SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

# Samsung Innovation Campus

## Artificial Intelligence Course

# King County House Prices  Prediction Model

**Team Members:**
**Mina Sameh Mahrous**
**Kerolos Monier**
**Galal Mohammed**

**Supervised by:**
**Dr.  Doaa Mahmoud**
**Eng/ Abdelrahman Sadory**

## Objectives

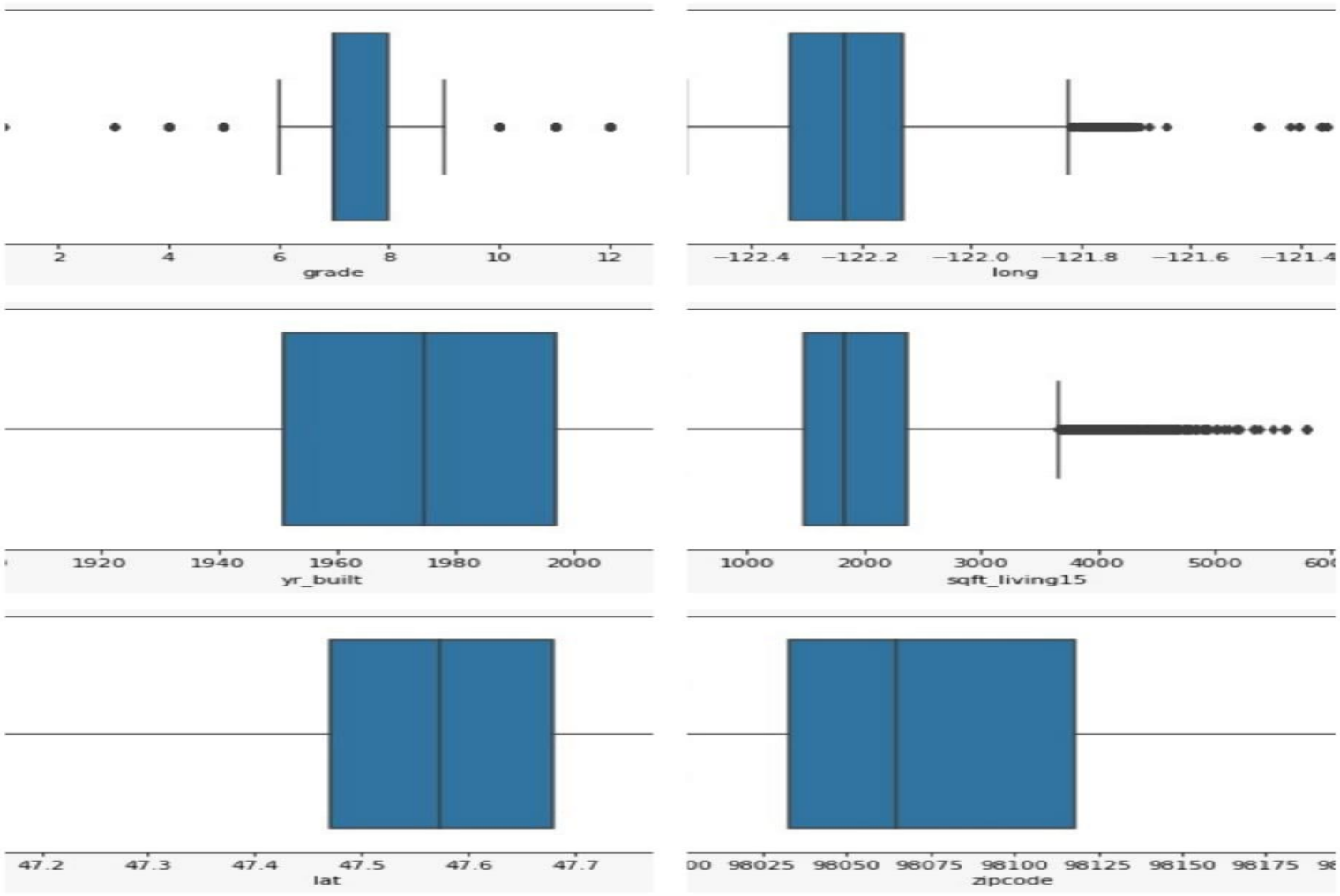
I. Overview of Data
II. Data pre-processing
III. Data visualization and pattern discovery
IV. Predictive Modeling
V. Model Implementation
VI. Plan for future upgrades

# I. Overview of Data

| Variable | Description |
|---|---|
| **Id** | Unique ID for each home sold |
| **Date** | Date of the home sale |
| **Price** | Price of each home sold |
| **Bedrooms** | Number of bedrooms |
| **Bathrooms** | Number of bathrooms, where .5 accounts for a room with a toilet but no shower |
| **Sqft_living** | Square footage of the apartments interior living space |
| **Sqft_lot** | Square footage of the land space |
| **Floors** | Number of floors |
| **Waterfront** | A dummy variable for whether the apartment was overlooking the waterfront or not |
| **View** | An index from 0 to 4 of how good the view of the property was |
| **Condition** | An index from 1 to 5 on the condition of the apartment, |
| **Grade** | An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design |
| **Sqft_above** | The square footage of the interior housing space that is above ground level |
| **Sqft_basement** | The square footage of the interior housing space that is below ground level |
| **Yr_built** | The year the house was initially built |
| **Yr_renovated** | The year of the house's last renovation |
| **Zipcode** | What zipcode area the house is in |
| **Lat** | Lattitude |
| **Long** | Longitude |
| **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 |

