Analyze and Predict House Price of Washington DC

Sweta Kumari

Department of Business Analytics, Golden Gate University

Dr. Heinz Joerg Schwarz

April 20, 2020

Contents

Abstract

House is one of the important elements in basic human needs. In the housing market, house price is analyzed and predicted for seller and buyer considering several characteristics of the house. This paper uses a dataset for house price of Washington DC, highlights the housing characteristics, suggest buyers/ investors understand how house price can be assessed with the help of attributes of the dataset. This paper analyzed that house price is positively correlated with the gross building area, number of bathrooms, bedrooms, and fireplaces. It highlights that most of the buyers do not look at houses with many bedrooms and fireplaces but preferred 2 story houses. This paper uses the regression technique and identifies that the availability of AC, good house condition, Quadrants, and Wards do help in house price estimation. The study also tried to seek insight into different areas and suggest that ward2 has expensive houses while ward 8 and ward7 have the cheapest one. Similarly, the Northeast quadrant where most of the house sold is the costliest. The analysis also emphasizes on a trend of the house price and house sold over the years and observed that house price and house count increased over the years. With the help of tools, house price is forecasted till 2023 to help buyer and sellers to decide on their financial planning. This paper also recommends avoiding Ward2 & Ward3 and look in Ward1 & Ward4 for bigger houses. It also recommends that buyers should start with Randle Heights, Georgetown, Deanwood if they look for newer houses in Washington DC. This analysis can help several sellers and buyers in buying houses in Washington DC and plan better for their lifetime financial investment.

Keywords: house prices, Washington DC house price, predict house price, descriptive statistics of house price.

Introduction

Rising house prices are becoming a major issue for young professionals and working-class people. Let us take an instance of house prices in Washington DC, United States. According to the Greater Capital Area Association of REALTORS, Washington DC has house sales price as $600000 which is the highest median price in the last 10 years reported in December 2018 (GCAAR, 2018). The city has attracted the wealthier residents and appeals to permanent residents due to its unique location with big-city amenities.

In the last decade, house sales have significantly increased by 34% and recorded a high sale of 5.51 million in 2018 in the United States (IEEE, 2019). Analysis of house prices has since invited widespread recognition because the results of these forecasts can help several buyers and real estate stakeholders to take informed actions. Extracting and analyzing the relevant traits that affect house prices would be profitable for buyers, sellers, and property investors. For instance, customers can utilize the house price prediction model to perceive information on the houses that meet their monetary capacities. In the same way, house owners would also want to have an opinion and would watch for the most suitable moment to sell. Real estate agents can also rely on this model to assist clients to find the most suitable homes as per market trends. Therefore, analysis and prediction should become an essential criterion for evaluating and understanding the house prices.

We have collected data about Washington DC house prices. To understand the various factors of increasing price, we will perform statistical analyses such as descriptive analytics and predictive analytics. We will offer few recommendations for house prices based on house age, gross building area, and location such as wards, neighborhood area. Washington DC area is divided into 4 Quadrants and 8 Wards. There are neighborhood area and sub-neighborhood area

as well. We will explore all attributes with the help of descriptive analytics and correlation analysis. It will be helpful in estimating the house price. In addition, correlation analysis can help in identifying the significant variables. We will perform a time series analysis to understand the variation of house prices and houses sold over the past two decades and forecast the house price. We will be using multiple regression techniques to create a model that can predict the price to help buyers and property investors. ANOVA will be used for model comparison and find the best-suited model.

This paper is divided into different sections. Section “Data Description” gives a brief introduction to the data and variables. Section “Literature Review” talks about the research made to substantiate the analysis. Section “Focus of Analysis” presents goals of analysis and a list of questions that can be addressed by the analysis. Section “Data Preparation” discusses the steps taken in data preparation, data cleansing, and data modification. Section “Descriptive Analytics” provides descriptive statistics, data visualization of housing attributes versus house prices, correlation analysis, and time series analysis. Section “Predictive Analytics” presents the regression technique to predict the house price, ANOVA method and time series analysis to forecast the house price until 2023. Section “Prescriptive Analytics” provides recommendations for homeowners and buyers. Section “Conclusion” discusses the results of the analysis and methodology used in this paper.

Business Problem

This paper analyzes and predicts the house price of Washington DC. This study will help real estate agents, sellers, buyers, and property investors to identify which attributes contribute more towards house prices and how the price will be varied until 2023. Various analytics and

technique used in the papers will help to make a more effective decision. In the future, this analysis can be used by real-estate companies like Zillow or Trulia.

Data Description

To perform the various analysis, we collected data from Kaggle website: https://www.kaggle.com/christophercorrea/dc-residential-properties#DC_Properties.csv. Due to a wide variety of datasets, we chose Kaggle, among others. This dataset has been originally collected from https://opendata.dc.gov/. Washington DC data was captured for the year 1982 to 2018. The attribute and description of the dataset is given below:

Number of Columns: 49; Number of observations: 1,58,958

Table1

Data Description of All Variables in the Dataset

ATTRIBUTES	DESCRIPTION	ATTRIBUTES	DESCRIPTION
ID	#serial number	KITCHENS	Number of Kitchens
BATHRM	Number of Full Bathrooms	FIREPLACES	Number of Fireplaces
HF_BATHRM	Number of Half Bathrooms (no bathtub or shower)	USECODE	Property use code
HEAT	Heating Type (For example: Forced Air, Hot water, etc.)	LANDAREA	Land area of property in square feet
AC	Cooling (Y or N)	GIS_LAST_MOD_DTTM	Last Modified Date
NUM_UNITS	Number of Units	CMPLX_NUM	Complex number
ROOMS	Number of Rooms	LIVING_GBA	Gross building area in square feet
BEDRM	Number of Bedrooms	SOURCE	Residential House
AYB	The earliest time the main portion of the building was built	FULLADDRESS	Full Street Address
YR_RMDL	Year structure was remodeled	STATE	State
EYB	The year an improvement was built more recent than actual year built	CITY	City
STORIES	Number of stories in primary dwelling	ZIPCODE	Zip Code
SALEDATE	Date of most recent sale	NATIONAL GRID	Address location national grid coordinate spatial address
PRICE	Price of most recent sale	LATITUDE	Latitude
QUALIFIED	Qualified	LONGITUDE	Longitude
SALE_NUM	Sale Number	ASSESSMENT_NBHD	Neighborhood
GBA	Gross building area in square feet	ASSESSMENT_SUBNBHD	Sub Neighborhood
BLDG_NUM	Building Number on Property (1 or 2)	CENSUS_BLOCK	Census Block
STYLE	Style (For example: 2 story or 3 story etc.)	CENSUS_TRACT	Census Tract
STRUCT	Structure (For example: Vacant Land, Town inside etc.)	WARD	Ward (District is divided into eight wards, each with approx. 75,000 residents)
GRADE	Grade (Good, Excellent, etc.)	SQUARE	Square (from SSL)
CNDTN	Condition (Good, Excellent, etc.)	Y	Latitude
EXTWALL	Exterior wall (For example: Aluminum, Concrete etc.)	X	Longitude
ROOF	Roof type (For example: Waterproof, Built-up etc.)	QUADRANT	City quadrant (NE, SE, SW, NW)
INTWALL	Interior Wall (For example: Ceramic Tile, Hardwood etc.)		

Literature Review

According to the U.S. Census Bureau (DC.gov, 2020), Washington D.C.'s population was 705,749 as of July 2019 and grew by 104,000 people since the 2010 Census. Washington DC is the capital of the United States of America and one of the six cities in the USA (other five cities are Chicago, San Francisco, Boston, Los Angeles, New York City) which have primary real estate markets. These cities showed a large number of transactions in the housing market (Young, 2019).

Also, according to BUSINESS INSIDER, the average home price of all properties in Washington DC was $545000 in 2015 as compared to $400,000 in 2010. This stat reveals a staggering 36% increase in the average house price (Josephson, 2015). This event triggered many interested buyers to re-think several times before looking for a house. Housing has become a big part of the cost of living due to its utility bill, progressive taxes, amazing healthcare industries, increasing transportation service. The analyst has started analyzing the house data for a good house price prediction that would help better prepare everyone before they think of one of the critical investments of their lives. The following literature review will delve into the house price estimation and the factors affecting the selling price.

According to the Journal of Real Estate Literature (Sirmans, Macpherson, & Zietz, 2005), many variables/characteristics affect the estimation of housing prices. It provides the study of 125 papers that estimate house prices and analyze the factors associated with it. For instance, the number of bathrooms has a positive correlation with the sale price. The bathroom is statistically significant 35 times out of 40 times as per analysis done on house selling. This analysis illustrates that the house price can increase by 10-12% even if 1 bathroom is increased. The other variables would be the number of bedrooms, number of fireplaces, lot size, etc. Similarly, our

dataset has the same attribute (number of bathrooms) to analyze the house price. It will help to determine the effects on selling prices in Washington DC.

Consistent with the approach in this proposal, according to Review of Regional Studies (Cebula, 2010), the number of bathrooms, bedrooms, fireplaces, number of stories, and the gross building area of the house help in estimating the house price and shown a positive correlation. The study was conducted on data of the City of Savannah, Georgia for 2000 and 2005 with the data of 2,888 single-family houses. Our dataset has predictive variables as the number of bathrooms, bedrooms, fireplaces, number of stories, and the gross building area of the house which will help to estimate the house price.

Until now we know that the internal factors of housing such as bathrooms, fireplaces, bedrooms, total square feet are the common factors to estimate the price of the house. There are other external factors also which buyers may think while purchasing the house. For example, House Age. According to the Journal of Facilities Management (Abidoye & Chan, 2016), the real estate property has a heterogeneous nature that made each stakeholder value the property differently. Among many internal and external factors of real estate property, this study found that there are many attributes that affect the property price in which the property area and age of house influences the overall estimation of the house price. In our dataset, we have variables such as gross building area, year of modeling done, and year at which house sold that help to analyze the house price.

Also, according to the Journal of Real Estate Economics, newer homes should have higher property prices (Coulson & Lahr, 2005). If house age is less or zero, then there is little, or no repair is needed and is generally equipped with modern appliances. House is cleaner and brighter that increases the price on the higher side. Similar research has been performed in the

Real Estate Finance and Economics Journal (Asabere & Huffman, 1993) where it says if house age is more, it establishes reverse correlation with house price. Year of modeling done and year at which house sold variables help to predict the age of the house and examine the effect on selling price.

Another example of an external factor impacting hosing prices would be the local neighborhood. According to the Journal of Housing Economics (Kiel & Zabel, 2008), the neighborhood is the most important factor affecting the house price. The consumer selects the neighborhood environment first by looking at the public facilities and services, then they look for a house site. The research also confirms that consumers are concerned about the characteristics of the broader region (such as the school district or crime rate) and the local community (such as street quality).

In addition, Research of Journal of Real Estate Literature (Sirmans, Macpherson & Zeitz, 2005) and Regional Science and Urban Economics (Brasington & Hite, 2005) states that house price is affected by neighborhood areas. If a property is located nearby the park or the place where amenities and facilities are better including healthcare, the price is on the higher side. In Washington DC, areas are divided into 8 different wards with different neighborhood areas. In our dataset, we have information about wards, neighborhood, and sub-neighborhood areas. We can extract the information about the address of each house from our data. This data will help to analyze to see which neighborhood/sub-neighborhood area has the highest pricing and highest number of houses.

Adding to the previous examples of external factors, supply and demand have an impact on housing prices. According to Australian Geographer-a peer-reviewed journal (Kim & Park, 2005), house price is affected by supply and demand. Price is increased due to the demand

increases and the supply cannot be met in a short time. It further argues that price is affected by the socio-economic condition and macro-economic variables. GDP, unemployment rate, income growth affects the consumer's ability to support housing price. The overall economy strength is significantly changing the real estate market. If the economy grows and wages increase, more people can afford the house, this again increases the demand of the house and in turn, increase the house price due to low supply or supply not met in time.

One of the important factors attributing to house prices is mortgages. According to the Federal Reserve Bank of New York (Khan, 2008), the mortgage rate affects the housing market on a large scale. Lower Federal Reserve interest rate generally leads to a lower mortgage rate. House is more affordable if the monthly payment amount is low which is decided by mortgage rate and mortgage amount. Lower mortgage rate also affects the supply and demand of the housing that results in fluctuating the house price.

It is becoming increasingly important that buyers know the estimation of house prices and how other factors affect it. Real estate sales agents can help buyers to give price estimation based on their budget. This will help them to make the most important decision in terms of finance. The analysis will be a benefit for homeowners, and they will not put an inappropriate price to delay the selling. In addition, house price analysis is also helpful for property investors to know the trend of prices in a certain location. Our focus is to find the correlation between house prices and all the factors, predict the house price and forecast it to give a brief idea about it.

