# REAL ESTATE PRICING PREDICTION

# KING COUNTY

Presentation by: Group 2

# GROUP 2 MEMBERS

1. Yvonne Mwangi
2. Charles Egambi
3. Thomas Otiende
4. Frida Oyucho
5. Myles Kiganane
6. Sonia Ojay

# Agenda Style

01 **Project Overview**

02 **Business and Data Understanding**

03 **Modeling**

04 **Conclusion and Recommendations**

# Project Overview

The project focuses on analysing and understanding the dynamics of the Real Estate market in King County, Washington DC. The County's real estate landscape has witnessed significant shifts due:

- Rapid population growth,
- Changing buyer preference
- Economic fluctuations

The main objective is to address the central challenge faced by stakeholders which is to **accurately predict the prices of houses in the county.**

This is crucial for stakeholders who in this case are:

- Real estate agents
- Property owners
- Investors

The long term goal is to sell properties efficiently, maximize returns, and stay competitively in the market. The analysis is a data-driven initiative to generate insights and predictive models that would be helpful in:

- Improving pricing decisions,
- Enhancing efficiency in property sales
- Boosting the overall competitiveness within King County's Real Estate market.

To achieve this, the analysis employed **Simple and Multiple Linear Regression Modelling** as well as **Train-Test Split method** to analyze house sales in King County.

# Business Understanding

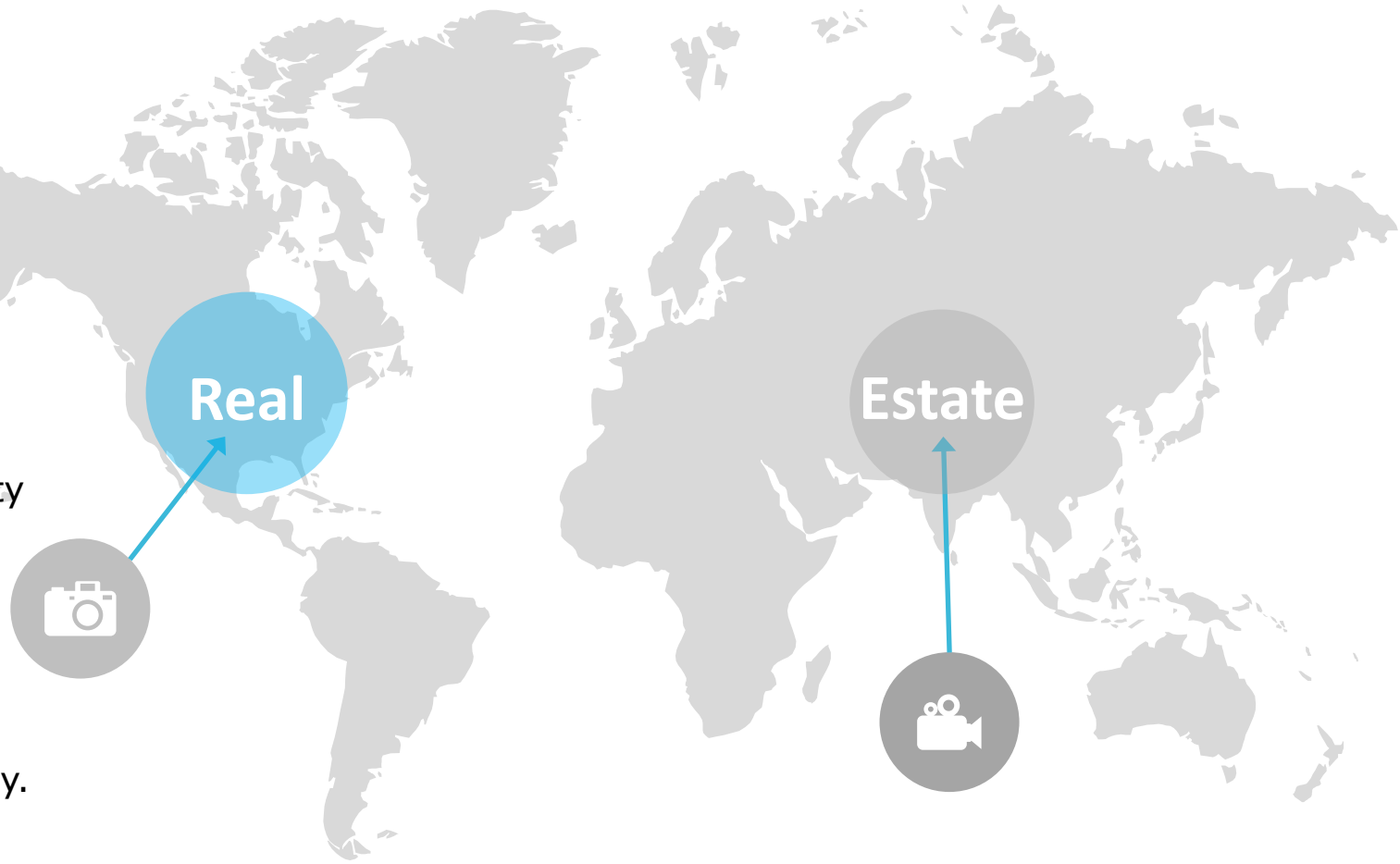

King County is located in Washington DC.

The County  faces significant housing challenges due to its rapidly increasing population.

The area is characterized by diverse neighborhoods, varying property values, and fluctuating market dynamics.

Real estate is a crucial sector in King County, influencing both economic growth and community development.

Understanding the intricacies of the housing market, including factors affecting property prices, buyer preferences, and market trends, is essential for making informed decisions and strategies in the real estate sector of King County.



# Business Problem

- The primary challenge facing stakeholders in King County's real estate market is **accurately predicting house prices.**
- The dynamic nature of the housing market, coupled with factors such as population growth, economic fluctuations, and changing buyer preferences, makes it difficult to determine optimal pricing strategies for properties.
- Real estate agents and homeowners often struggle to set competitive prices that reflect the true value of their properties and meet market demand.
- In the absence of accurate price predictions, stakeholders may encounter difficulties in selling properties efficiently, maximizing returns on investments, and maintaining competitiveness in the market.
- Addressing this business problem requires developing a robust predictive models and leveraging on data-driven insights to guide pricing decisions effectively in the County's real estate market.

# Research Questions?

**01** What are the key factors influencing house prices in King County, Washington DC?

**02** How do factors such as the number of bedrooms, bathrooms, and overall grade of the property influence house prices in King County?

**03** How accurate is the price prediction when a single feature is considered as compared to multiple housing features?

# Data Cleaning and Exploratory Data Analysis

# DATA Understanding

## King County

- For purposes of this project, the ***'kc_house_csv***' dataset was used.
- Data preparation was conducted including the identification and handling:

✓ Missing values

✓ Duplicates

✓ Outliers

✓ Data types for specific variables

- The original data set contained 21,597 houses.
- After Data preparation, a total of 16,856 houses were adopted for further analysis.
- The following Housing features were used in conducting the analysis: price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, lat, long, sqft_living15, sqft_lot15, grading.

