Project Proposal

Group Name

INSIGHT SQUAD

Groupmates Name

Kalappa, Sharath Kumar
Sahu, Nikita

The real estate market is a dynamic and influential sector impacting millions of individuals. Housing decisions play a crucial role in people's lives, shaping their overall well-being. Understanding the complexities of the real estate market is essential for making informed decisions and navigating the challenges associated with buying or selling a property. In this context, the dataset offers a valuable opportunity to explore and analyze various factors influencing housing prices, providing insights that can benefit both homeowners and industry professionals.

The specific problem addressed in this project revolves around understanding the determinants of housing prices within the dataset. The dataset provides a rich collection of variables such as bedrooms, bathrooms, and square footage, offering a comprehensive view of the features that contribute to the value of a house and the prices of houses are influenced by these factors.Identifying these factors and their respective impacts on housing prices is our main objective.

This project aims to investigate the determinants of housing prices using regression analysis techniques. By analyzing the dataset, we seek to develop predictive models that can accurately estimate housing prices based on various features. Through regression analysis, we can identify the key factors driving housing prices and gain insights into how different variables interact to influence property values.

To facilitate this investigation, we will utilize the housing dataset, which contains detailed information on housing attributes such as number of bedrooms, bathrooms, square footage, location, and other relevant features. This dataset, sourced from Kaggle.com , provides a comprehensive and diverse set of variables for analyzing housing prices and understanding the underlying factors affecting property values in the region.

Analysis:

The initial phase of our analysis involves conducting Exploratory Data Analysis on the dataset where we explored the distribution, summary statistics, and relationships between various variables in the dataset. This process helped us to gain a deeper understanding of the data and identify any patterns or trends that might inform our subsequent analyses.

summary(house):

The prices of houses vary widely, with a minimum value of $75,000 and a maximum of $7,70,000. The average price is $5,40,008 and the median is $4,50,000 , suggesting a right-skewed distribution.Another take away from the summary statistics is year of renovation, the average year of renovation is 84.4, suggesting that many houses have not been renovated. Most houses have 1 to 2 floors with average being approximately 1.494 and a very small percentage of houses about 0.75% have a waterfront view. The majority of houses have 3 bedrooms, and the average of bedrooms is approximately 3.37.

```
##       price             bedrooms        bathrooms       sqft_living
##  Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290
##  1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427
##  Median : 450000   Median : 3.000   Median :2.250   Median : 1910
##  Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080
##  3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
##  Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540
##     sqft_lot           floors        waterfront           view
##  Min.   :     520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:    5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :    7618   Median :1.500   Median :0.000000   Median :0.0000
##  Mean   :   15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343
##  3rd Qu.:   10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000
##    condition         grade           sqft_above   sqft_basement
##  Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0
##  1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0
##  Median :3.000   Median : 7.000   Median :1560   Median :   0.0
##  Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5
##  3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0
##  Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0
##     yr_built     yr_renovated       zipcode           lat
##  Min.   :1900   Min.   :   0.0   Min.   :98001   Min.   :47.16
##  1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033   1st Qu.:47.47
##  Median :1975   Median :   0.0   Median :98065   Median :47.57
##  Mean   :1971   Mean   :  84.4   Mean   :98078   Mean   :47.56
##  3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118   3rd Qu.:47.68
##  Max.   :2015   Max.   :2015.0   Max.   :98199   Max.   :47.78
##       long         sqft_living15    sqft_lot15
##  Min.   :-122.5   Min.   : 399   Min.   :   651
##  1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100
##  Median :-122.2   Median :1840   Median :  7620
##  Mean   :-122.2   Mean   :1987   Mean   : 12768
##  3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083
##  Max.   :-121.3   Max.   :6210   Max.   :871200
```

Correlation:

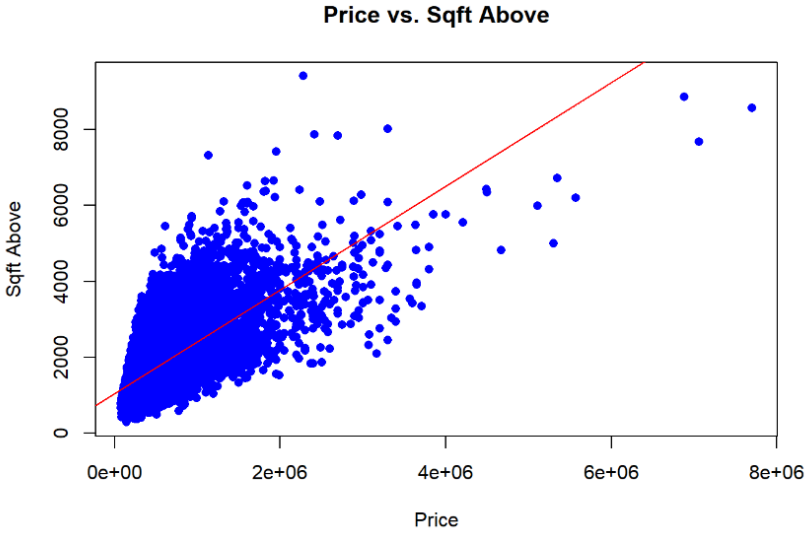
```
> cor(house$sqft_living,house$sqft_above)
[1] 0.8765966
>
> cor(house$sqft_living,house$bathrooms)
[1] 0.7546653
>
> cor(house$sqft_living,house$price)
[1] 0.7020351
```