**Outlier Detection:** Outliers were detected and analyzed using the Outlier Boxplots.

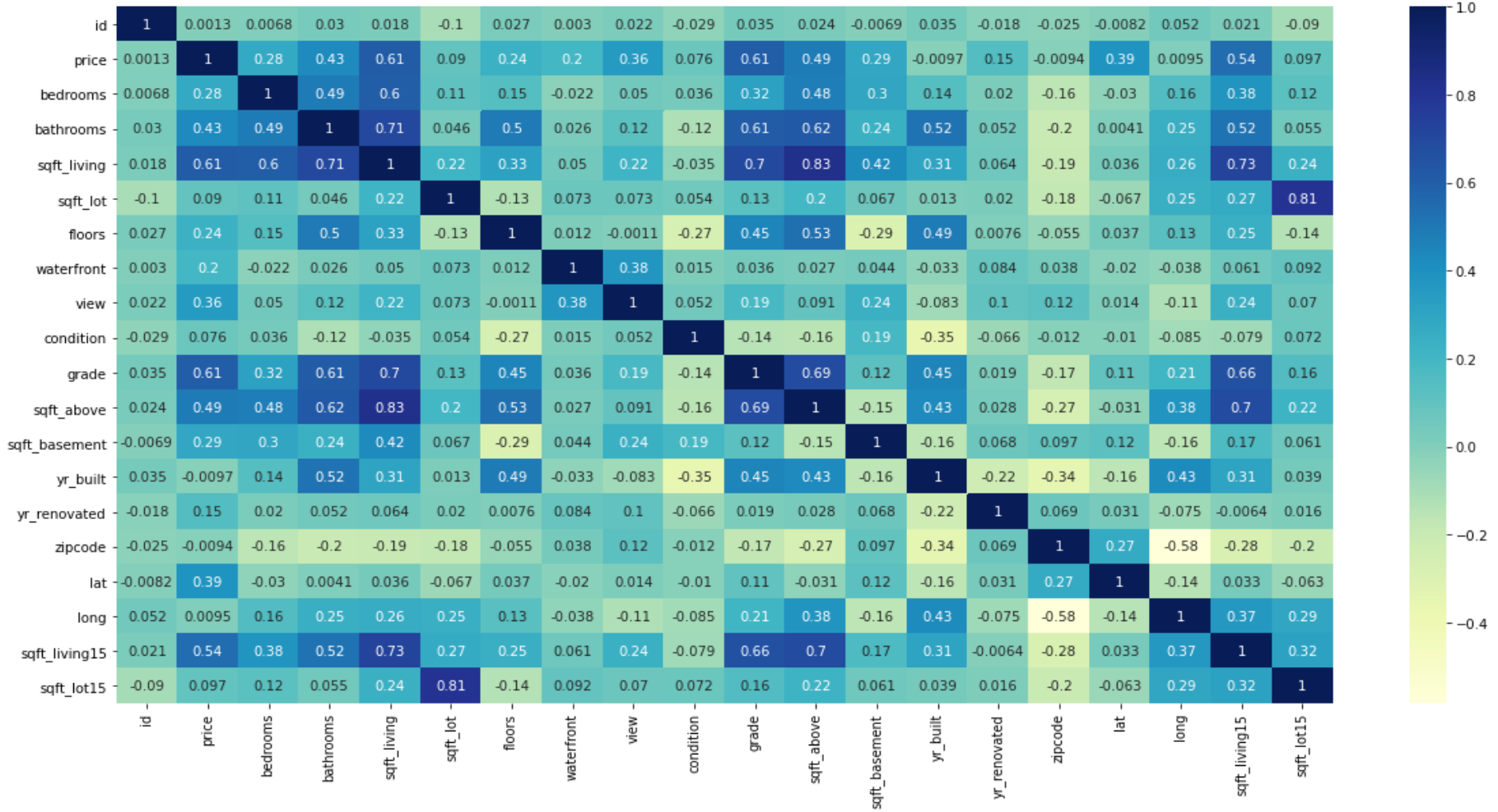
**Missing Values Detection:** Missing data pattern was used to identify the missing data in the dataset. From the table below it can be observed that the data does not consist of any missing data for any of the variables.


2
4
6
8
10
12
grade
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
long
1920
1940
1960
1980
2000
yr_built
1000
2000
3000
4000
5000
sqft_living15
47.2
47.3
47.4
47.5
47.6
47.7
lat
98025
98050
98075
98100
98125
98150
98175
zipcode

The objective of data visualization and pattern discovery was to reveal relationships between the house features and the response variable, price. We wanted to identify house features that affect price variable and could be potential predictors. Through visualization, we gathered the following information about the data.

**Correlation Table:** The below correlation table provides a summary of correlation between the continuous variables in the data. The objective behind analyzing correlation between the continuous variables in the data was to identify variables that have significant linear relationship with price and those who don't. Further, the table helps to identify relationship between potential predictors. If two predictors are highly correlated with each

other they may explain the same variation in the price variable, leading to over fitting.