Focus of Analysis

With the statistical analysis on Washington DC house price, we would like to address the below questions:

1. Which factors affect the house price in Washington DC? For example, the number of bathrooms, living area, etc.?
2. How/ Which are the numerical data correlated with house prices in DC? For example: is the number of bedrooms highly correlated with house prices? How the number of fireplaces helps to predict the house price?
3. How house price is varying concerning categorical data such as types of apartment, condition, heat type, house age?
4. Which neighborhood and area have the maximum number of houses and maximum house price?
5. How will house prices be varied in the next 5 years? Dataset has house data till 2018, we will forecast the price till 2023 (for 5 years).
6. What are the recommendations can be given to homeowners and buyers?

Above analysis covers all three types of analysis: Descriptive, Predictive and Prescriptive Analytics.

Data Preparation

As part of Data Preparation, we performed Data Cleansing and Data Modification steps to make sure data has good quality.

Data Cleansing

As part of the data cleansing process, we have performed the below steps. We checked the number of null values in the dataset to avoid bias results.

a. At first step, we dropped null values of columns named 'PRICE', 'LONGITUDE', 'ASSESSMENT_NBHD', 'ZIPCODE', 'WARD', 'LATITUDE'.

b. We dropped null values of columns 'STRUCT', 'STYLE', 'FULLADDRESS', 'QUADRANT', ‘AYB’ in the second step.

c. We dropped columns Unnamed: 0, CENSUS_BLOCK, CENSUS_TRACT, CITY, STATE, NATIONALGRID, GIS_LAST_MOD_DTTM, QUALIFIED, SQUARE, CMPLX_NUM, LIVING_GBA, X, Y, SALE_NUM, STORIES, NUM_UNITS, BLDG_NUM . These are the columns either related to Census data or columns that not required for our analysis. We also deleted duplicate columns. For instance: GBA and LIVING_GBA are the same; X and Y represent latitude and longitude respectively which are already present in the dataset.

d. We dropped those rows that have

    a. a value of ‘0’ in column AC.

    b. a value of ‘No Data’ in column HEAT.

    c. a value of ‘Default’ in column CNDTN.

Apart from these, we dropped a value of 44 for KITCHENS, 12 for BATHRM, 11 for HF_BATHRM, 30 & 31 for ROOMS, 13, 14,15 &20 BEDRM, 10,11, 12 & 13 for FIREPLACES. These values had very fewer entries in the column (number of rows < 3).

Data Modification

We also modified a few elements in the dataset for enhancing user experience and extract more meaningful data quickly.

a. We renamed values in Quadrant column:

   NE: Northeast, NW: Northwest, SE: Southeast, SW: Southwest

b. We extracted the sale year from SALEDATE and calculated house age for each house. We added column SALEYEAR and HOUSEAGE in the dataset.

SALEYEAR=Year of sale extracted from SALEDATE column

HOUSEAGE=SALEDATE- EYB

After the above modifications, the number of columns is 34, the number of rows is 57341.

Descriptive Analysis

Descriptive Analysis includes descriptive statistics, distribution of the variables and how it varied with Price. Apart from this, correlation analysis and time series analysis are done.

Tools Used: Python for descriptive statistics and correlation. Power BI for time series analysis.

Descriptive Statistics

Table 2

Numerical Variables: Descriptive Statistics

Statistics	BATHRM	HF_BATHRM	ROOMS	BEDRM	EYB	PRICE	GBA	KITCHENS
mean	2.20	0.65	7.44	3.42	1970	578281	1723.25	1.2
std (standard deviation)	1.06	0.61	2.29	1.11	17	582999	817.44	0.6
min	0	0	0	0	1915	1	252	0
1st Quartile:25%	1	0	6	3	1957	240000	1215	1
2nd Quartile/ Median:50%	2	1	7	3	1967	440000	1504	1
3rd Quartile:75%	3	1	8	4	1975	750000	1980	1
max	11	7	28	12	2018	25100000	15902	6

Table 3

Numerical Variables: Descriptive Statistics continued

Statistics	FIREPLACES	USECODE	LANDAREA	ZIPCODE	LATITUDE	LONGITUDE	SALEYEAR	HOUSEAGE
mean	0.64	13.28	3163.50	20011.64	39	-77	2008.90	39.1
std (standard deviation)	0.90	4.18	3111.52	7.74	0	0	7.00	17.0
min	0	11	216	20001	38.819731	-77	1982	-21
1st Quartile:25%	0	11	1520	20003	38.892965	-77	2004	33
2nd Quartile/ Median:50%	0	12	2210	20011	38.916101	-77	2010	43
3rd Quartile:75%	1	13	3997	20018	38.942314	-77	2015	50
max	9	39	187301	20052	38.995435	-77	2018	92

We will analyze the target variable “PRICE” followed by “STYLE” (House style such as “2 story”), GBA (Gross Building Area) and LANDAREA (Land Area for house construction).

PRICE

Figure 1. Histogram of House Price

Interpretation: The distribution of price is right-skewed (mean> median). Most of the houses have price within 2.5million.

STYLE

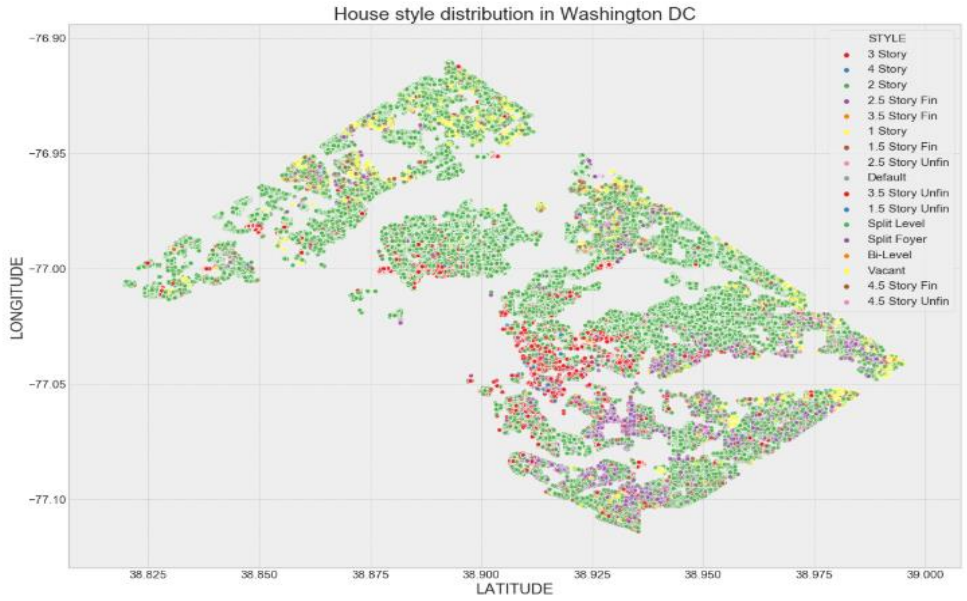

Figure 2. House style distribution in Washington DC

Interpretation: Most of the houses in Washington DC have 2 story. Top 3 styles are shown below.

Table 4

Top3 House Style with number of houses

Style	House Count
2 story	43804
3 story	5541
2.5 Story Fin	3881

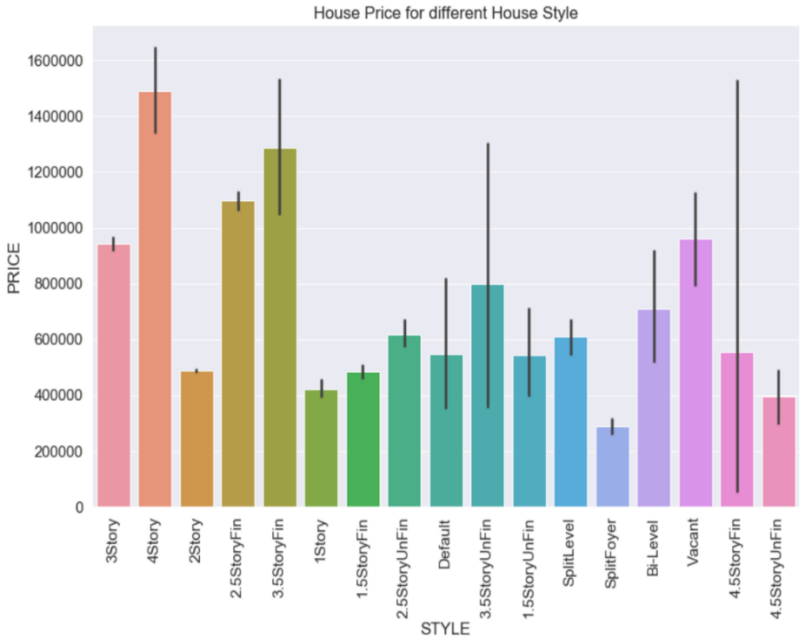

Figure 3. House Price for different House style

Interpretation: The cost of the house is high with 4 story and 3.5 story. Cheapest houses are with 2 story.

GBA

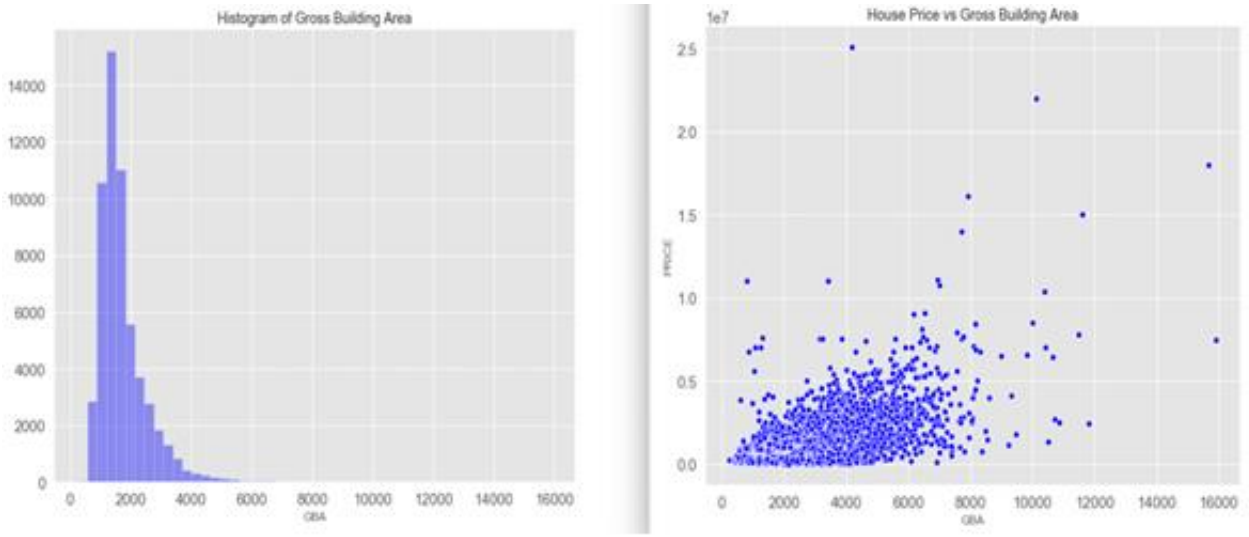

Figure 4. Histogram of Gross Building Area & House Price vs Gross Building Area

Interpretation: The distribution of the gross building area is right-skewed (mean> median). Most of the houses have gross building area within 4000 square feet. Cost of house increases when GBA increases. Cheapest houses are with small GBA.

LANDAREA

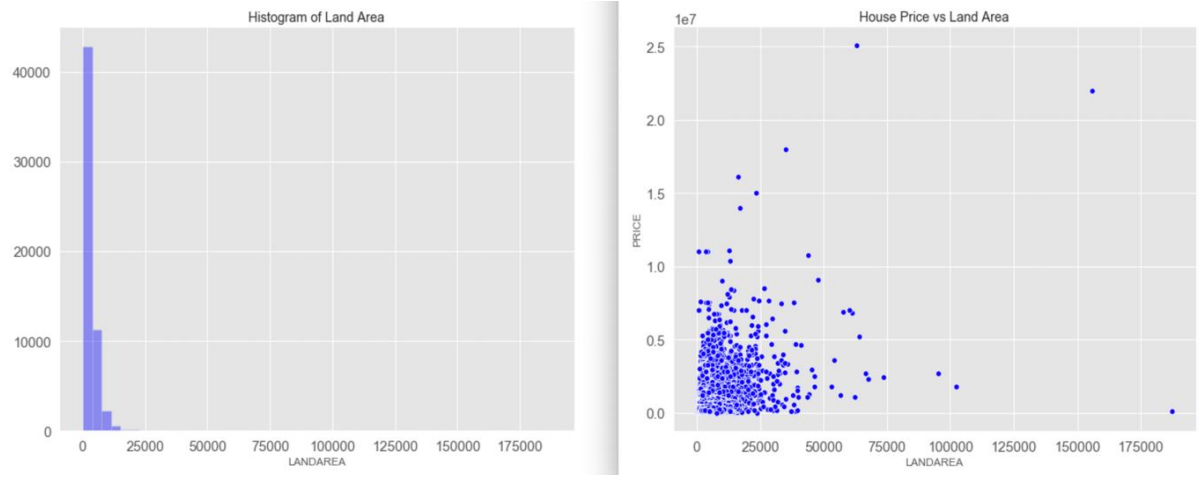

Figure 5. Histogram of Land Area & House Price vs Land Area

Interpretation: The distribution of land area is right-skewed (mean> median). Most of the houses have a land area within 25000 square feet. Cost of house increases when land area increases. Cheapest houses are with small land area.

