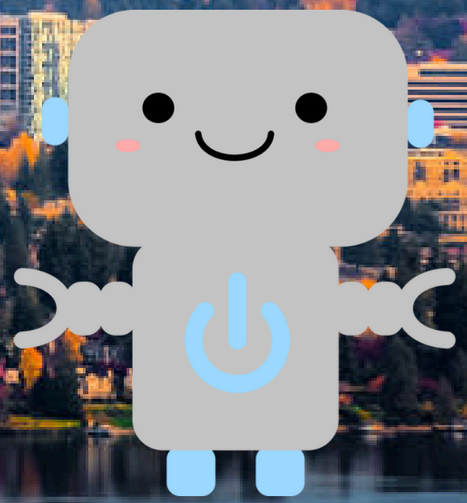
SC1015: Mini Project

Team 5:

Jumana, Aung, Yichi

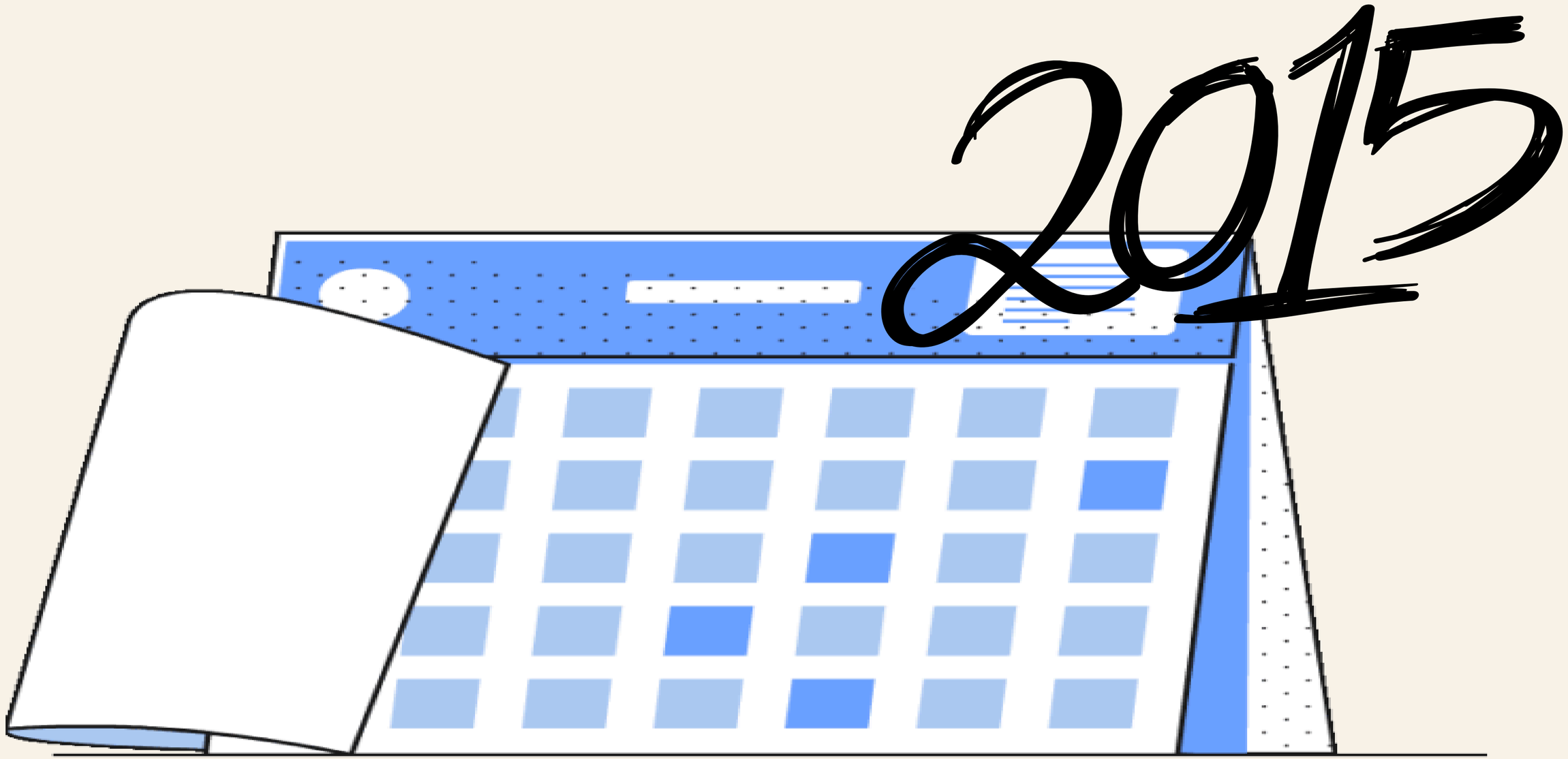


01 - Problem formulation

02 - Exploratory Data Analysis

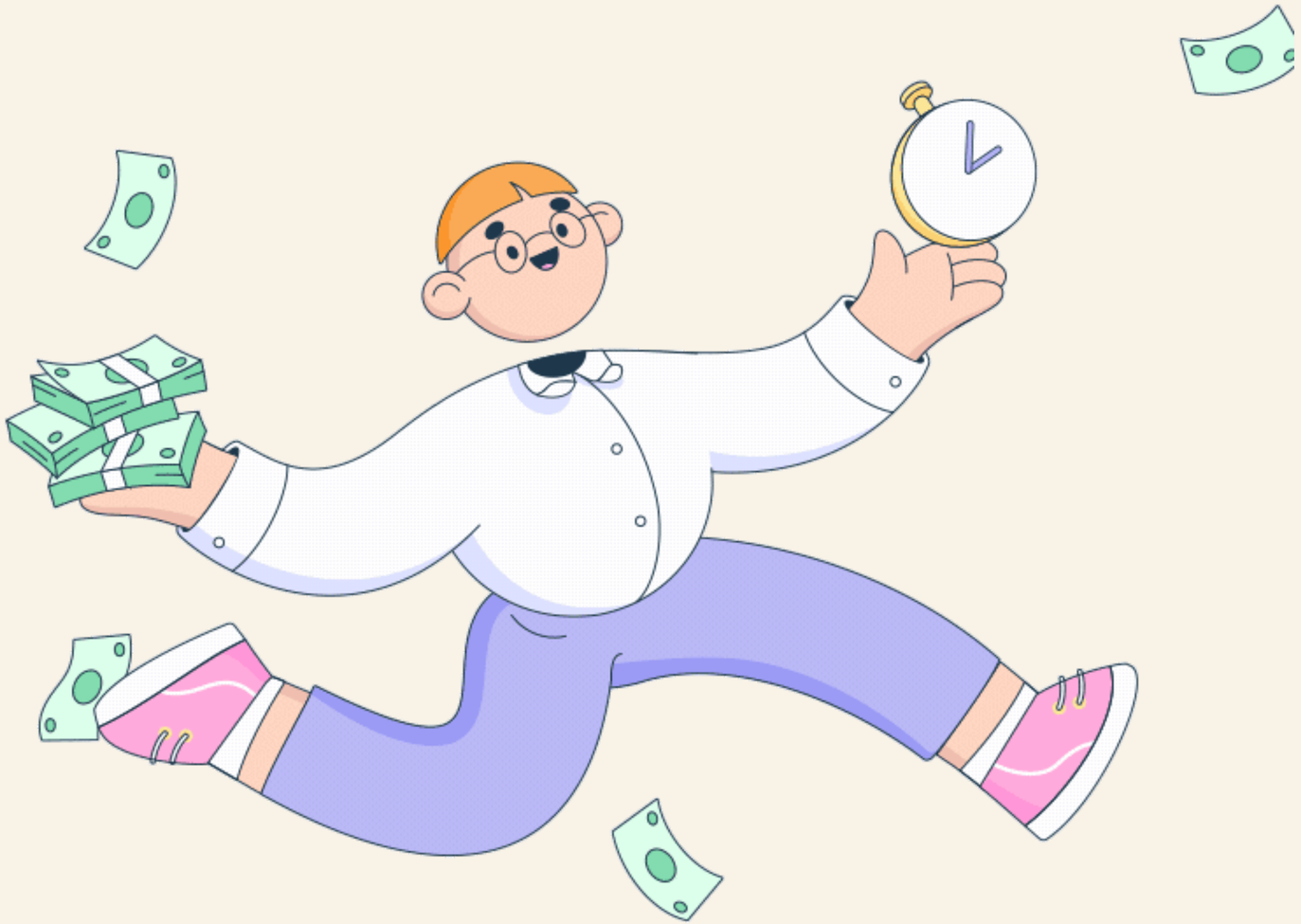
03 - Machine Learning

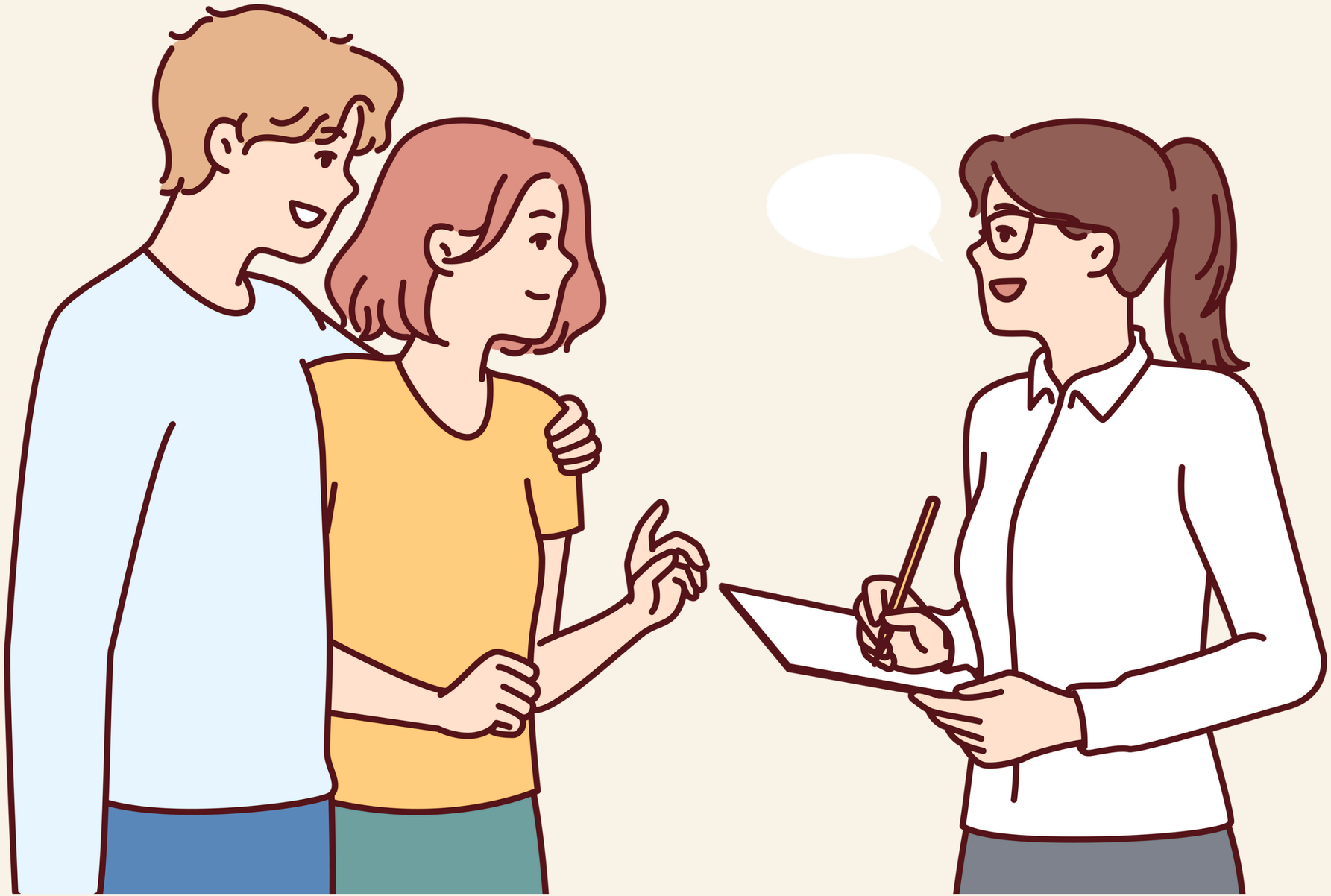
04 - Insights & Recommendations


2015

Lucas

Lucas





Problem formulation

Problem formulation

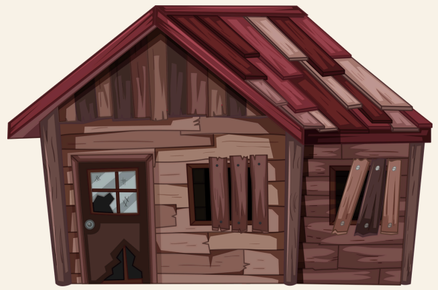
Problem formulation

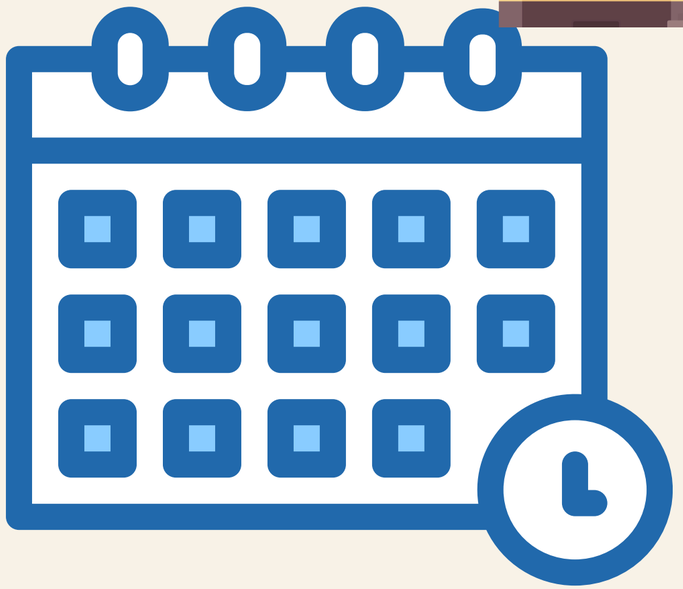
kaggle

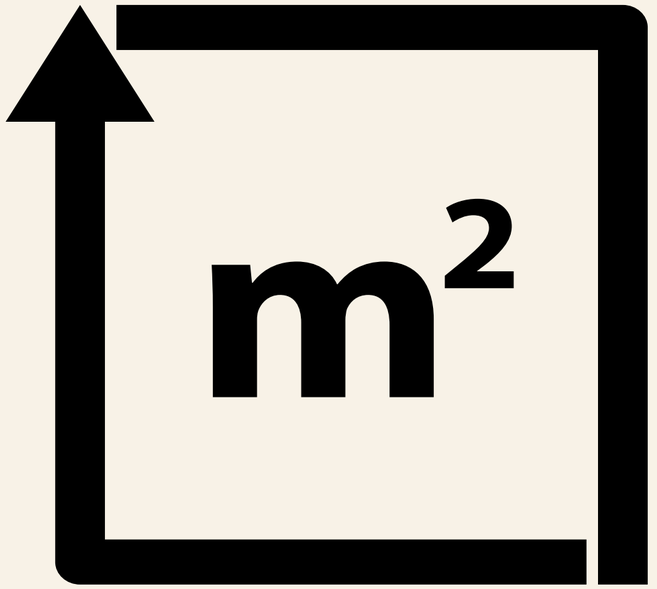
Variable	Description
id	Unique ID per house sale
date	Date of the house sale
price	Price of house sale in currency of USD
bedrooms	Number of bedrooms
bathrooms	Number of Bathrooms, where 0.5 represents a bathroom with a toilet but with no shower
sqft_living	Square footage of the apartments interior living space
sqft_lot	Square footage of the land space
floors	Number of floors
waterfront	An index to indicate if the house was overlooking the waterfront or not. 0 represents no waterfront, 1 represents with waterfront.
view	An index from 0 to 4 of how good the view of the property was. 0 represents no good view, 4 represents very good view.
