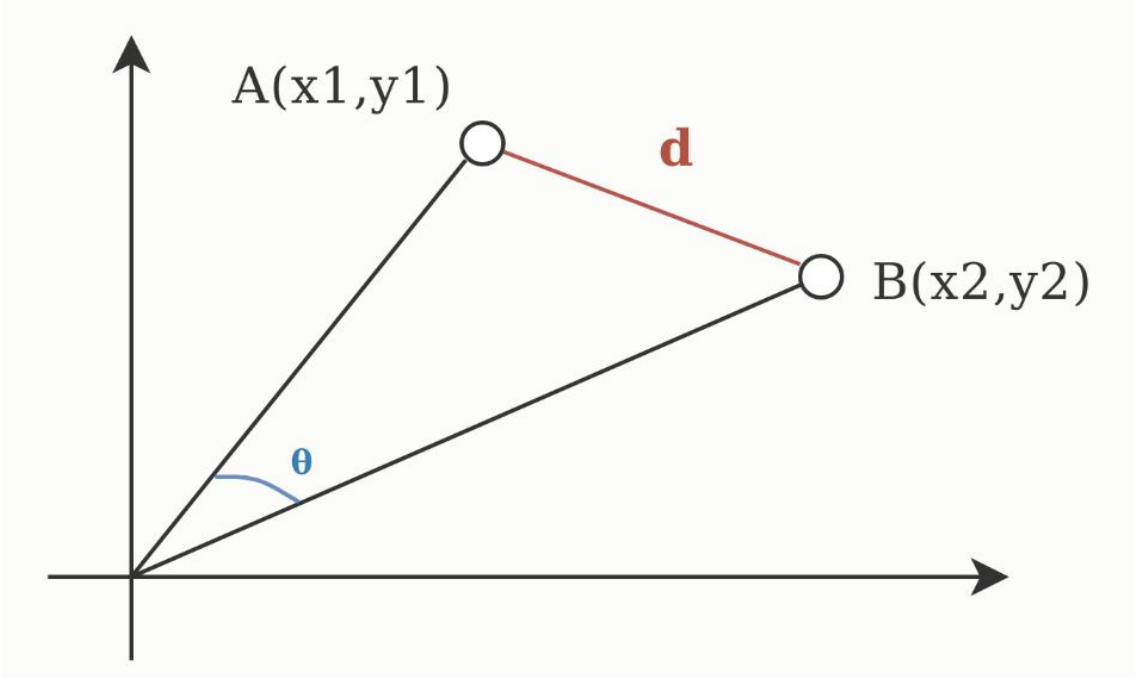
Further, a knn algorithm was applied to a small subset of the data within the clusters to predict new data points (rental prices) based on the rental prices of known properties. The knn algorithm uses 'feature similarity' to predict the value of a new data point by assigning it a value correlated to how closely it resembles the points in the training set. In order to assign a value to the new data point, the distance between the new point and each training point is calculated, then the closest k data points are selected. The average of these data points is the final prediction for the new data point.

Next, the output of the feature similarity calculation was coupled with a separate similarity analysis to create an adjustment index for the final rental price. The second similarity analysis was run within each of the clusters to establish how similar or dissimilar the modelled properties were to the properties for which CDP real data. Similarity is always measured in a vector space, which makes it simpler than measuring objects in nD space.

There are a number of common methods for similarity analysis, two of which are euclidean similarity and cosine distance. Euclidean distance measures the distance between two vectors. Cosine similarity is calculated using the cosine of the angle

between two vectors of interest. Cosine similarity measures direction (and not magnitude). (Figure 4).

Figure 4. *Cosine similarity (θ) and Euclidean distance (d)*



Euclidean distance is analogous to using a ruler to measure the distance between two vectors. Cosine distance, on the other hand, is not affected by vector length or magnitude. Each of the properties was given a similarity score based on its similarity to the properties for which we had real data.

Using the similarity analysis, we developed a weighted index and each property rent price was adjusted based on the index. Unlike the model for condominiums, single-family and mixed properties, the model for the multi-residential properties distinguished between lower and higher prices, so as to accommodate the full spectrum of apartment types within a particular building.