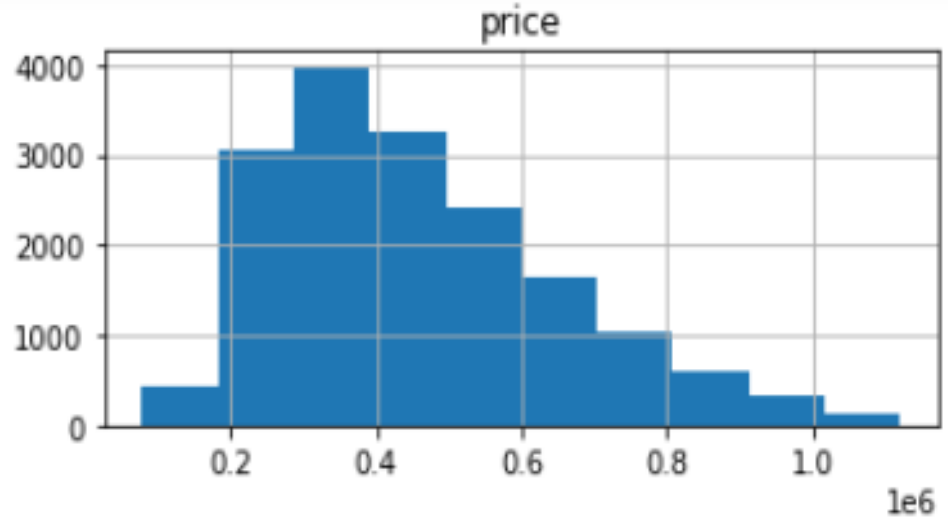
# Exploratory Data Analysis

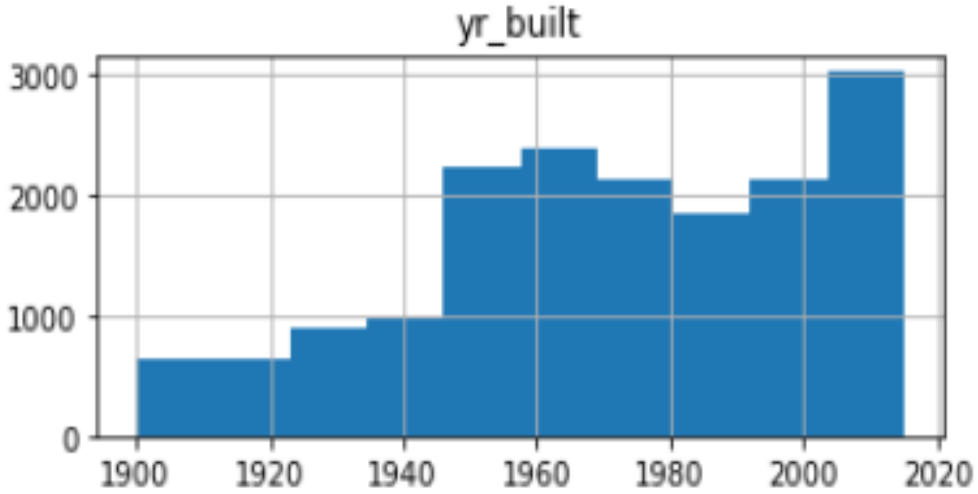
| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | sqft_above | yr_built | lat | long | sqft_living15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 | 16856.00000 |
| mean | 457700.71879 | 3.28156 | 1.99867 | 1874.11925 | 7156.29005 | 1.47209 | 1622.11171 | 1971.21440 | 47.55905 | -122.22670 | 1847.11936 |
| std | 196215.26781 | 0.78376 | 0.65870 | 671.83018 | 3437.87453 | 0.54561 | 641.49376 | 29.37420 | 0.13916 | 0.13269 | 548.65798 |
| min | 78000.00000 | 2.00000 | 0.75000 | 540.00000 | 520.00000 | 1.00000 | 480.00000 | 1900.00000 | 47.15590 | -122.50300 | 620.00000 |
| 25% | 305498.75000 | 3.00000 | 1.50000 | 1370.00000 | 4800.00000 | 1.00000 | 1150.00000 | 1952.00000 | 47.46850 | -122.33500 | 1440.00000 |
| 50% | 420000.00000 | 3.00000 | 2.00000 | 1780.00000 | 7155.00000 | 1.00000 | 1460.00000 | 1974.00000 | 47.56855 | -122.26200 | 1740.00000 |
| 75% | 575000.00000 | 4.00000 | 2.50000 | 2300.00000 | 9138.25000 | 2.00000 | 1960.25000 | 1998.00000 | 47.68070 | -122.14900 | 2180.00000 |
| max | 1120000.00000 | 5.00000 | 3.50000 | 4230.00000 | 19141.00000 | 3.50000 | 4190.00000 | 2015.00000 | 47.77760 | -121.31900 | 3660.00000 |

From the above descriptive output, its revealed that:

- The number of houses are 16,856.
- The mean price of the houses is $ 457,700.7.
- The minimum price of a house is $78,000.
- The max maximum price of a house is $1,120,000.
- The lower percentile of the price is $305,498.8.
- The upper percentile is $575,000.
- The output also reveals the description of other house features such as bedrooms, bathroom, sqft_living, sqft_lot,etc

# EDA - Univariate Analysis




## Insights

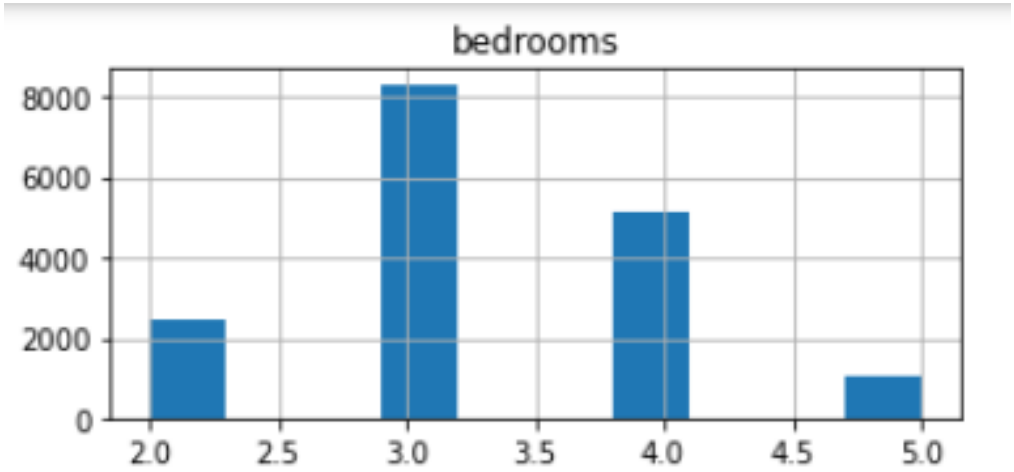
### Price

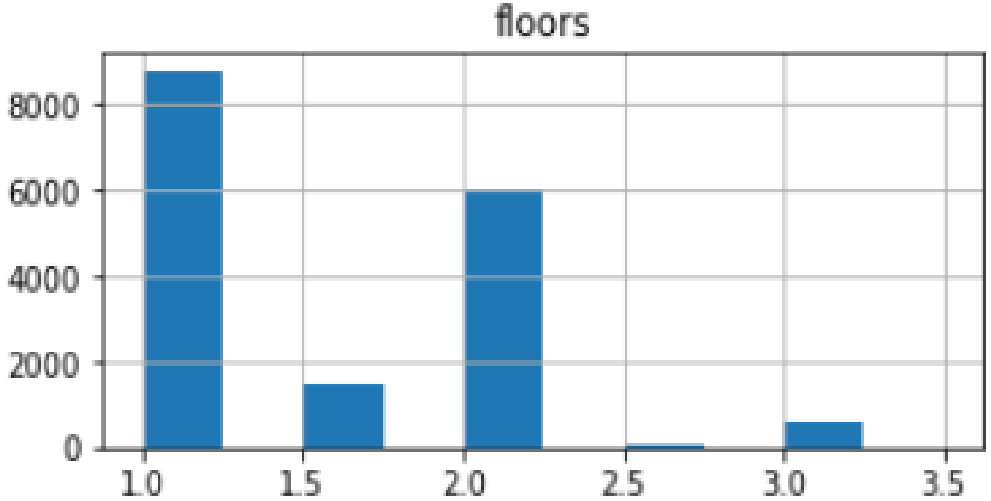
- Price is positively skewed; this means that majority of the prices of houses are clustered between $0.2M and $0.6M.
- As the price increases, the number of houses reduces.