condition	An index from 1 to 5 on the condition of the house. 1 represents poorer condition, and 5 represents superb condition.
grade	An index from 1 to 13. 1 to 3 falls short of building construction and design, 7 has an average level of construction and design, and 11 to 13 have higher quality level of construction and design.
sqft_above	The square footage of the interior housing space that is above the ground level
sqft_basement	The square footage of the interior housing space that is below the ground level
yr_built	The year of house built
yr_renovated	The year of the house's last renovation
zipcode	The zipcode is the postal code to indicate the area the house is in
lat	Latitude
long	Longitude
sqft_living15	The average square footage of interior housing living space for the nearest 15 neighboring houses
sqft_lot 15	The average square footage of land space for the nearest 15 neighboring houses

Investopedia

Investopedia







Questions to ask

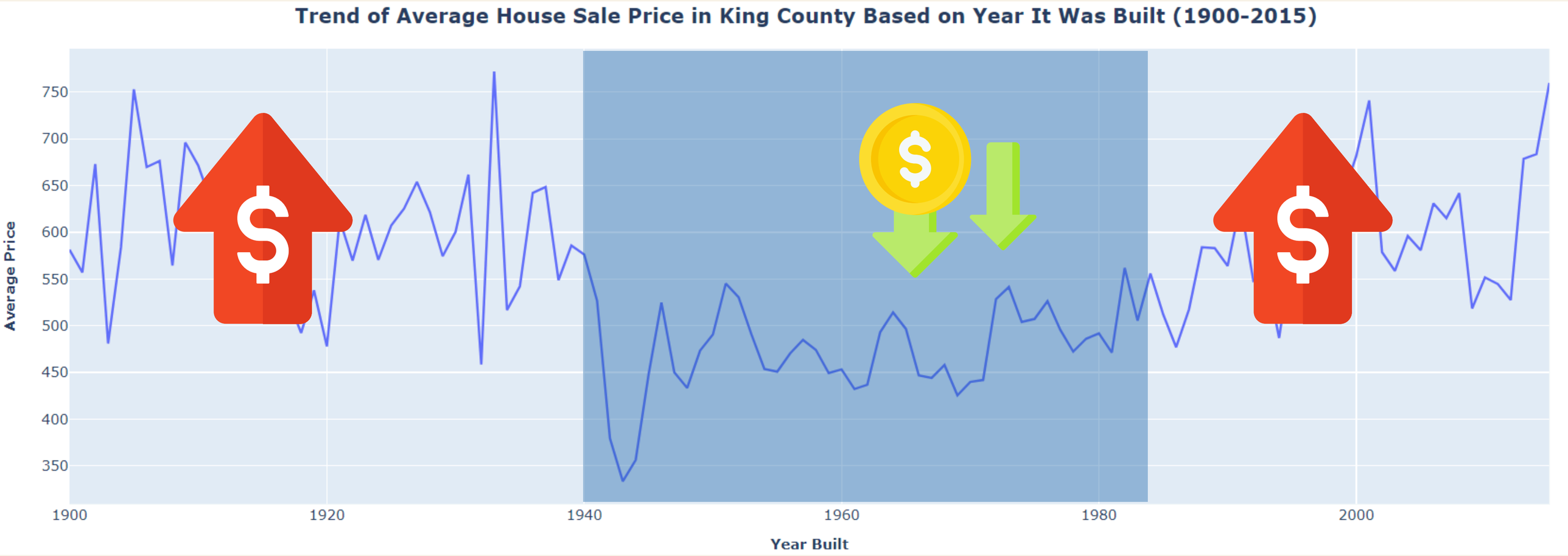
- What factors affect the changes in house sale price?
- How do house sale prices vary in King County?
- What might be the pattern of houses sold in different parts of King County?
- Is there any specific time period where houses are more expensive/cheaper?