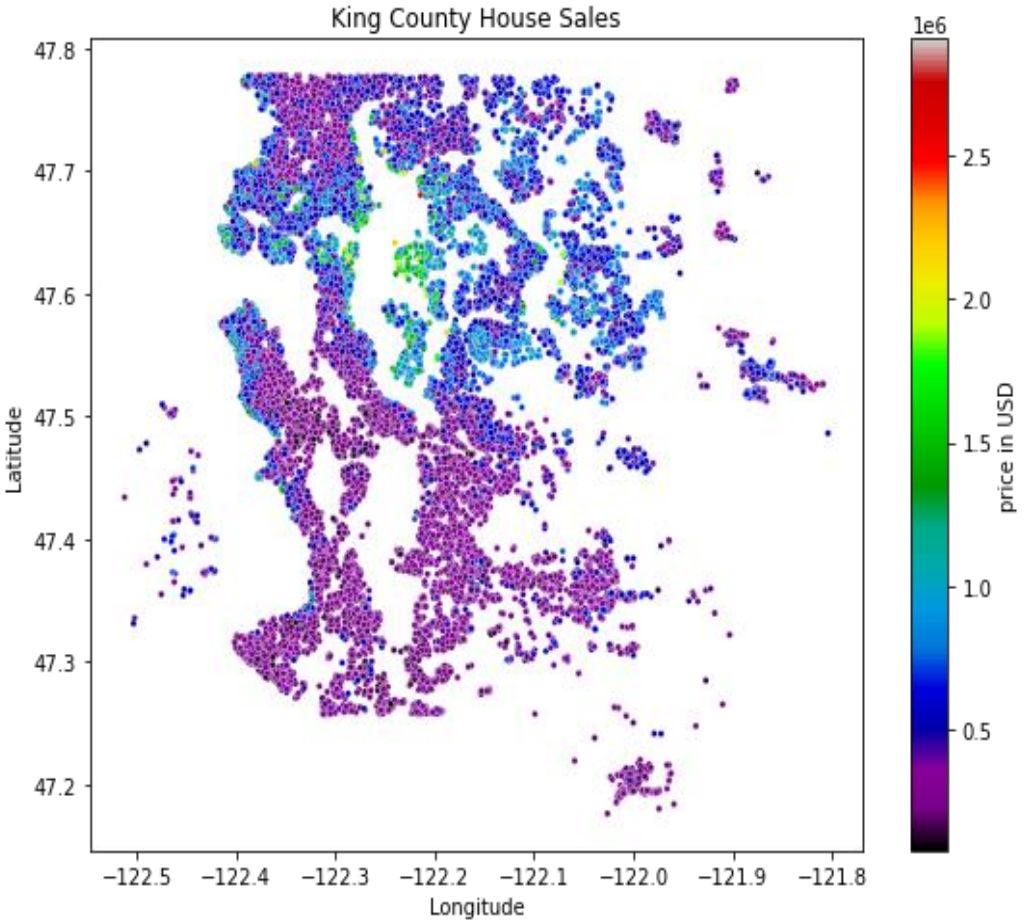

id 1 0.0013 0.0068 0.03 0.018 -0.1 0.027 0.003 0.022 -0.029 0.035 0.024 -0.0069 0.035 -0.018 -0.025 -0.0082 0.052 0.021 -0.09
price 0.0013 1 0.28 0.43 0.61 0.09 0.24 0.2 0.36 0.076 0.61 0.49 0.29 -0.0097 0.15 -0.0094 0.39 0.0095 0.54 0.097
bedrooms 0.0068 0.28 1 0.49 0.6 0.11 0.15 -0.022 0.05 0.036 0.32 0.48 0.3 0.14 0.02 -0.16 -0.03 0.16 0.38 0.12
bathrooms 0.03 0.43 0.49 1 0.71 0.046 0.5 0.026 0.12 -0.12 0.61 0.62 0.24 0.52 0.052 -0.2 0.0041 0.25 0.52 0.055
sqft_living 0.018 0.61 0.6 0.71 1 0.22 0.33 0.05 0.22 -0.035 0.7 0.83 0.42 0.31 0.064 -0.19 0.036 0.26 0.73 0.24
sqft_lot -0.1 0.09 0.11 0.046 0.22 1 -0.13 0.073 0.073 0.054 0.13 0.2 0.067 0.013 0.02 -0.18 -0.067 0.25 0.27 0.81
floors 0.027 0.24 0.15 0.5 0.33 -0.13 1 0.012 -0.0011 -0.27 0.45 0.53 -0.29 0.49 0.0076 -0.055 0.037 0.13 0.25 -0.14
waterfront 0.003 0.2 -0.022 0.026 0.05 0.073 0.012 1 0.38 0.015 0.036 0.027 0.044 -0.033 0.084 0.038 -0.02 -0.038 0.061 0.092
view 0.022 0.36 0.05 0.12 0.22 0.073 -0.0011 0.38 1 0.052 0.19 0.091 0.24 -0.083 0.1 0.12 0.014 -0.11 0.24 0.07
condition -0.029 0.076 0.036 -0.12 -0.035 0.054 -0.27 0.015 0.052 1 -0.14 -0.16 0.19 -0.35 -0.066 -0.012 -0.01 -0.085 -0.079 0.072
grade 0.035 0.61 0.32 0.61 0.7 0.13 0.45 0.036 0.19 -0.14 1 0.69 0.12 0.45 0.019 -0.17 0.11 0.21 0.66 0.16
sqft_above 0.024 0.49 0.48 0.62 0.83 0.2 0.53 0.027 0.091 -0.16 0.69 1 -0.15 0.43 0.028 -0.27 -0.031 0.38 0.7 0.22
sqft_basement -0.0069 0.29 0.3 0.24 0.42 0.067 -0.29 0.044 0.24 0.19 0.12 -0.15 1 -0.16 0.068 0.097 0.12 -0.16 0.17 0.061
yr_built 0.035 -0.0097 0.14 0.52 0.31 0.013 0.49 -0.033 -0.083 -0.35 0.45 0.43 -0.16 1 -0.22 -0.34 -0.16 0.43 0.31 0.039
yr_renovated -0.018 0.15 0.02 0.052 0.064 0.02 0.0076 0.084 0.1 -0.066 0.019 0.028 0.068 -0.22 1 0.069 0.031 -0.075 -0.0064 0.016
zipcode -0.025 -0.0094 -0.16 -0.2 -0.19 -0.18 -0.055 0.038 0.12 -0.012 -0.17 -0.27 0.097 -0.34 0.069 1 0.27 -0.58 -0.28 -0.2
lat -0.0082 0.39 -0.03 0.0041 0.036 -0.067 0.037 -0.02 0.014 -0.01 0.11 -0.031 0.12 -0.16 0.031 0.27 1 -0.14 0.033 -0.063
long 0.052 0.0095 0.16 0.25 0.26 0.25 0.13 -0.038 -0.11 -0.085 0.21 0.38 -0.16 0.43 -0.075 -0.58 -0.14 1 0.37 0.29
sqft_living15 0.021 0.54 0.38 0.52 0.73 0.27 0.25 0.061 0.24 -0.079 0.66 0.7 0.17 0.31 -0.0064 -0.28 0.033 0.37 1 0.32
sqft_lot15 -0.09 0.097 0.12 0.055 0.24 0.81 -0.14 0.092 0.07 0.072 0.16 0.22 0.061 0.039 0.016 -0.2 -0.063 0.29 0.32 1
id price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above sqft_basement yr_built yr_renovated zipcode lat long sqft_living15 sqft_lot15
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

# # Plots for understanding or Analysis

**A plot is a graphical technique for representing a data set, usually as a graph showing the relationship between two or more variables. Graphs are a visual representation of the relationship between variables, which are very useful for humans who can then quickly derive an understanding which may not have come from lists of values.**