HOUSEAGE

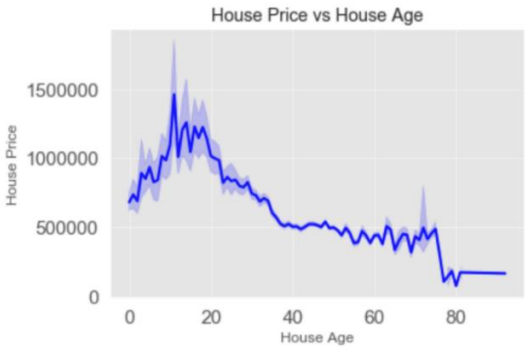

Figure 6. House Price vs House Age

Interpretation: House Price is reaching its highest peak if the house age varies from 0 to 20 years and its decreasing as house age is increasing. This also supports the research of the Real Estate Economics Journal (Coulson & Lahr, 2005) and the Real Estate Finance and Economics Journal (Asabere & Huffman, 1993), which states that newer homes have higher property prices.

Let's check how house price is varied with respect to other variables. We categorized our variables in numerical and categorical variables. We understand the house price variation with numerical variables first.

Numerical variables: BATHRM, HF_BATHRM, ROOMS, BEDRM, KITCHENS, FIREPLACES.

BATHRM

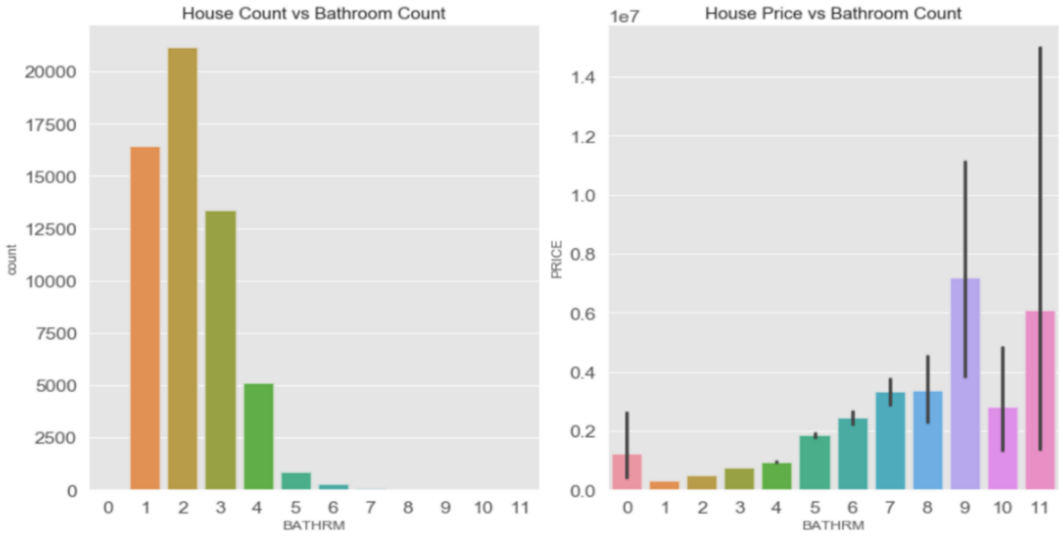

Figure 7. House Count vs Bathrooms Count & House Price vs Bathroom Count

HF_BATHRM

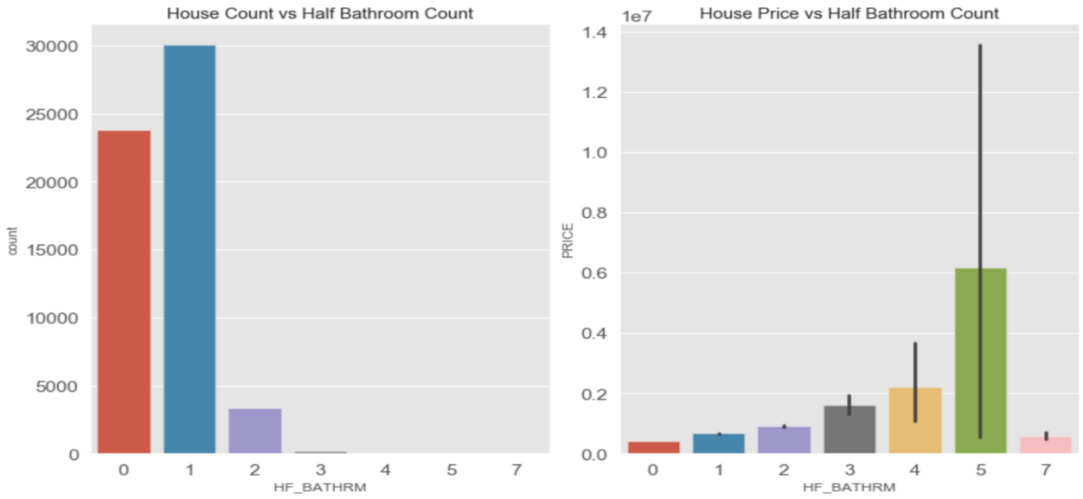

Figure 8. House Count vs Half Bathroom & House Price vs Half Bathroom Count

ROOMS

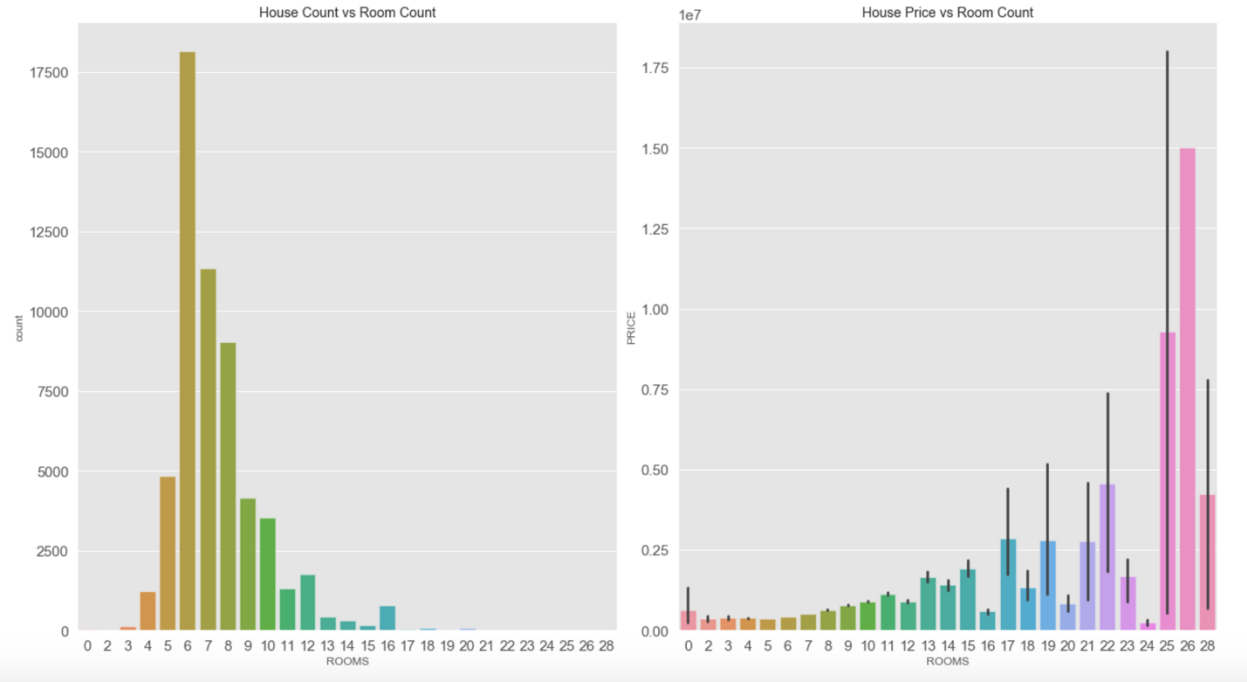

Figure 9. House Count vs Number of Rooms & House Price vs Number of Rooms

BEDRM

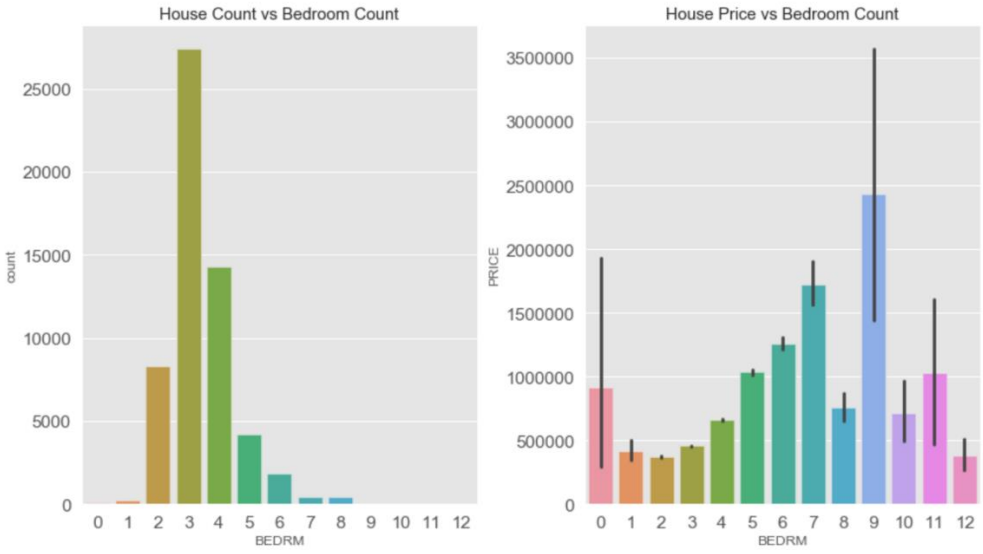

Figure 10. House Count vs Bedroom Count & House Price vs Bedroom Count

KITCHENS

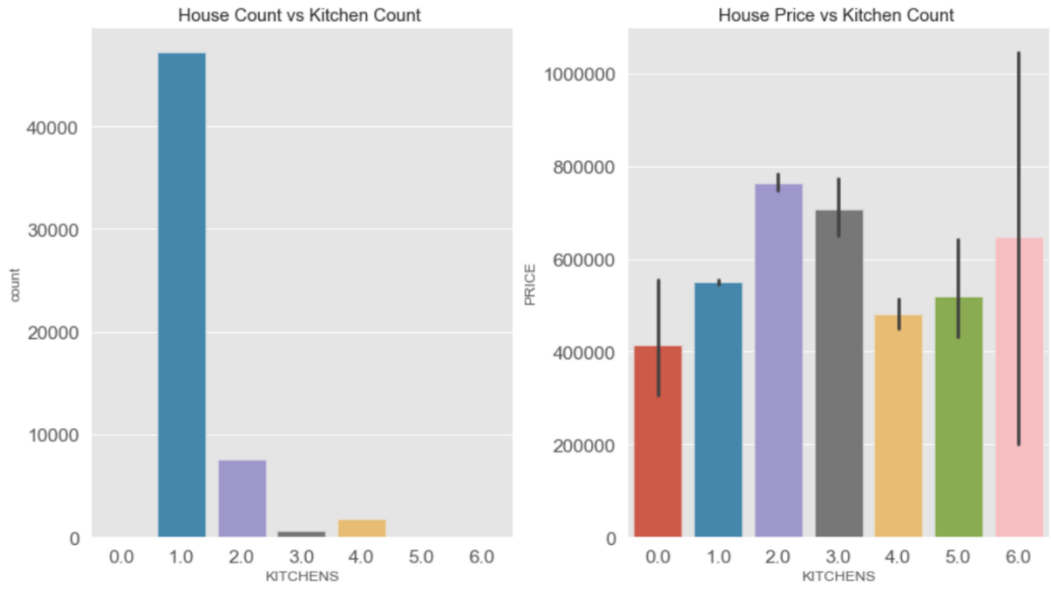

Figure 11. House Count vs Kitchen Count & House Price vs Kitchen Count

FIREPLACES

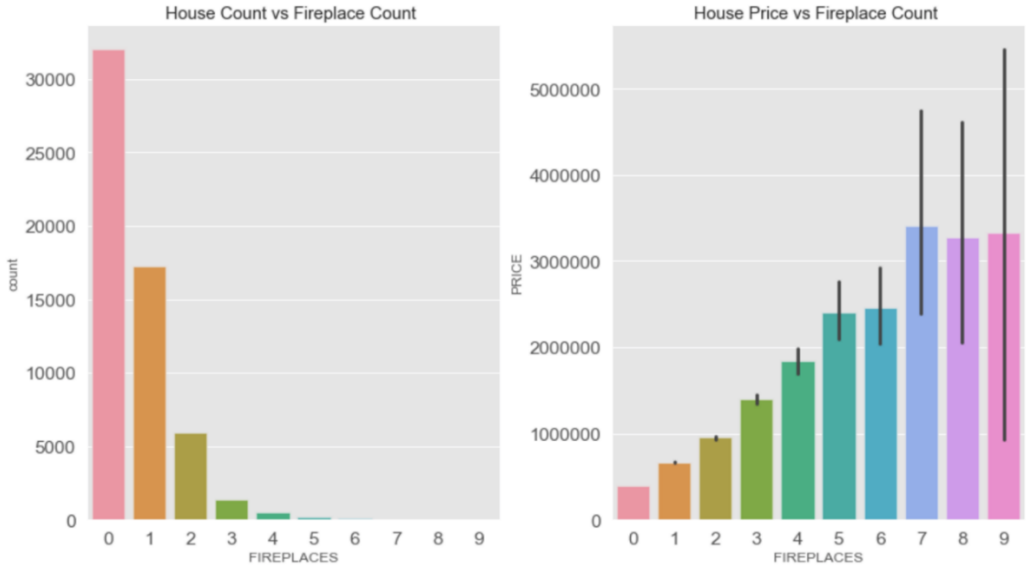

Figure 12. House Count vs Fireplace Count and House Price vs Fireplace Count

Table 5

Interpretation of Housing Internal Characteristics (Numerical Variables)

Housing Internal Characteristic	Number of Houses are higher with	House Price is higher with	Cheap houses are with	Less houses have
BATHRM	2,1 and 3	9,11 and 8	1, 2 and 3	More than 4
HF_BATHRM	1 and 0	5, 4	0, 1	More than 2
ROOMS	6, 7 and 8	26, 25 and 22	2, 3 and 4	More than 12
BEDRM	3, 4 and 2	9, 7 and 6	2, 1 and 3	More than 5
KITCHENS	2 and 1	2 and 3	0 and 4	More than 2
FIREPLACES	0 and 1	7 and 8	0 or 1	More than 2

Categorical variables: AC, HEAT, STRUCT, GRADE, EXTWALL, ROOF, INTWALL, CNDTN.

