Phase 2 Project

Title: King County Project

Introduction

With nearly 2.2 million residents, King County is the largest county in Washington State. Nationally, it is the 13th largest by population and ninth largest by total employment. Two million of its residents live in one of the 39 cities in the county and the remaining 200,000 in the unincorporated area. Seattle, the largest city in the county, is home to 730,000 residents

It is home to Seattle, the state's largest city, and is a hub of innovation, technology, and creative industries. The housing market in King County has been one of the fastest-growing in the country, with strong demand and limited supply leading to a highly competitive market. Several nationally-known businesses are collectively the major economic drivers for the region: Amazon, Boeing Commercial Airplanes, Microsoft, Starbucks and the University of Washington. These large businesses, and along with smaller enterprises, have led King County out of the Great Recession and into a period of overall economic growth.

Challenges

As a result of this strong economy, the population has increased, attracting new employees for burgeoning businesses, and wages for higher-income households have increased. King County has experienced some of the fastest growing housing prices in the nation. The average King County home value is USD 914,300, an 88 percent jump since 2015.

At its core, the housing crisis is driven by a supply and demand challenge that is two-fold. First, since 2012, King County’s population has grown faster than new homes have been built, creating a growing gap between supply and demand. Second, King County’s population has not grown evenly across the income spectrum. Sixty percent of the new households in King County between 2006 and 2016 earned USD 125,000 or more per year, while 18 percent earned less than USD 50,000. Middle income earners constituted only 22 percent of new households

In response to demand for housing by high-earner households, housing developers have focused new projects to serve the upper end of the market and many of what were once existing affordable units have increased in price beyond what many middle- and low-income working families can afford.

Since 2012, both rent and home purchase prices have increased faster than income, placing intense pressure on middle- and low-income households throughout King County and forcing many to relocate far from where they work or to struggle with paying more than 30 percent or even 50 percent of their income on housing.

Problem Statement

New real estate developers are planning to build new housing facilities and there has been a problem in evaluating homes in King County.

Objectives

Throughout the course of the project, we will answer several questions:

1. To develop a model that will help in identifying the attributes that bring more value to the houses, hence bringing maximum profit. Specifically, to uncover:
    - Which features have the biggest impact on the sale price of a house?
    - How much does location affect the sale price of a house?
2. To create a regression model to advice developers on how to accurately price a property

Data Understanding

For this project, we were tasked with analyzing the King County House Data dataset. This dataset included 21,597 observations on the housing market in King County, WA. This includes 20 feature columns and 1 target column (price). We also learn that there is some missing data in the waterfront, view and yr_renovated columns and that sqft_basement is set as an object datatype when it should probably be set as an integer or float. Looking at the min and max statistics immediately tells that one house has 33 bedrooms which is outrightly an outlier! Also, while the most expensive property is 7.7 million dollars, 75% of the properties fall below $645,000 and that the maximum number of bathrooms is 8 but 75% of the properties are at 2.5 bathrooms or below.

Data Cleaning

Data cleaning involved changing floors from float into integers, changing dates into months then to quarters, cleaning the square foot basement column, dealing with missing values, removing duplicates, checking outliers, turning zip codes into regions. Our dataset consists of 21,597 rows. The total number of records with null values (2339+61 + 3754 = 28%) is currently far greater than 5%, meaning that simply dropping these records would likely significantly impact our analyses.

Waterfront and View are both categorical features. Waterfront has 2339 missing values (roughly 17% of all records) while view has only 61. The great news is that we don't have much heavy lifting to do here — all we have to do is fill the missing values with the mode. Values for Basement square footage are currently stored as strings — and these include question marks. since the missing values are significantly small, we drop them. We also got a sense of the outliers in our data using boxplots and removed outliers for price, bedrooms, bathroom, square footage of living room and floors.

Data Analysis

We explored the following areas to set context for the presentation

1. The average price of property based on number of bedrooms was one thing we considered It was clear that as the bedrooms increase, price increases, up to 8 bedrooms afterwards the price decreases.