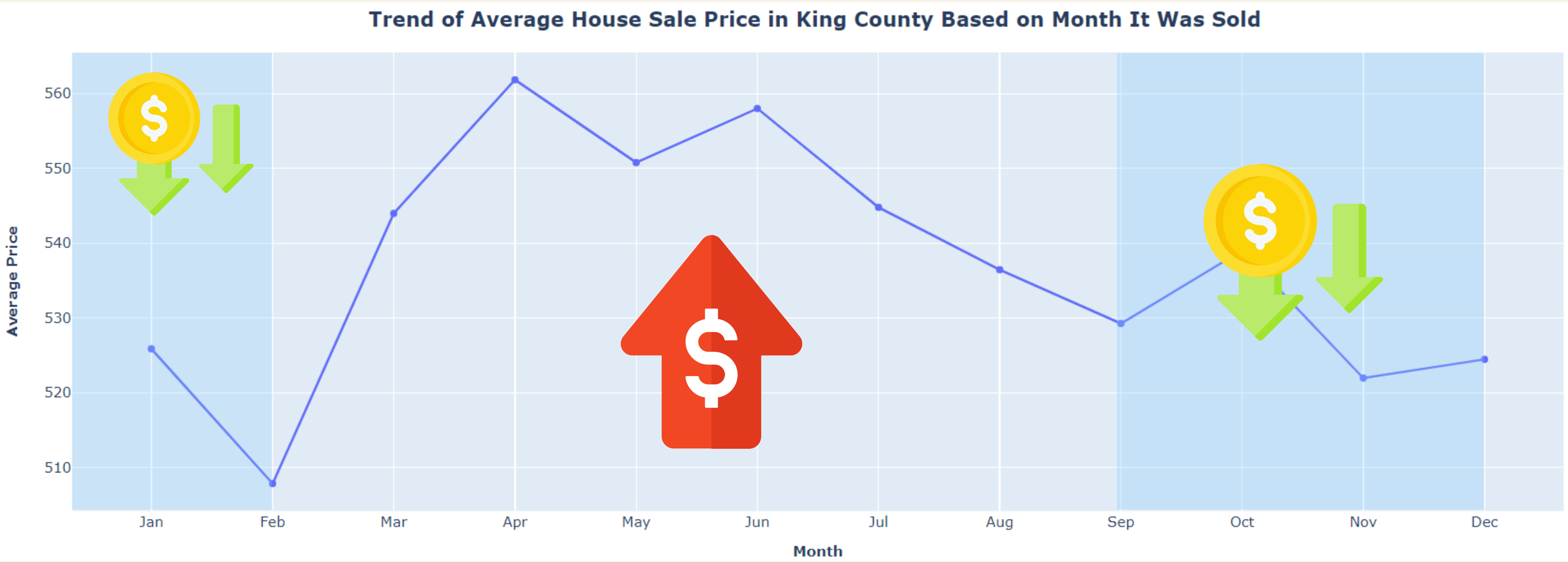
Exploratory Data Analysis

seaborn

matplotlib

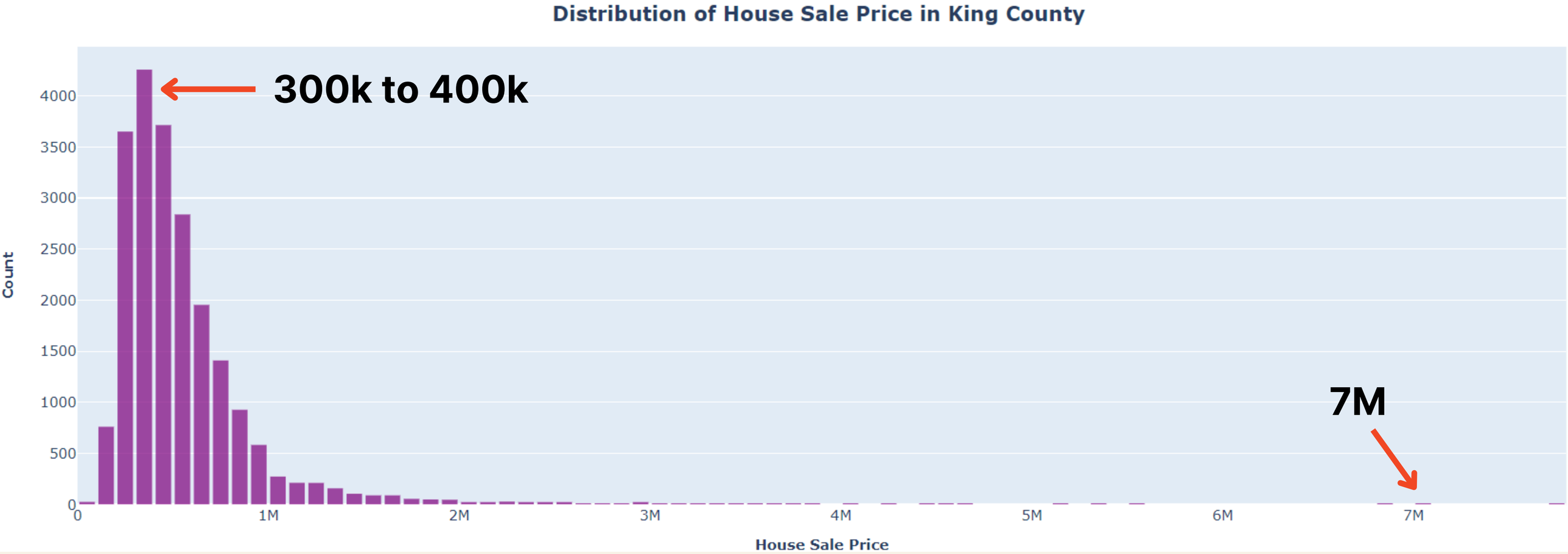
plotly


Trend of Average House Sale Price in King County Based on Year It Was Built (1900-2015)
Average Price
750
700
650
600
550
500
450
400
350
1900
1920
1940
1960
1980
2000
Year Built

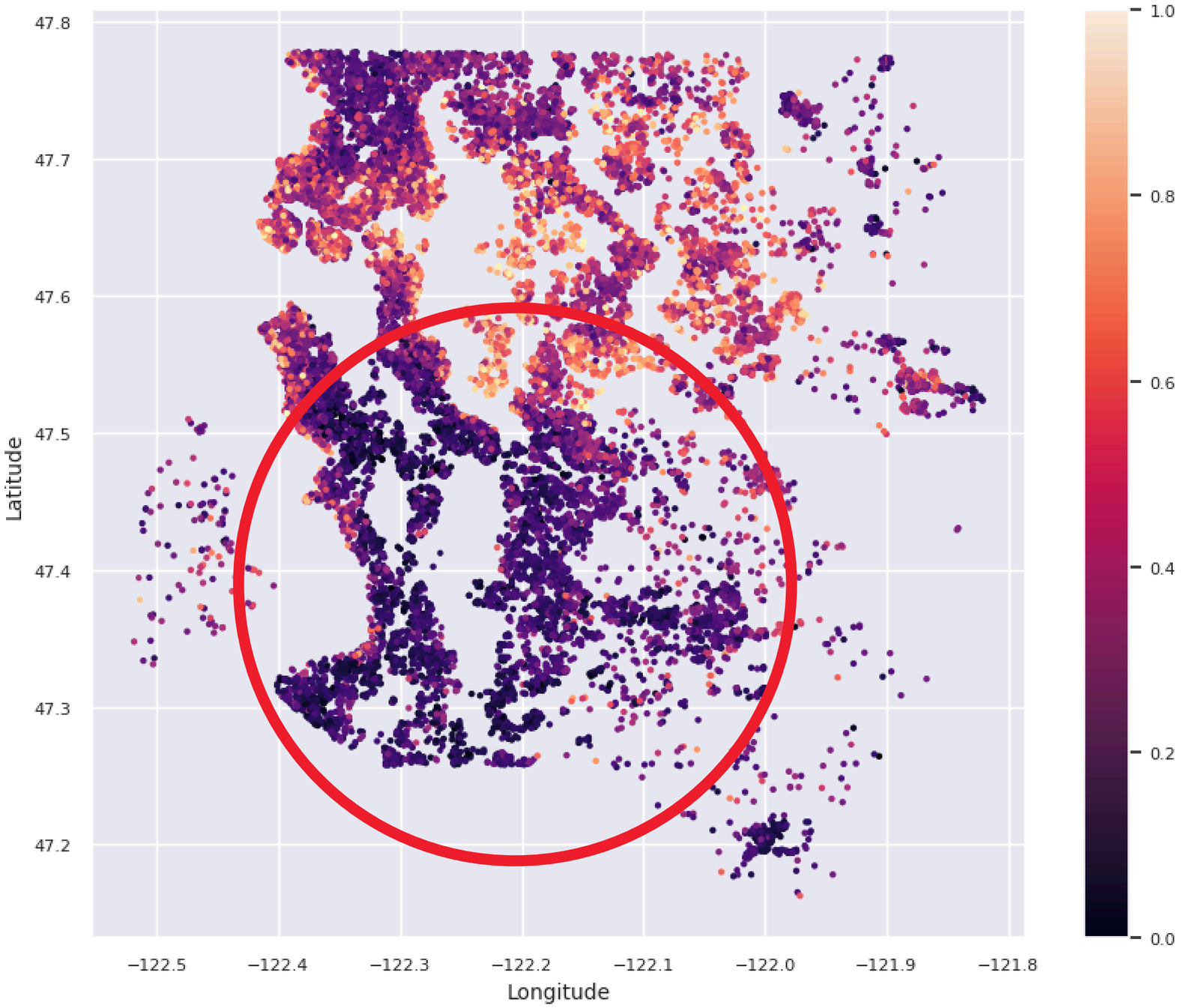

Trend of Average House Sale Price in King County Based on Month It Was Sold
Average Price
560
550
540
530
520
510
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

Correlation of Sale Price And Different Measures of Houses in King County

	price	sqft_above	sqft_basement	sqft_lot	sqft_living
price	1.00	0.61	0.32	0.09	0.70
sqft_above	0.61	1.00	-0.05	0.18	0.88
sqft_basement	0.32	-0.05	1.00	0.02	0.44
sqft_lot	0.09	0.18	0.02	1.00	0.17
sqft_living	0.70	0.88	0.44	0.17	1.00