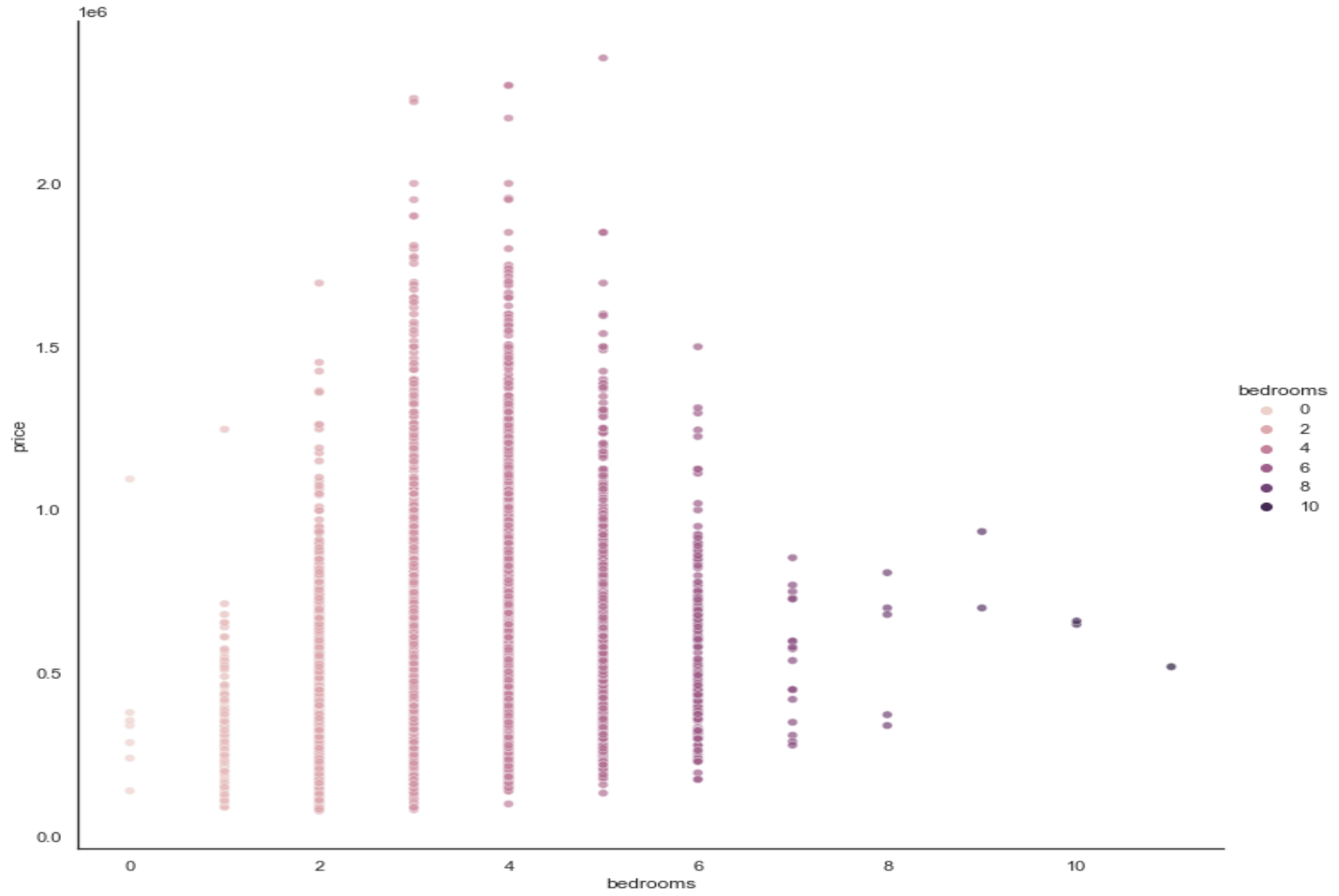
**Location**

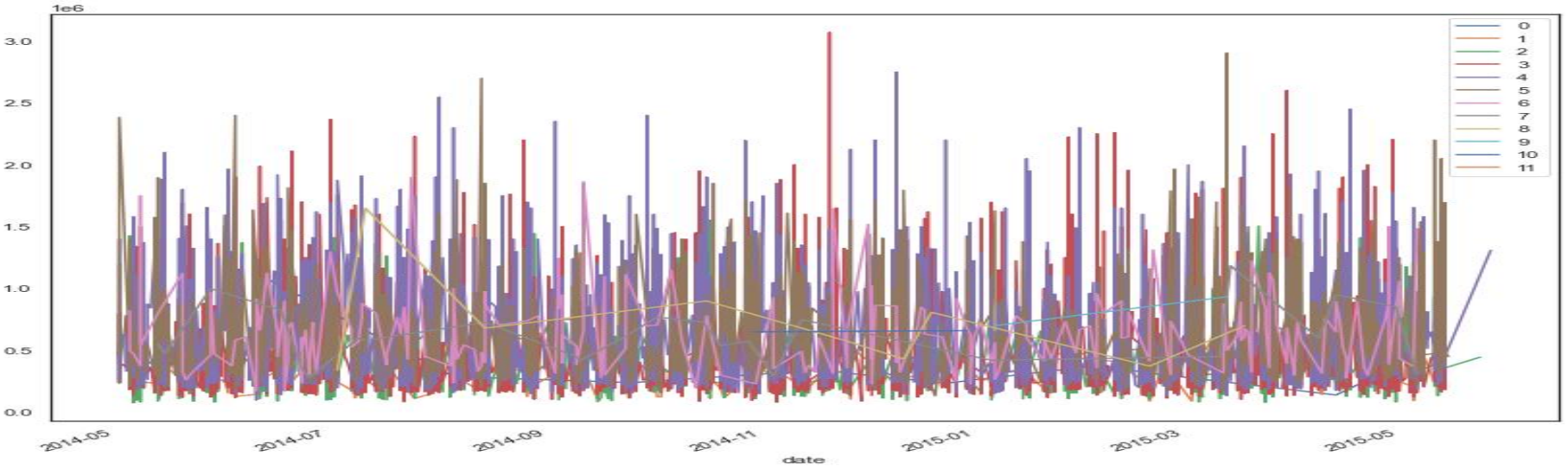
Location is key when it comes to real estate. Our first question seeks to understand the geographical distribution of the homes in our dataset and determine where the highest house sales were recorded. As a starting point, let us create a scatterplot using latitude and longitude features.

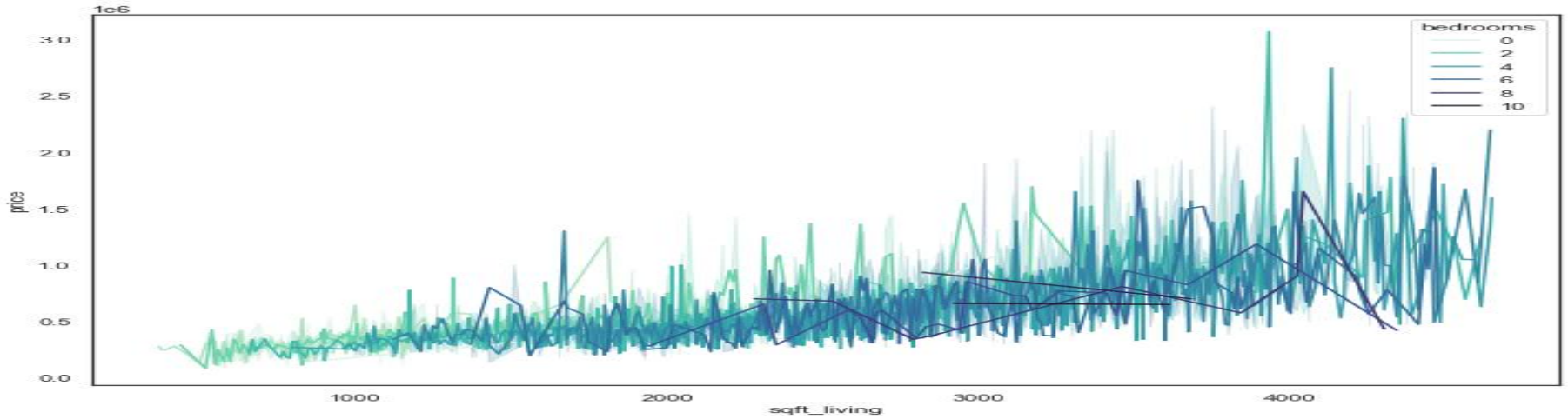

King County House Sales
Latitude
Longitude
47.8
47.7
47.6
47.5
47.4
47.3
47.2
−122.5
−122.4
−122.3
−122.2
−122.1
−122.0
−121.9
−121.8
1e6
2.5
2.0
1.5
1.0
0.5
price in USD

