1.0 MODELING HOUSE PRICES USING DIFFERENT HOUSE FEATURES IN KING COUNTY.

1.1 Background/Overview

The purchase, sale, leasing, and development of real estate, including residential, commercial, and industrial properties, are all included in the wide variety of operations that make up the real estate sector. Properties meant for individual use as dwellings are included in the category of residential real estate. Single-family homes, townhouses, apartments, condos, and vacation homes are all included in this sector.

Properties utilized for business, such as office buildings, retail stores, industrial facilities, and hospitality establishments (hotels, restaurants), are included in the category of commercial real estate. The third type of real estate is industrial, which comprises buildings used for logistics, distribution, manufacturing, and warehousing. Manufacturing sites, distribution hubs, warehouses, and industrial parks are all included in this category.

Stakeholders in the real estate industry include homeowners, buyers, sellers, real estate agents, developers, lenders financiers like banks and contractors.

The real estate industry faces several challenges that impact various stakeholders. These include:

1. Limited availability of developable land or high cost of land to develop.
2. Technological advancements, such as digital platforms, data analytics, and automation present challenges for traditional real estate businesses and industry practices.
3. Changing demographics, including urbanization trends and migration patterns, are reshaping housing preferences, demand for different types of properties.
4. High price of houses which presents difficulties for potential tenants and homeowners.
5. Economic uncertainty, initiating reduced demand for real estate, challenges in securing financing, and heightens risk for both investors and developers.
6. Inadequate infrastructure investment and planning such as poor social amenities, traffic congestion, environmental degradation affects property values and development opportunities.
7. Vulnerability of infrastructure and structures to risks associated with climate change

Various solutions that can be used to address these challenges include:

1. Introducing measures and regulations to improve developers' and purchasers' access to cheap financing options. These include low mortgage rates or interest rates by the financial institutions.
2. Improving the infrastructure to facilitate urbanization and real estate developments. This involves making investments in infrastructure related to public transportation, roads, water supply, sewage, and energy.

3. Embracing and adapting to new modern or innovative designs. These cover a broad spectrum of architectural characteristics and styles that correspond to modern tastes, technological developments, and lifestyle trends.
4. Utilizing storm-resistant building materials, construction methods that withstand natural disasters and climate-related risks.

1.2 Conclusion

Using King County House Data to predict house prices is challenging and exciting task. By using data on different features and characteristics of a house, we will develop a regression model that can predict house prices. This will be a systematic process that entails steps such as preprocessing, model training and model tuning. The insights that will be gained from the model can help buyers and sellers to make well informed decisions to their advantage.

1.3 Problem Statement

Home sellers and buyers often struggle to determine the optimal pricing and value of a property. This project aims to analyze a dataset of King County house sales, focusing on features like square footage, number of bedrooms, bedrooms and many more.

Through this analysis, we aim to provide valuable insights to assist sellers, whether homeowners or developers, in implementing data-driven strategies to enhance their properties for maximum sale price. Additionally, we aim to equip buyers with a clearer understanding of the factors influencing home value, empowering them to make well-informed decisions throughout the buying process.

1.4 Main Objective

To build and evaluate models using various combinations of the available features in the King County.

1.5 Specific Objectives

1. To evaluate how the number of floors impact the price of a house in King County.
2. To determine how the number of bedrooms impact the price of a house in King County.
3. To examine the impact the number of bathrooms has on the price of a house in King County.
4. To assess the impact of renovations on the price of a house in King County.
5. To determine how the square footage of living space of a house impacts house price in King County.
6. To evaluate which combinations of the available features in the dataset are the most impactful features for predicting sale price.

2.0 DATA UNDERSTANDING

This project analyzes data about homes sold in King County, Washington between May 2014 and May 2015 in order to make recommendations to relevant stakeholders.

This dataset is housed in the kc_house_data.csv file within the project's data folder and the columns outlined in the accompanying column_names.md file. The original dataset contains records of 21,597 home sales.

The columns in the dataset are:

id - Unique identifier for a house

date - Date house was sold

price - Sale price (prediction target)

bedrooms - Number of bedrooms

bathrooms- Number of bathrooms

sqft_living - square footage of living space in the home

sqft_lot - Square footage of the lot

floors - Number of floors (levels) in house

waterfront - Whether the house is on a waterfront

view - Quality of view from house

condition - How good the overall condition of the house is. Related to maintenance of house.

grade - Overall grade of the house. Related to the construction and design of the house.

sqft_above - Square footage of house apart from basement

sqft_basement - Square footage of the basement

yr_built - Year when house was built

yr_renovated - Year when house was renovated

zip code - ZIP Code used by the United States Postal Service

lat - Latitude coordinate

long - Longitude coordinate

sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors

sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors.

The dataset encompasses a diverse range of data types and categories. The dataset predominantly consists of numeric variables, which represent quantitative measurements or counts. These include essential attributes such as the property's sale price (**price**), the number of bedrooms (**bedrooms**), bathrooms (**bathrooms**), total living area (**sqft_living**), lot area **(sqft_lot**), number of floors (**floors**).

It also has other size-related features such as **sqft_above, sqft_basement, sqft_living15**, and **sqft_lot15**. These numeric variables provide valuable insights into the physical characteristics and dimensions of the properties.

There are several categorical variables, which represent qualitative attributes or characteristics. These include features such as **waterfront**, indicating whether the property has a waterfront view (categorized as 'Yes' or 'No'), **view**, indicating the rating of the property's view, **condition**, representing the overall condition rating of the property, and **grade**, denoting the overall grade assigned to the housing unit based on a grading system established by King County.

Moreover, the dataset includes **yr_renovated**, although it's represented as a numeric variable, its discrete nature implies it holds categorical significance, marking the year of the property's last renovation.

2.0 DATA PREPARATION AND CLEANING.

We dropped the original column ‘yr_renovated’ and replaced it with ‘house_renovation’ which would return yes or no if the house was renovated or not.

Checked the null and duplicate values in the dataframe and found out that ‘waterfront’ has 2376 null values and the ‘view’ column has 63 null values, and all other columns have no null values. Our dataframe appears to have no duplicates.

Dropped the null values.

We also changed the output of waterfront column from NaN to None.

Went ahead and checked for outliers. Outliers are simply points that differ from the rest in the dataset and they may distort statistical measures leading to misinterpretation of the data. Because of this we will focus on the numerical columns.

Dropped the outliers that were detected.

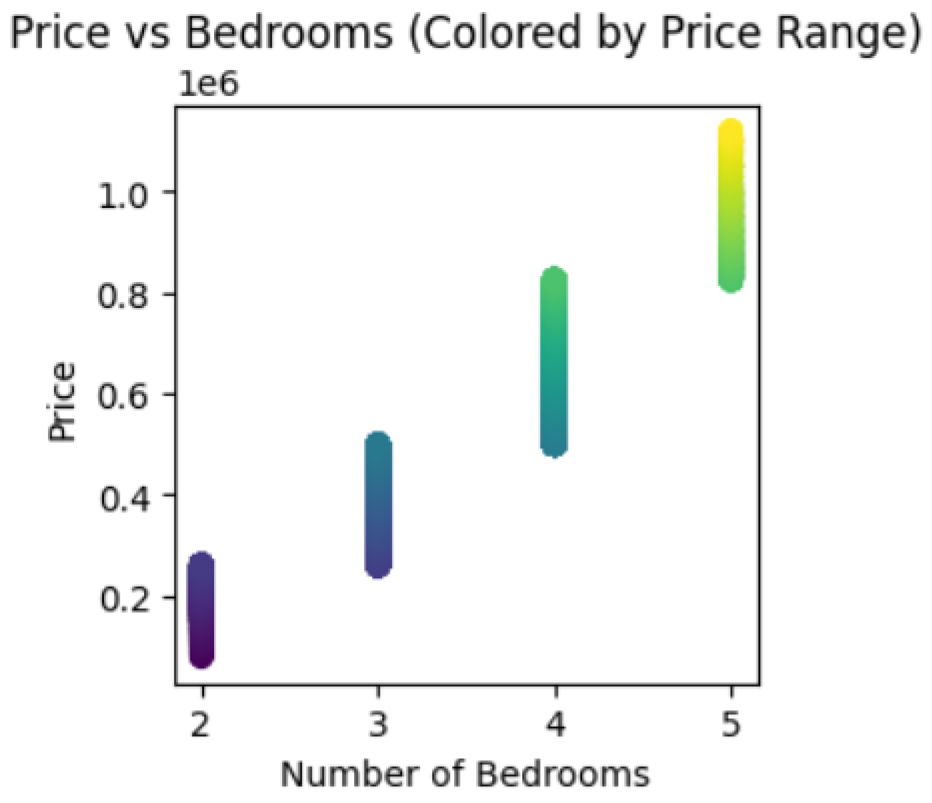
We also checked at the shape of original data frame and compared it with the shape of the cleaned data frame and we now have 17702 rows and initially we had 21597 rows meaning we managed to remove outliers.