AC

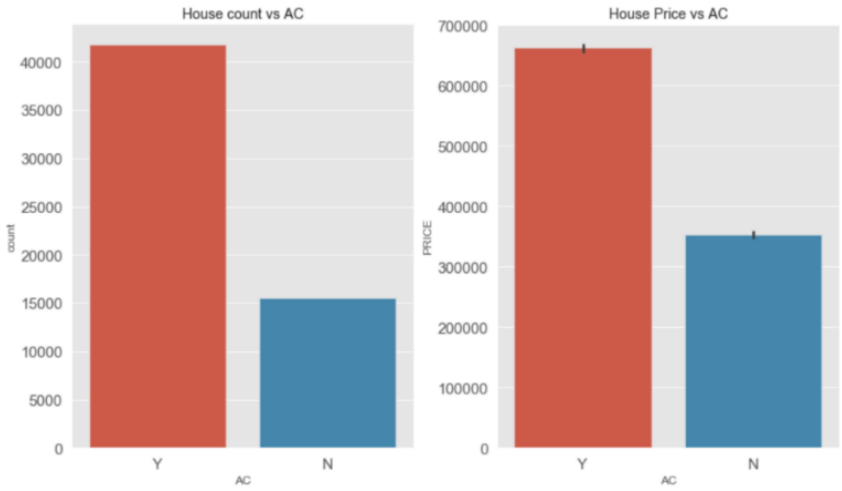

Figure 13. House Count vs AC & House Price vs AC

HEAT

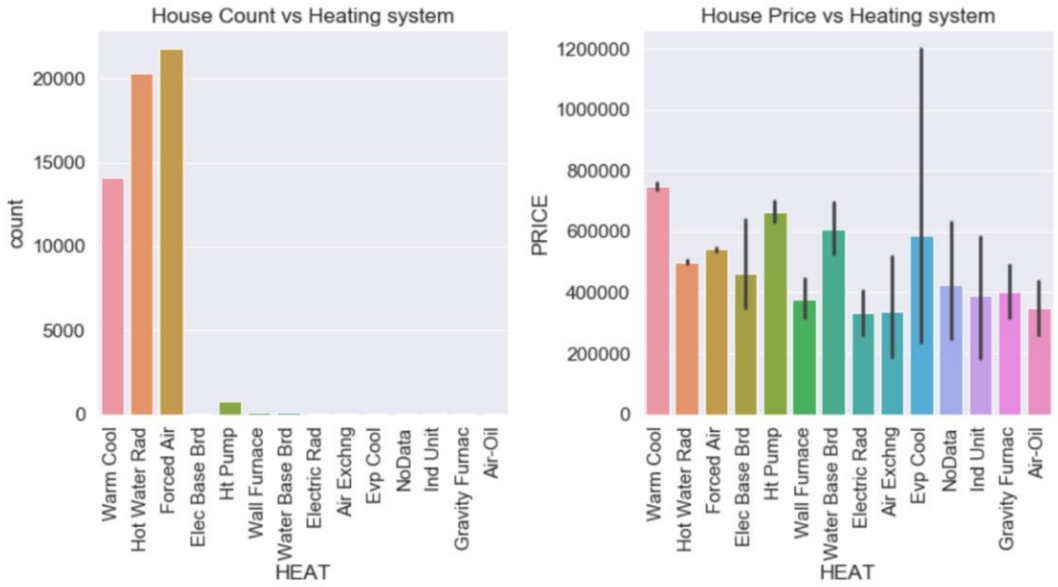

Figure 14. House Count vs Heating System & House Price vs Heating System

STRUCT

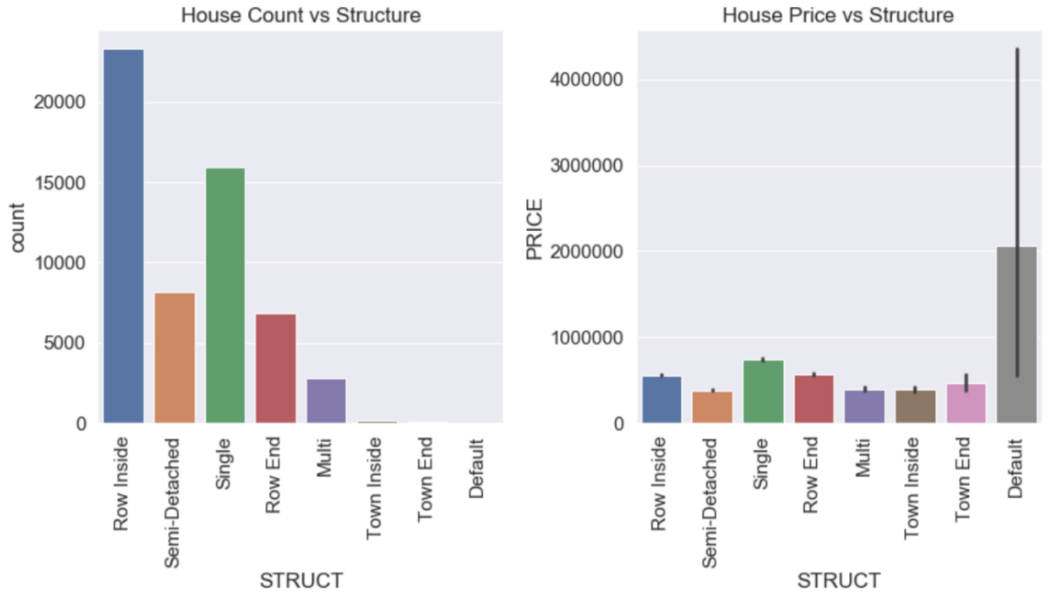

Figure 15. House Count vs Structure & House Price vs Structure

GRADE

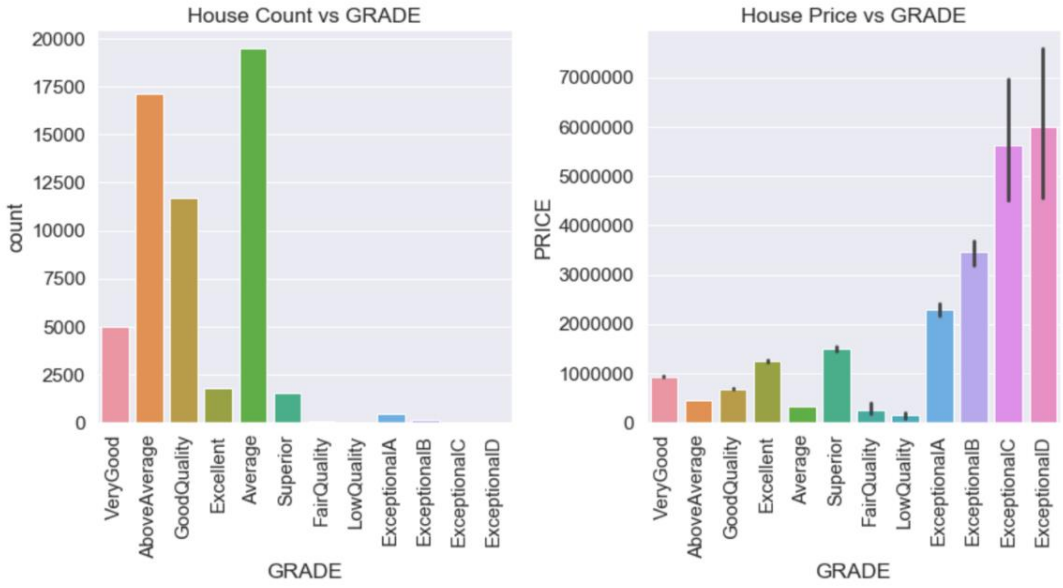

Figure 16. House Count vs Grade & House Price vs Grade

EXTWALL

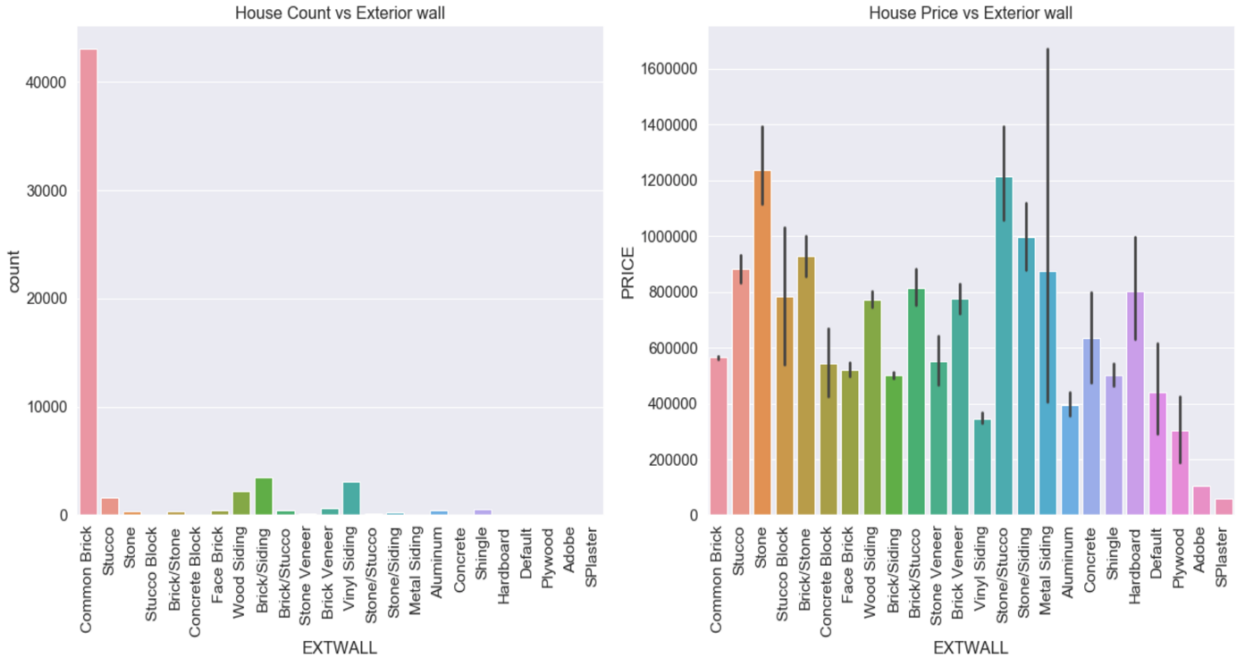

Figure 17. House Count vs Exterior Wall & House Price vs Exterior Wall

ROOF

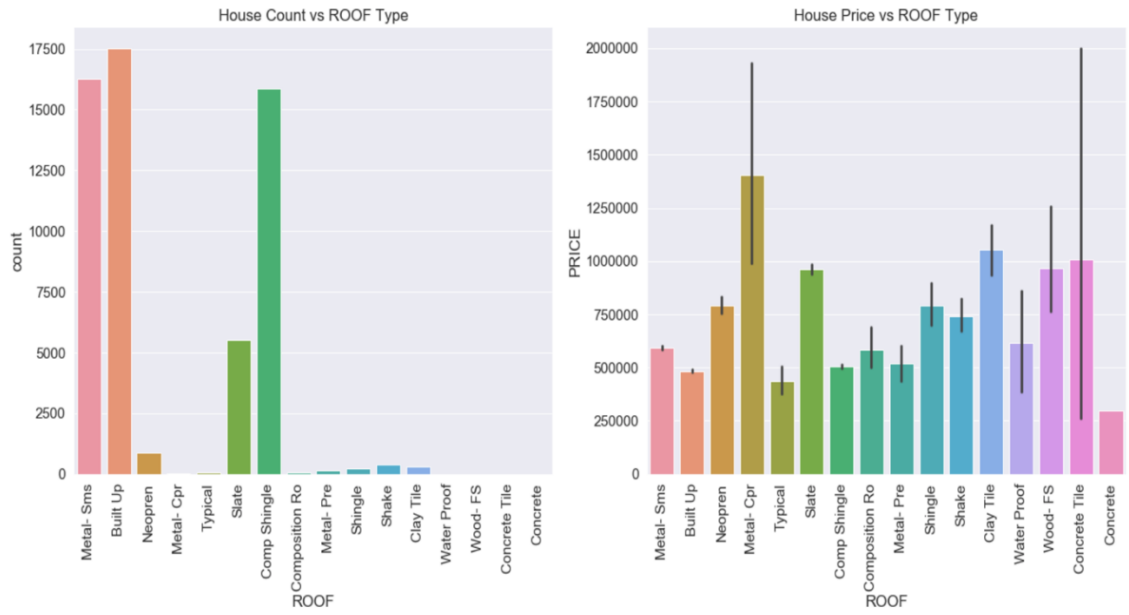

Figure 18. House Count vs Roof Type & House Price vs Roof Type

INTWALL

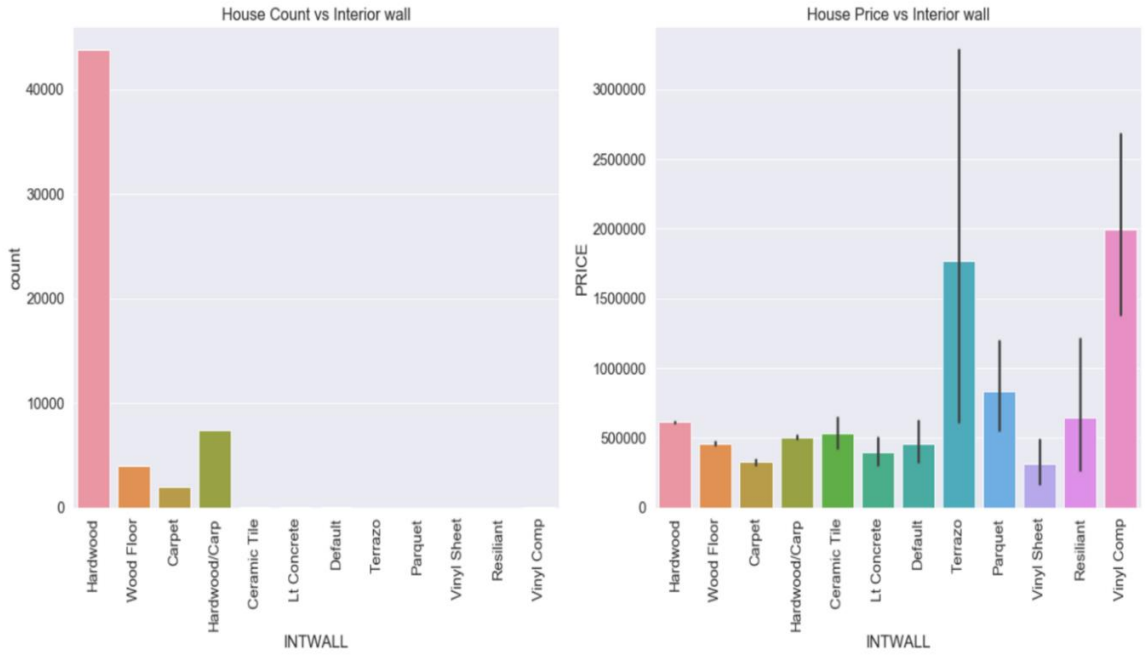

Figure 19. House Count vs Interior Wall & House Price vs Interior Wall

CNDTN

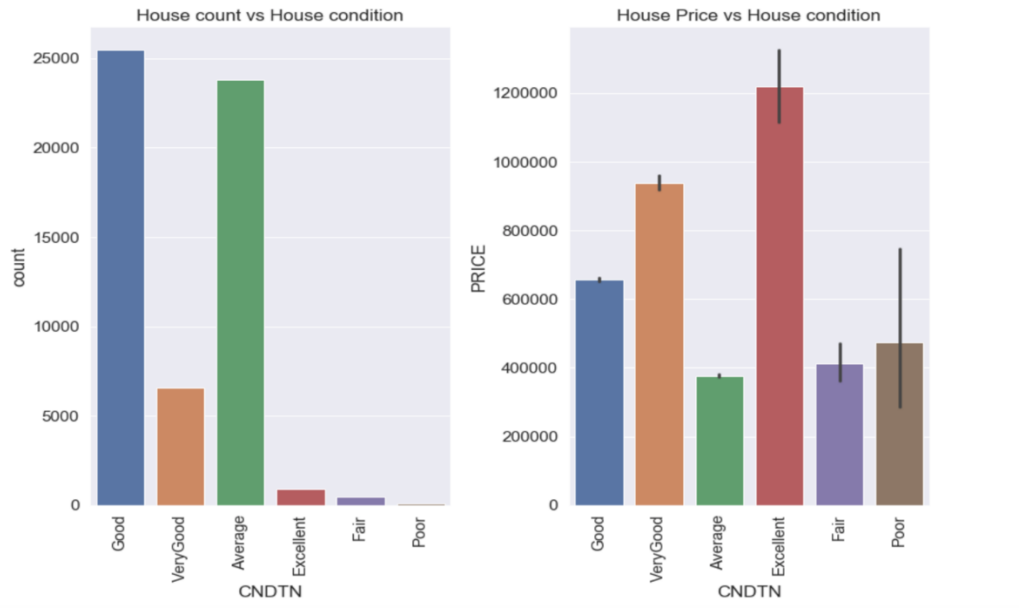

Figure 20. House Count vs House Condition & House Price vs House Condition

Table 6

Interpretation of Housing Internal Characteristics (Categorical Variables)

Housing Internal Characteristic	Number of Houses are higher with	House Price is higher with	Cheap houses are with