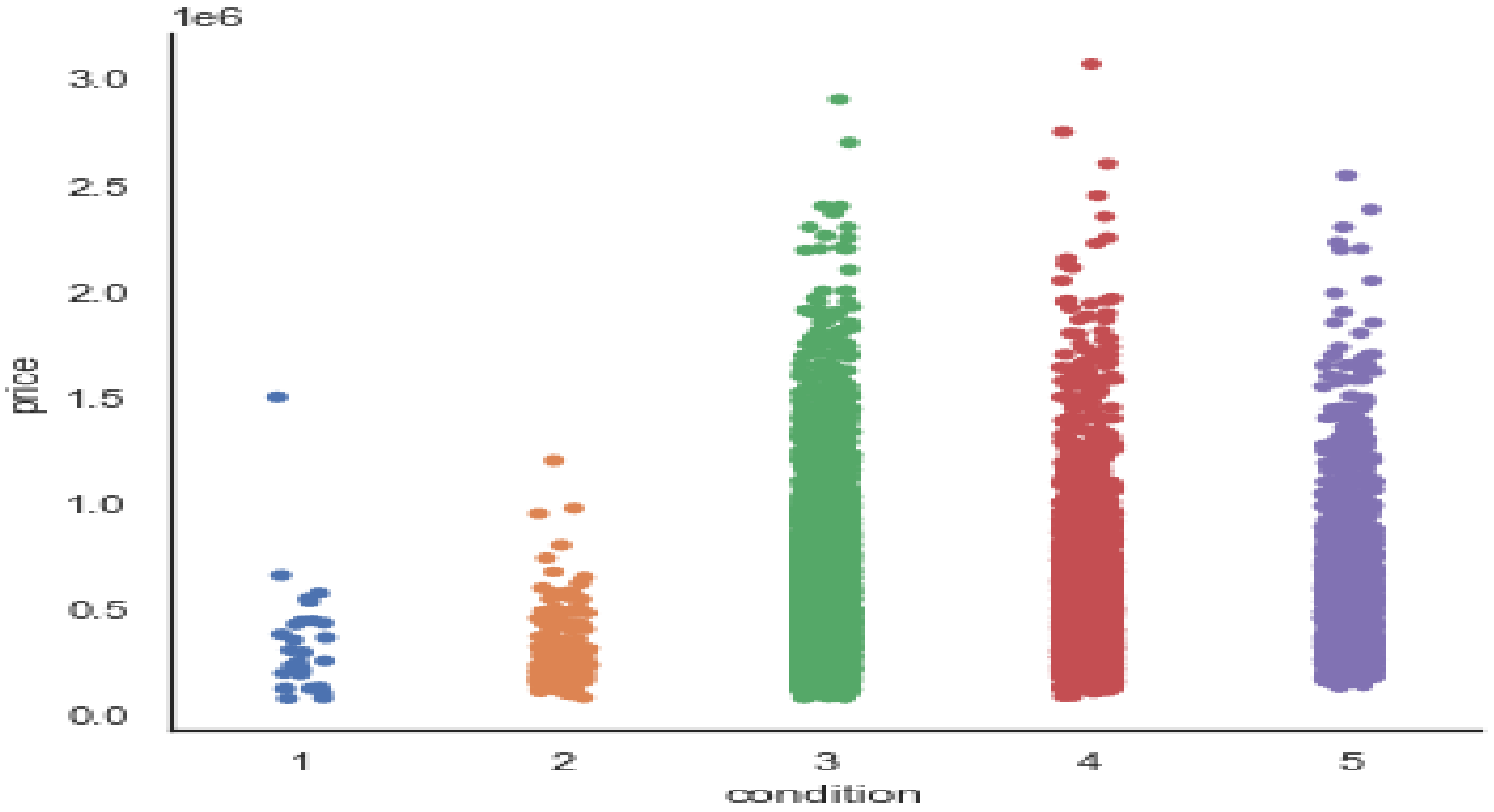

SNOHOMISH COUNTY
LAKE FOREST PARK
BOTHELL
SHORELINE
KENMORE
WOODINVILLE
DUVALL
KIRKLAND
YARROW POINT
REDMOND
HUNTS POINT
CLYDE HILL
SEATTLE
BELLEVUE
MEDINA
CARNATION
KITSAP COUNTY
PUGET SOUND
BEAUX ARTS
SAMMAMISH
MERCER ISLAND
NEWCASTLE
ISSAQUAH
SNOQUALMIE
NORTH BEND
BURIEN
RENTON
TUKWILA
SEATAC
VASHON ISLAND
NORMANDY PARK
DES MOINES
KENT
COVINGTON
MAPLE VALLEY
FEDERAL WAY
AUBURN
BLACK DIAMOND
ALGONA
MILTON
PACIFIC
TACOMA
ENUMCLAW
KING COUNTY
WASHINGTON
PIERCE COUNTY

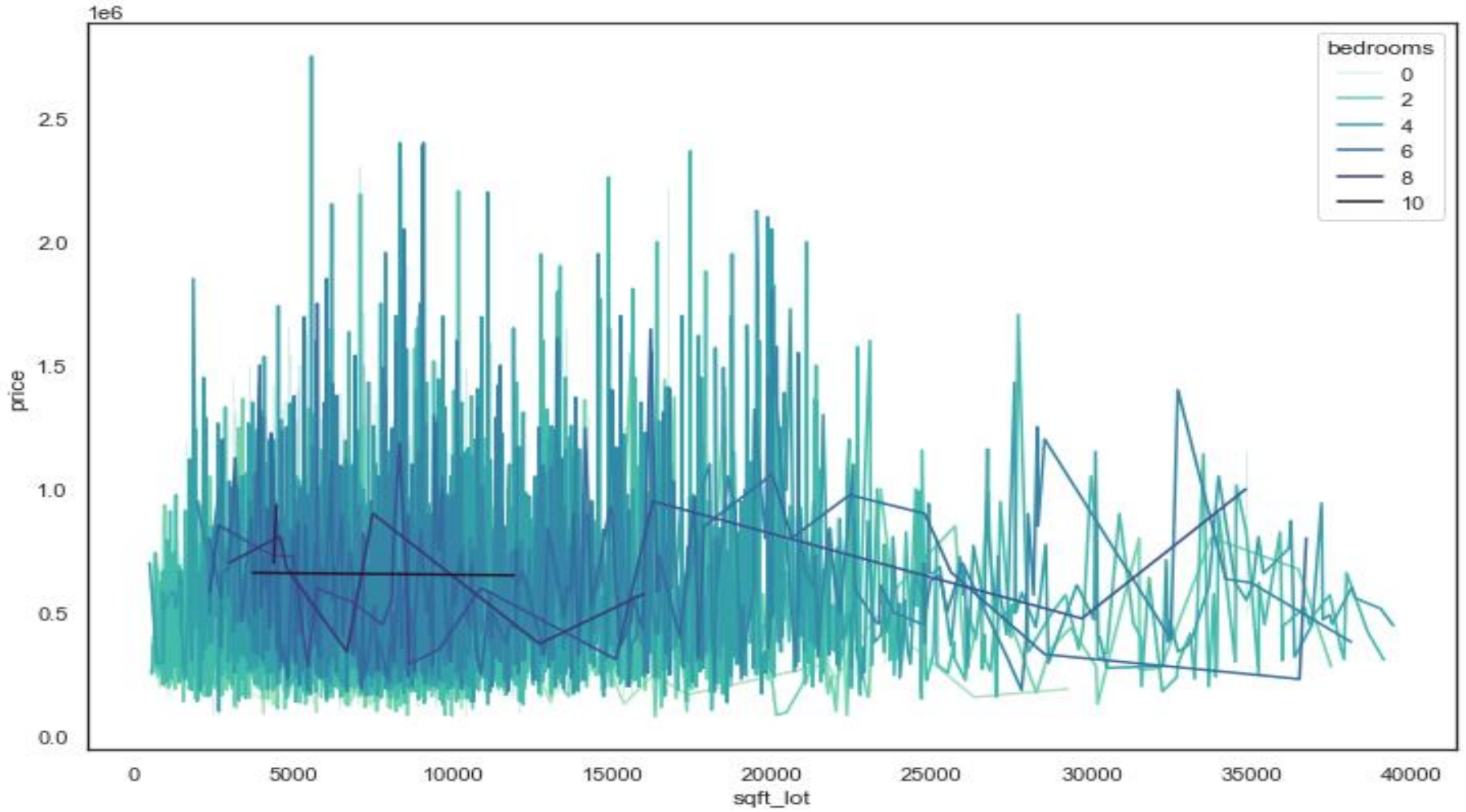
From this visualization we can already draw initial insights based on the houses' geographical locations. The highest house prices are concentrated in the area with latitude around 47.6 and longitude around -122.25. There is a disparity with southern locations achieving lower house prices.