We also checked for null values in our cleaned data frame and found the cleaned_dataframe to have 53 null values in the’view’ column which we then dropped. After that we looked at the information of the cleaned data frame.

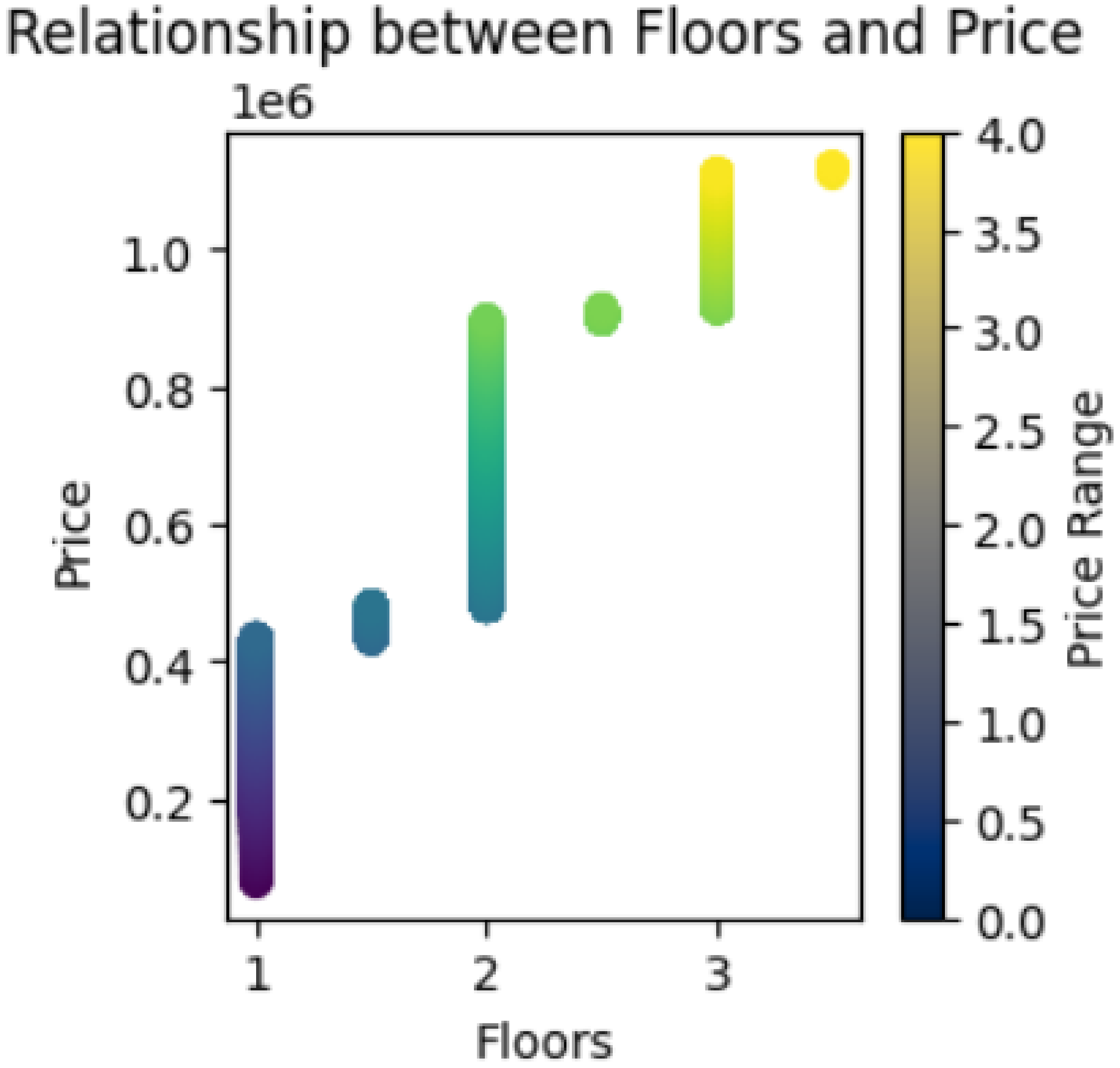
3.0 DATA ANALYSIS

In this comprehensive data analysis, we delved into the intricate web of factors influencing house prices in King County through Exploratory Data Analysis (EDA). Our objective was to uncover the relationships between various features and house prices to gain valuable insights for stakeholders in the real estate market.

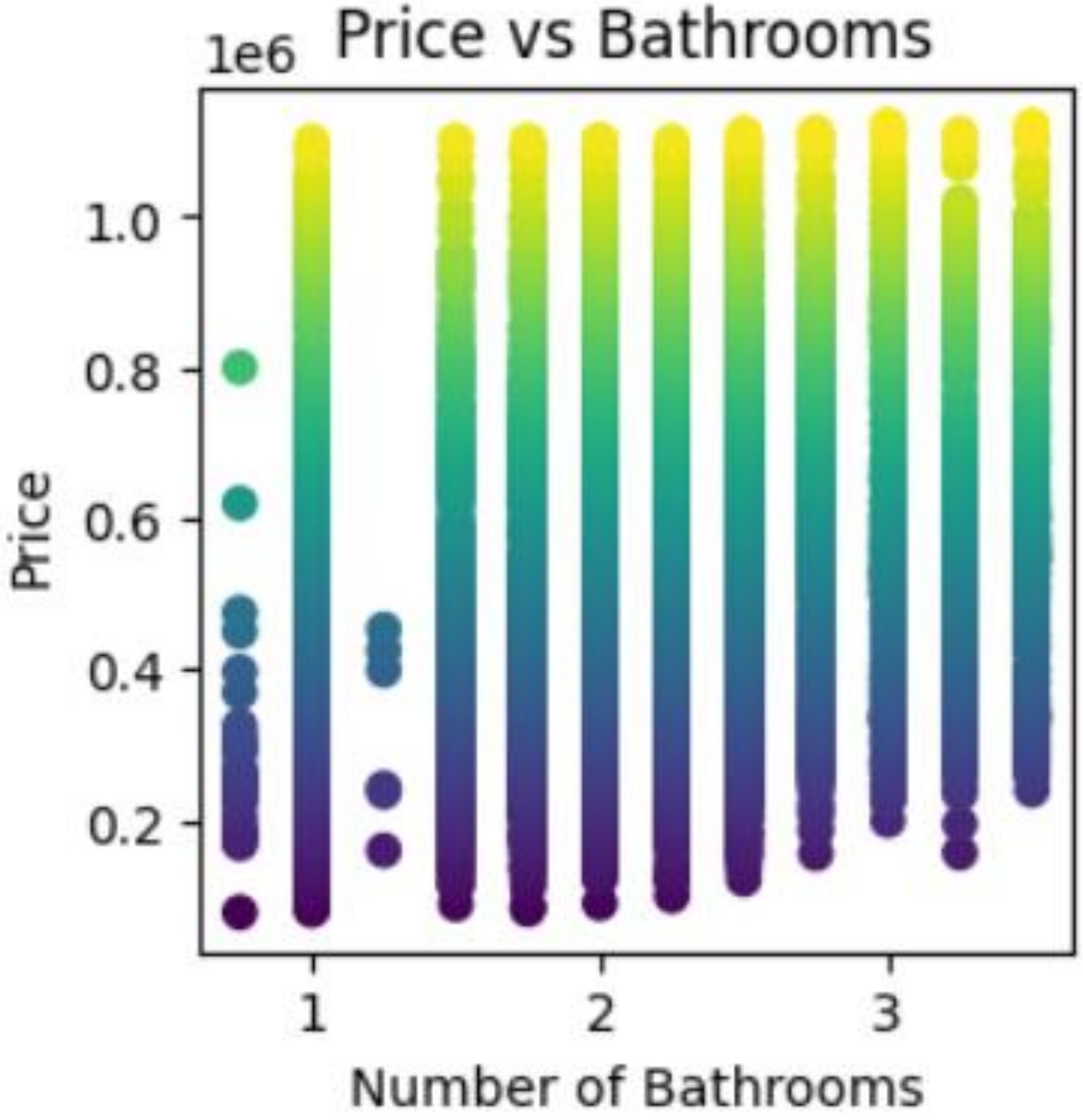
Firstly, we explored the impact of the number of floors on house prices. Analyzing the distribution of floor counts revealed that the majority of houses boast either 1.0 or 2.0 floors, with the latter being the preferred choice among buyers. Notably, as the number of floors increases, so does the price, indicating a positive correlation between floor count and house prices.