### yr_built

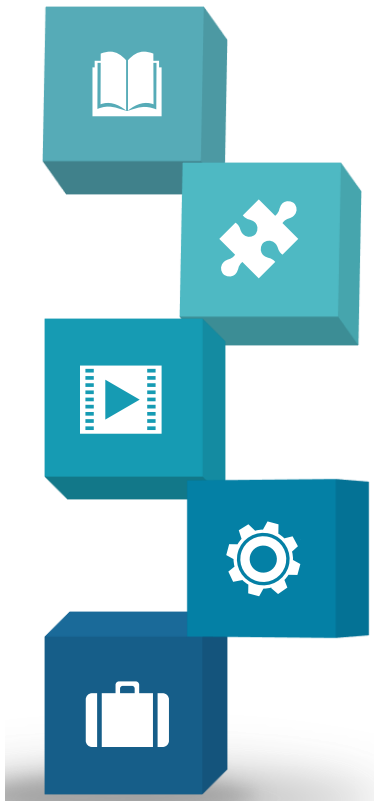
- The yr_built is negatively skewed.
- This means that majority of the houses are built are as they year progresses.
- Specifically, the highest number of houses were built around 2010.
- As time advances, the number of houses built also increases.

# EDA - Univariate Analysis




## Insights



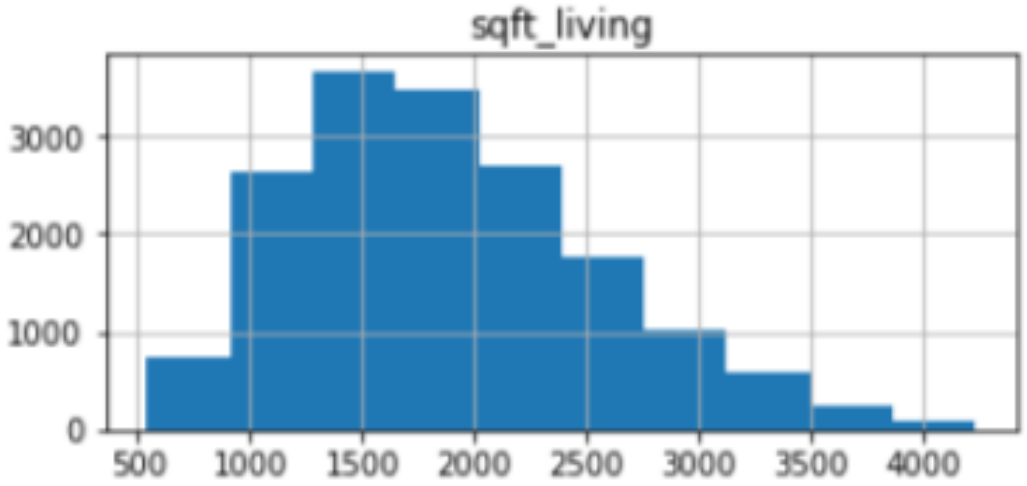
### Bedrooms

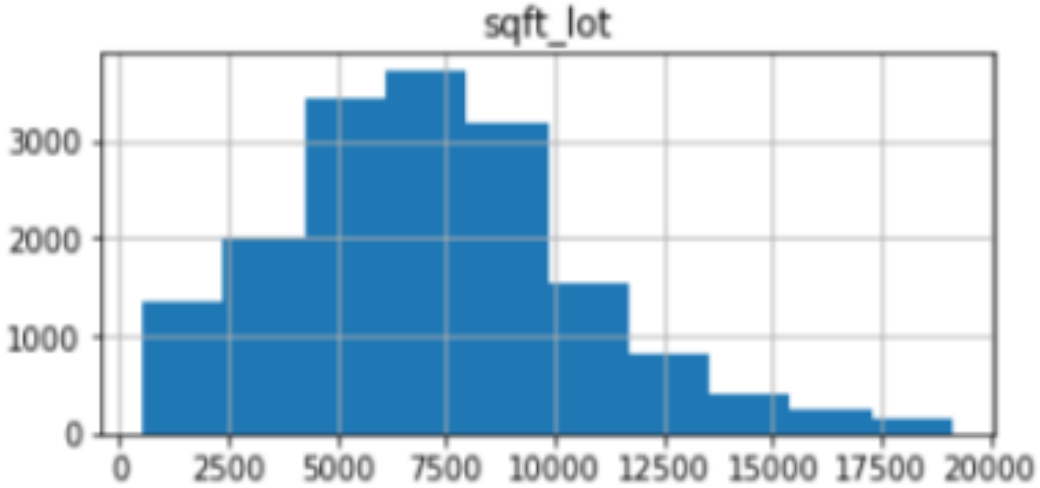
- The distribution of bedrooms is nearly normal.
- Majority of the houses found nearly at the centre of the distribution.
- Specifically, 3-bedroomed houses are the highest.
- 5 bedroomed houses are the least in number.

### Floors

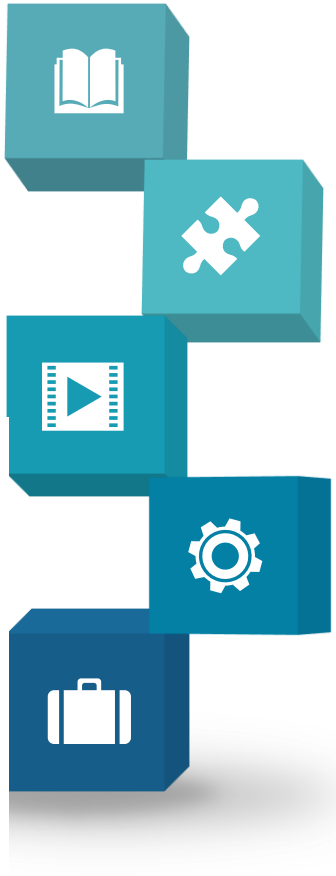
- The number of floors is positively skewed.
- Majority of the houses have fewer floors.
- Specifically, houses with 1 floor are the highest at about 9000 houses.
- This is followed by 2-floor houses at about 6000.