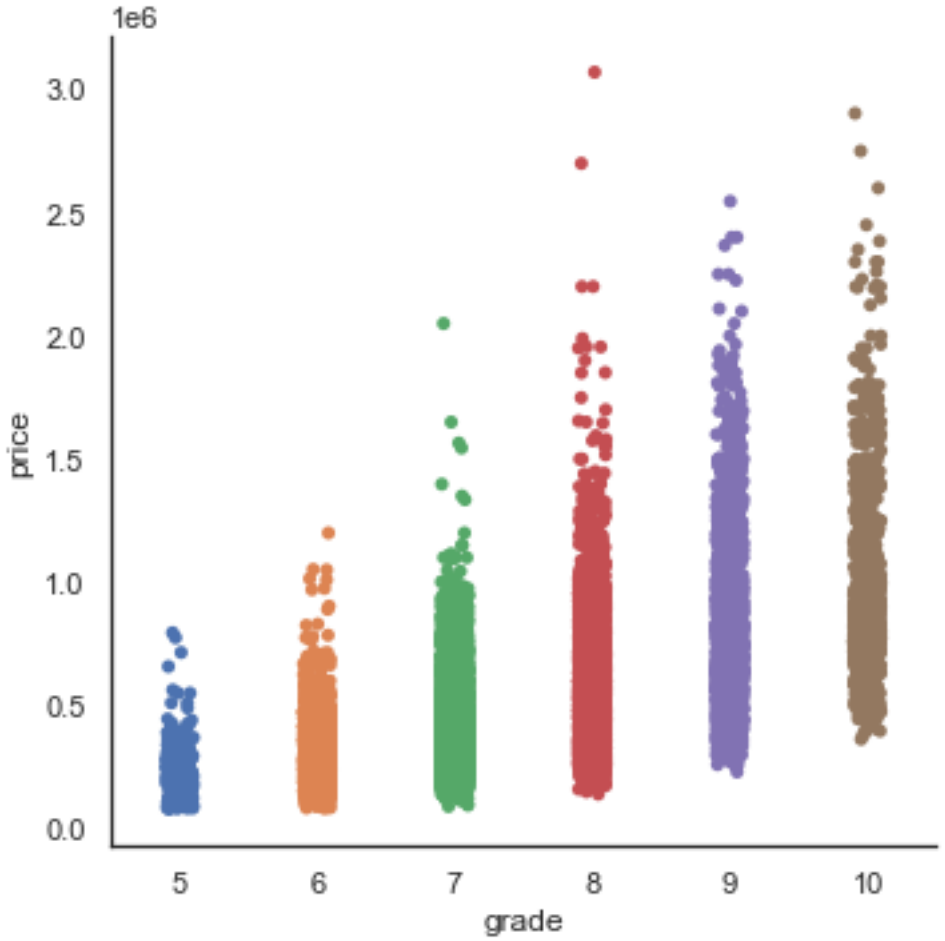

1e6
2.0
1.5
1.0
0.5
0.0
price
0
2
4
6
8
10
bedrooms
bedrooms
0
2
4
6
8
10

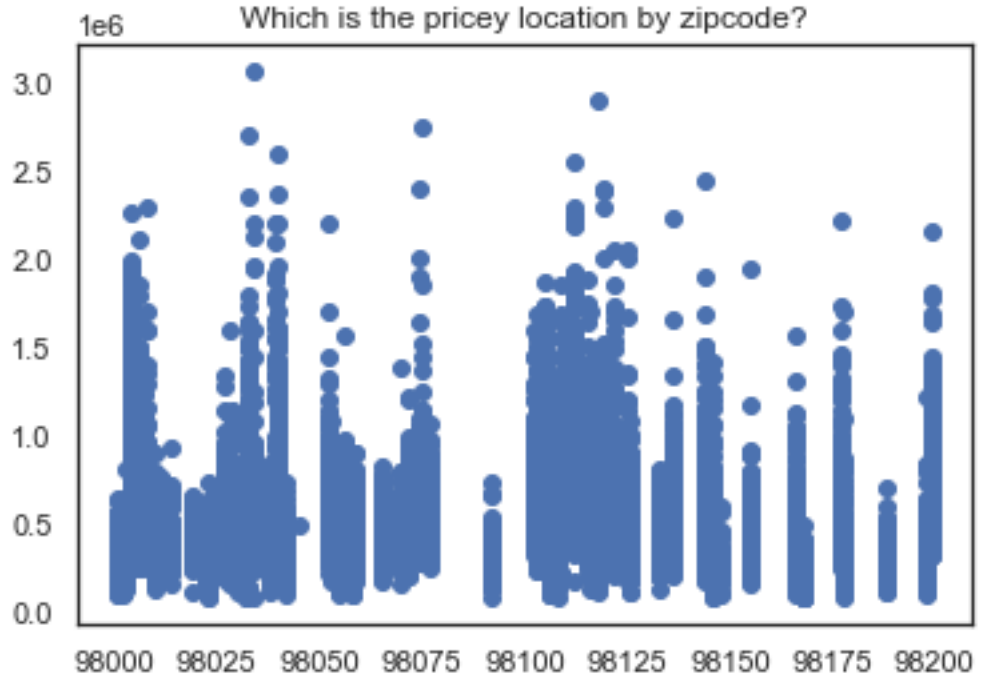

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2014-05
2014-07
2014-09
2014-11
2015-01
2015-03
2015-05
date
0
1
2
3
4
5
6
7
8
9
10
11

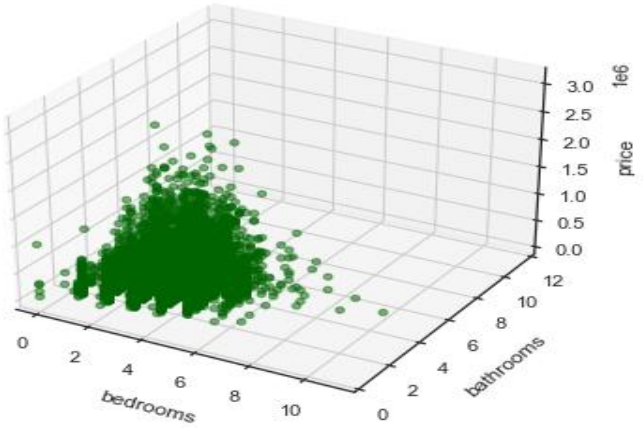

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
price
1000
2000
3000
4000
sqft_living
bedrooms
0
2
4
6
8
10

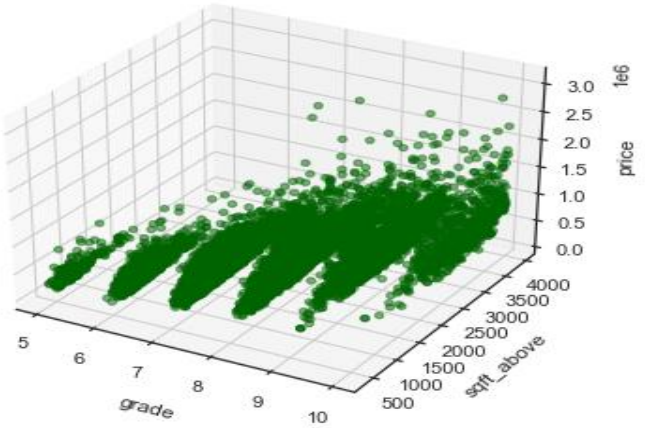

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
price
1
2
3
4
5
condition

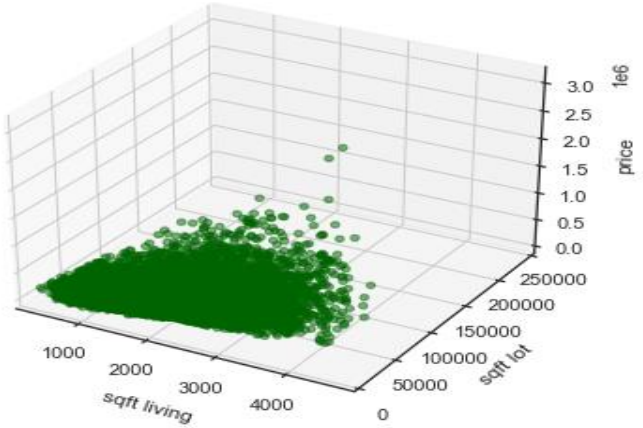

1e6
2.5
2.0
1.5
1.0
0.5
0.0
price
0
5000
10000
15000
20000
25000
30000
35000
40000
sqft_lot
bedrooms
0
2
4
6
8
10