Moving on, we investigated the relationship between the number of bedrooms and house prices. Our analysis indicated that houses with a higher number of bedrooms tend to command higher prices. The majority of houses fell within the 3 to 4 bedrooms range, reflecting the premium placed on spacious accommodation.

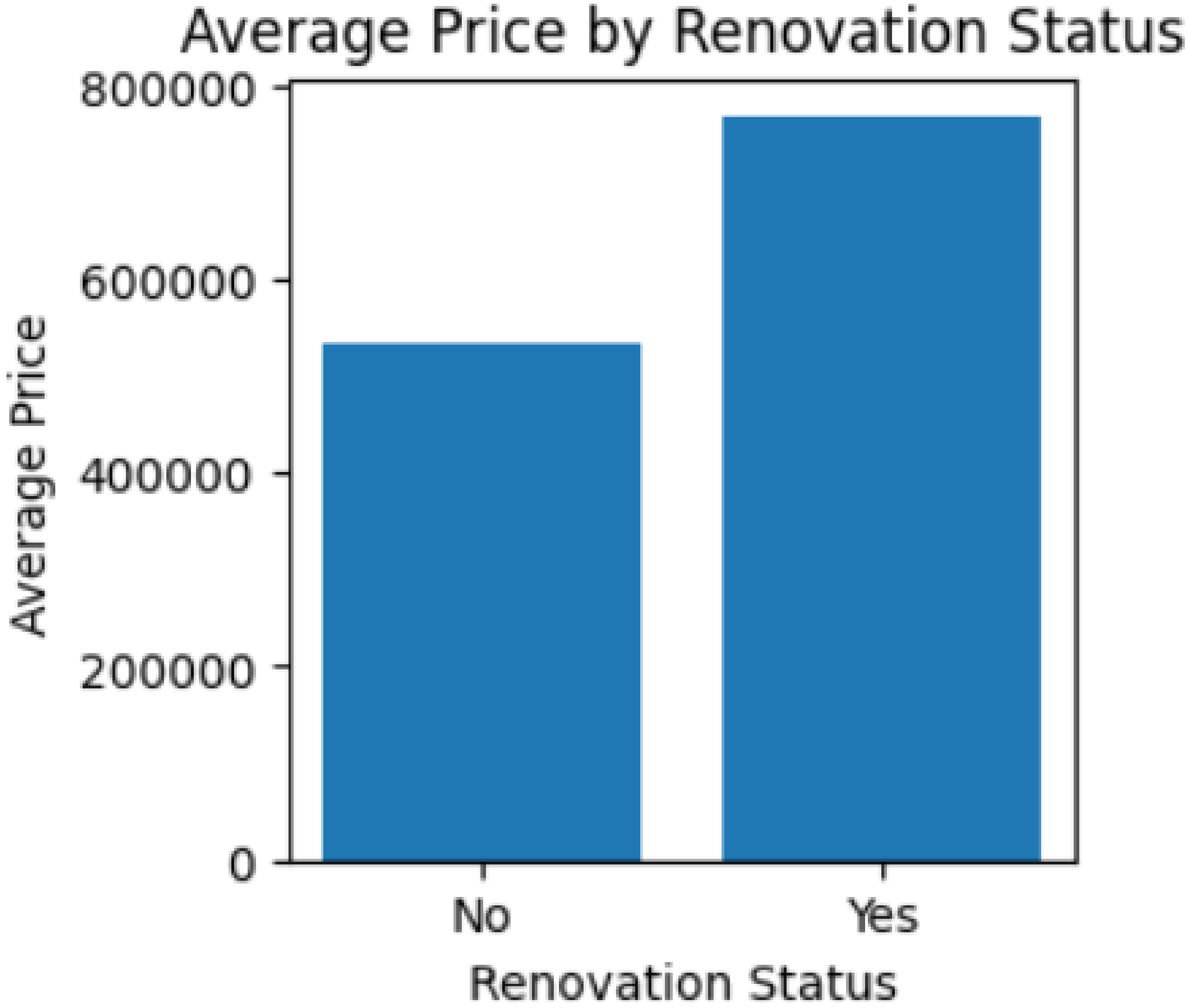


Surprisingly, our examination of the impact of the number of bathrooms on house prices revealed minimal differences in pricing based on bathroom count. Even houses with as few as 0.75 bathrooms exhibited comparable prices to those with more lavish amenities, suggesting that bathrooms have a limited impact on house prices in King County.