AC	with AC	with AC	without AC
HEAT	“Forced Air”, “Hot Water Rad” and “Warm Cool”	“Warm Cool”, “Hit Pump” and “Water Base Brd”.	Electric Rad and "Air Exchanging"
STRUCT	“Row Inside”, “Single” and "Semi-Detached"	“Default” and "Single"	“Semi-Detached” and "Multi"
GRADE	“ExceptionalD” and “ExceptionalC”	“Average” and “AboveAverage”	“ExceptioanlA”
EXTWALL	“Common Brick”	“Stone”	“Splaster”
ROOF	“Built Up” or “Metal sms”	“Metal-cpr”	“Concrete”
INTWALL	“HARDWOOD”	“Vinyl Comp”	“Vinyl sheet”
CNDTN	“Good” and “Average”	“Excellent”	“Average”

Other Categorical variables: Let’s talk about the neighborhood area, ward and quadrants in Washington DC and how the price has varied in these areas. The area is divided into 4 Quadrant.

Each Quadrant is further divided into 8 Ward and each Ward is divided into Assessment Neighborhood. Assessment Neighborhood is further divided into Assessment Sub-Neighborhood. Variables are QUADRANT, WARD, ASSESSMENT_NBHD, ASSESSMENT_SUBNBHD. MicroStrategy is used for ASSESSMENT_NBHD, ASSESSMENT_SUBNBHD analysis.