Table 5. *Rental unit dataset for each model (city, county, municipality).*

Table 5a. *City rental unit dataset*

Variable	Description
slosh_cat2_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat3_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat4_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat5_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
old_city_district	It shows a name of old city district if an address point belongs to it
number_of_road_closures	How many times the closest road was closed
fld_zone_2004	Categorical variable for flooding zone type in 2004. It shows flood zone code where point location belongs
legal_acreage	Legal acreage
city_limits_dist	Distance to the city limits (feet)
grade	Assessor's rating of the condition of the structure
type_of_foundation	Type of foundation
number_of_half_bathrooms	Number of half bathrooms

number_of_living_units	Number of living units
type_Commercial	No of commercial permits at a subdivision
type_Residential	No of residential permits at a subdivision
LNB_STR_range	Range of the number of nights booked at a zipcode since Sep 2019 until Aug 2020
playground_dist_nan	Binary indicator to show no data was available for playground_dist and was imputed with the median
school_dist_nan	Binary indicator to show no data was available for school_dist and was imputed with the median
year_annexed_nan	Binary indicator to show no data was available for year_annexed and was imputed with the median
elevation_nan	Binary indicator to show no data was available for elevation and was imputed with the median
building_area_nan	Binary indicator to show no data was available for building_area and was imputed with the median
heated_space_nan	Binary indicator to show no data was available for heated_space and was imputed with the median
eff_year_built_nan	Binary indicator to show no data was available for eff_year_built and was imputed with the median
number_of_living_units_nan	Binary indicator to show no data was available for number_of_living_units and was imputed with the mode
imp_DWELL_nan	Binary indicator to show no data was available for imp_DWELL and was imputed with 0

PR_STR_range_nan	Binary indicator to show no data was available for PR_STR_range and was imputed with 0
type_Total_per_nan	Binary indicator to show no data was available for type_Total_per and was imputed with 0

Table 5b. *County rental unit dataset*

Variable	**Description**
city_limits	Binary indicator if the property is away from the city limits
slosh_cat2_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat3_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat4_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat5_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
number_of_road_closures	How many times the closest road was closed
fld_zone_2004	Categorical variable for flooding zone type in 2004. It shows flood zone code where point location belongs
type_of_foundation	Type of foundation
type_of_roof	Type of roof
number_of_half_bathroom	Number of half bathrooms

s	
number_of_living_units	Number of living units
type_Total_per	Proportion of the total of Bed_breakfast, Commercial and Residential permits at a Subdivision
PR_STR_max	Max number of private room advertisements at a Zipcode since Sep 2019 until Aug 2020
school_dist_nan	Binary indicator to show no data was available for school_dist and was imputed with the median
university_dist_nan	Binary indicator to show no data was available for university_dist and was imputed with the median
elevation_nan	Binary indicator to show no data was available for elevation and was imputed with the median
building_area_nan	Binary indicator to show no data was available for building_area and was imputed with the median
eff_year_built_nan	Binary indicator to show no data was available for eff_year_built and was imputed with the median
number_of_living_units_nan	Binary indicator to show no data was available for number_of_living_units and was imputed with the mode
imp_DWELL_nan	Binary indicator to show no data was available for imp_DWELL and was imputed with 0

Table 5c. *Municipality Rental Unit Dataset*

Variable	**Description**

slosh_cat3_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat4_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
slosh_cat5_2011	Storm Surge by Hurricane Category. It shows true if an address point is within a category
number_of_road_closures	How many times the closest road was closed
fld_zone_2004	Categorical variable for flooding zone type in 2004. It shows flood zone code where point location belongs
condition_x	Code for the assessor's rating of the condition of the structure fetched from the building table
type_of_foundation	Type of foundation
type_of_roof	Type of roof
number_of_floors	Number of floors
number_of_half_bathrooms	Number of half bathrooms
number_of_living_units	Number of living units
PR_STR_max	Max number of private room advertisements at a Zipcode since Sep 2019 until Aug 2020
university_dist_nan	Binary indicator to show no data was available for university_dist and was imputed with the median
elevation_nan	Binary indicator to show no data was available for

	elevation and was imputed with the median
building_area_nan	Binary indicator to show no data was available for building_area and was imputed with the median
eff_year_built_nan	Binary indicator to show no data was available for eff_year_built and was imputed with the median
number_of_living_units_nan	Binary indicator to show no data was available for number_of_living_units and was imputed with the mode
imp_DWELL_nan	Binary indicator to show no data was available for imp_DWELL and was imputed with 0

Variables present in all subsets

- slosh_cat3_2011
- slosh_cat4_2011
- slosh_cat5_2011
- number_of_road_closures
- fld_zone_2004
- type_of_foundation
- number_of_half_bathrooms
- number_of_living_units
- elevation_nan
- building_area_nan
- eff_year_built_nan
- number_of_living_units_nan
- imp_DWELL_nan

County only

- city_limits
- type_Total_per

City only

- old_city_district
- legal_acreage
- city_limits_dist
- grade
- type_Commercial
- type_Residential
- LNB_STR_range
- playground_dist_nan
- year_annexed_nan
- heated_space_nan
- PR_STR_range_nan
- type_Total_per_nan