# EDA - Univariate Analysis




## Insights



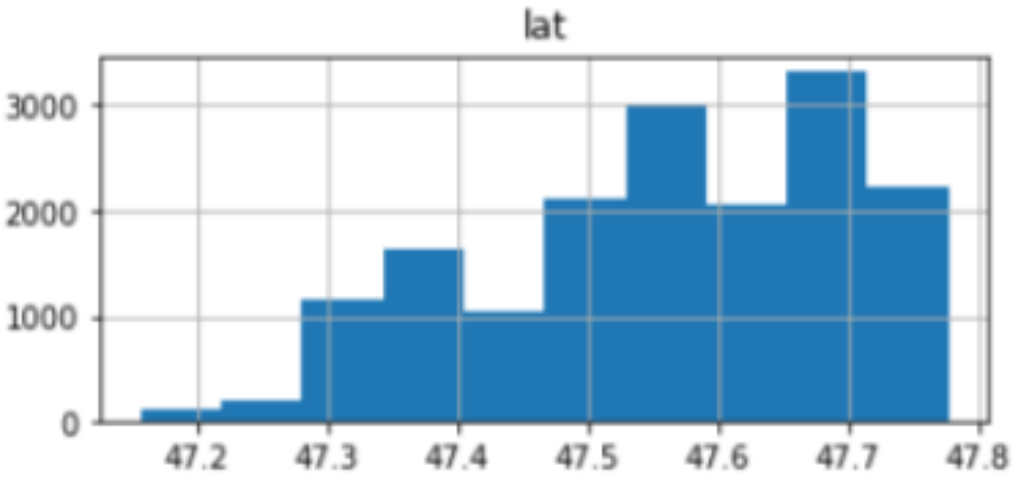
### Sqft_living

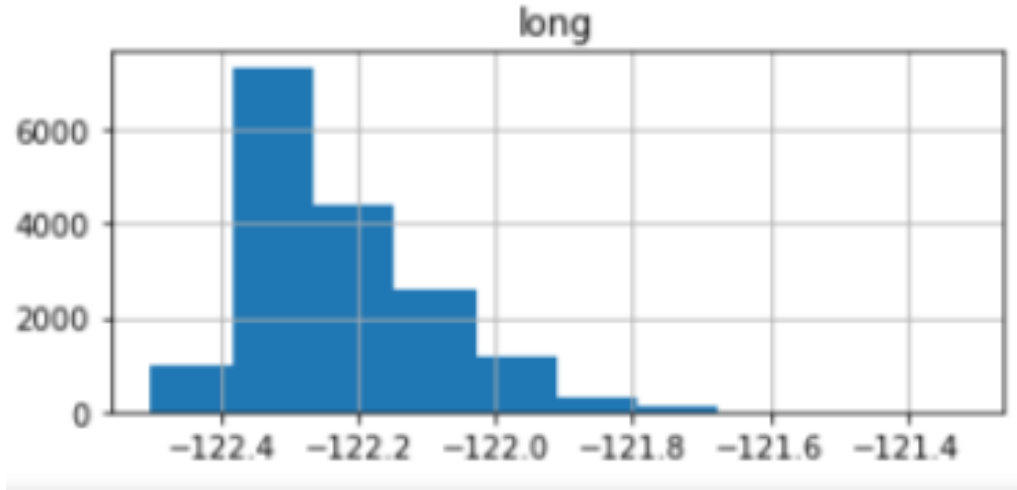
- The distribution of sqft_living is positively skewed.
- This means that majority of house square foot living are between approximately 900-2400 sqr feet.
- As the the size of the square feet increase, the number of houses reduces.

### Sqft_lot

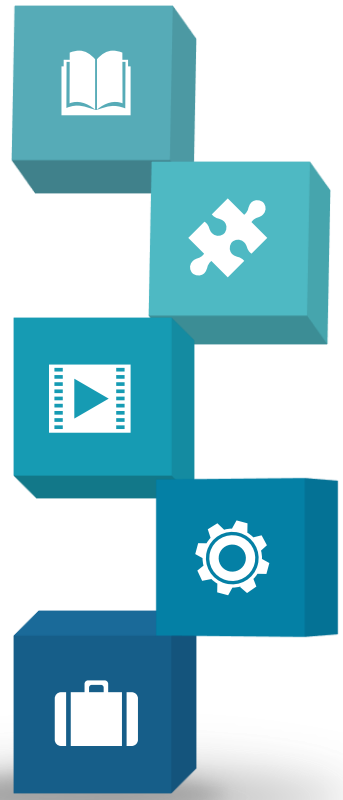
- The distribution of the sqft_lot is also positively skewed.
- Majority of the houses have sqft_lot falling between approximately 4500 – 9800 sqr feet.
- As sqft lot area increase, the number of houses reduces.

# EDA - Univariate Analysis




## Insights



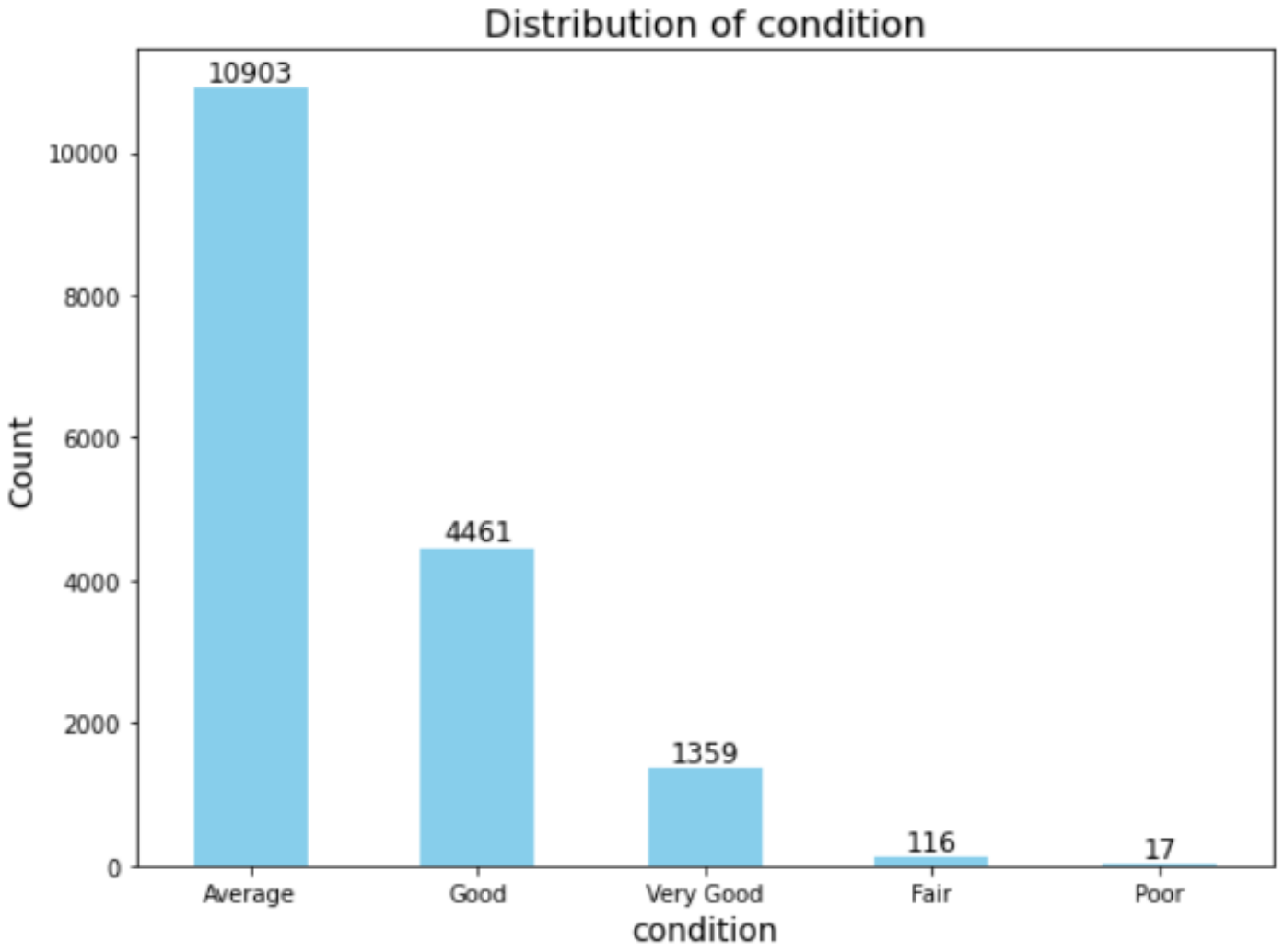
### Lat

- The distribution of the latitude is negatively skewed.
- The number of houses increase as the latitude increases.
- Majority of the houses are located between approximately latitude 47.48-47.72 degrees above sea-level.