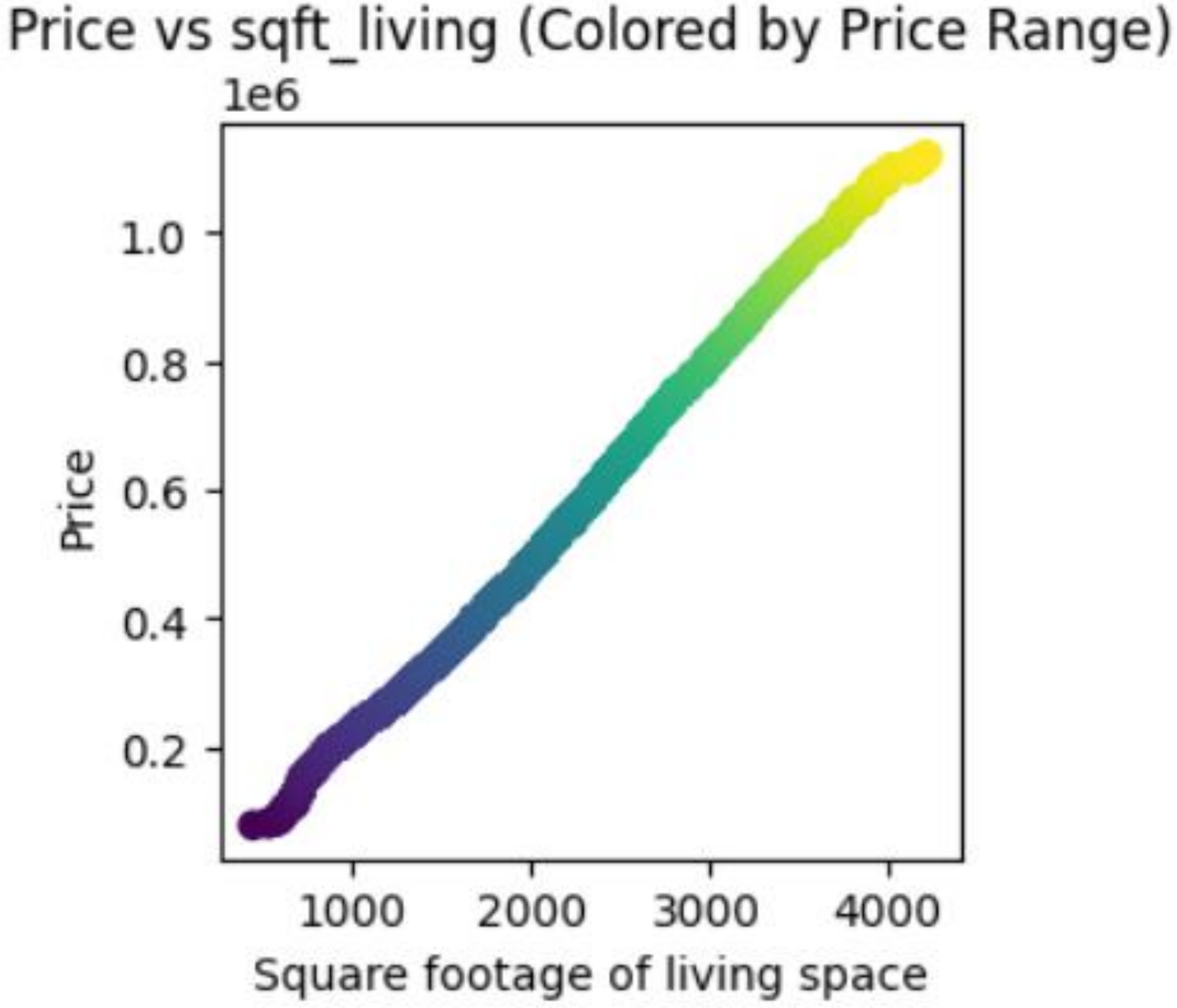


Next, we turned our attention to the influence of renovations on house prices. Our findings indicated that renovated houses fetch higher prices compared to their non-renovated counterparts.

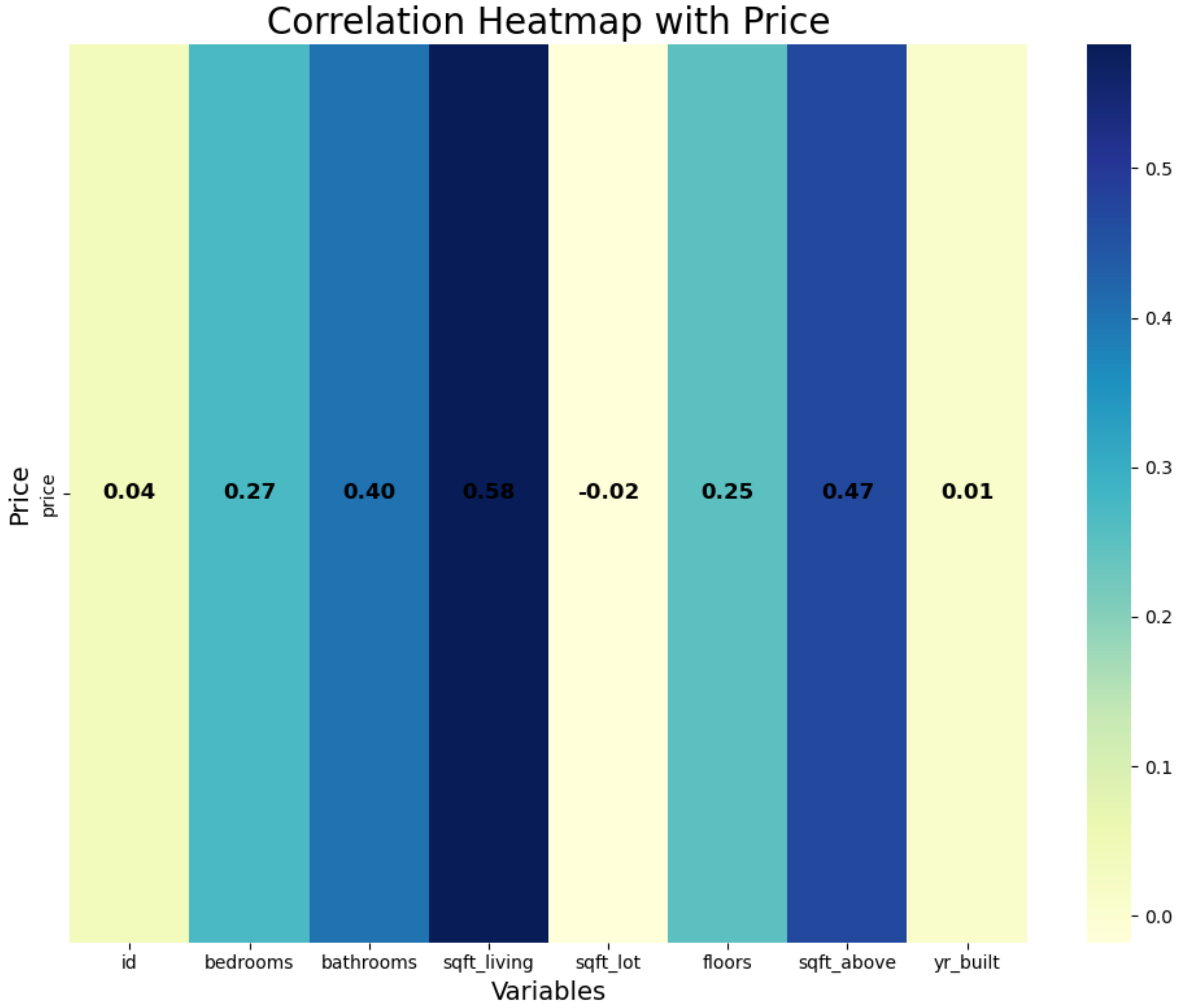
This underscores the importance of property enhancements in augmenting perceived value and commanding premium prices in the market.



Size also emerged as a significant determinant of house prices, with larger living spaces correlating positively with higher prices. This reaffirms the notion that spacious living environments are highly valued by buyers in King County.



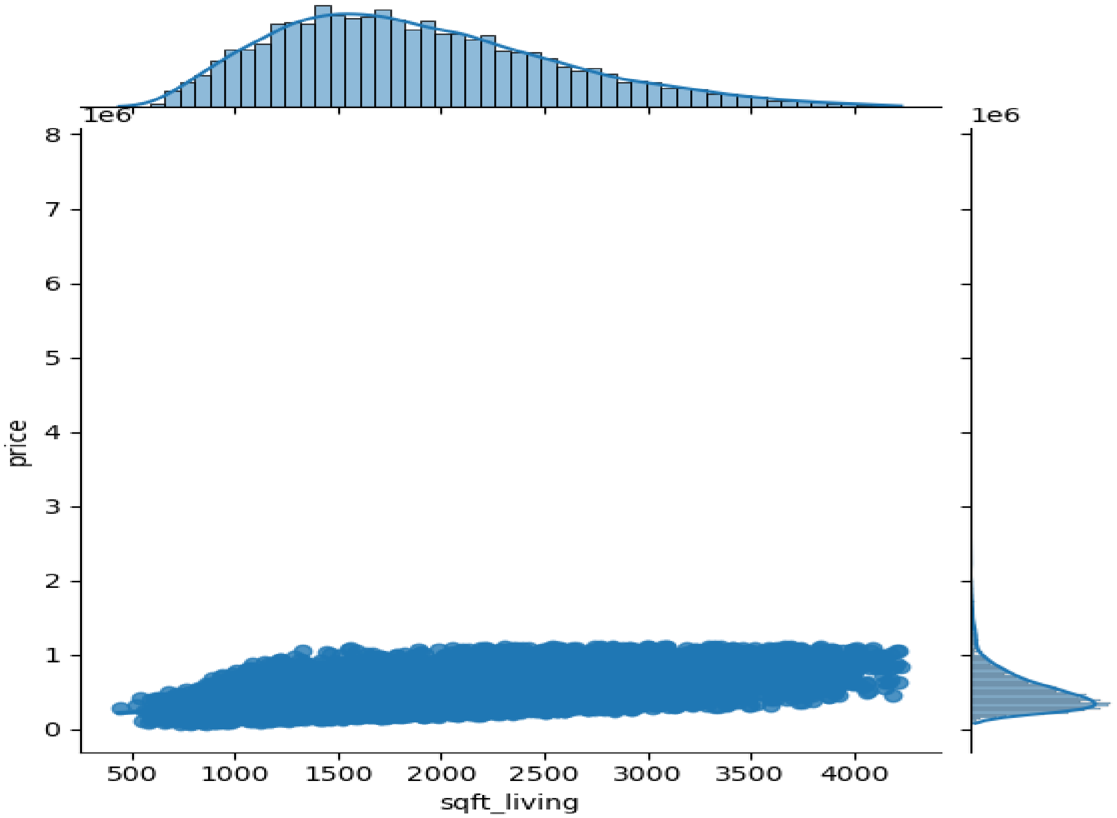
Finally, we sought to identify the most impactful features for predicting sale prices through correlation analysis. Square footage of living space (sqft_living) emerged as the most influential variable, followed closely by sqft_above and the number of bathrooms. Leveraging these insights, stakeholders can develop robust models for predicting house prices and making informed decisions in the dynamic real estate landscape of King County.