2. The average price of property based on the condition of the property was also a pointer. As the condition of the house improves, price increases - houses with a very good condition have higher prices. Thus, keeping property in a good condition will result to a higher valuation

3. The Regions against the number of houses sold. The King County is divided into 4 regions, Northwest, Northeast, Southwest and Southeast. Northwest region which is the region surrounding Seattle has the highest number of houses on sale.

4. Price distribution. The prices were normally distributed largely because we remove price outliers.

5. The Houses sold per quarter was also a key pointer. Houses sold in quarter 2 (April, May and June) have the highest price and the trend shows seasonality of house sales.

Modeling

We did a correlation analysis and found that the column `sqft_living` has the strongest correlation to price. We therefore built our baseline model with this variable. The scatter plot below shows a linear relationship between our predictor and target variable, thus a good candidate for the baseline model. The square footage of the living room only explained 30.6% of variation in price and thus we tried to use standardization to improve it but it didn't change the R^2. We used RSQUARED and MAE to meet 2 objectives:

I. Rsquared will help us asses model improvements in prediction variation in house prices
II. Mean absolute error will help us to measure the distance of the predicted prices from the actual prices. From the code below, we have chosen MAE over RMSE because it is less influenced by outliers

We had 4 model after the baseline model

1. For our first multiple regression model we begin with the numerical variables.
2. In this second model we will combine the first model columns plus the encoded categorical variables: `months_sold`, `condition` & `region`
3. We added the view column to the model to improve accuracy. We one hot encoded, merged, then fitted the third model.
4. We noticed a strong correlation between bathrooms and the square footage of the living room and thus dropped the bathroom column and performed our final multiple linear regression.

Conclusion

1. Feature Importance:

- The number of bedrooms, number of bathrooms, and square footage of the living area are all important features that significantly affect the price of the house. In particular, an increase of 1 sq foot in sqft_living will result in an increase in house price by USD 169.85, and an increase in 1 bathroom results in an increase in price by USD 15120.0. However, an increase in 1 bedroom results in a decrease in price by USD 22130.0.

2. Location:

- The location of the house also appears to be important, with different regions having different associated increases or decreases in price. In particular, the north-west region has an associated increase in price of about USD 107500.0, while the south-west has an associated decrease in price of about USD 102100.0.

3. Condition and View:

- The condition of the house and the view from the house are also important factors that affect the price. For example, having an "EXCELLENT" view is associated with an increase in price of about USD 142900.0, while having a "NONE" view is associated with a decrease in price of about USD 76080.0. In terms of condition, having a "GOOD" condition is associated with an increase in price of about USD 36110.0, while having a "FAIR" condition is associated with a decrease in price of about USD 44500.0.

Recommendations

Based on the results provided, here are some recommendations that you could provide to realtors:

1. Focus on important features: Emphasize to realtors the importance of the number of bedrooms, number of bathrooms, and square footage of the living area when determining a property's value. In particular, an increase of 1 sq foot in sqft_living will result in an increase in house price by USD 169.85, and an increase in 1 bathroom results in an increase in price by USD 15120.0. However, an increase in 1 bedroom results in a decrease in price by USD 22130.0. Realtors should consider these features when pricing a property or advising sellers on what features to highlight in a property.
2. Highlight desirable locations: Encourage realtors to highlight the location of a property, especially in terms of the region it is located in. Realtors should inform clients that different regions have different associated increases or decreases in price. In particular, the north-west region has an associated increase in price of about USD 107500.0, while the south-west has an associated decrease in price of about USD 102100.0.
3. Emphasize condition and view: Realtors should also emphasize the importance of the condition of the property and the view from the property. For example, having an "EXCELLENT" view is associated with an increase in price of about USD 142900.0, while having a "NONE" view is associated with a decrease in price of about USD 76080.0. Similarly, having a "GOOD" condition is associated with an increase in price of about USD 36110.0, while having a "FAIR" condition is associated with a decrease in price of about USD 44500.0. Realtors should ensure that the condition of a property is highlighted in any marketing materials or listings, and they should take steps to ensure that the view from the property is presented in the best possible light.

Next steps

1. Improve the model accuracy by adding more data.
2. Create a dashboard for easy data analysis.