### Long

- The distribution of longitudes are positively skewed.
- Most houses are located between -approximately -122.38 to -122.08 degrees of the Western Hemisphere.
- As the number of degrees reduce the number of houses tend to reduces.

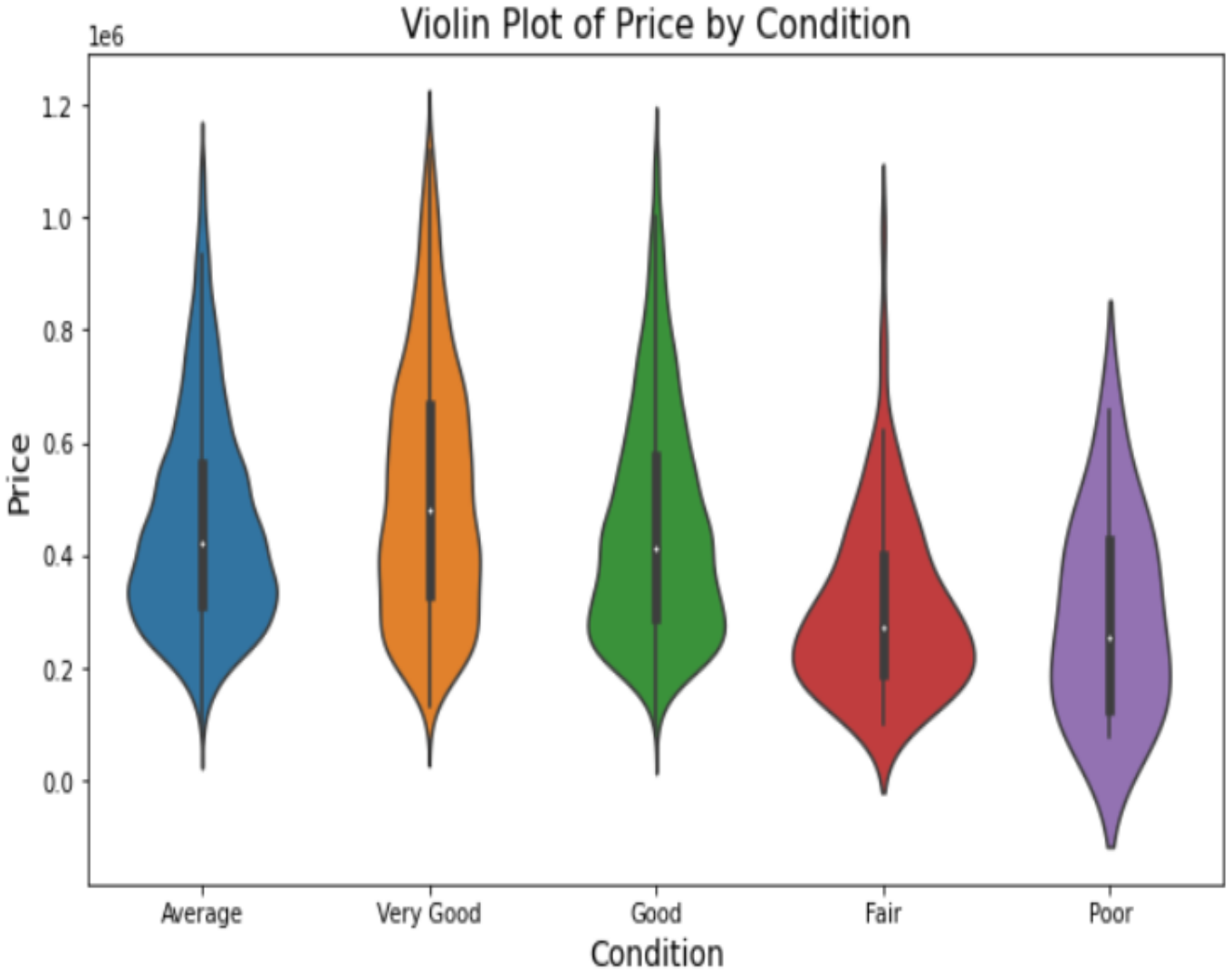
# EDA - Univariate Analysis




## Condition

- The analysis reveals that houses with average condition have the highest count at 10,903.
- Further, about 4,461 houses have a good condition while there are only 17 houses in poor condition