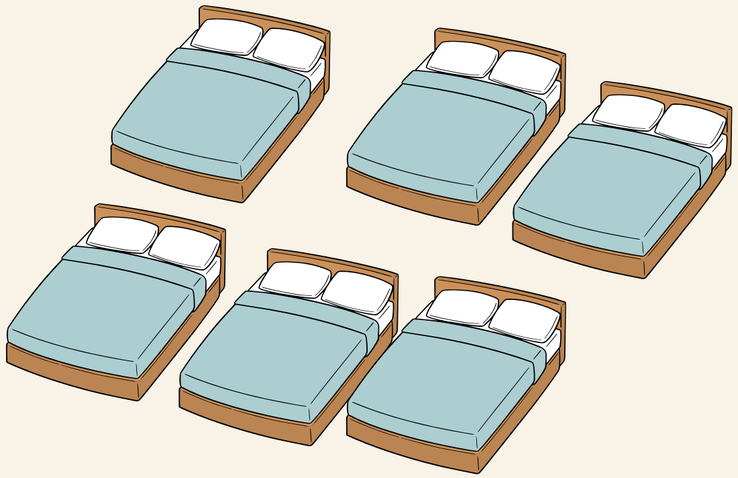

Distribution of House Sale Price in King County
300k to 400k
7M
Count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1M
2M
3M
4M
5M
6M
7M
House Sale Price

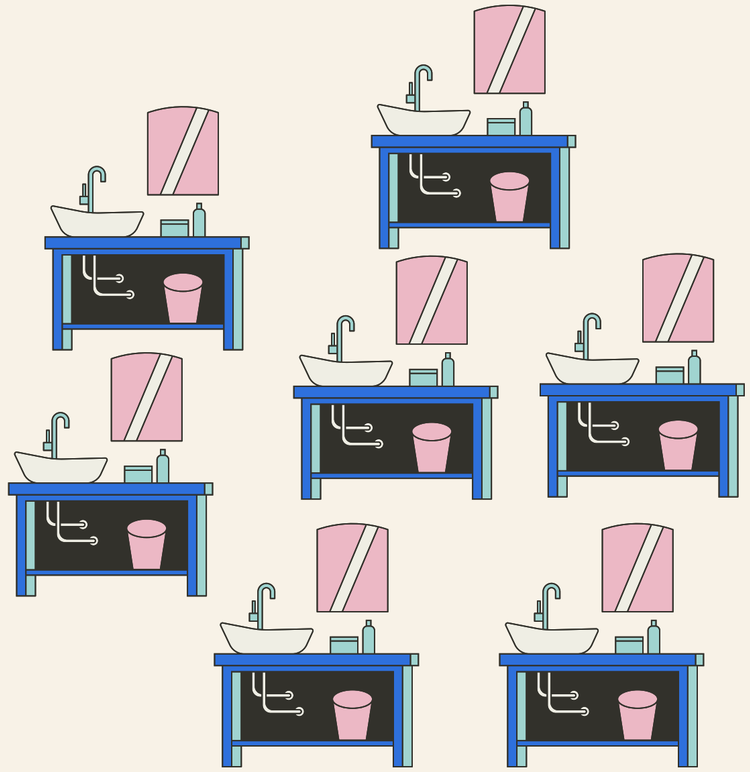
General Changes in House Sale Price Across Different Parts of King County