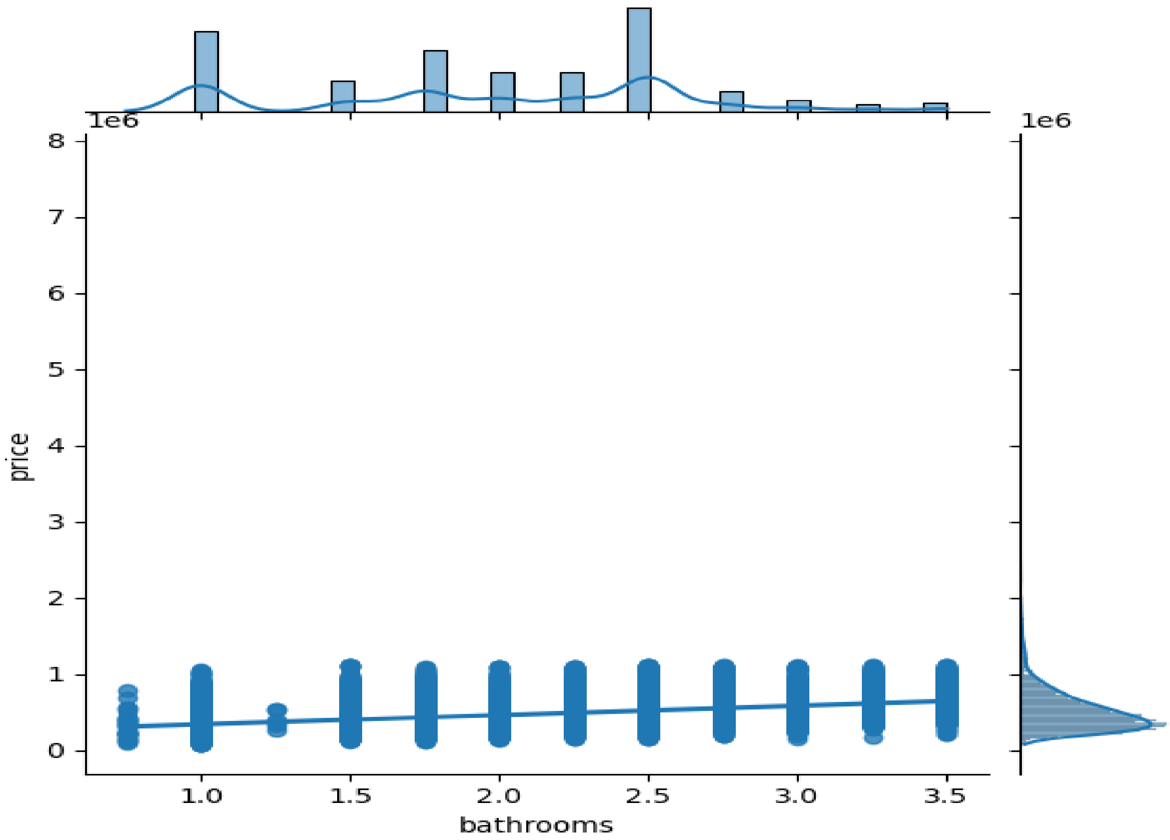


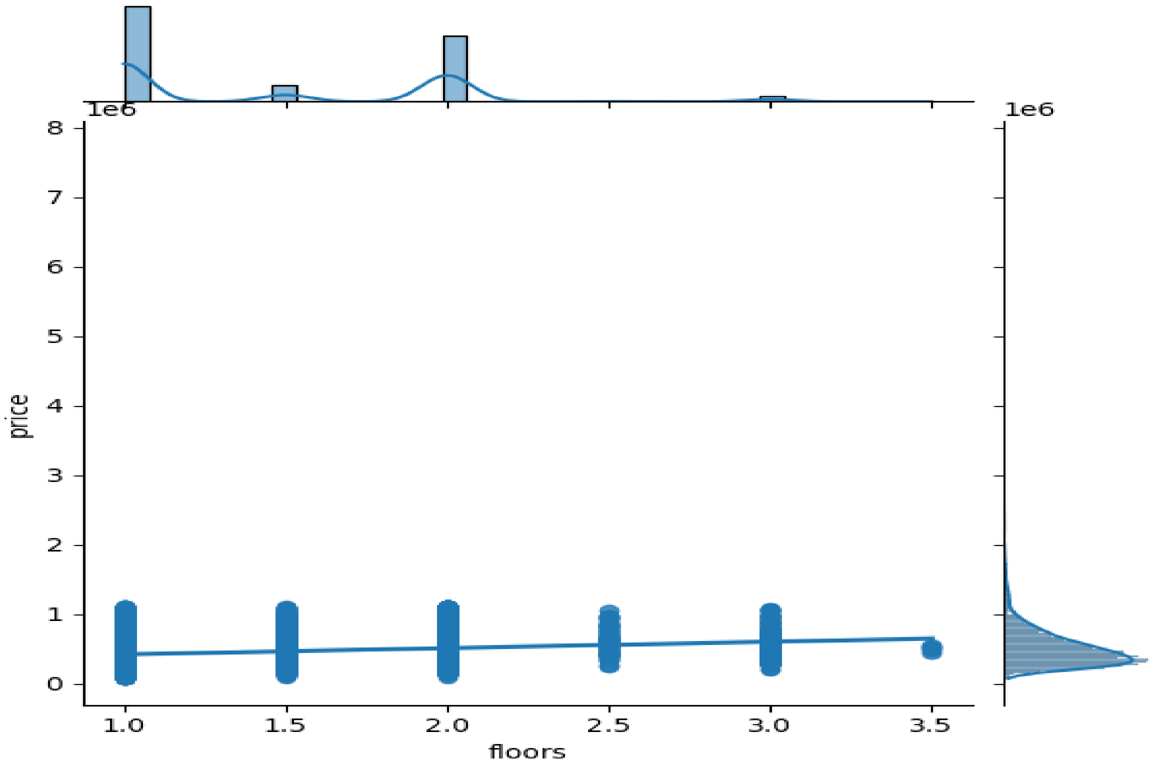
4.0 MODELING AND EVALUATION

4.1 Checking for linearity

The success of any statistical or machine learning model depends greatly on the underlying assumption of linearity between the dependent variable and the independent variables. In the context of our objective, where we aim to predict housing prices based on several features such as the number of floors, bedrooms, bathrooms, sqft_living, and renovation status, it becomes imperative to ascertain the linearity between these predictors and the target variable, 'Price'.