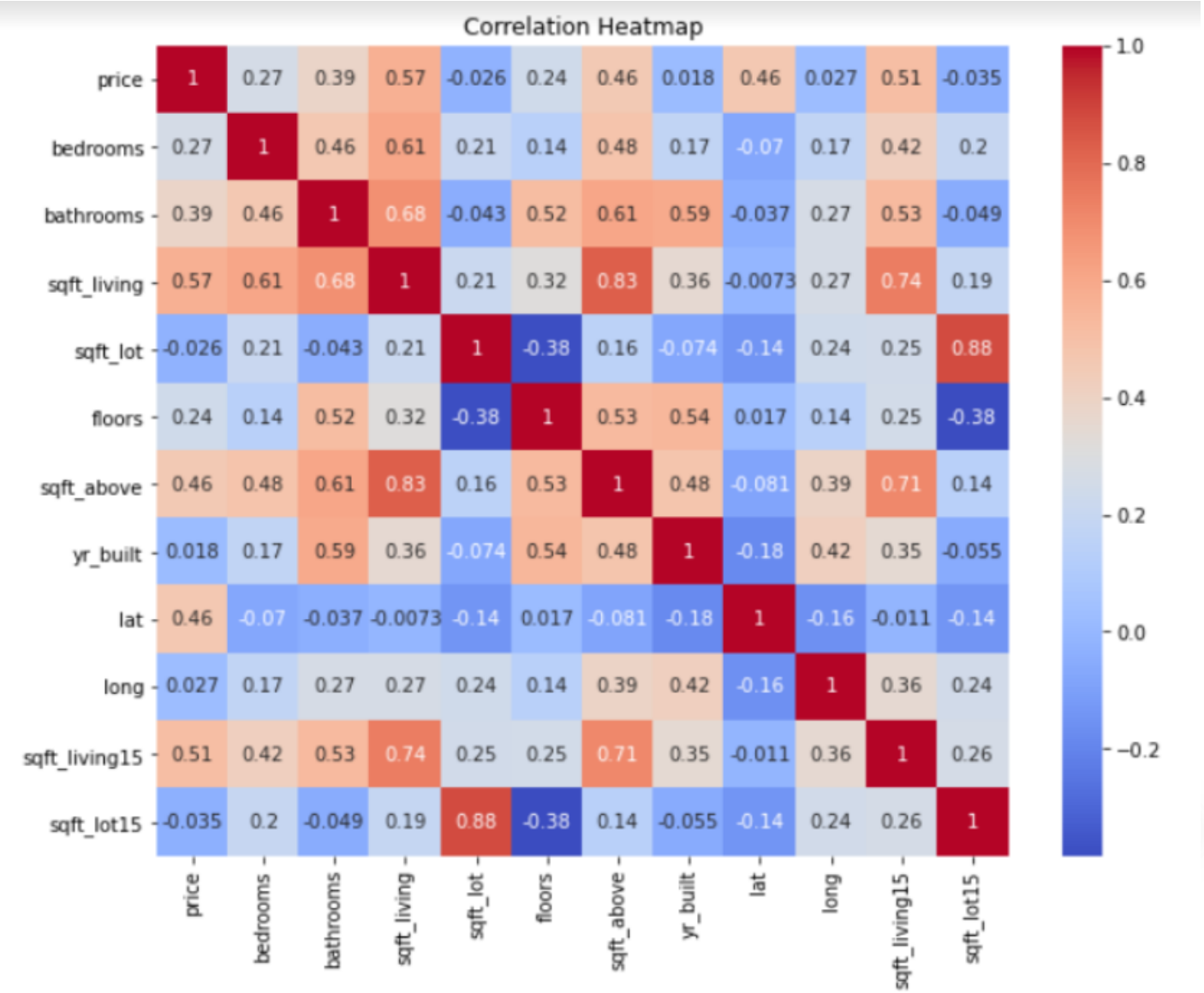
# EDA - Bivariate Analysis




## Price by Condition

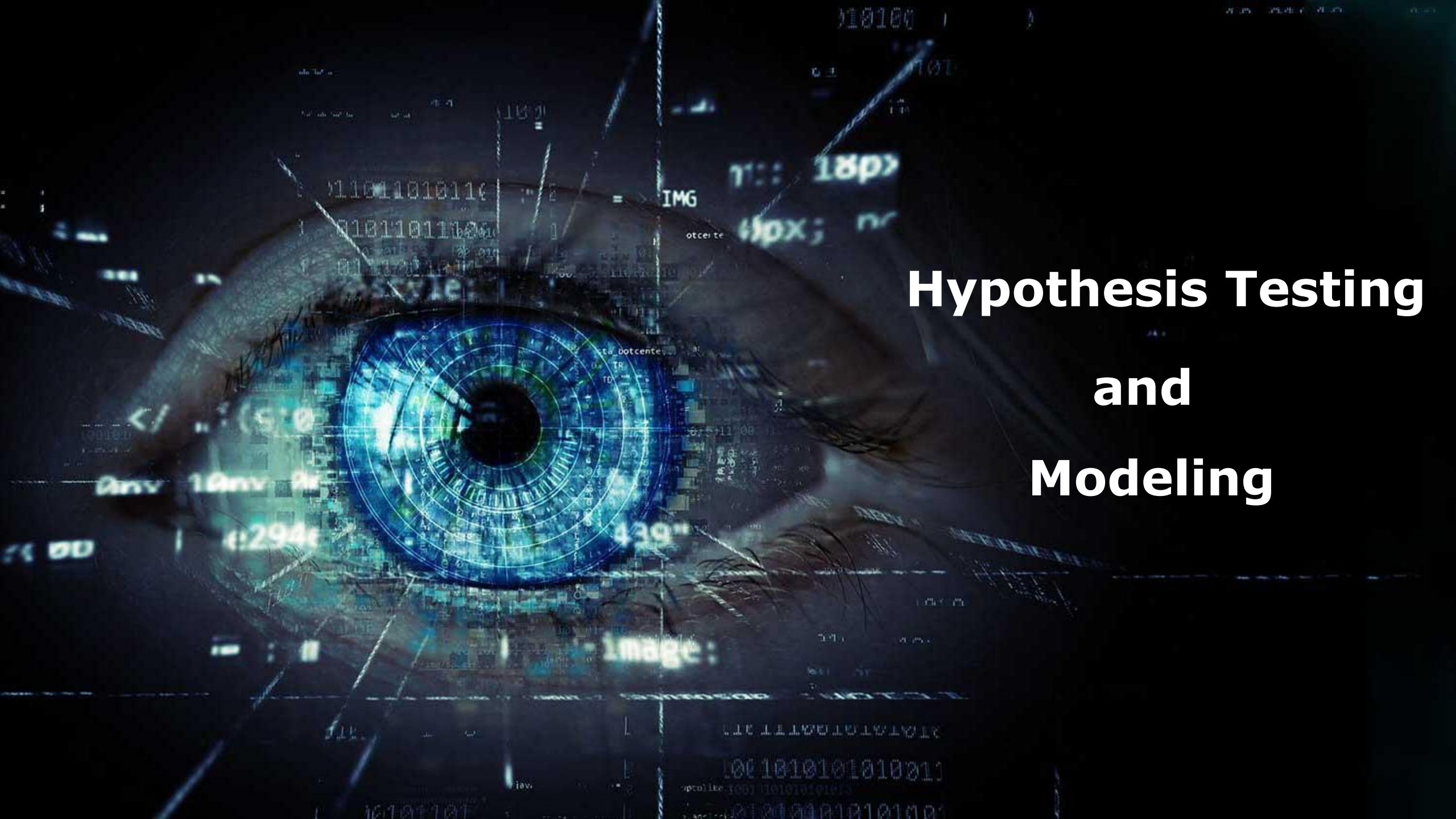
- Houses with very good condition are highly priced than houses with condition.
- This is explained by the median(indicated by the dot inside the violin) where 'very good' has the highest median.
- On the other hand 'poor' condition has the lowest median meaning that they are least priced.

# EDA - Correlation



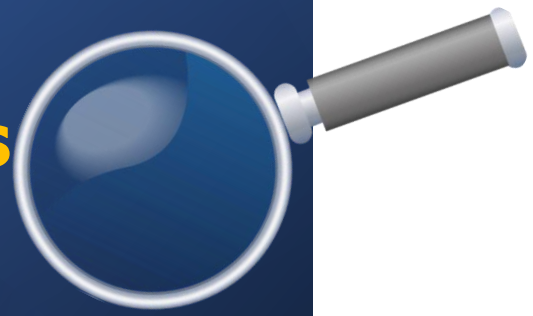

## Findings

- From the correlation matrix, **sqft_living** has the moderate positive correlation with price at **0.57.**
- This suggests a moderately positive linear relationship between the living area size (in square feet) and the price of the property.
- This is in line with the expectation of the houses with larger sqft living commanding higher prices.
- Sqft lot has the weakest negative correlation with price at -0.026.
- This implies that the size of the lot (in square footage) is not a strong determinant of the price.



# Hypothesis Testing and Modeling

# Hypothesis Testing



| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| waterfront | 464634044691.03497 | 1.00000 | 39.50206 | 0.00000 |
| view | 5045252365375.09375 | 4.00000 | 107.23378 | 0.00000 |
| condition | 5711310620672.99219 | 4.00000 | 121.39044 | 0.00000 |
| grade | 37361076591544.27344 | 7.00000 | 453.76400 | 0.00000 |
| sqft_basement | 5311068772914.99316 | 226.00000 | 1.99794 | 0.00000 |
| grading | 10625895908500.87500 | 2.00000 | 451.69393 | 0.00000 |
| bedrooms | 282515277209.50568 | 1.00000 | 24.01876 | 0.00000 |
| bathrooms | 809171913682.13672 | 1.00000 | 68.79383 | 0.00000 |
| sqft_living | 264262708738.09708 | 1.00000 | 22.46697 | 0.00000 |
| sqft_lot | 176957102365.67294 | 1.00000 | 15.04446 | 0.00011 |
| floors | 279388340678.79602 | 1.00000 | 23.75292 | 0.00000 |
| sqft_above | 10474531365.69360 | 1.00000 | 0.89052 | 0.34535 |
| yr_built | 17688915110987.56641 | 1.00000 | 1503.86860 | 0.00000 |
| lat | 78457892140789.29688 | 1.00000 | 6670.29941 | 0.00000 |
| long | 301507451228.52209 | 1.00000 | 25.63343 | 0.00000 |
| sqft_living15 | 4119115357437.15234 | 1.00000 | 350.19718 | 0.00000 |
| sqft_lot15 | 939594228177.91821 | 1.00000 | 79.88202 | 0.00000 |
| Residual | 195277279937173.96875 | 16602.00000 | nan | nan |

**Null Hypothesis ($H_0$):** There is no significant relationship between the various housing features and house prices in King County's real estate market.