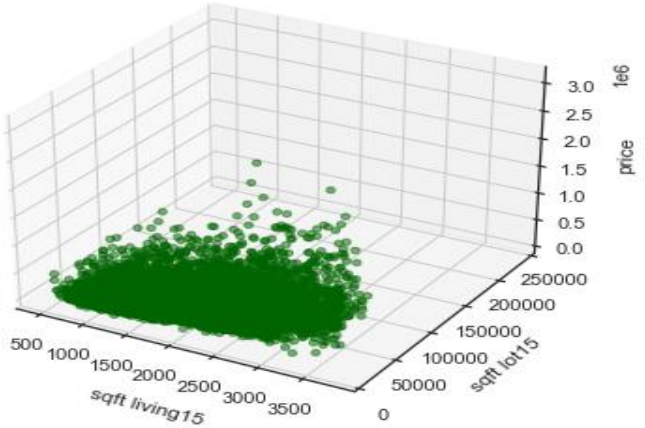

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
price
5
6
7
8
9
10
grade


Which is the pricey location by zipcode?
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
98000
98025
98050
98075
98100
98125
98150
98175
98200


price
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
bedrooms
0
2
4
6
8
10
bathrooms
0
2
4
6
8
10
12


price
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
grade
5
6
7
8
9
10
sqft_above
500
1000
1500
2000
2500
3000
3500
4000


price
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
sqft living
1000
2000
3000
4000
sqft lot
0
50000
100000
150000
200000
250000


price
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
sqft living15
500
1000
1500
2000
2500
3000
3500
sqft lot15
0
50000
100000
150000
200000
250000

# Preliminary Observation

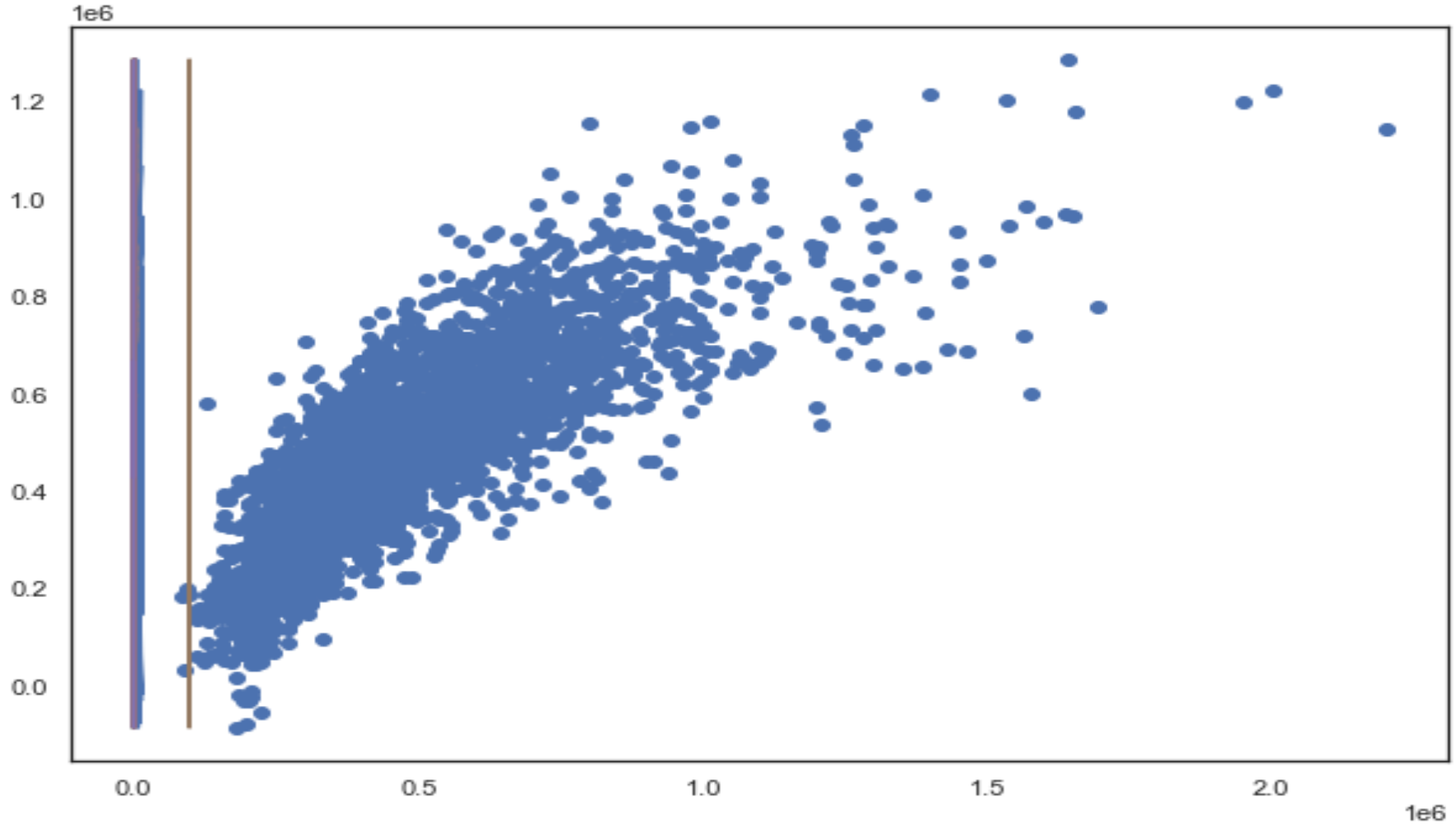
(1) The frequency of no of bedrooms=3,4 is higher than any other bedrooms. The price of these are mostly similar, but some have higher price than usual because of the other features like bathrooms, location, etc.
(2) The heatmap identifies the correlations between the features which help us in identifying how the features are dependent on each other which cannot be known by seeing the data. (Example: sqft living is dependent on grade of the house)
(3) The highest priced houses are sold in months: 9th to 11th. This shows people tend to spend more money on houses which are having more comforts in winter.
(4) Most of the houses have sqft living in between 500 to 6000 irrespective of no of bedrooms. The higher the living space, the higher is the cost.
(5) Price of the house is also dependent on sqft of lot (parking) as most people own their own car.
(6) People are tending to pay less if the condition of the house is bad. They are spending more if the house is in good condition.
(7) The 3d plot gives relationship between multiple features.