Municipality only

- Condition_x
- number_of_floors

Common variables in city and county

- slosh_cat2_2011
- slosh_cat3_2011
- slosh_cat4_2011
- slosh_cat5_2011
- number_of_road_closures
- fld_zone_2004
- type_of_foundation
- number_of_half_bathrooms
- number_of_living_units
- school_dist_nan
- elevation_nan
- building_area_nan

- eff_year_built_nan
- number_of_living_units_nan
- imp_DWELL_nan

Unique variables in city and county (do not intersect)

- city_limits
- city_limits_dist
- grade
- heated_space_nan
- legal_acreage
- LNB_STR_range
- old_city_district
- playground_dist_nan
- PR_STR_max
- PR_STR_range_nan
- type_Commercial
- type_of_roof
- type_Residential
- type_Total_per
- type_Total_per_nan
- university_dist_nan
- year_annexed_nan

Common variables in city and municipality

- slosh_cat3_2011
- slosh_cat4_2011
- slosh_cat5_2011
- number_of_road_closures
- fld_zone_2004
- type_of_foundation
- number_of_half_bathrooms
- number_of_living_units

- elevation_nan
- building_area_nan
- eff_year_built_nan
- number_of_living_units_nan
- imp_DWELL_nan

Unique variables in city and municipality (do not intersect)

- city_limits_dist
- condition_x
- grade
- heated_space_nan
- legal_acreage
- LNB_STR_range
- number_of_floors
- old_city_district
- playground_dist_nan
- PR_STR_max
- PR_STR_range_nan
- school_dist_nan
- slosh_cat2_2011
- type_Commercial
- type_of_roof
- type_Residential
- type_Total_per_nan
- university_dist_nan
- year_annexed_nan

Common variables in county and municipality

- slosh_cat3_2011
- slosh_cat4_2011
- slosh_cat5_2011
- number_of_road_closures

- fld_zone_2004
- type_of_foundation
- type_of_roof
- number_of_half_bathrooms
- number_of_living_units
- PR_STR_max
- university_dist_nan
- elevation_nan
- building_area_nan
- eff_year_built_nan
- number_of_living_units_nan
- imp_DWELL_nan

Unique variables in county and municipality (do not intersect)

- city_limits
- condition_x
- number_of_floors
- school_dist_nan
- slosh_cat2_2011
- type_Total_per

Historical valuation

Findings

As an additional piece of the valuation process, CDP conducted a valuation of Charleston properties between 2006 and 2019. In order to conduct the historical valuation, CDP used data including property improvements, condition, size, and year built.

To assess property improvements, CDP relied on three primary sources of data

- Internal Revenue Service (IRS) publication 523
- National rental association of realtors and affiliated agencies

- Research institutions

Using IRS publication 523, the following set of capital improvements were identified as those which are both permanent and generate an increase in a property's market value.

Additions

- Bedroom
- Bathroom
- Deck
- Garage
- Porch
- Patio

Lawn & Grounds

- Landscaping
- Driveway
- Walkway
- Fence
- Retaining wall
- Swimming pool

Systems

- Heating system
- Central air conditioning
- Furnace
- Duct work
- Central humidifier
- Central vacuum

- Air/water filtration systems
- Wiring
- Security system
- Lawn sprinkler system

Exterior

- Storm windows/doors
- New roofing
- Wall-to-wall carpeting
- Fireplace
- New siding
- Satellite dish

Insulation

- Attic
- Walls
- Floors
- Pipes and ductwork

Plumbing

- Septic system
- Water heater
- Soft water system
- Filtration system

Interior

- Built-in appliances
- Kitchen modernization
- Flooring

- Wall-to-wall carpeting
- Fireplace

In order to estimate the cost of various property improvements, we gathered data on cost and recovery from the National Association of the Remodeling Industry (NARC) for the 20 most popular types of remodeling projects. The NARC projections are based on a 2,495 square foot house of average quality (structure and material) with no hidden issues. Actual cost of each remodeling project and cost recovery are influenced by many factors, including project design, material quality, location and homeowner preferences, and thus can vary greatly. To scale the data properly, CDP adjusted the costs according to the size and condition of each property. There are 11 categories available in the rating scale and, based on CDP's R&D, the following value upgrades or degrades are applicable (Table 6a).

Table 6a. *Value upgrades and degrades for each rating*

Code	Meaning	Adjustment (percent)
EX	Excellent	15
VG	Very good	10
GD	Good	5
AG	Average to Good	2.5
AV	Average	0
FA	Fair to average	-2.5
FR	Fair	-5
PR	Poor	-10
VP	Very poor	-12.5
DL	Dilapidated	-15

Cost and recovery of various renovation projects are detailed in Tables 7a-7d.