Positive correlation: In the above correlation analysis, we can observe that the square footage of the living area (sqft_living) exhibits a strong positive relationship with the number of bathrooms, the size of the above ground living area (sqft_above) and the price of the house. This suggests that as the square footage of the living area increases, there tends to be a corresponding increase in the number of bathrooms, the size of the above ground living area, and ultimately, the price of the house.

```
> cor(house$condition, house$floors)
[1] -0.2637679
>
> cor(house$long, house$zipcode)
[1] -0.5640716
```

Negative correlation: There is a negative correlation between the overall condition of the house and number of floors. Similarly, the negative relationship between the longitude and zip code suggests that houses located farther west(lower longitude values) tend to have higher zip codes, indicating a geographical trend where lower numerical zip codes are found in areas with higher longitude values.

Scatter plot:



The scatter plots of “price” against “sqft_above” and “sqft_living” reveal a positive, strong and linear relationship between the variables. As “price” increases, there is a noticeable upward trend in both variables, indicating that the size of the above ground living area (sqft_above) and the overall living space (sqft_living) positively influences the “price” of the house, with a linear pattern in their association.

Pivot table:

pivot_table <- house %>%group_by(bedrooms,bathrooms) %>%summarise(count = n())%>%spread(bedrooms, count, fill = 0)

	Number of bedrooms											
Number of bathrooms	0	1	2	3	4	5	6	7	8	9	10	11	33
0	7	3											
0.5		1	2		1								
0.75	1	27	26	16	2								
1	1	138	1558	1780	325	43	6	1					
1.25		2	3	4									
1.5	1	12	294	829	254	48	6	2					
1.75		4	304	1870	719	134	16						1
2		6	216	1048	525	110	24				1		
2.25		4	118	1082	709	116	15	3					
2.5	3	2	197	2357	2502	287	29	2	1				
2.75			20	275	639	214	31	3	3				
3			13	197	326	163	45	3	2	2	1	1	
3.25			8	184	254	129	12	1	1				
3.5			1	143	395	169	17	5	1				
3.75				17	78	44	13	2	1				
4				11	58	48	11	5	2	1			
4.25				6	38	25	8	2					
4.5				5	32	35	23	3		2			
4.75					7	11	3	2					
5					7	7	6		1				
5.25					5	4	3				1		
5.5					5	4		1					
5.75					1	2		1					
6						4	1		1				
6.25						2							
6.5						1	1						
6.75						1		1					
7.5										1			
7.75							1						
8							1	1					

As we can see in the above pivot table, properties with 3 bedrooms and 2.5 bathrooms have the highest count which is 2357. Also, the majority of properties have 3 bedrooms and 2 bathrooms, followed by 4 bedrooms and 2 bathrooms. As the number of bedrooms and bathrooms increases, the count generally decreases, indicating that properties with more bedrooms or bathrooms are less common.

Our primary goals include identifying the most significant predictors of housing prices and understanding how different factors contribute to variations in property values. Specifically, we aim to investigate the impact of location, size, condition, and other features on housing prices. By addressing these research questions, we aim to provide valuable insights for homeowners, real estate professionals, and policymakers.

Team Member roles:

We have 2 active members in the team and both members' efforts are crucial for achieving project success.

Sharath will lead the visualization and model building, ensuring alignment with the project goals. He will focus on logistic regression and oversee the analysis's progress. Sharath will play a crucial role in drafting the final report, synthesizing the findings, and presenting them in a clear and concise manner.

Nikita will specialize in exploratory data analysis, building predictive models to enhance accuracy. She will employ advanced techniques to optimize model performance.Nikita's expertise will be instrumental in fine-tuning the models and optimizing their performance to achieve the highest level of accuracy possible.

Conclusion:
In conclusion, the analysis of the housing dataset reveals significant insights into the determinants of housing prices. The wide price variation, ranging from $75,000 to $770,000, suggests diverse pricing dynamics within the market.
The data also highlights the prevalence of non-renovated properties, with the average year of renovation being 84.4, indicating potential implications for pricing based on property condition. Additionally, the distribution of the number of floors, with most houses having 1 to 2 floors, and the scarcity of waterfront views, present notable factors influencing housing prices. Moreover, the predominance of 3-bedroom houses underscores the importance of this feature in pricing considerations. The scatter plots depicting the relationship between "price" and variables such as "sqft_above" and "sqft_living" demonstrated a strong, positive, and linear association. As the "price" of the house increased, there was a noticeable upward trend in both the size of the above ground living area (sqft_above) and the overall living space (sqft_living), suggesting that larger living spaces contribute to higher house prices in a linear fashion. These findings underscore the importance of property size in influencing housing prices and highlight the significance of square footage as a crucial factor for prospective buyers and sellers to consider when evaluating real estate investments.Overall, these findings underscore the complexity of housing price determination and provide valuable insights for stakeholders navigating the real estate market. Further research and modeling could enhance our understanding of these dynamics and inform more informed decision-making processes for buyers and sellers alike.