## 1. Models

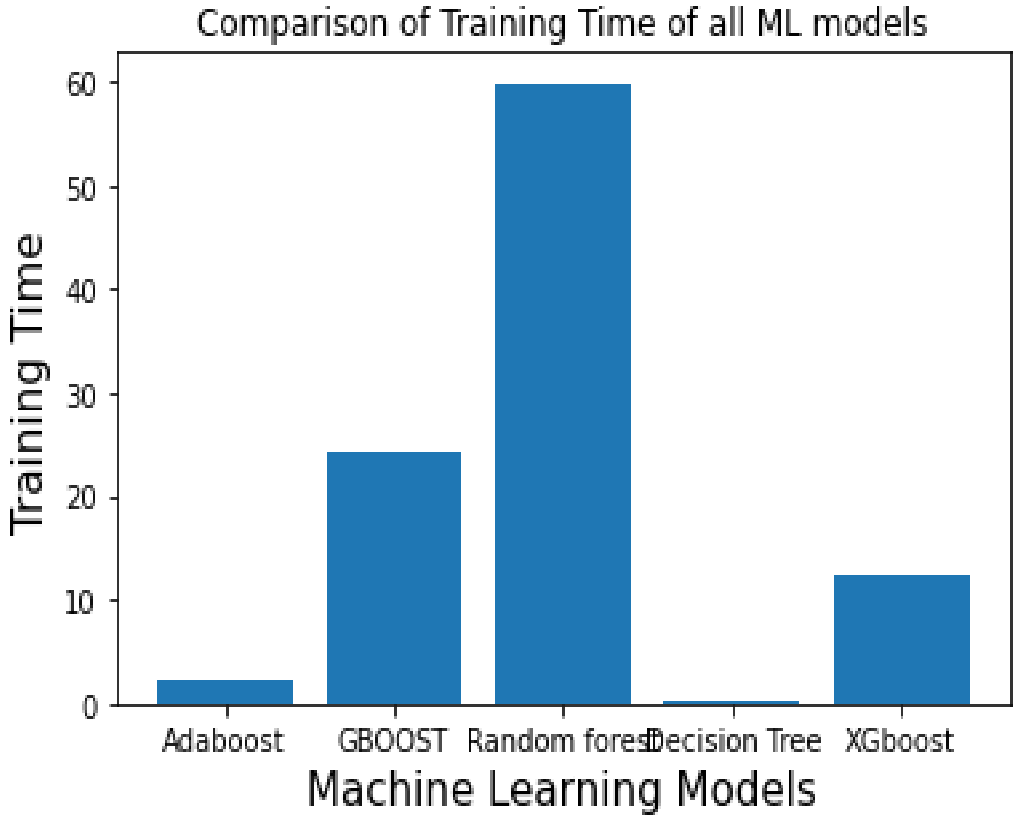
**We developed the following models:**

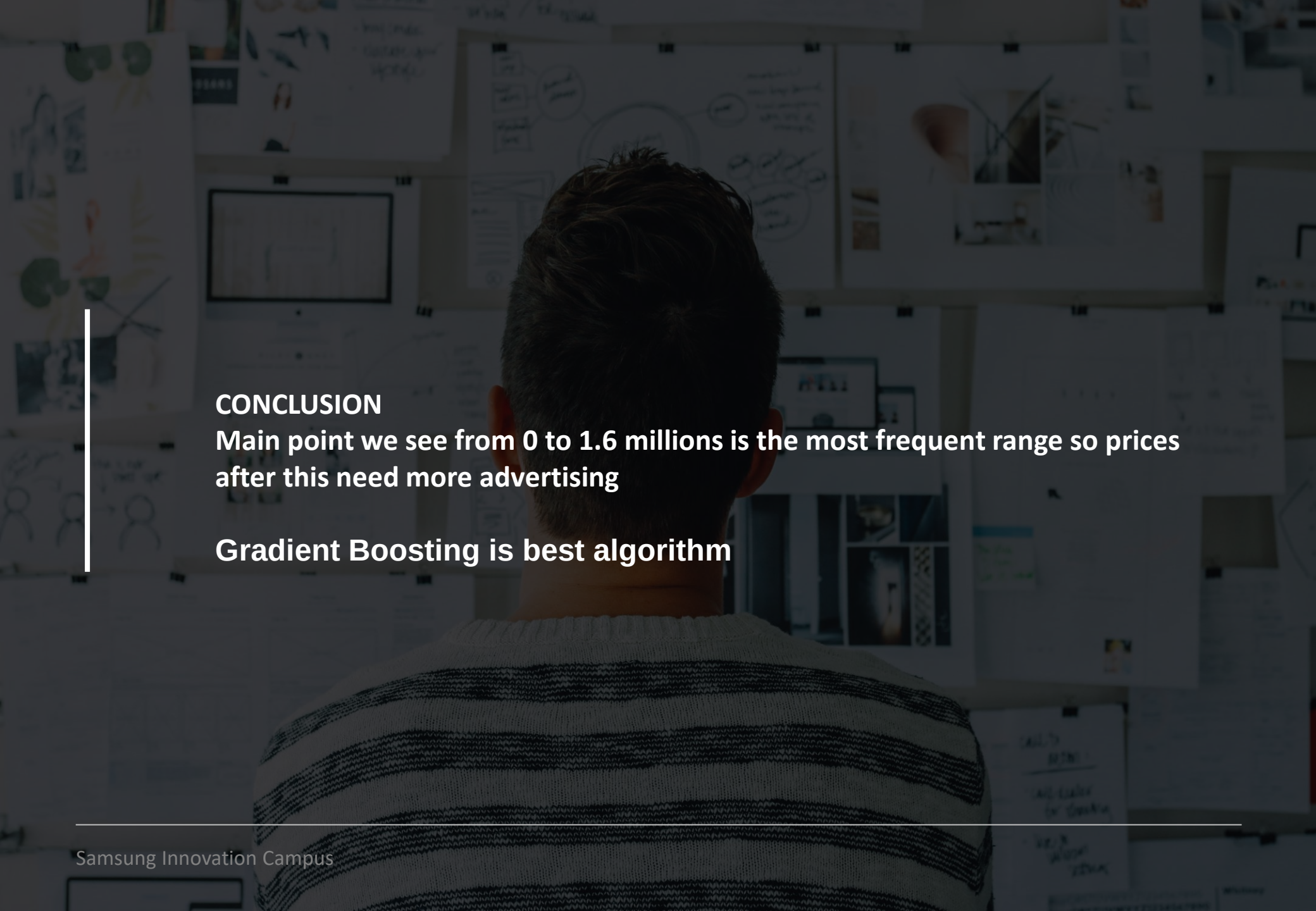
1. **Linear regression**
2. **Gradient BoostingRecursive Partitioning Model**
3. **AdaBoost**
4. **Random Forest**
5. **'Decision Tree**
6. **XGBoost**

# Linear Regression



| | Model | Score | Variance Score | R2 Score | Mean Squared Error | |
|---|---|---|---|---|---|---|
| **0** | Gradient Boosting | 0.872013 | 0.872022 | 0.872013 | 7.623835e+09 | |
| **4** | XGBoost | 0.870216 | 0.870242 | 0.870216 | 7.730887e+09 | |
| **2** | Random Forest | 0.854403 | 0.854413 | 0.854403 | 8.672788e+09 | |
| **3** | Decision Tree | 0.714179 | 0.714331 | 0.714179 | 1.702558e+10 | |
| **1** | AdaBoost | 0.675359 | 0.696531 | 0.675359 | 1.933796e+10 | |


Comparison of Training Time of all ML models
Training Time
0
10
20
30
40
50
60
Adaboost
GBOOST
Random forest
Decision Tree
XGboost
Machine Learning Models



**CONCLUSION**

**Main point we see from 0 to 1.6 millions is the most frequent range so prices after this need more advertising**

**Gradient Boosting is best algorithm**

SAMSUNG

Together for Tomorrow!

# Enabling People

Education for Future Generations