QUADRANT

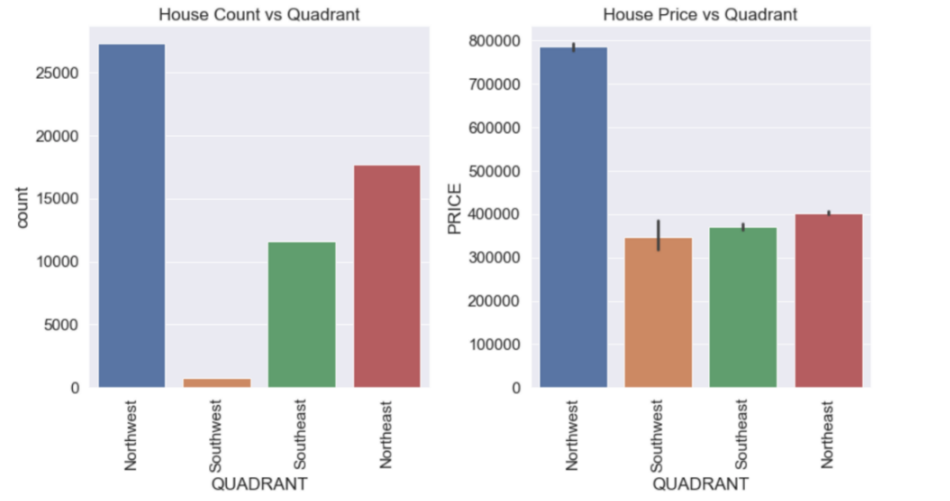

Figure 21. House Count vs Quadrant & House Price vs Quadrant

WARD

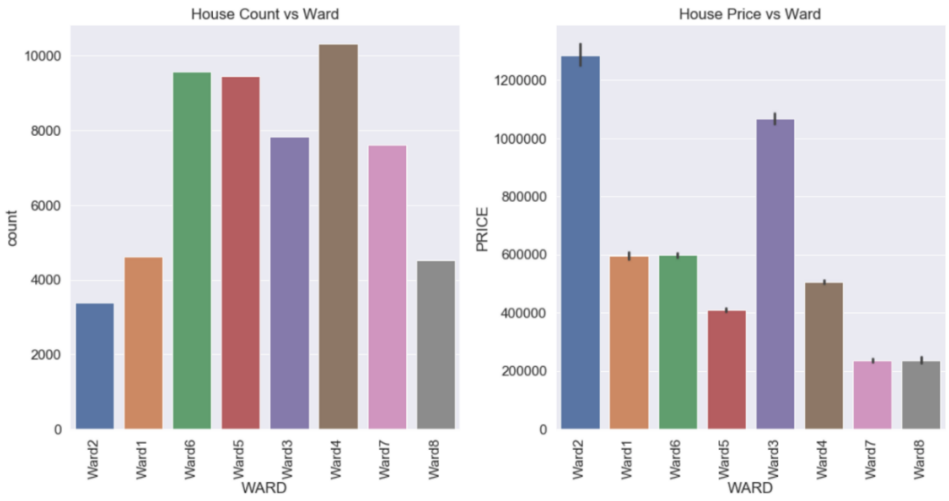

Figure 22. House Count vs Ward & House Price vs Ward

ASSESSMENT_NBHD

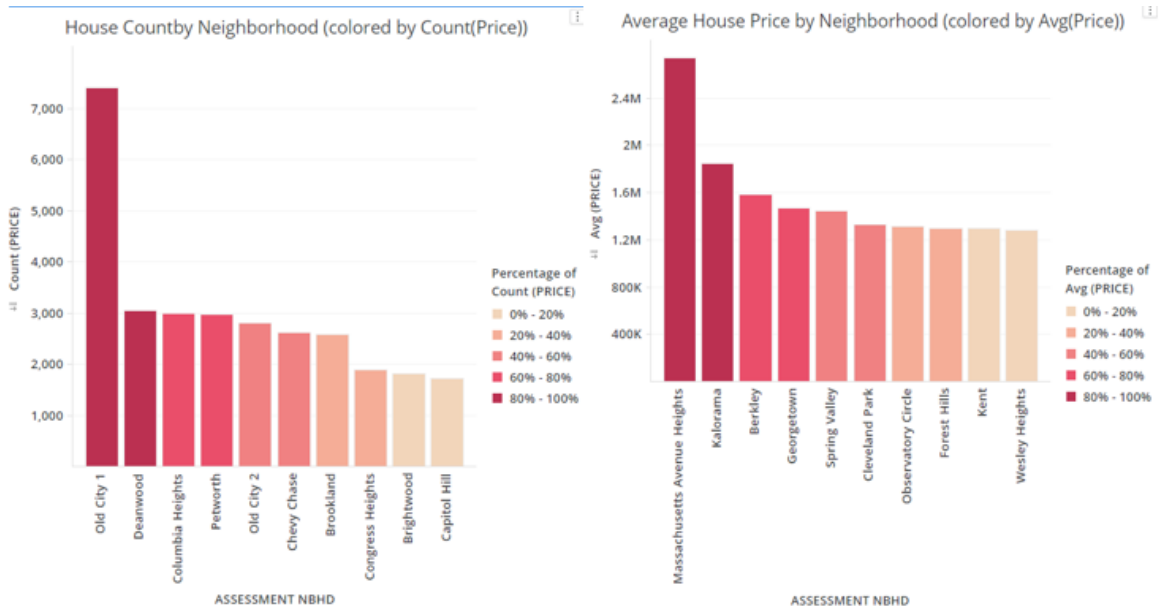

Figure 23. Top 10 House Count by Neighborhood and House Price by Neighborhood

ASSESSMENT_SUBNBHD

Figure 24. Top 10 House Count by Sub-Neighborhood and Top 10 Sub-Neighborhood by Highest Avg House Price

Table 7

Interpretation of Housing External Characteristics (Categorical Variables)

Housing External Characteristic	Number of Houses are higher in	House Price is higher in	Cheap houses are in
Quadrant	Northwest, Northeast	Northwest, Northeast	Southwest, Southeast
WARD	Ward4, Ward6, Ward5	Ward2, Ward3	Ward8, Ward7
ASSESSMENT_NBHD	Old City1, Deanwood, Columbia Heights	Massachusetts Avenue Heights, Kalorama, Berkley	Wersley Heights (in top 10)
ASSESSMENT_SUBNBHD	039k Old City1, 042 B Petworth, 039L Old City1	039k Old City1, 007E Brookland, 039F Old City1,	018 B Deanwood (in top 10)

Correlation

To predict the house price, we need to identify which variables are highly correlated with house prices. Let's check the correlation heatmap. Correlation heatmap considers only numerical variables. We considered all numerical variables.

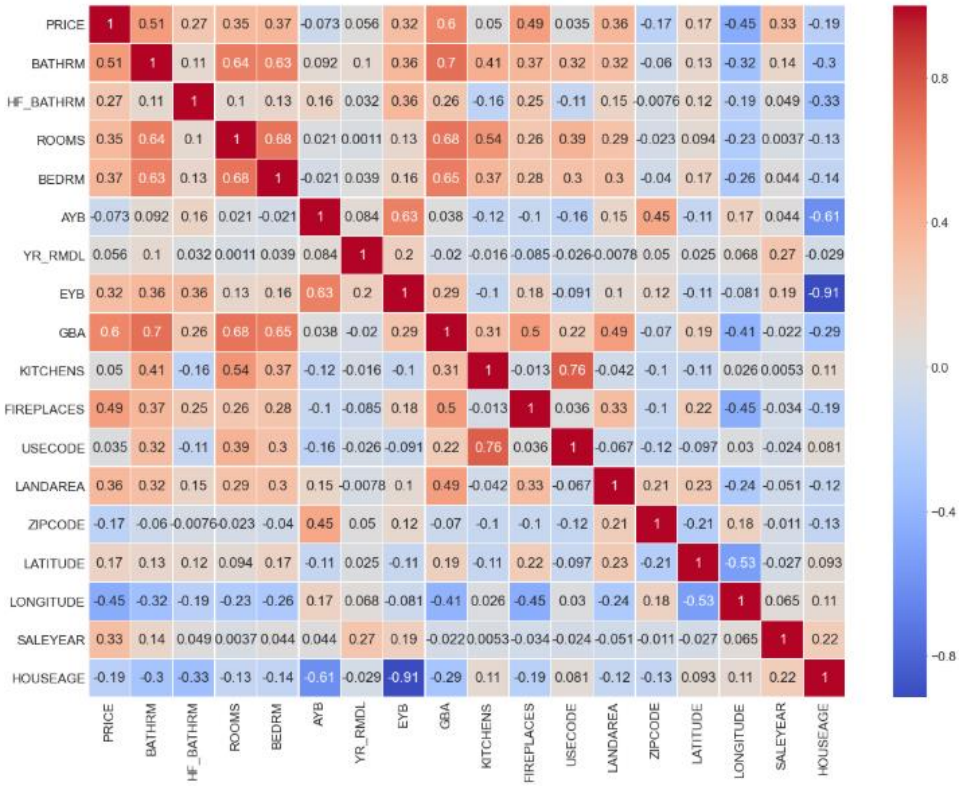

Figure 26. Correlation heatmap

Interpretation: Not all variables are correlated with house prices. There are few variables such as “HOUSEAGE” is negatively correlated (-0.19) with House price. This trend is also confirmed in Figure 6 where we saw house prices is decreasing with an increase in house age. Longitude is also negatively correlated with house prices. Other variables “KITCHEN”, “HF_BATHRM”, “USECODE” have a very low correlation (less than 0.3) with the target variable. From the above heatmap, we have extracted the variables which have a correlation with a price greater than 0.35 (either positive or negative).

Table 8

Variables with Correlation (>0.35)

VARIABLES	CORRELATION
BATHRM	0.513
BEDRM	0.365
GBA	0.604
FIREPLACES	0.495
LANDAREA	0.357
LONGITUDE	-0.453

We also observed that the gross building area (GBA) and the number of bathrooms (BATHRM) are highly correlated (greater than 0.5) with house prices. These variables are used to predict house prices. This also satisfies the study of Journal of Real Estate Literature (Sirmans, Macpherson, & Zietz, 2005) and Review of Regional Studies (Cebula, 2010) where it states that the number of bathrooms, bedrooms, fireplaces, and the gross building area of the house help in estimating the house price and shown a positive correlation.

Time Series Analysis

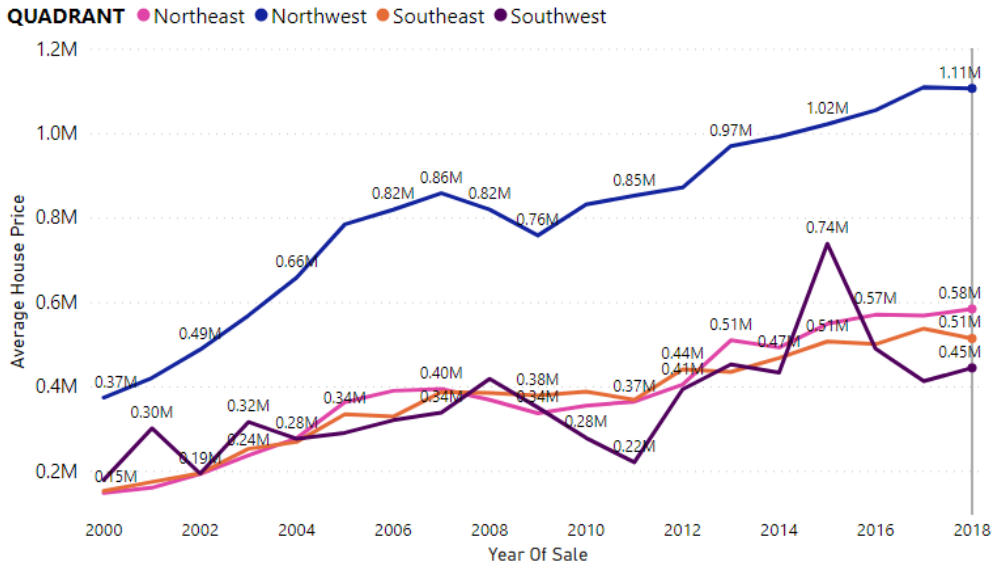

Figure 27. House price variation over years for all quadrant

Interpretation: House Price is increased from 2000 to 2007. The graph shows a decline in House prices between 2007 to 2009. House Price variation is almost the same for all quadrants except Southwest. Southwest shows huge variation throughout the years. There is no seasonal pattern and cyclic movements observed in the graph.

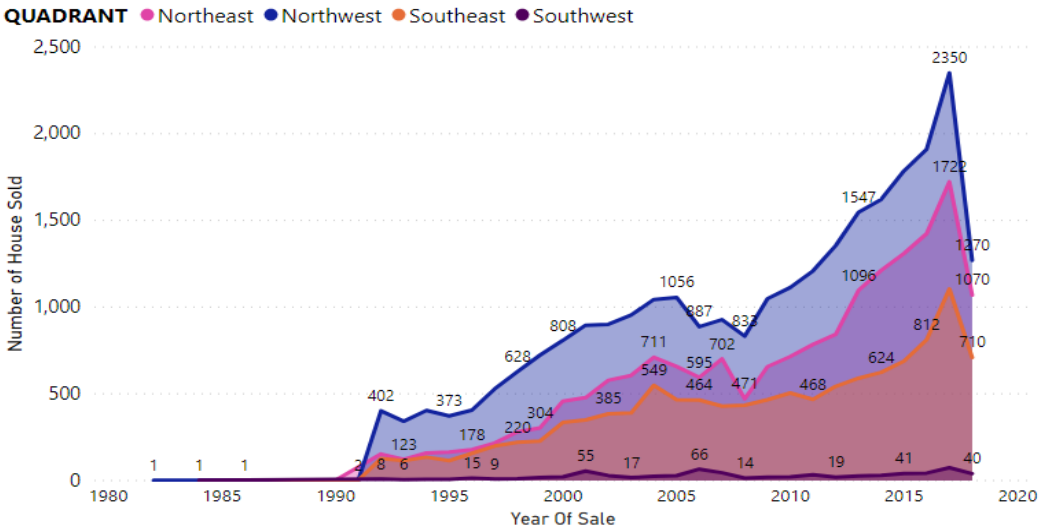

Figure 28. Number of houses sold for all quadrant over years

Interpretation: Maximum and minimum number of houses sold in the Northwest and Southwest respectively. This trend continued throughout the years. The number of houses sold is increased from 2010 to 2017 and shows a steep incline in 2017. The graph shows a decline in the house sold in 2018. House sold variation is almost the same for all quadrants except Southwest. Southwest does not show huge variation throughout the years. There is no seasonal pattern and cyclic movements observed in the graph.

Predictive Analysis

Predictive analysis includes regression analysis, model improvement, model comparison with the help of ANOVA, Prediction using machine learning technique and forecasting of house price.

Tool used: Python for regression analysis and Power BI for forecasting the price.

Regression Analysis

One of the causes of the chaotic results of our statistical analysis is due to data outliers. We removed the outlier from the PRICE before performing the regression analysis.

```
def check_outlier(data, col):
    q1 = data[col].quantile(0.25)
    q3 = data[col].quantile(0.75)
    iqr = q3-q1
    upper_limit = q3 + (1.5 * iqr)
    lower_limit = q1 - (1.5 * iqr)
    return data[(data[col] < lower_limit) | (data[col] > upper_limit)].index ,upper_limit,lower_limit
index_to_drop,upper,lower= check_outlier(df,'PRICE')
df.drop(index_to_drop,inplace=True)
df = df.reset_index().drop('index',axis=1)
df.count()
```

Figure 29. Removing Outliers for regression analysis

We already dealt with numerical variables and their correlation with house prices. To deal with the categorical variable, we used the ordinary least square method to run regression against each categorical variable with house price.