1e6
8
7
6
5
4
3
2
1
0
price
1.0
1.5
2.0
2.5
3.0
3.5
bathrooms
1e6


1e6
8
7
6
5
4
3
2
1
0
price
1.0
1.5
2.0
2.5
3.0
3.5
floors
1e6

Our results showed that all our features appear linear.

4.2 Checking for multicollinearity

Multicollinearity refers to the situation where two or more independent variables in a regression model are highly correlated with each other. When multicollinearity exists, it can cause issues in the regression analysis, such as unstable coefficient estimates or inflated standard errors. To assess multicollinearity, we used the Variance Inflation Factor (VIF), a statistical measure that quantifies the severity of multicollinearity among predictor variables.

Variance Inflation Factor (VIF)

The VIF measures how much the variance of an estimated regression coefficient increases if independent variables are correlated. It quantifies the extent to which the variance of the coefficient estimate is inflated due to multicollinearity. A VIF of 1 indicates no multicollinearity, while values greater than 1 suggest increasing levels of multicollinearity. Generally, a VIF greater than 10 is considered problematic and may warrant further investigation.

Calculation of VIF

To calculate the VIF for each independent variable, we fit a separate regression model for each predictor variable, using all other predictors as independent variables. The VIF for each predictor is then computed as the reciprocal of the tolerance, where tolerance is defined as 1 minus the coefficient of determination (R-squared) of the regression model.

Interpretation of VIF

A high VIF indicates that the variance of the estimated regression coefficient is inflated due to multicollinearity. In practical terms, this means that the presence of multicollinearity makes it difficult to determine the true effect of the independent variable on the dependent variable. Therefore, it is essential to identify and address multicollinearity to ensure the reliability and interpretability of the regression results.

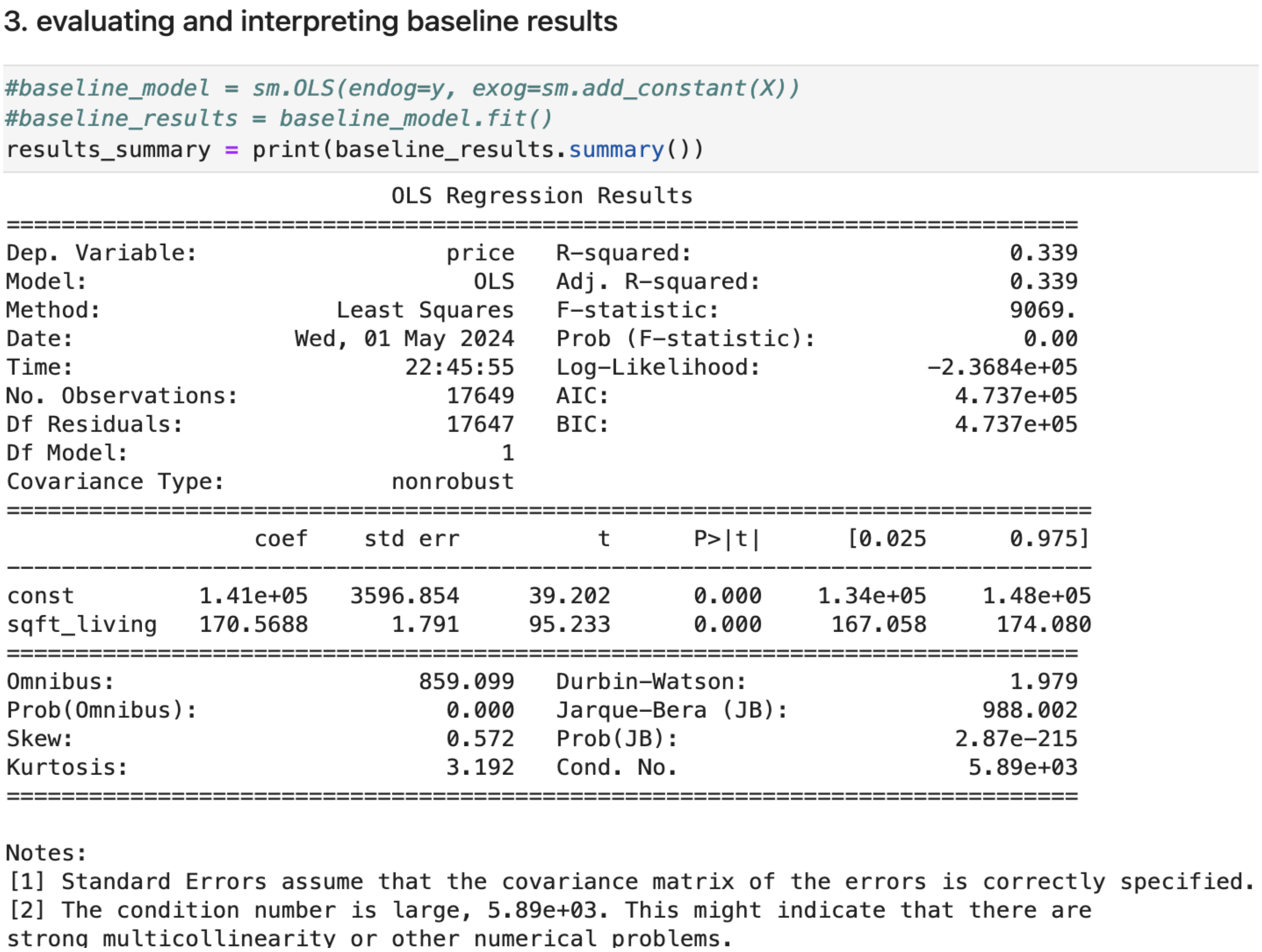
From our findings: features with VIF values greater than 5 (sqft_living and sqft_above) are concerning because they exhibit a high multicollinearity. The others exhibit a multicollinearity effect that is not problematic.

4.3 MODELING

A) Simple linear regression

1.Baseline Model

- Analyzed how the size of living space affected the price.
- Square foot living only predicted price with a 33.9% accuracy
- 1 unit change in soft living led to an increase in house prices