**Alternate Hypothesis ($H_1$):** There is a significant relationship between the various housing features and house prices in King County's real estate market.

## Findings

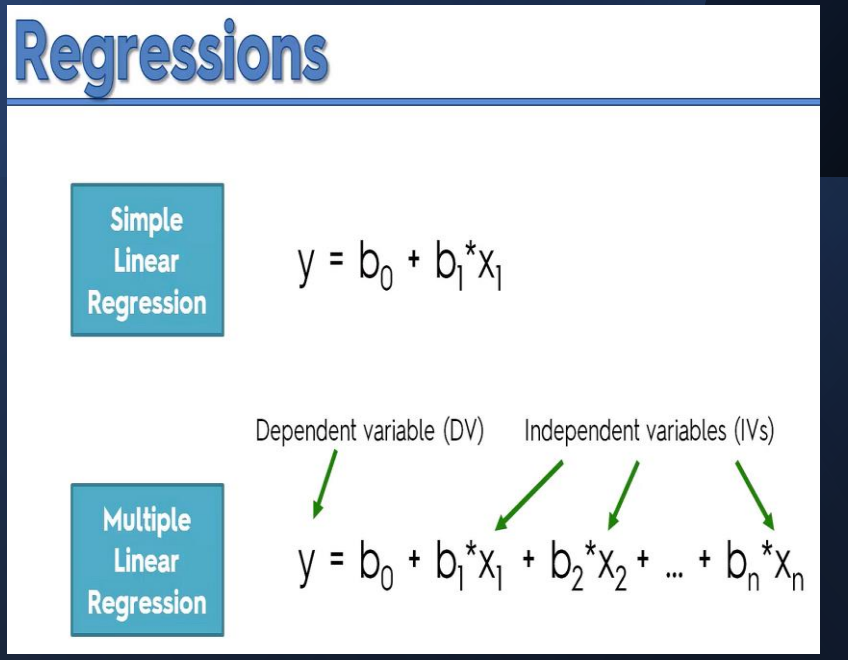
The ANOVA test reveals that collectively, the p-values are below alpha of 0.05.

**Conclusion:**

- Based on the provided ANOVA table, we reject the null hypothesis ($H_0$) that there is no significant relationship between the various housing features and house prices in King County's real estate market.

- Therefore, we accept the alternate hypothesis ($H_1$) that there is a significant relationship between the various housing features and house prices in King County's real estate market

# Data Modelling




The following models were applied for the analysis.

a) Simple linear regression model

b) Multiple linear regression model.

Price was set as the dependent variable and the house features as predictors or independent variables.

The general formula applied for both models included:

$$y=mx+c$$

Where;

Y is the dependent variable,Price

m is the slope of the gradient.

X is independent variable(s)

C is a coefficient

<u>The interpretation of the model summary was based on:</u>

1. The **r-squared** metric to identify how the model fits the data, and
2. The model **parameters** (intercept and coefficients) to infer how the model is using the housing feature(s) to predict the Price.

# Simple Linear Regression Model

$y=mx+b$

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.327 |
| Model: | OLS | Adj. R-squared: | 0.327 |
| Method: | Least Squares | F-statistic: | 8184. |
| Date: | Wed, 03 Apr 2024 | Prob (F-statistic): | 0.00 |
| Time: | 21:56:49 | Log-Likelihood: | -2.2600e+05 |
| No. Observations: | 16856 | AIC: | 4.520e+05 |
| Df Residuals: | 16854 | BIC: | 4.520e+05 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 1.448e+05 | 3674.675 | 39.395 | 0.000 | 1.38e+05 | 1.52e+05 |
| sqft_living | 166.9787 | 1.846 | 90.467 | 0.000 | 163.361 | 170.597 |

| | | | |
|---|---|---|---|
| Omnibus: | 867.752 | Durbin-Watson: | 1.972 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 1005.835 |
| Skew: | 0.588 | Prob(JB): | 3.85e-219 |
| Kurtosis: | 3.226 | Cond. No. | 5.90e+03 |

According to a correlation matrix ,**sqft_living** feature was found to have the highest correlation with **Price**, thus was used to build the simple linear regression model.

$y=mx+c$

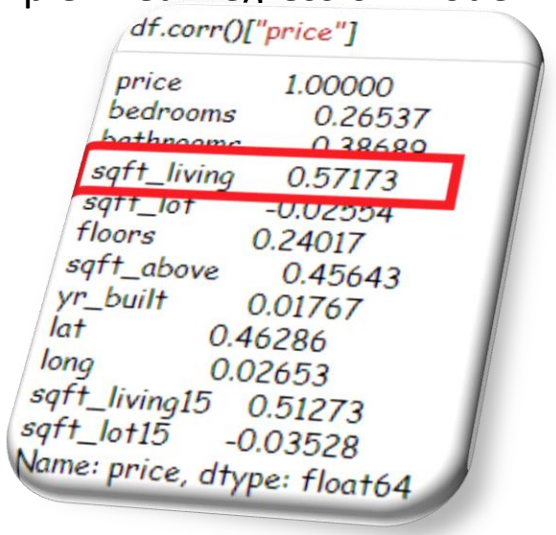
Where:

$y$ is the **Price**,

$m$ is the slope of sqft_living,

$x$ is **sqft_living**, and

c is the y-intercept (the value of $y$ when $x$ is 0).




## Results:

The **r-squared** value of the Simple Linear regression model was **32.7%**. This value indicates that approximately 32.7% of the variability in the dependent variable (**Price**) is accounted for by the independent variable(**sqft_living**) in our model. Meaning the remaining **67.3%** of the variability is not accounted for.

The r-squared value of 32.7% suggests a **moderate** level of explanatory power of our regression model. This result highlights the need for further research to explore additional variables and factors that may influence the dependent variable (**Price**) and to improve the predictive accuracy and explanatory power of the model.

**Is the model statistically significant at $\alpha$=0.05?**

The p-value obtained is 0.000 which is less than the $\alpha$=0.05 thereby concluding that the model is statistically significant.

Further results shows that a y-intercept of $144,762.76 was achieved, meaning for every increase of 1 square foot living area, the price increases by $166.98

# Multiple Linear Regression Model

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | price | **R-squared:** | 0.691 |
| **Model:** | OLS | **Adj. R-squared:** | 0.690 |
| **Method:** | Least Squares | **F-statistic:** | 1393. |
| **Date:** | Sat, 06 Apr 2024 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 17:42:55 | **Log-Likelihood:** | -2.1945e+05 |
| **No. Observations:** | 16856 | **AIC:** | 4.389e+05 |
| **Df Residuals:** | 16828 | **BIC:** | 4.392e+05 |
| **Df Model:** | 27 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -8.354e+06 | 5.13e+05 | -16.279 | 0.000 | -9.36e+06 | -7.35e+06 |
| **X[0]** | -7509.5608 | 1419.856 | -5.289 | 0.000 | -1.03e+04 | -4726.493 |
| **X[1]** | 2.003e+04 | 2238.944 | 8.945 | 0.000 | 1.56e+04 | 2.44e+04 |
| **X[2]** | 81.1639 | 3.197 | 25.388 | 0.000 | 74.898 | 87.430 |
| **X[3]** | -2.0159 | 0.523 | -3.853 | 0.000 | -3.041 | -0.990 |
| **X[4]** | 1.179e+04 | 2555.412 | 4.614 | 0.000 | 6782.148 | 1.68e+04 |
| **X[5]** | 2.7217 | 3.200 | 0.851 | 0.395 | -3.550 | 8.993 |
| **X[6]** | -1808.5333 | 46.434 | -38.949 | 0.000 | -1899.548 | -1717.519 |
| **X[7]** | 5.32e+05 | 6475.032 | 82.159 | 0.000 | 5.19e+05 | 5.45e+05 |
| **X[8]** | 3.607e+04 | 7823.334 | 4.610 | 0.000 | 2.07e+04 | 5.14e+04 |
| **X[9]** | 50.1800 | 2.596 | 19.329 | 0.000 | 45.091 | 55.269 |

In the case Multiple Linear Regression Model, all 11 features were used to identify their level of significance on the Price.