Remove outliers

- Bigger houses are cheaper
- Depends on the owner

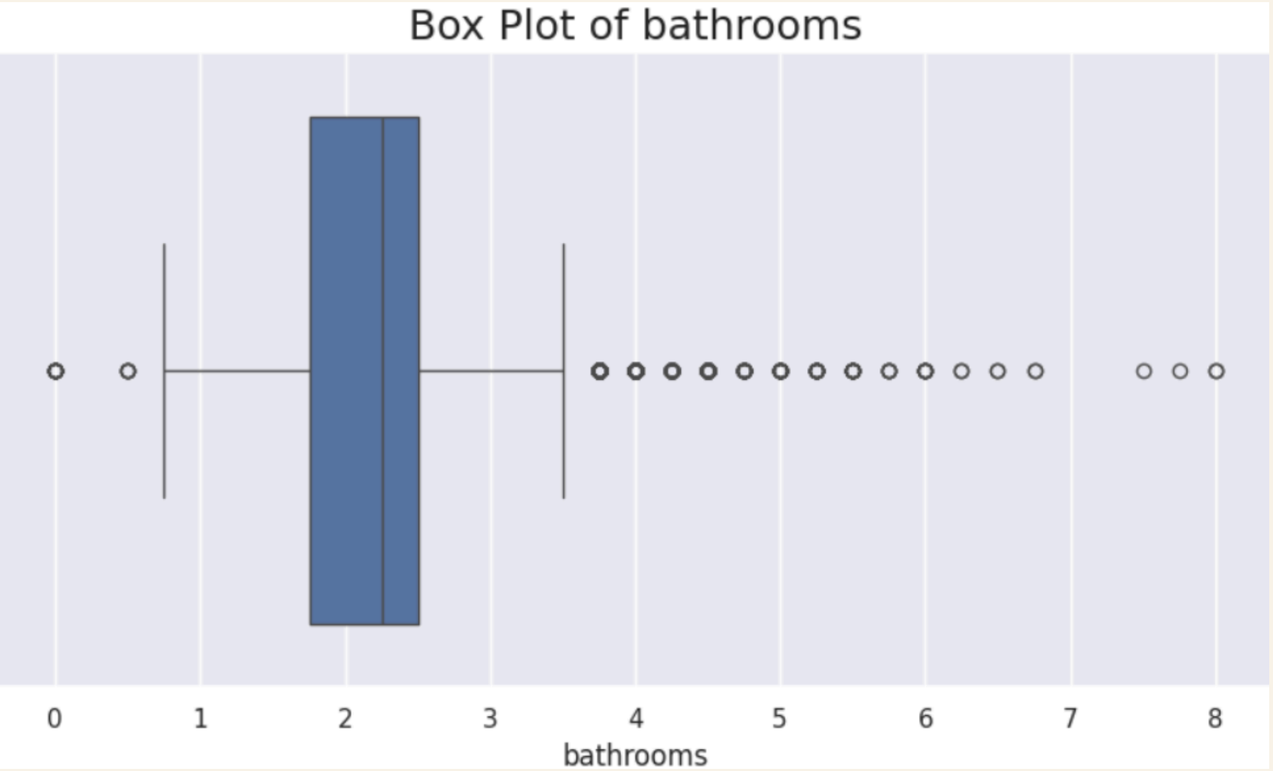




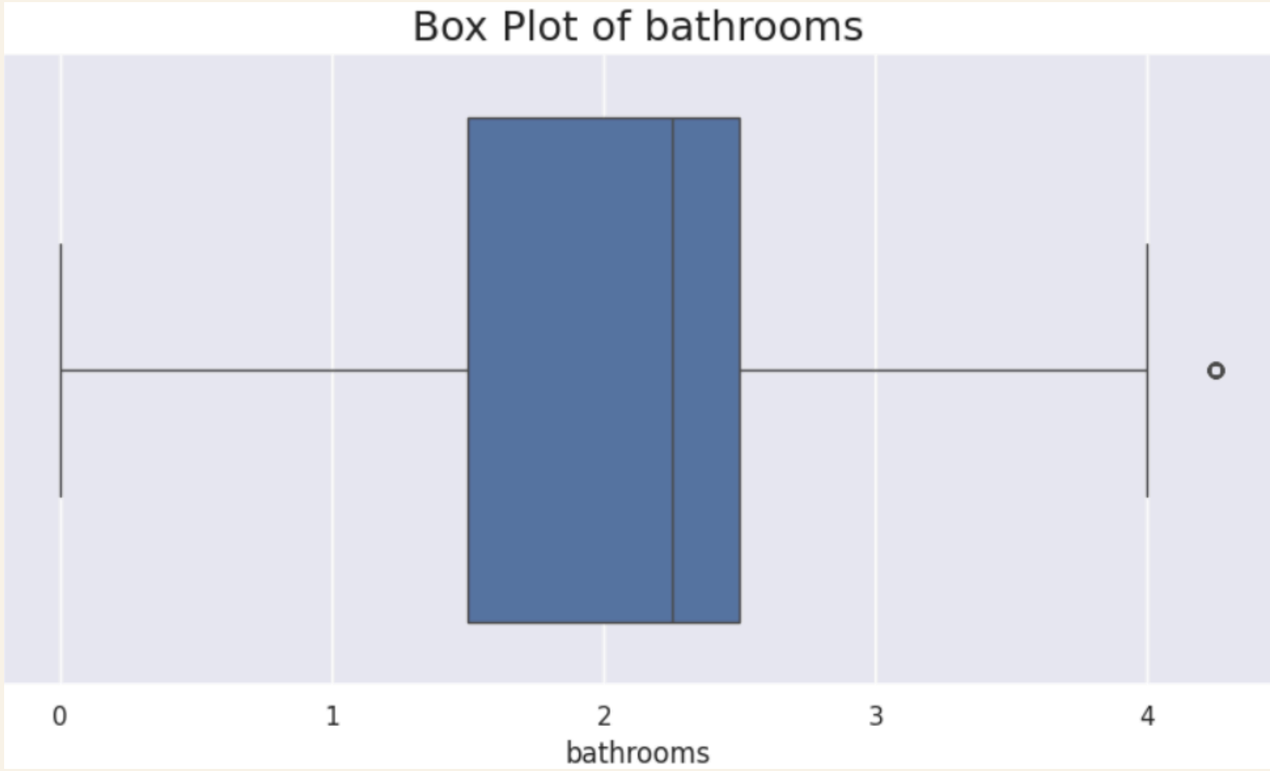
Before

After

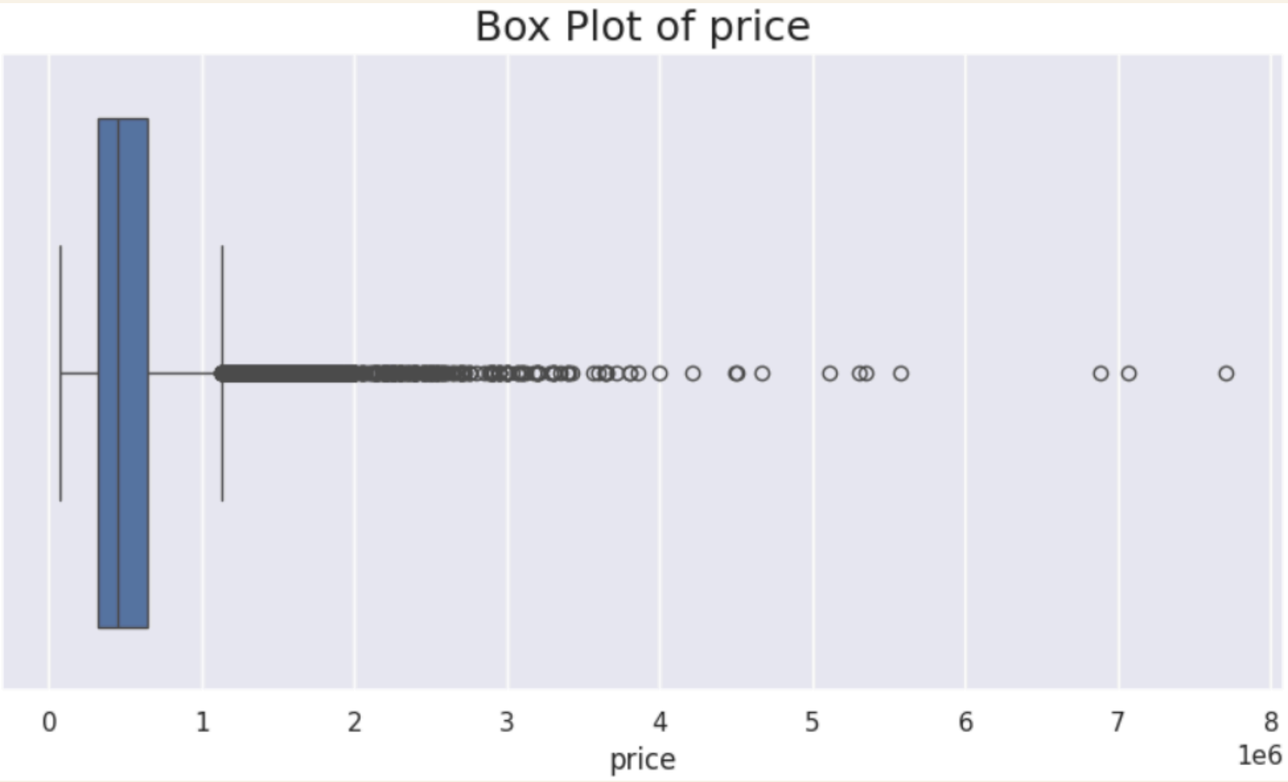
Before



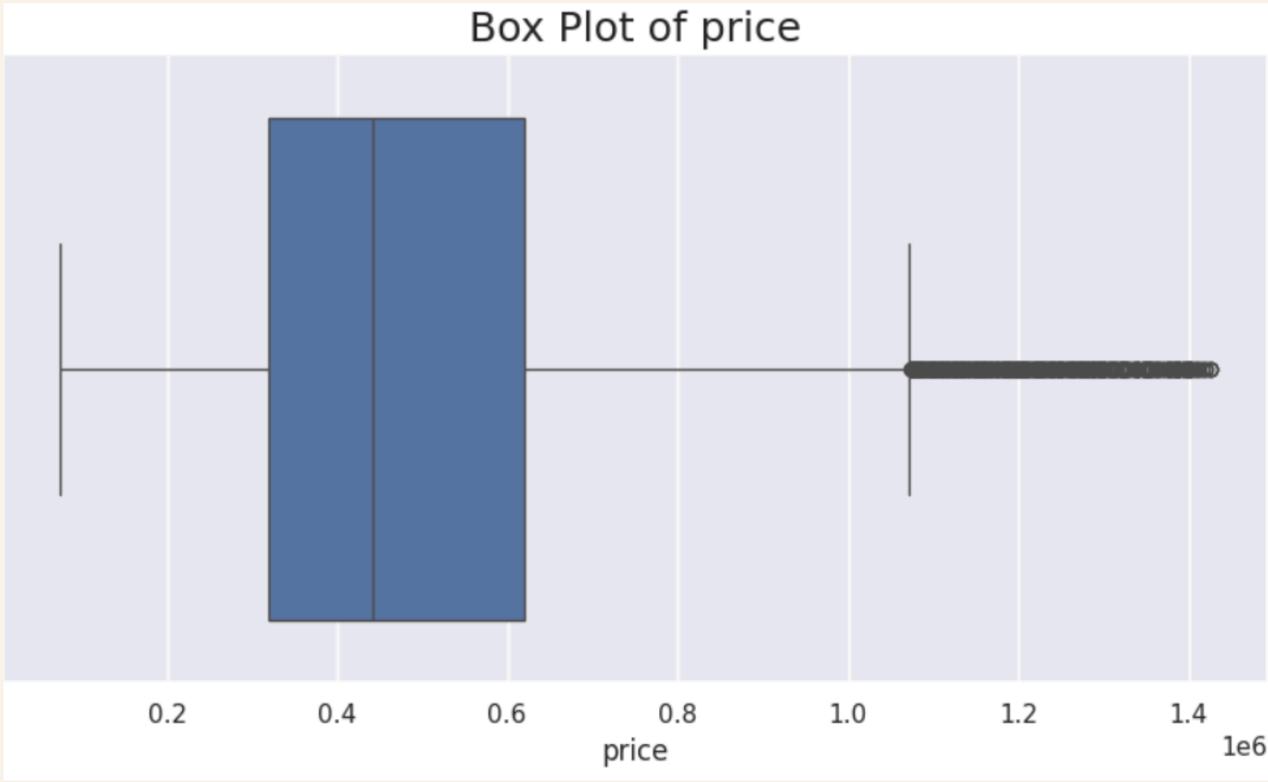
After



Before



After

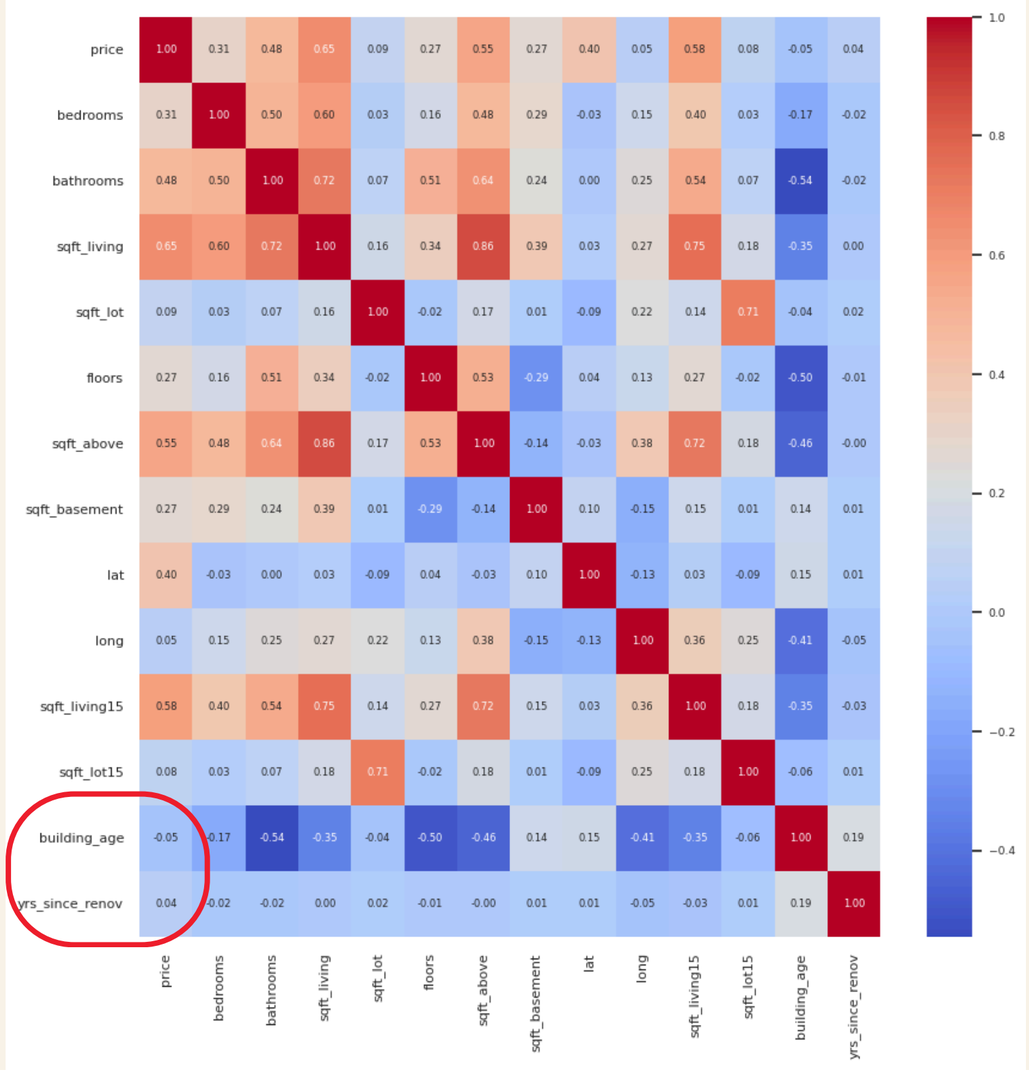


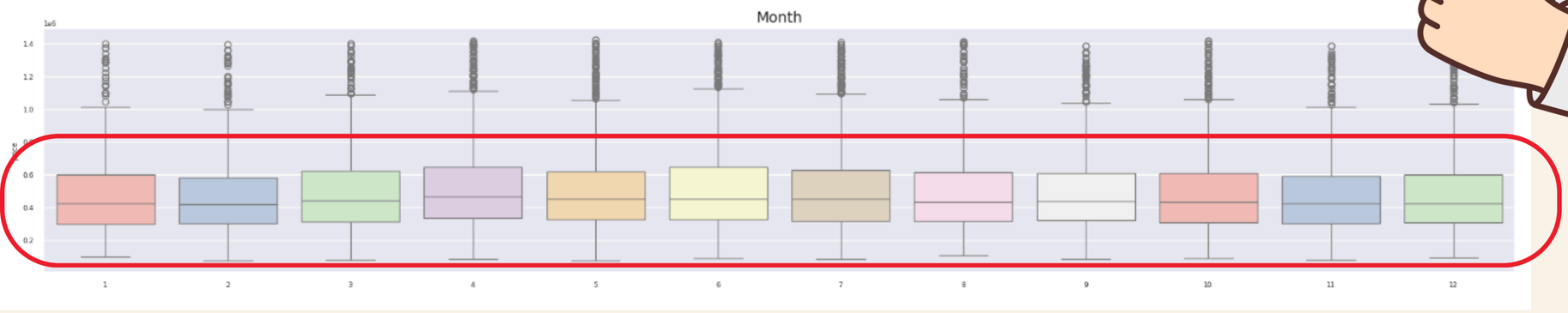
Feature Engineering

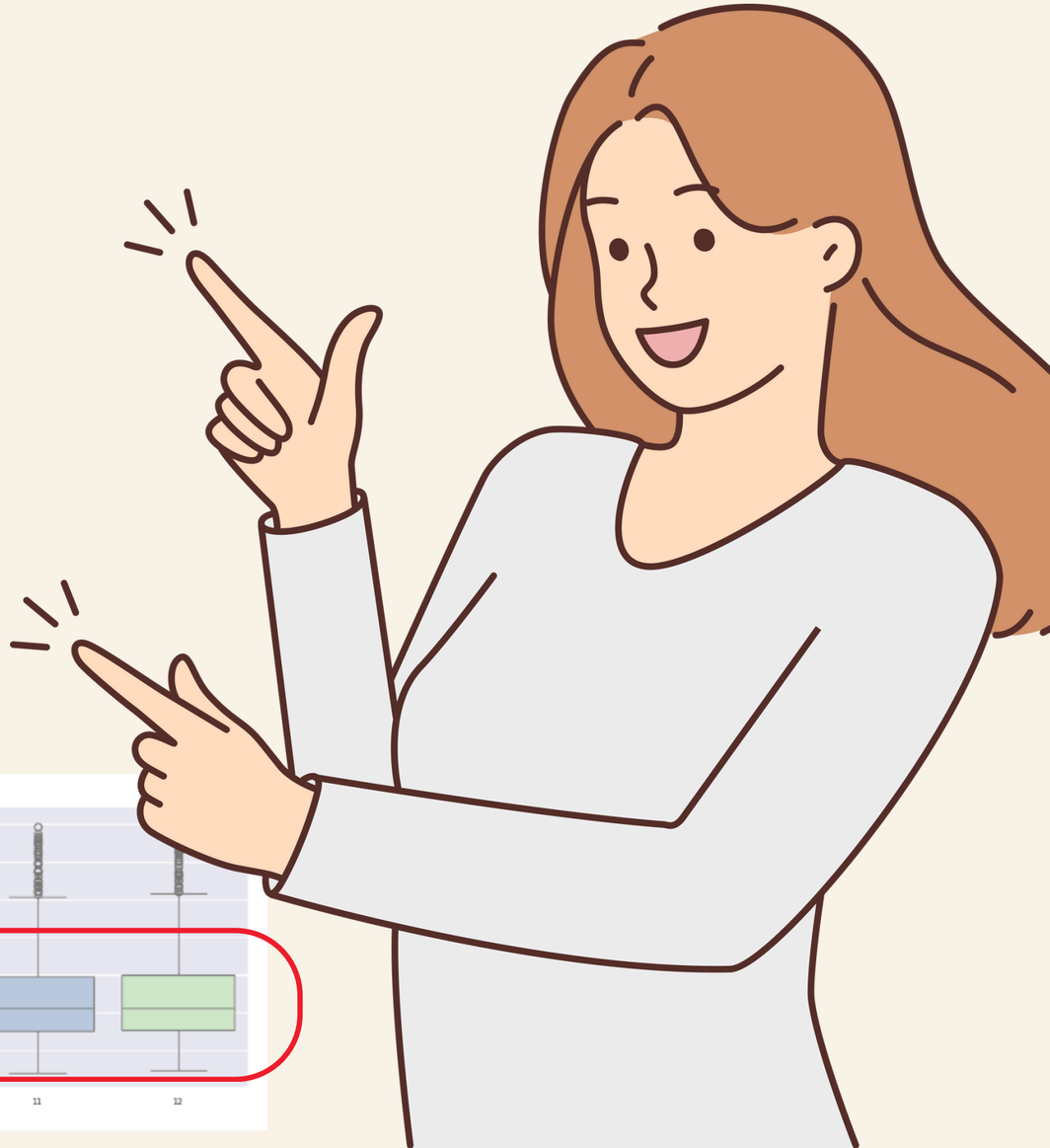
Age of property = Year from “date” - “yr_built”

No. of years since renovated = Year from “date” - “yr_renovated”

Month sold = Month from “date”


price 1.00 0.31 0.48 0.65 0.09 0.27 0.55 0.27 0.40 0.05 0.58 0.08 -0.05 0.04
bedrooms 0.31 1.00 0.50 0.60 0.03 0.16 0.48 0.29 -0.03 0.15 0.40 0.03 -0.17 -0.02
bathrooms 0.48 0.50 1.00 0.72 0.07 0.51 0.64 0.24 0.00 0.25 0.54 0.07 -0.54 -0.02
sqft_living 0.65 0.60 0.72 1.00 0.16 0.34 0.86 0.39 0.03 0.27 0.75 0.18 -0.35 0.00
sqft_lot 0.09 0.03 0.07 0.16 1.00 -0.02 0.17 0.01 -0.09 0.22 0.14 0.71 -0.04 0.02
floors 0.27 0.16 0.51 0.34 -0.02 1.00 0.53 -0.29 0.04 0.13 0.27 -0.02 -0.50 -0.01