3. evaluating and interpreting baseline results

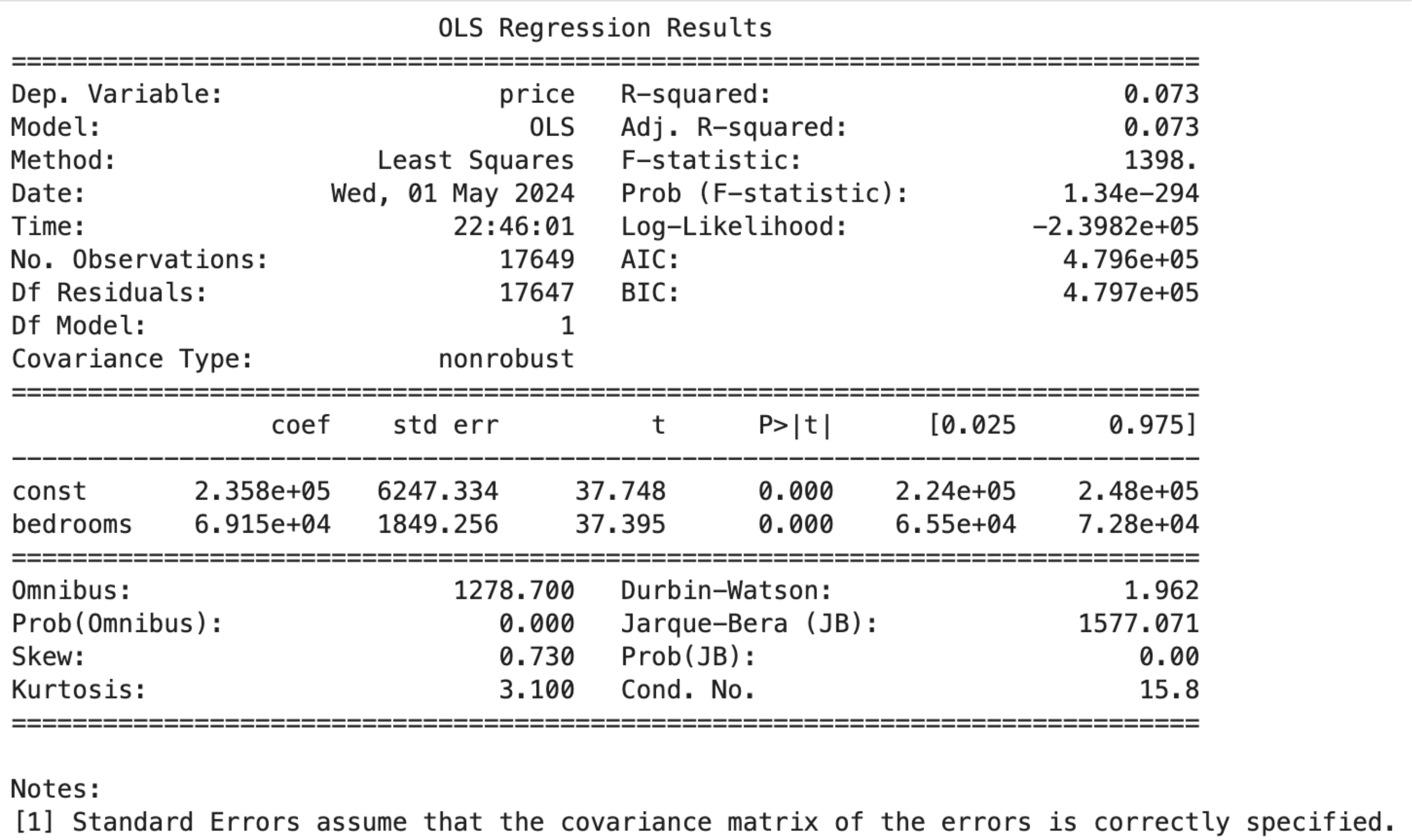
```
#baseline_model = sm.OLS(endog=y, exog=sm.add_constant(X))
#baseline_results = baseline_model.fit()
results_summary = print(baseline_results.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.339
Model:                            OLS   Adj. R-squared:                  0.339
Method:                 Least Squares   F-statistic:                     9069.
Date:                Wed, 01 May 2024   Prob (F-statistic):               0.00
Time:                        22:45:55   Log-Likelihood:            -2.3684e+05
No. Observations:               17649   AIC:                         4.737e+05
Df Residuals:                   17647   BIC:                         4.737e+05
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const         1.41e+05   3596.854     39.202      0.000    1.34e+05    1.48e+05
sqft_living   170.5688      1.791     95.233      0.000     167.058     174.080
==============================================================================
Omnibus:                      859.099   Durbin-Watson:                   1.979
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              988.002
Skew:                           0.572   Prob(JB):                    2.87e-215
Kurtosis:                       3.192   Cond. No.                     5.89e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.89e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Caption

2nd Model: Price vs Bedrooms

- Analyzed how the number of bedrooms affected the price.
- Number of bedrooms could only account for 7.3 % accuracy



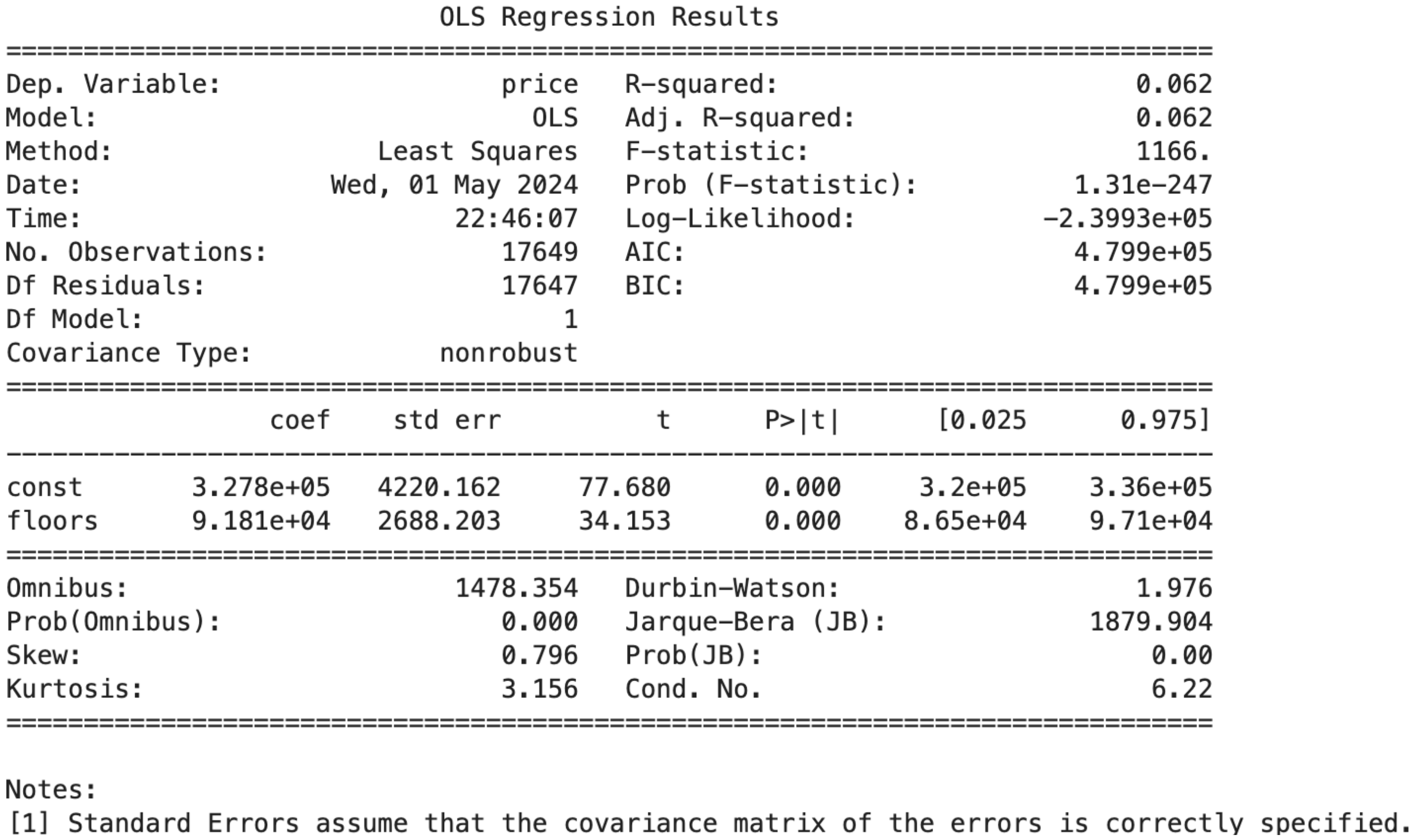
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.073
Model:                            OLS   Adj. R-squared:                  0.073
Method:                 Least Squares   F-statistic:                     1398.
Date:                Wed, 01 May 2024   Prob (F-statistic):          1.34e-294
Time:                        22:46:01   Log-Likelihood:            -2.3982e+05
No. Observations:               17649   AIC:                         4.796e+05
Df Residuals:                   17647   BIC:                         4.797e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.358e+05   6247.334     37.748      0.000    2.24e+05    2.48e+05
bedrooms    6.915e+04   1849.256     37.395      0.000    6.55e+04    7.28e+04
==============================================================================
Omnibus:                     1278.700   Durbin-Watson:                   1.962
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1577.071
Skew:                           0.730   Prob(JB):                         0.00
Kurtosis:                       3.100   Cond. No.                         15.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Caption

- 1 unit change in number of bedrooms led to a decrease in house prices

3rd Model: Price vs number of floors

- Checked if number of floors could have any effect on house prices.
- It had a minimum effect on the price
- Number of floors could only account for 6.2 % accuracy
- Performed poorly compared to previous model.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.062
Model:                            OLS   Adj. R-squared:                  0.062
Method:                 Least Squares   F-statistic:                     1166.
Date:                Wed, 01 May 2024   Prob (F-statistic):          1.31e-247
Time:                        22:46:07   Log-Likelihood:            -2.3993e+05
No. Observations:               17649   AIC:                         4.799e+05
Df Residuals:                   17647   BIC:                         4.799e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3.278e+05   4220.162     77.680      0.000     3.2e+05    3.36e+05
floors      9.181e+04   2688.203     34.153      0.000    8.65e+04    9.71e+04
==============================================================================
Omnibus:                     1478.354   Durbin-Watson:                   1.976
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1879.904