Table 7a. *Cost and recovery of 20 major renovation projects*

Item	Cost	Cost recovered	Percentage
Adding master suite	150000	75000	50
Kitchen upgrade	38300	20000	52
Complete kitchen renovation	68000	40000	59
Bathroom renovation	35000	20000	57
Adding new bathroom	60000	30000	50
Basement conversion to living area	46900	30000	64
Attic conversion to living area	80000	45000	56
Insulation upgrade	2400	2000	83
Closet renovation	6300	2500	40
New wood flooring	4700	5000	106
Hardwood flooring refinish	2600	2600	100
HVAC replacement	8200	7000	85
New steel front door	2000	1500	75
New fiberglass front door	2700	2000	74
New garage door	2100	2000	95
New fiber cement siding	19700	15000	76
New vinyl siding	15800	10000	63
New roofing	7500	8000	107
New vinyl windows	22500	16000	71
New wood windows	35000	20000	57

Table 7b. *Cost and recovery of adding a garage to a property*

Type	Attached garage	Detached garage	Percentage

boundary	min	max	min	max	
1 car	6150	11650	7500	14200	65
2 cars	15650	22500	19600	28200	65
3 cars	22800	34500	28200	42700	65
1 car	average	8900	average	10850	65
2 cars	average	19075	average	23900	65
3 cars	average	28650	average	35450	65

Table 7c. *Property value added by building a pool*

Price of the property	**Min value added**	**Max value added**
<150000	7000	8600
150000-250000	14000	17100
250000-350000	21000	25700
350000-450000	28000	34300
450000-550000	35100	42900
550000-650000	42100	51400
650000-750000	49100	60000
>750000	56100	68600

Table 7d. *Property value added by other projects*

Item	**Value increase (%)**
Adding a walkway	11.3

Carport	1
Installing boat lift	1
Adding residential Hot Tub	1.5
Adding hay cover	2.2
Adding waterfront Bulkhead	1
Adding or improving feed barn	2.2
Adding general purpose building	5
Adding gazebo	11.3
Adding boat ramp	1
Adding residential utility room	5
Adding a barn	2.2
Adding pool enclosure	1.5
Adding spa	1.5

Further, CDP extended the data on improvements and value increases based on work emerging from various research institutions, including the Virginia Cooperative Extension and the University of Michigan. Values are adjusted for inflation and are calculated historically based on the output of the valuation models for 2020.

After the above mentioned calculations, CDP ran a set of experiments to establish how the economic and financial situation of the city was reflected in the overall real estate market. CDP used a set of macroeconomic indicators from FED, BEA and affiliated agencies. However, due to the complexity of these hedonic models, CDP decided to evaluate existing methodologies and indices and incorporate the ones which showed the highest performance. The index which showed the highest performance was S&P/Case-Shiller U.S. National Home Price Index. Additionally, unlike the HPI index, which only includes houses with mortgages within the conforming amount limits, S&P/Case-Shiller U.S. National Home Price Index measures prices monthly and tracks repeat sales of houses using a modified

version of the weighted-repeat sales methodology. It is thus able to adjust for the quality of the homes sold, unlike simple averages.

Visualizations

Affordability analysis

Zoning Density

Purpose: Evaluate whether the current City zoning laws allow enough density to accommodate population growth.

Process: In order to complete the Zoning Density analysis, we used the following steps:

1. In order to evaluate zoning density, we first counted the number of single family and condo units per the assessor.
2. Because the number of units within multifamily apartment complexes is not well tracked by the assessor or address point data, we researched the multifamily apartment complexes in order to populate the total number of units contained in each. For some multifamily apartment complexes, there was no information available regarding the number of units contained within the complexes. In such cases, we inferred the number of units based on the total square footage of the complexes. Specifically, we found the average square footage of an apartment in the city of Charleston, and calculated how many apartments of that size would fit within a certain building's total square footage of heated space. Where the assessor had no information about square footage of heated space within a building, we assumed the parcel was undeveloped.
3. Next, we used the maximum density values provided by Charleston Planning to determine how much land in Charleston is developable. Because optimal development is not practical, we used the following logic to calculate the number of units allowable by zoning.

a. We ruled out the following land areas as undevelopable
    - City_Parks: these are mostly city parks. County, state and other parks are contained within Public_Owned.
    - National_Wetland_Inventory: this is the closest thing Charleston has to a critical line layer. We used all wetland types, in conjunction with the water layer, in order to indicate all areas that are undevelopable.
    - Private_Conserved: this layer consists mostly of private lands encumbered by conservation easements or other conservation instruments. We considered everything in this layer undevelopable.
    - Public_Owned: this layer contains all publicly-owned properties including many parks. Except for a handful, we considered these properties off-limits for development.

b. We assumed that 80% of the remaining land was developable -- 20% was excluded to account for rights of way and other infrastructure.