sqft_above 0.55 0.48 0.64 0.86 0.17 0.53 1.00 -0.14 -0.03 0.38 0.72 0.18 -0.46 -0.00
sqft_basement 0.27 0.29 0.24 0.39 0.01 -0.29 -0.14 1.00 0.10 -0.15 0.15 0.01 0.14 0.01
lat 0.40 -0.03 0.00 0.03 -0.09 0.04 -0.03 0.10 1.00 -0.13 0.03 -0.09 0.15 0.01
long 0.05 0.15 0.25 0.27 0.22 0.13 0.38 -0.15 -0.13 1.00 0.36 0.25 -0.41 -0.05
sqft_living15 0.58 0.40 0.54 0.75 0.14 0.27 0.72 0.15 0.03 0.36 1.00 0.18 -0.35 -0.03
sqft_lot15 0.08 0.03 0.07 0.18 0.71 -0.02 0.18 0.01 -0.09 0.25 0.18 1.00 -0.06 0.01
building_age -0.05 -0.17 -0.54 -0.35 -0.04 -0.50 -0.46 0.14 0.15 -0.41 -0.35 -0.06 1.00 0.19
yrs_since_renov 0.04 -0.02 -0.02 0.00 0.02 -0.01 -0.00 0.01 0.01 -0.05 -0.03 0.01 0.19 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
sqft_above
sqft_basement
lat
long
sqft_living15
sqft_lot15
building_age
yrs_since_renov


Month
1e6
1.4
1.2
1.0
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9
10
11
12



Feature Engineering

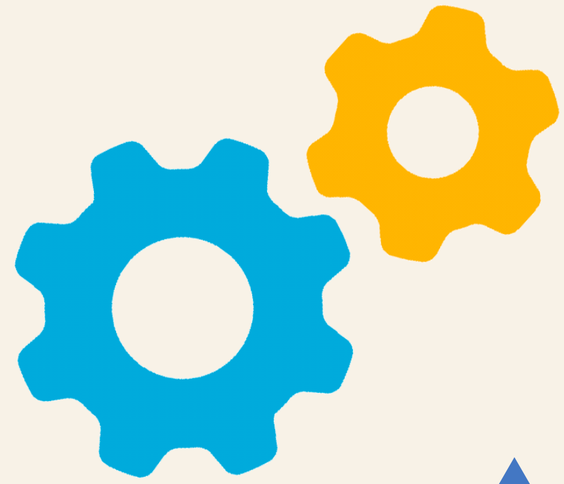
Age of property = Year from “date” - “yr_built”

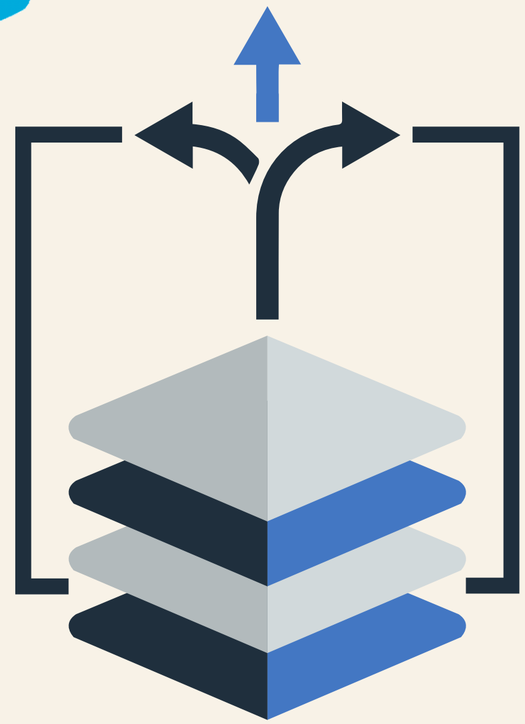
No. of years since renovated = Year from “date” - “yr_renovated”