Skew:                           0.796   Prob(JB):                         0.00
Kurtosis:                       3.156   Cond. No.                         6.22
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

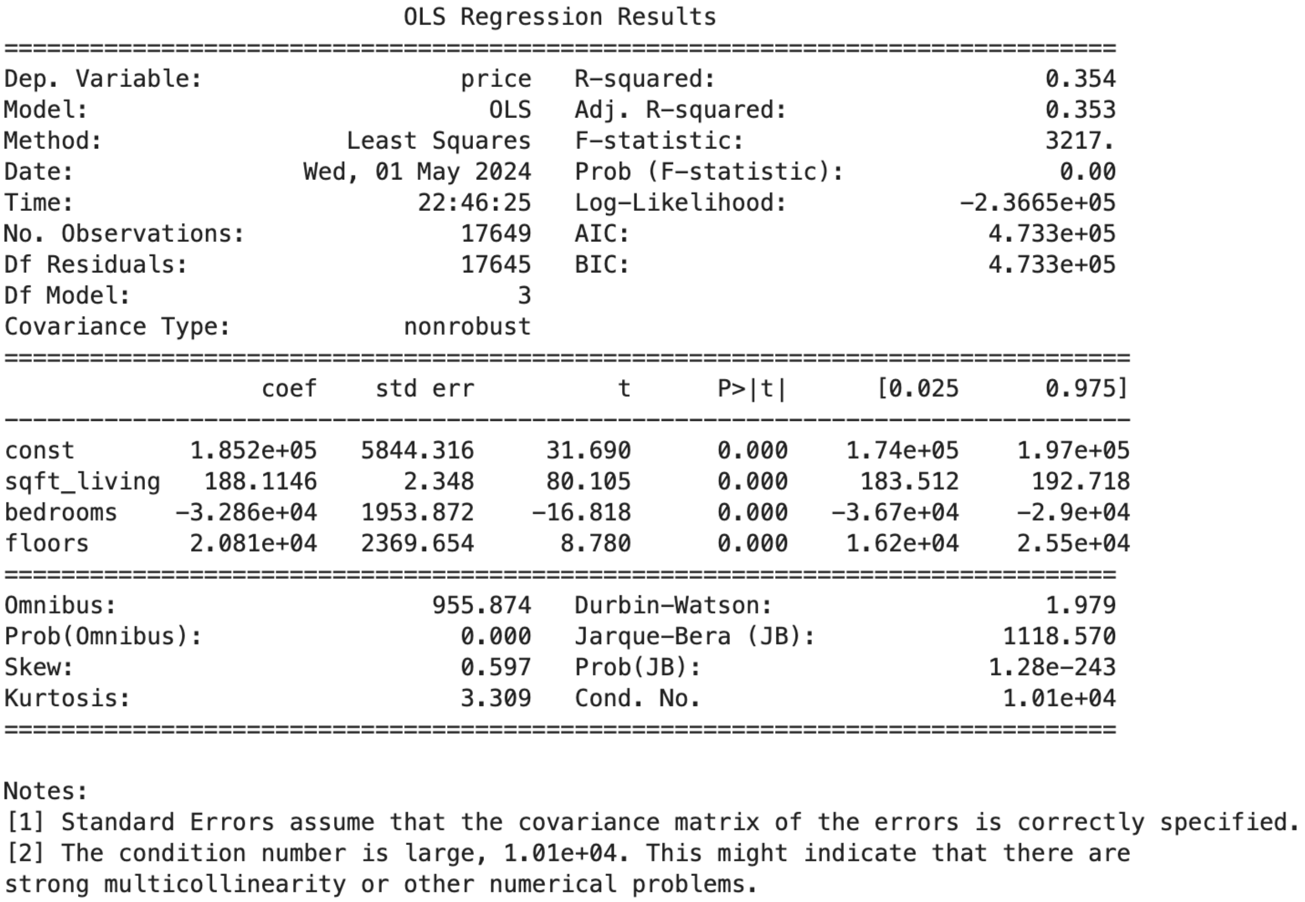
Caption

- 1 unit change in number of floors led to an increase in house prices

b) **Multiple linear regression**

1.price vs no of bedrooms, sqft_living, floors

- Combined all the feature to see if I could get a better model.
- Model price vs the other features(sqft_living, bedrooms, floors)
- Noticed I had a better model overall that predicted price with a 35.3% (adjusted r-squared) accuracy
- Also, for every one unit change in sqft_living ,then the price of the house increases
- Also, for every one unit change in the number of bedrooms, then the price of the house decreases



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.354
Model:                            OLS   Adj. R-squared:                  0.353
Method:                 Least Squares   F-statistic:                     3217.
Date:                Wed, 01 May 2024   Prob (F-statistic):               0.00
Time:                        22:46:25   Log-Likelihood:            -2.3665e+05
No. Observations:               17649   AIC:                         4.733e+05
Df Residuals:                   17645   BIC:                         4.733e+05
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const        1.852e+05   5844.316     31.690      0.000    1.74e+05    1.97e+05
sqft_living   188.1146      2.348     80.105      0.000     183.512     192.718
bedrooms    -3.286e+04   1953.872    -16.818      0.000   -3.67e+04    -2.9e+04
floors       2.081e+04   2369.654      8.780      0.000    1.62e+04    2.55e+04
==============================================================================
Omnibus:                      955.874   Durbin-Watson:                   1.979
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1118.570
Skew:                           0.597   Prob(JB):                    1.28e-243
Kurtosis:                       3.309   Cond. No.                     1.01e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.01e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Caption

- Also, for every one unit change in the number of floors, then the price of the house increases

5.0 CONCLUSION

From our models we noted the following

1.With less bedrooms or 0 bedrooms the price of a house is higher, that is the more the number of bedrooms the lower the price of that house

2.For every additional square foot of living space the price of a house increases

3.For every additional number of floors then the price of a house increases

4.After carrying out the multiple linear regression of all the variables, the number of bedrooms and number of floors seemed to have a higher impact on the price of a house as opposed to the sqft_living, ,i.e., the less the number of bedrooms the higher the price of a house and the more the number of floors the higher the price of the house.

6.0 RECOMMENDATIONS

1.Bedrooms: Considering the impact of the number of bedrooms on the house price. I would advise the homeowners to focus on properties with a lower bedroom count so that in return it will lead to an increase in price of the house.

2.Square Foot living: Pay attention to the square footage of the living space. Increasing the living space generally increases the house price.

3.Floors: Houses with multiple floors tend to have higher prices. If feasible, explore opportunities to add or emphasize multiple floors in properties to increase their price value.

NEXT STEPS

1. Refine the multiple linear regression model by incorporating additional variables that may influence house prices, they include the condition of the house and the age of the house.
2. Explore the relationship between the number of bedrooms and house prices in more detail. Investigate whether there are specific bedroom configurations that have a greater impact on house prices, e.g., master suites.
3. Validate the findings and recommendations using different datasets to ensure that the recommendations and findings are solid and broadly applicable.