Table 9

Regression Analysis of Categorical variables

OLS method with Confidence Interval=95%, Alpha =0.05		
Script ran	**R-square**	**Adjustes R square**
results = ols('PRICE ~ HEAT', data=df).fit()	0.013	0.013
results = ols('PRICE ~ AC', data=df).fit()	0.092	0.092
results = ols('PRICE ~ STYLE', data=df).fit()	0.089	0.089
results = ols('PRICE ~ STRUCT', data=df).fit()	0.056	0.056
results = ols('PRICE ~ GRADE', data=df).fit()	0.336	0.336
results = ols('PRICE ~ CNDTN', data=df).fit()	0.143	0.143
results = ols('PRICE ~ EXTWALL', data=df).fit()	0.033	0.033
results = ols('PRICE ~ ROOF', data=df).fit()	0.063	0.063
results = ols('PRICE ~ INTWALL', data=df).fit()	0.019	0.019
results = ols('PRICE ~ WARD', data=df).fit()	0.341	0.341
results = ols('PRICE ~ QUADRANT', data=df).fit()	0.157	0.157

From the above analysis, based on R square, Adjusted R square, and p-value, we choose AC, GRADE, CNDTN, QUADRANT, WARD. Other variables show less value of R square. Also, the p-value was greater than 0.05 (alpha) for a few variables.

model1:

```
model1 = ols('PRICE ~BATHRM+BEDRM+GBA+FIREPLACES+LANDAREA+LATITUDE', data=df).fit()
model1.summary()
```

OLS Regression Results

Dep. Variable:	PRICE	**R-squared:**	0.334
Model:	OLS	**Adj. R-squared:**	0.334

Figure 30. Regression Analysis of Model1

Interpretation: We started our regression with the variables which a showed correlation greater than 0.35. model1 did not help much to predict house prices and showed low R square as 0.334. The p-value for all variables is less than alpha (0.05) which shows all variables are significant.

model2:

```
model2 = ols('PRICE ~BATHRM+HF_BATHRM+ROOMS+BEDRM+AYB+EYB+GBA+KITCHENS+FIREPLACES+USECODE+LANDAREA+ZIPCODE+LATITUDE+LONGITUDE+SA
LEYEAR+HOUSEAGE', data=df).fit()
model2.summary()
```

OLS Regression Results

Dep. Variable:	PRICE	**R-squared:**	0.721
Model:	OLS	**Adj. R-squared:**	0.721

Figure 31. Regression Analysis of Model2

Interpretation: We tried to improve R square for better prediction. For model2, we considered all numerical variables that improved R square as 0.721. The p-value for all variables is less than alpha (0.05) which shows all variables are significant.

Model Improvement (Model3):

```
model3 = ols('PRICE ~ BATHRM+HF_BATHRM+ROOMS+BEDRM+AYB+EYB+GBA+KITCHENS+FIREPLACES+USECODE+LANDAREA+ZIPCODE+LATITUDE+LONGITUDE+S
ALEYEAR+HOUSEAGE+AC__N+AC__Y+GRADE__AboveAverage+GRADE__Average+GRADE__Excellent+GRADE__ExceptionalA+GRADE__ExceptionalB+GRADE__
ExceptionalC+GRADE__ExceptionalD+GRADE__FairQuality+GRADE__GoodQuality+GRADE__LowQuality+GRADE__Superior+GRADE__VeryGood+CNDTN__
Average+CNDTN__Excellent+CNDTN__Fair+CNDTN__Good+CNDTN__Poor+CNDTN__VeryGood+QUADRANT__Northeast+QUADRANT__Northwest+QUADRANT__S
outheast+QUADRANT__Southwest+WARD__Ward1+WARD__Ward2+WARD__Ward3+WARD__Ward4+WARD__Ward5+WARD__Ward6+WARD__Ward7+WARD__Ward8', d
ata=df).fit()
model3.summary()
```

OLS Regression Results

Dep. Variable:	PRICE	**R-squared:**	0.774
Model:	OLS	**Adj. R-squared:**	0.774

Figure 32. Regression Analysis of Model3

Interpretation We further tried to improve the R square by adding categorical variables in the model. In the next step, to consider these variables in the regression equation, we converted it into dummy variables.

```
df= pd.get_dummies(df, prefix='AC_', columns=['AC'])
df= pd.get_dummies(df, prefix='GRADE_', columns=['GRADE'])
df= pd.get_dummies(df, prefix='CNDTN_', columns=['CNDTN'])
df= pd.get_dummies(df, prefix='QUADRANT_', columns=['QUADRANT'])
df= pd.get_dummies(df, prefix='WARD_', columns=['WARD'])
```

Figure 33. Convert variables to Dummy Variables

The model3 includes all the categorical (AC, GRADE, CNDTN, QUADRANT, WARD) which we choose after running individual regression (Table 9). It showed R square is improved from 0.721 to 0.774. With R square 0.774, the total variation explained by the model4 is 77.4%.

ANOVA

We used ANOVA method to compare all models to see which one is significantly better.

```
anova_compare_1_2 = sm.stats.anova_lm(model1, model2)
anova_compare_1_2
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	54858.0	3.939103e+15	0.0	NaN	NaN	NaN
1	54849.0	1.649883e+15	9.0	2.289220e+15	8455.912062	0.0

Figure 34. ANOVA Compare of model1 and model2

Interpretation: ANOVA result shows F-statistics as 8455.91 with p-value as 0. We can interpret that Model 2, based on ANOVA results, is statistically an improvement over Model 1.

```
anova_compare_2_3= sm.stats.anova_lm(model2, model3)
anova_compare_2_3
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	54849.0	1.649883e+15	0.0	NaN	NaN	NaN
1	54822.0	1.336928e+15	27.0	3.129553e+14	475.297221	0.0

Figure 35. ANOVA Compare of model2 and model3

Interpretation: ANOVA result shows F-statistics as 475.297 with p-value as 0. We can interpret that Model 3, based on ANOVA results, is statistically an improvement over Model 2.

Prediction Using ML technique

We also tried to use machine learning technique and divided the data (considering the variables used in model3) into test and train to see how our model predict the house price.

X=df[['BATHRM','HF_BATHRM','ROOMS','BEDRM','AYB','EYB','GBA','KITCHENS','FIREPLACES','USECODE','LANDAREA','ZIPCODE','LATITUDE','LONGITUDE','SALEYEAR','HOUSEAGE','AC__N','AC__Y','GRADE__AboveAverage','GRADE__Average','GRADE__Excellent','GRADE__ExceptionalA','GRADE__ExceptionalB','GRADE__ExceptionalC','GRADE__ExceptionalD','GRADE__FairQuality','GRADE__GoodQuality','GRADE__LowQuality','GRADE__Superior','GRADE__VeryGood','CNDTN__Average','CNDTN__Excellent','CNDTN__Fair','CNDTN__Good','CNDTN__Poor','CNDTN__VeryGood','QUADRANT__Northeast','QUADRANT__Northwest','QUADRANT__Southeast','QUADRANT__Southwest','WARD__Ward1','WARD__Ward2','WARD__Ward3','WARD__Ward4','WARD__Ward5','WARD__Ward6','WARD__Ward7','WARD__Ward8']]

y=df['PRICE']

```
print(lm.intercept_)
```

```
-170750298.37391582
```

	Coeff
BATHRM	3.094700e+04
HF_BATHRM	2.689237e+04
ROOMS	-1.306612e+03
BEDRM	9.501265e+03
AYB	-1.175325e+03
EYB	8.085562e+03
GBA	5.808881e+01
KITCHENS	-2.451772e+04
FIREPLACES	3.797273e+04
USECODE	1.212762e+03
LANDAREA	6.435225e+00
ZIPCODE	-7.526787e+02
LATITUDE	-2.247512e+05
LONGITUDE	-1.954203e+06
SALEYEAR	1.514475e+04
HOUSEAGE	7.059187e+03
AC__N	-9.191337e+03
AC__Y	9.191337e+03

	Coeff
GRADE__AboveAverage	-3.629830e+04
GRADE__Average	-5.536839e+04
GRADE__Excellent	1.293653e+05
GRADE__ExceptionalA	1.287594e+05
GRADE__ExceptionalB	1.375155e+05
GRADE__ExceptionalC	-4.060729e+05
GRADE__ExceptionalD	1.164153e-09
GRADE__FairQuality	-6.772955e+04
GRADE__GoodQuality	2.575380e+04
GRADE__LowQuality	-7.926739e+04
GRADE__Superior	1.241309e+05
GRADE__VeryGood	9.921162e+04
CNDTN__Average	-3.297006e+04
CNDTN__Excellent	7.614716e+04
CNDTN__Fair	-5.741985e+04
CNDTN__Good	1.336006e+04
CNDTN__Poor	-7.487663e+04
CNDTN__VeryGood	7.575931e+04

	Coeff
QUADRANT__Northeast	-1.335606e+03
QUADRANT__Northwest	-1.292511e+03
QUADRANT__Southeast	2.717068e+04
QUADRANT__Southwest	-2.454256e+04
WARD__Ward1	1.388874e+04
WARD__Ward2	1.387953e+05
WARD__Ward3	6.024219e+04
WARD__Ward4	-3.692653e+04
WARD__Ward5	-3.715379e+04
WARD__Ward6	7.867353e+04
WARD__Ward7	-4.384059e+04
WARD__Ward8	-1.736788e+05

Figure 36. Intercept and Co-efficient of each predictor

Interpretation: Above figure shows the intercept and co-efficient of each predictor used to predict the house price.

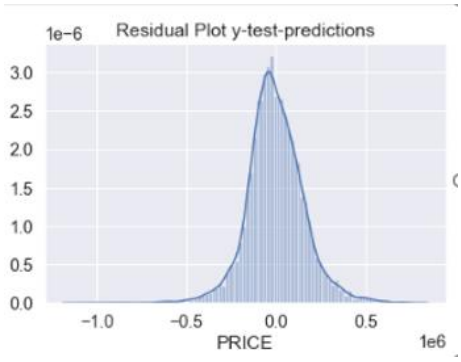

Figure 37. Residual Plot y-test-predictions

Interpretation: The residual plot shows error is normally distributed which satisfy the condition of regression.

```
from sklearn import metrics
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train,y_train)
accuracy = regressor.score(X_test,y_test)
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
df1 = pd.DataFrame({'Actual': y_test, 'Predicted': predictions})
print('Mean of Predicted Price:', df1['Predicted'].mean())
print('Mean of Actual Price:', df1['Actual'].mean())
print('Median of Predicted Price:', df1['Predicted'].median())
print('Median of Actual Price:', df1['Actual'].median())
print('Accuracy rate of the model:', accuracy*100,'%')
```

```
MAE: 117647.3111506675
MSE: 24450452675.43499
RMSE: 156366.40520084545
Mean of Predicted Price: 495682.8143896634
Mean of Actual Price: 494811.57167593186
Median of Predicted Price: 490664.9728253782
Median of Actual Price: 415000.0
Accuracy rate of the model: 77.4645145335255 %
```

Figure 38. Regression statistics

We also calculated mean absolute error, mean squared error, root mean square error to express average model prediction error. Mean of Actual Price and Predicted Price is close and

not show much difference. There is a gap between the median price of actual and predicted price. The model accuracy rate is 77.46 %.

From the above analysis, we see that housing internal characteristics (bathroom, bedroom, fireplaces, AC, etc.) help to determine the house price. We can also interpret that Quadrants and Wards do help to measure the house price which also supports the Research of Journal of Real Estate Literature (Sirmans, Macpherson & Zeitz, 2005), Regional Science and Urban Economics (Brasington & Hite, 2005) and the Journal of Housing Economics (Kiel & Zabel, 2008) states that house price is affected by areas.

Forecast House Price

House Price has been forecasted for five years. Dataset had data from 1982 to 2018. We used forecasting tool Power BI to forecast the price till 2023 (for five years from 2018) at confidence interval 95%. We also used the filter to show the sale year data from the year 2000 and the price has been forecasted for each quarter.