~~**Month sold** = Month from “date”~~

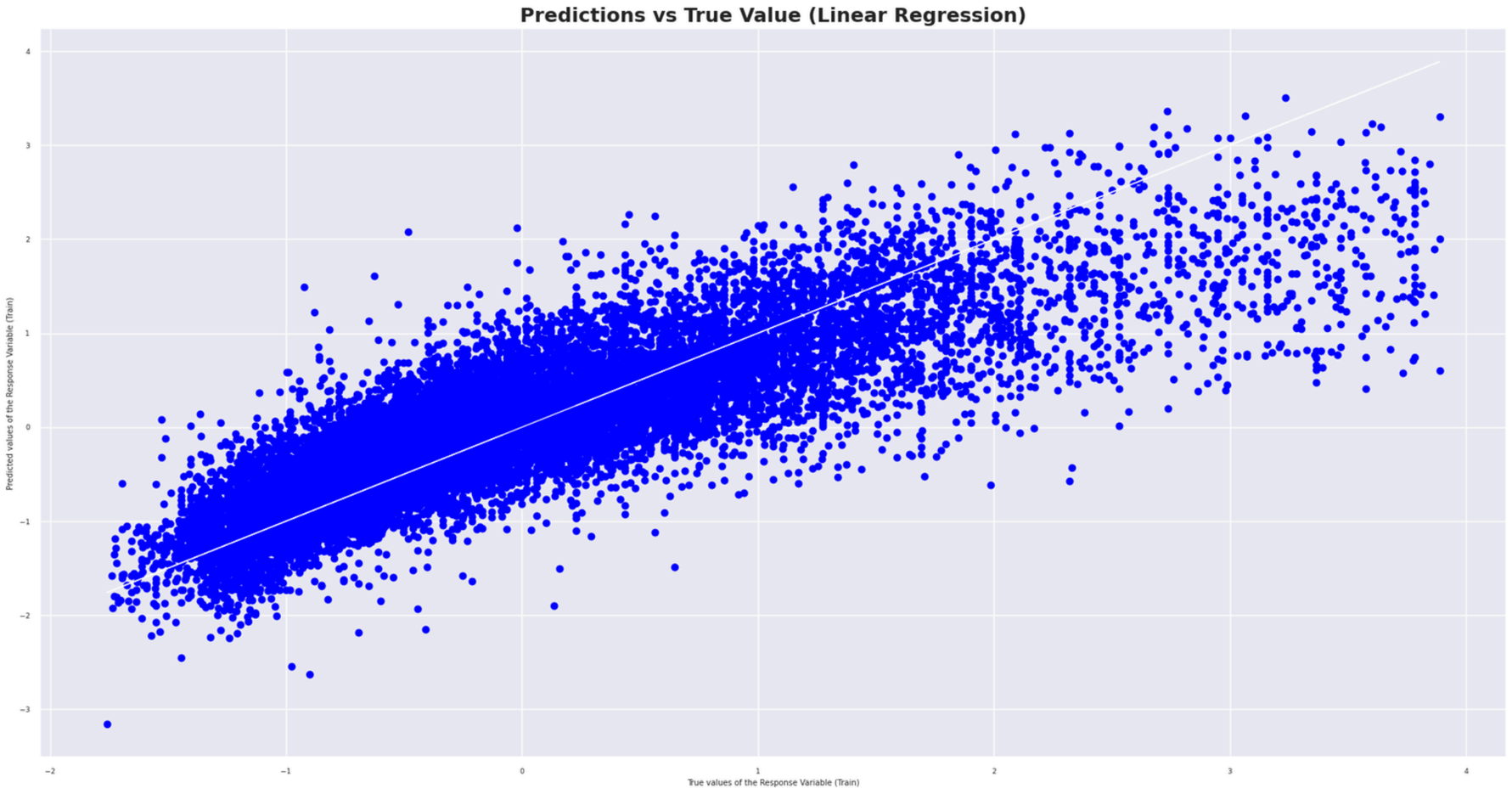
Machine Learning

- Linear Regression
- Support Vector Regressor
- Decision Tree Regressor
- Random Forest Regressor.