The formula included:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta p x p + \varepsilon$$

Where;

Y – is the dependent variable, the price

$B_0$ is the y-intercept

$\beta_1 x_{1\ldots}$ Predictors (housing features)

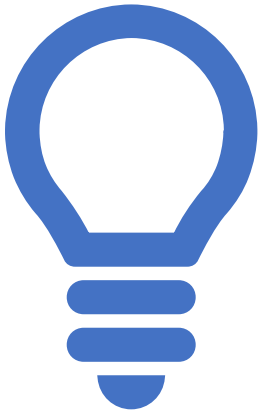
Results:

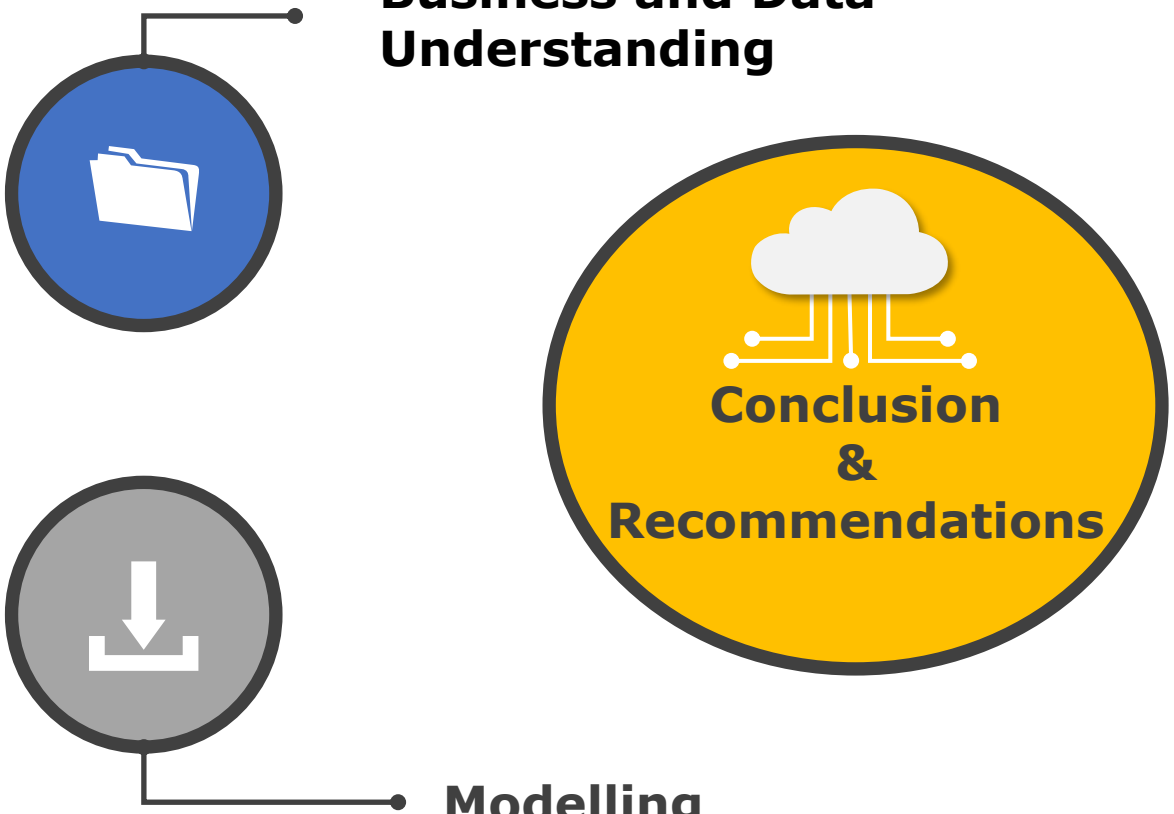
The Multiple Regression model had an r-squared value of **69.1%** demonstrating substantially improved explanatory power compared to the simple linear regression model which had an r-squared value of **32.7%**.

The inclusion of additional independent variables in the multiple regression model significantly enhanced the model's ability to explain the variability in the price.

The r-squared values and the comparison between the two models highlight the importance of considering multiple factors and variables in regression analysis to develop a more comprehensive and accurate understanding of the relationship between the independent variables and the price.

# Conclusion






**Key Factors Influencing House Prices in King County**
According to the analysis, sqft living had a stronger relationship with price as compared to the other housing features with a positive moderate correlation of 0.57. This was followed by sqft living 15 and sqft above.

**How factors such as bedrooms, bathrooms and the overall grade influence house prices in King County.**
According to the hypothesis tested, features such as a bedrooms, bathrooms and the overall grade are statistically significant in relation to price with a p-value of  0.000 which is less than the alpha value of 0.05.

**How accurate is the price prediction when a single feature is considered as compared to multiple housing features?**
According to the analysis, the r-squared of the simple linear regression was 32.7%, on the other hand the inclusion of additional independent variables in the multiple regression model significantly enhanced the model's ability to explain the variability in the price with an improved r-squared of 69.1%. This shows that the more the more variables, the accurate the price prediction.

# Recommendations

**Key Factors Influencing House Prices in King County**

For home sellers in King County, it is advisable to optimize the living area square footage through renovations or extensions, emphasize features like the square foot of the living space for neighbouring houses and overall square footage apart from the basement of property listings, and market all property features effectively to appeal to a wider range of buyers.

For home buyers, prioritizing properties with larger living areas, evaluating the value of additional square footage metrics, and conducting a comprehensive assessment of all property aspects, including location and amenities is recommended. Real estate professionals should educate clients about the significance of living area square footage and additional metrics in determining property prices.

**How  factors such as the number of bedrooms, bathrooms, and overall grade of the property influence house prices.**

Based on the tested hypothesis which found that features like bedrooms, bathrooms, and overall grade to be statistically significant in relation to house prices in King County (with a p-value of 0.000, less than the alpha value of 0.05), stakeholders are recommended to integrate these significant features into their pricing strategies and marketing campaigns for properties in the area. This includes emphasizing the value of these features in property listings, creating tailored marketing campaigns to attract potential buyers, and educating real estate agents and homebuyers about their impact on house prices. Additionally, it is crucial to continuously monitor and analyze market trends in King County to adjust pricing strategies and marketing efforts accordingly, based on the significance of these identified features. By implementing these recommendations, stakeholders can optimize their pricing strategies, enhance marketing efforts, and make informed decisions to maximize the value of properties in the King County real estate market.

**How accurate is the price prediction when a single feature is considered as compared to multiple housing features?**

Based on the analysis comparing the accuracy of price prediction using a single feature versus multiple housing features, which resulted in an R-squared of 32.7% for the simple linear regression and an improved R-squared of 69.1% for the multiple regression model, stakeholders are advised to employ a data driven approach by sourcing data expertise that would help in conducting relevant modeling to bring out insights and accurate price prediction to guide construction and the general real estate business. Implementing these recommendations, stakeholders can optimize their pricing strategies, improve decision-making processes, and align their strategies more effectively with the factors influencing house prices in King County.


Thank you!