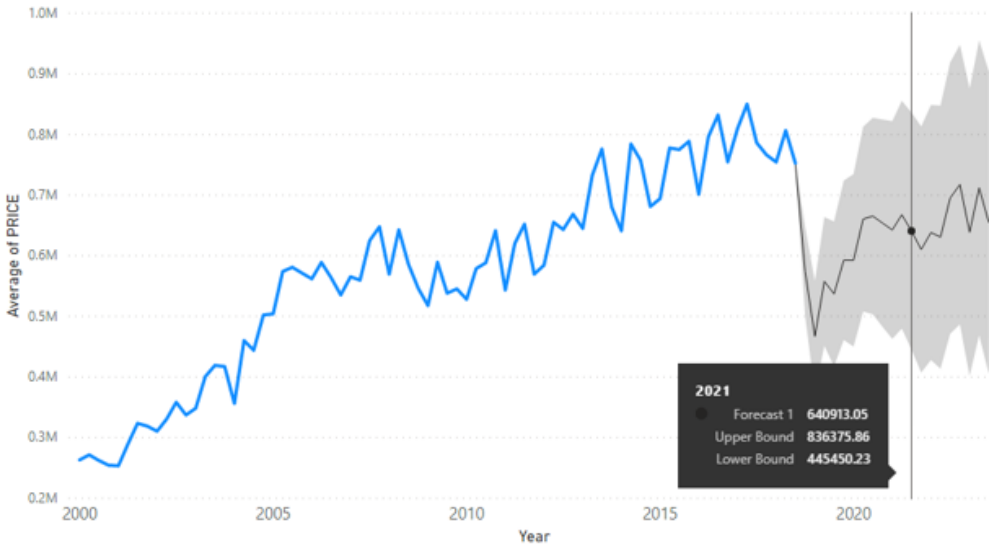

Figure 39. Forecast of House Price from 2019-2023 (for 5 years)

Table 10

Forecasted value for houses in Washington DC

Years and Quarters	Forecast Value (in Million)	Upper Bound (in Million)	Lower Bound (in Million)
2019 Qtr1	0.58	0.65	0.5
2019 Qtr2	0.47	0.56	0.38
2019 Qtr3	0.56	0.66	0.45
2019 Qtr4	0.54	0.65	0.42
2020 Qtr1	0.6	0.72	0.46
2020 Qtr2	0.6	0.73	0.45
2020 Qtr3	0.67	0.8	0.51
2020 Qtr4	0.67	0.83	0.5
2021 Qtr1	0.65	0.83	0.48
2021 Qtr2	0.64	0.82	0.46
2021 Qtr3	0.67	0.86	0.48
2021 Qtr4	0.64	0.84	0.45
2022 Qtr1	0.61	0.82	0.41
2022 Qtr2	0.64	0.85	0.43
2022 Qtr3	0.63	0.85	0.42
2022 Qtr4	0.695	0.92	0.47
2023 Qtr1	0.72	0.95	0.48
2023 Qtr2	0.64	0.88	0.4
2023 Qtr3	0.71	0.96	0.47
2023 Qtr4	0.65	0.9	0.4

Interpretation

House price shows a decline in 2019 and started increasing with little fluctuation each quarter from 2020 till 2023. The highest price observed is $ 0.72 million in the 2023 quarter1 while the lowest price observed is $0.47 million in the 2019 quarter 2. Since 2023 will be the worse year for the home buyer and 2022 can be planned for investment in housing, buyers, seller, and property investors can plan accordingly.

Prescriptive Analysis

Based on descriptive and predictive analysis, we would like to recommend the homeowners and buyers about the house price. Keeping the literature review in focus, we discussed that bathroom, bedroom, gross building area, location, and house age are dominant factors to buy houses. We would like to prescribe the users based on our statistical analysis.

Tools used: Power BI

Recommendation based on Gross Building Area:

Figure 40. House Price in Wards with Gross Building Area (in Groups)

The analysis shows that the average house prices are comparatively very high in Ward 2 and Ward 3 for different ranges of building areas. It would be recommended for homebuyers to invest in homes located in other Wards which would help them save a significant amount of money on similar building areas. For example: House price for houses with GBA (2001-3000 sqft) in Ward 2 is 1.74M as compared to 874K and 853K in Ward 1 and Ward 4 respectively which is almost half of the prices in Ward 2.

Recommendation based on House Age:

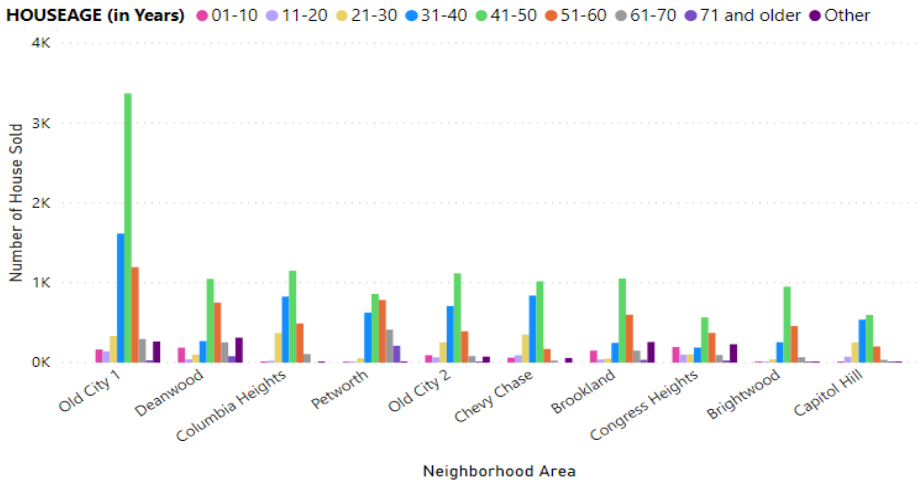

Figure 41. Top 10 House Count Based on House Age in Washington Neighborhood

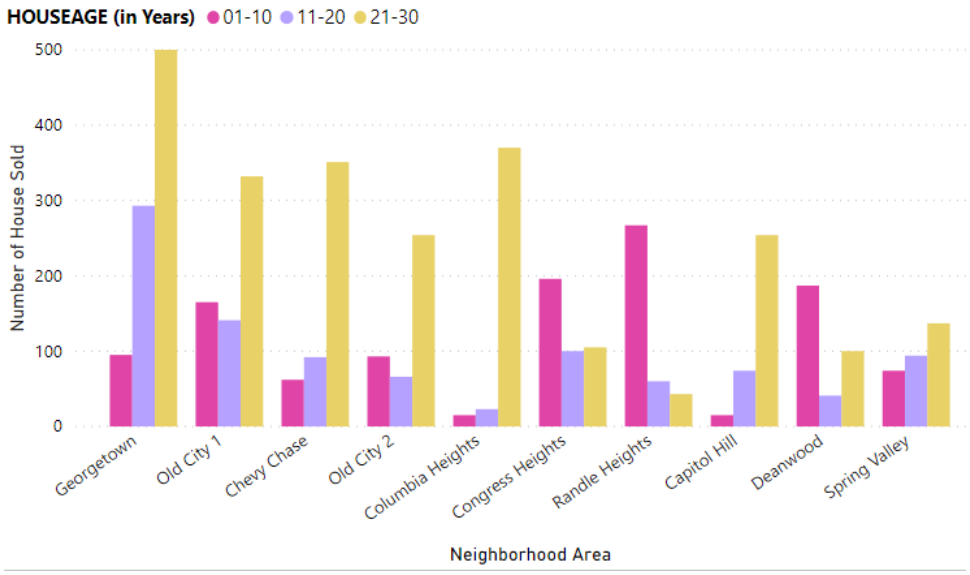

Figure 42. Top 10 House Count Based on House Age (<=30) in Washington Neighborhood

The above two graphs indicate the top 10 house counts based on house age. The first graph suggests that most of the properties are older than 30 years. If one has a preference for old houses, “Old City 1” has the maximum houses older than 30 years.

If someone needs to buy a new house, “Randle Heights” would be a good start with the maximum number of properties aged less than 10 years. “Georgetown” is also a good option for houses less than 30 years old. This is justified by the data shown in the second graph.

Conclusion

We have provided a detailed picture of the house price of Washington DC using a dataset that covers attributes to analyze and predict the house price. The statistical analysis shows that house price has a positive correlation with the gross building area, number of bathrooms, bedrooms, and fireplaces. Also, House age is negatively correlated with house prices. Besides, the availability of AC, good house conditions, Quadrants, and Wards do help to measure the house price. In Washington DC, most of the people have bought houses with 2 or 1 bathrooms, 1

or 0 half bathrooms, 6 or 7 rooms, 3 or 4 bedrooms, 2 or 1 kitchen, 1 or no fireplaces. This indicates that most of the buyers do not look at houses with many bedrooms and fireplaces. People prefer those houses which have AC, with good/average house condition with exceptional grading. People also prefer 2-story houses over other house styles. These facts highlight the housing internal characteristics and suggest buyers/ investors understand how house prices can be assessed with the help of these attributes. On the other side, price is not much affected by the type of heating system, exterior, and interior wall.

As the local neighborhood is the most important factor affecting the house price (Kiel & Zabel, 2008), we tried to seek insight into the neighborhood areas' house price index. At the geographic level, the Northeast quadrant where most of the house sold is the costliest. Ward 4 and Ward 6 have a large number of houses sold, ward2 is the area which buyer can avoid if they don’t want expensive houses and can look for cheap houses in ward8 and ward7. At the fine level, a buyer can also avoid Massachusetts Avenue Heights, Kalorama neighborhood area. Besides, Old City1 is an area that has a large number of houses. But, its sub-neighborhood area 039k Old City1 has the most expensive houses.

This paper also tried to check the trend of the house price and house sold for the past two decades and observed that house prices and house count increased over the years in every quadrant with multiple fluctuations in house prices in the southwest quadrant. As per the forecast of the house price, quarter1 and quarter 3 of 2023 will be the worse year for a home buyer and property investor with forecast value shows 0.72 Million. Quarter1 and Quarter 3 of 2022 can be planned for investment in housing. As part of the prediction, our regression model used 21 features of houses that predicts the price in Washington DC. The performance of the model shows 0.774 as R square and accuracy rate as 77.4%. House price variation depends on multiple

factors such as supply and demand of the house, demographic data, mortgage rate, unemployment rate, GDP, and other socio-economic factors. These factors also affect house prices on a large scale. Considering these facts can lead to further improvement of the model. We would also like to recommend buyers and investors that they can choose the area wisely when they look for bigger houses. It would be recommended to avoid Ward2 and Ward3 and look in Ward1 and Ward4 for bigger houses. It can save a significant amount of money. We would also like to recommend that if buyers look for newer houses, they can avoid Old City1 and start with Randle Heights, Deanwood, Congress heights. These are the area where newer homes exist.

Going ahead, it would be exciting to know how model, recommendation, description of housing internal characteristics would be helpful to buyers, sellers, real estate agents and property investors in Washington DC areas. It would be nice if this analysis can be helpful for Zillow, Trulia, or any real estate company which can assist the interested buyers and sellers.

References

Abidoye, R.B. & Chan, A.P.C. (2016). Critical determinants of residential property value: professionals' perspective. *Journal of Facilities Management*. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/JFM-02-2016-0003/full/html

Asabere, P.K.& Huffman, E.F(1993). "Price Concessions, Time of the Market, and the Actual Sale Price of Homes". In: Journal of Real Estate Finance and Economics 6pp. 167–174. Retrieved from : https://link.springer.com/article/10.1007%2FBF01097024

Brasington, D. M. & Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis. *Regional Science and Urban Economics*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0166046203000929

Cebula, R.J. (2010). The Review of Regional Studies. Retrieved from https://www.researchgate.net/publication/254447595_The_Hedonic_Pricing_Model_Applied_to_the_Housing_Market_of_the_City_of_Savannah_and_Its_Savannah_Historic_Landmark_District

Coulson, N. Edward & Michael L. Lahr. (2005). Gracing the Land of Elvis and Beale Street: Historic Designation and Property Values in Memphis. *Real Estate Economics.* Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1540-6229.2005.00127.x

DC.gov. (2020). Washington, DC. Retrieved from https://dc.gov/release/new-population-new-year-new-housing

GCAAR. (2018). Greater Capital Area Association of REALTORS®. Retrieved from https://gcaar.com/docs/default-source/dc-market-reports/gcaar-dc-housing-market-update---december-2018.pdf?sfvrsn=4f1df393_2

IEEE. (2018). Location-Centered House Price Prediction: A Multi-Task Learning Approach. Retrieved from https://arxiv.org/pdf/1901.01774.pdf

Josephson, A. (2015). The true cost of living in Washington, DC. Retrieved from https://www.businessinsider.com/the-true-cost-of-living-in-washington-dc-2015-11

Kiel, K.A. and J.E. Zabel, 2008. Location, location, location: The 3l approach to house price determination. *Journal of Housing Economics*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S105113770800003X

Kaggle. (n.d.) Washington DC House price. Retrieved from https://www.kaggle.com/christophercorrea/dc-residential-properties#DC_Properties.csv.

Kim, K. & Park, J. (2005). Segmentation of the housing market and its determinants: Seoul and its neighboring new towns in Korea. *Australian Geographer*. Retrieved from https://www.researchgate.net/publication/248999926_Segmentation_of_the_Housing_Market_and_its_Determinants_Seoul_and_its_neighbouring_new_towns_in_Korea

Sirmans, G., Macpherson, D.A., & Zietz, E.N. (2005). "The Composition of Hedonic Pricing Models". *Journal of Real Estate Literature*. Retrieved from: http://www.jstor.org/stable/44103506?seq=1#page_scan_tab_ contents.

Young, R. (2019). Why rent growth is stagnating in New York City, San Francisco & Boston—and how to survive. Retrieved from https://www.buildium.com/blog/primary-markets-2018/