4. We evaluated the total number of housing units and the total developable land in order to establish whether the City is allowing enough density to accommodate population growth.

Conclusion: The City is allowing plenty of density to accommodate population growth. Zoning is not a constraining factor at this time.

Current Housing Needed

Purpose: Evaluate whether households in Charleston with incomes associated with varying percentages of the Area Median Income (AMI) can find housing that is affordable within their budgets.

Process: In order to complete the analysis of Current Housing Needed, we used the following steps:

1. For this analysis, we based calculations on 4-person households with incomes associated with varying percentages of the current Area Median Income (AMI) provided by the Charleston Housing Department ($65-81K).

2. We assumed that, if given the choice, residents would choose housing valued at 30% of their gross household income. Households with high incomes, capped at 120% AMI in the visualization, would choose housing commensurate with their ability to pay.
3. We applied all of the deed-restricted affordable housing to the least wealthy households in Charleston, starting at the lowest income and moving upward.

Conclusion: The deed restricted units are likely not sufficient to house everyone requiring assistance. Current deed-restricted housing is covering a little under half of the <30% AMI cohort. This is only an estimate given the simplicity of presenting all households as 4-person. However, for planning purposes, the results clearly show the need for more affordable housing units.

Future Housing Needed

Purpose: Evaluate whether households in Charleston with future incomes associated with varying percentages of the Area Median Income (AMI) will be able to find housing that is affordable within their budgets.

Process: In order to complete the analysis of Future Housing Needed, we used the following steps:

1. We applied even population growth to all cohorts within the City of Charleston. We assumed that all income levels would grow uniformly and that all households would be in need of housing.
2. We assumed that, if given the choice, future residents would choose housing valued at 30% of their gross household income. Households with high incomes would choose housing commensurate with their ability to pay.
3. We applied all of the deed-restricted affordable housing to the least wealthy households in Charleston and evaluated whether the deed-restricted affordable housing in Charleston will be sufficient to cover all future households.

Conclusion: The data show that the Peninsula and West Ashley have the greatest need for new affordable housing. Additionally, there are many market value

options that remain available for a 4-person family earning 80-120% AMI. There is opportunity for growth in the housing market >120% AMI.

Limitations

The Affordability Analysis for Charleston is a quick diagnostic of overall housing statistics, and not a deep-dive into true demand. One limitation of this analysis is that it does not match household size to number of bedrooms. It also does not take into account the condition of the housing units. While according to our cost analysis, there is housing available to >100% AMI, this study does not consider whether this housing is safe or appropriate.

It was not possible to calculate all the AMI groups to specific housing needs. While CDP curates data on household size, income and address, the Charleston and Berkeley County Assessors do not track all of the unit designators thoroughly. CDP could place households in buildings but not determine the size of the unit in multi-family apartment situations. The assessors also do not keep track of the number of bedrooms or the number of apartment units in each building. CDP gathered some data from commercial sources on the number of units. For what was unknown, the number was estimated. Nonetheless there was no way to determine which household was in any given apartment unit of any given size. With these limitations, CDP made average assumptions for 4-person households based on the guidance of Charleston Housing Department.

- Given that many residents of Charleston currently live in housing units that are priced higher than 30% of their gross household income, it is clear that the actual decisions of households are more nuanced than this analysis would suggest. CDP believes that a housing survey for Charleston would reveal insights on true demand, how housing decisions differ by household size, and which types of housing stock should be supported by policy interventions.

Housing and Transportation Analysis

[Housing and Transportation Analysis - Visualizations | Tableau Public]

Purpose: From our discussions with the City of Charleston, we understood that in order to capture the cost burdens faced by residents, it would be necessary to analyze both the housing and the transportation costs incurred by residents of Charleston.

Process: In order to complete the Housing and Transportation analysis, we used the following steps:

1. We matched households with their actual housing, comparing income to cost.
2. Because rent is a relatively good measure for the true costs of owning a home (considering mortgage and repairs), and because determining which residents have a mortgage was outside of the scope of this analysis, we assumed every household was renting for the purpose of this analysis.
3. We used transportation cost data provided by Charleston City Planning.

Conclusion: Many people are already living in housing they cannot afford even before transportation is taken into account.