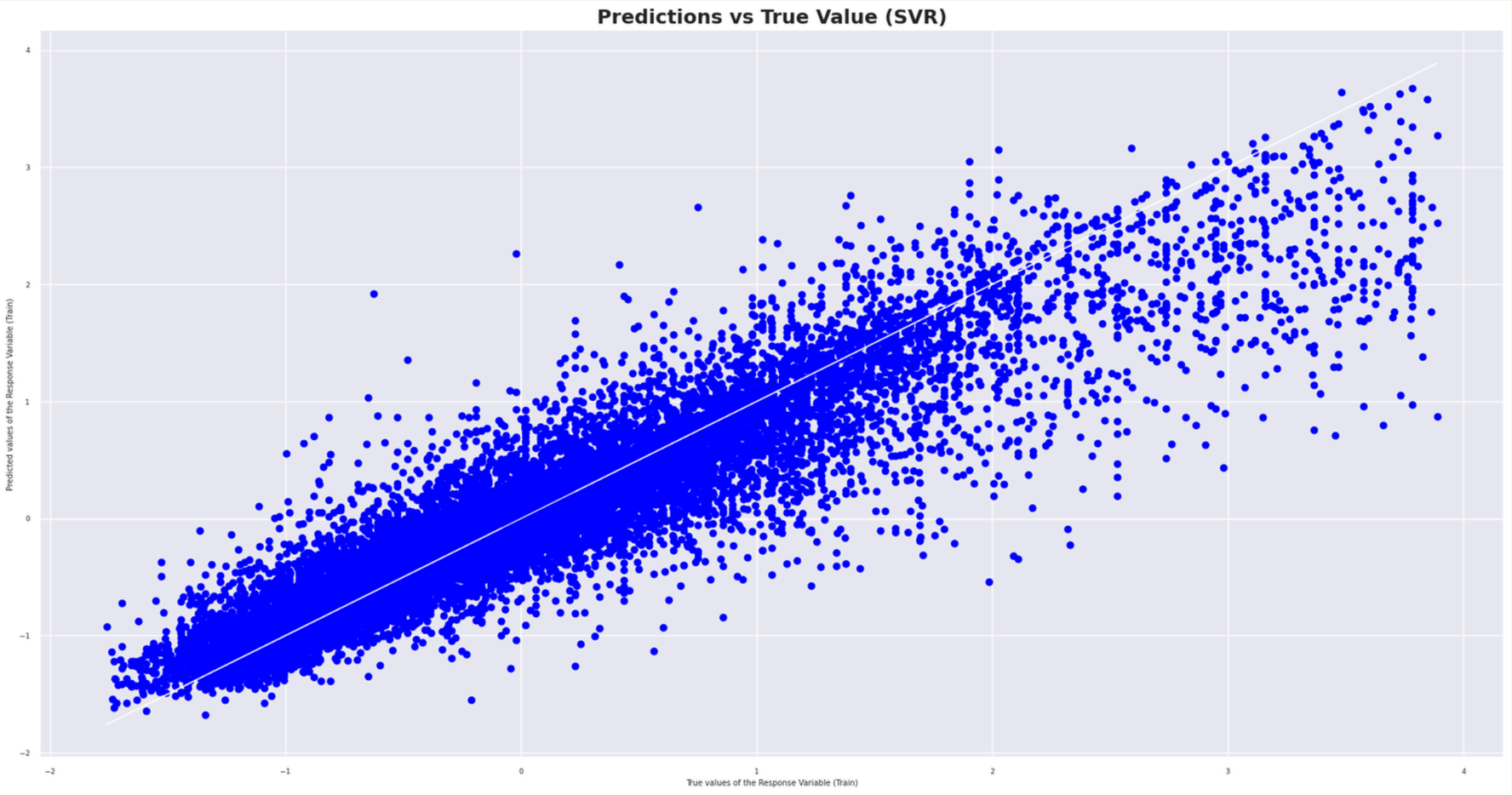
Linear Regression



Training accurancy =0.708

Testing accurancy = 0.697

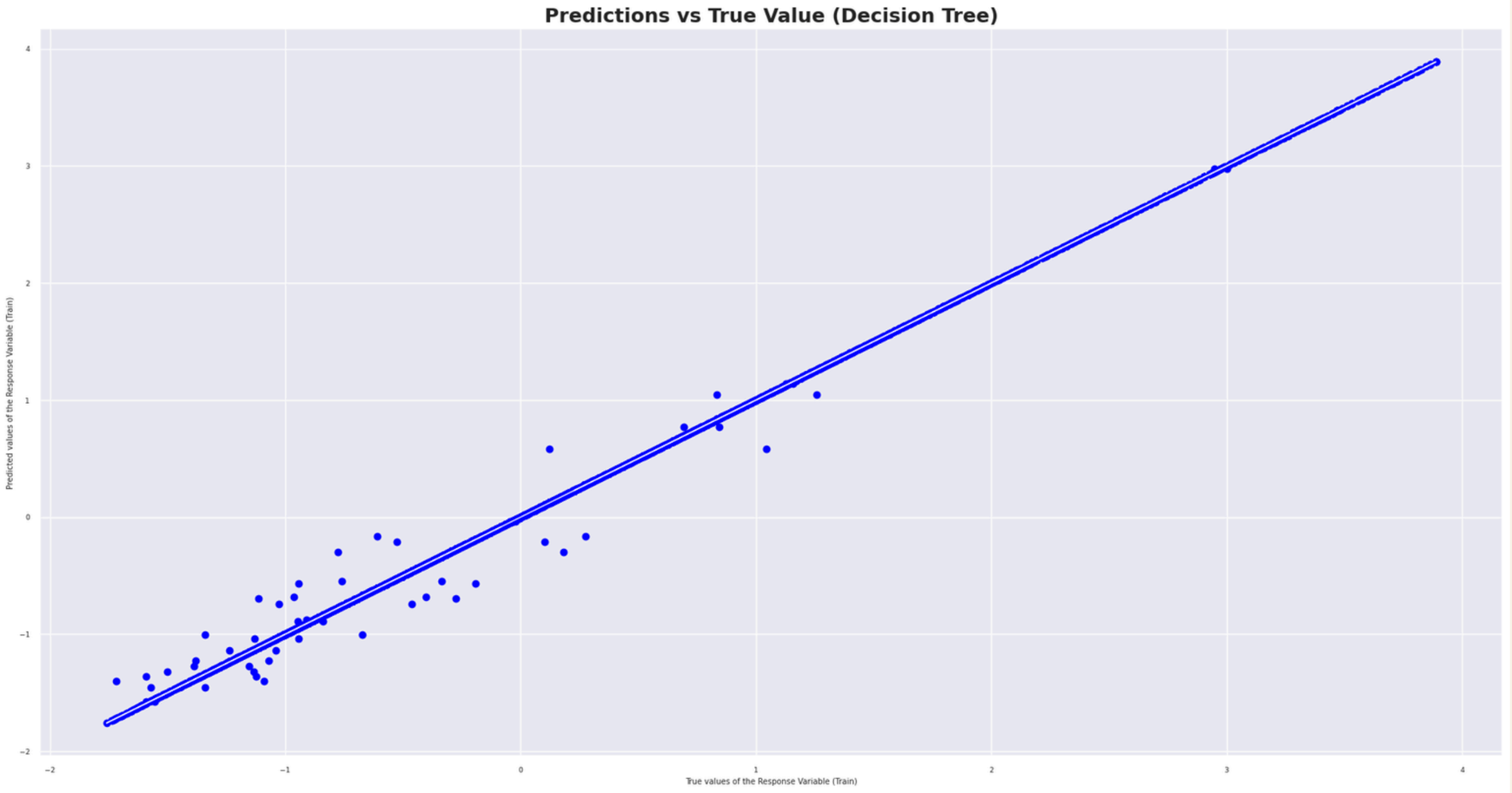
Support Vector Regressor



Training accurancy =0.848

Testing accurancy = 0.808

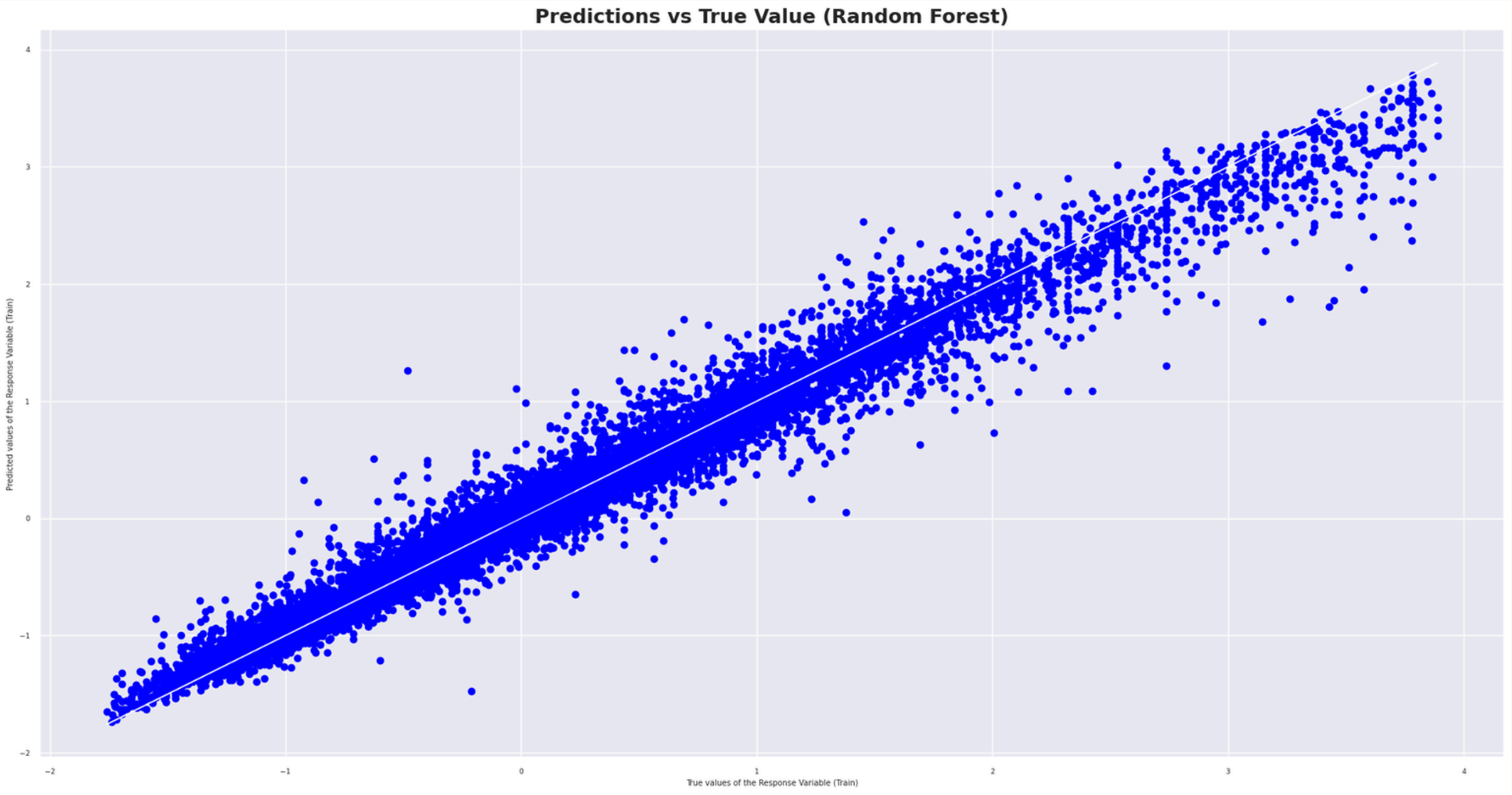
Decision Tree Regressor



Training accurancy =0.999

Testing accurancy = 0.719

Random Forest Regressor



Training accurancy =0.972

Testing accurancy = 0.845

R-squared

Mean squared error

Mean absolute percentage error

	Accuracy (R^2 score)	MSE	MAPE
Linear Regression	0.6967	0.3031	203.3716
SVR	0.8080	0.1919	149.7679
Decision Tree	0.7189	0.2810	243.3641
Random Forest	0.8453	0.1546	157.5226

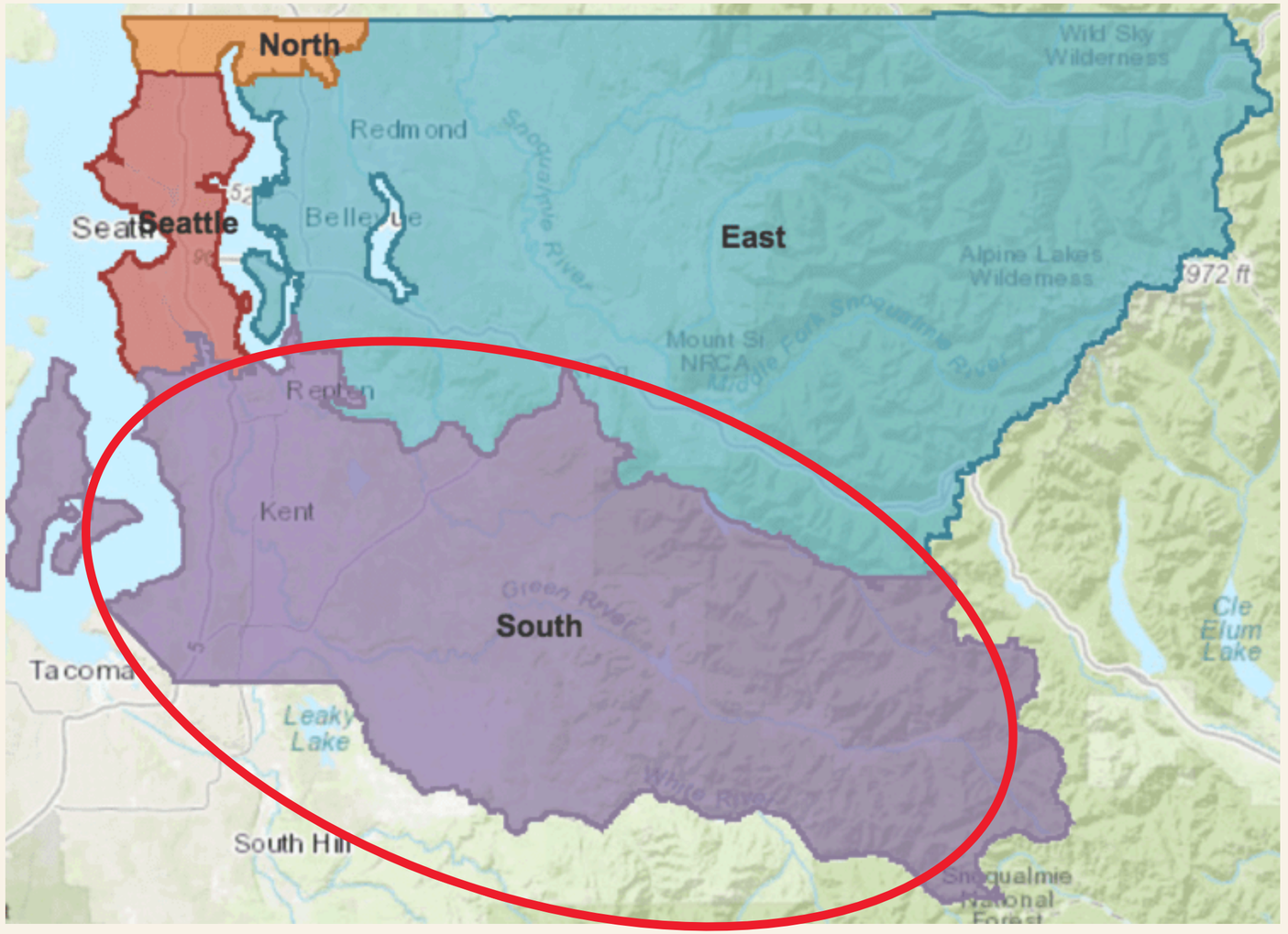
	Training Error	Testing Error
SVR	0.1515	0.1919
Random Forest	0.0275	0.1546

```
Random Forest Validation R^2: 0.8507727698622572
Random Forest Validation MSE: 0.15209237961780145
Random Forest Testing R^2: 0.8676280784009862
Random Forest Testing MSE: 0.13410534487871575
```

Random Forest Regressor

Insights and recommendations

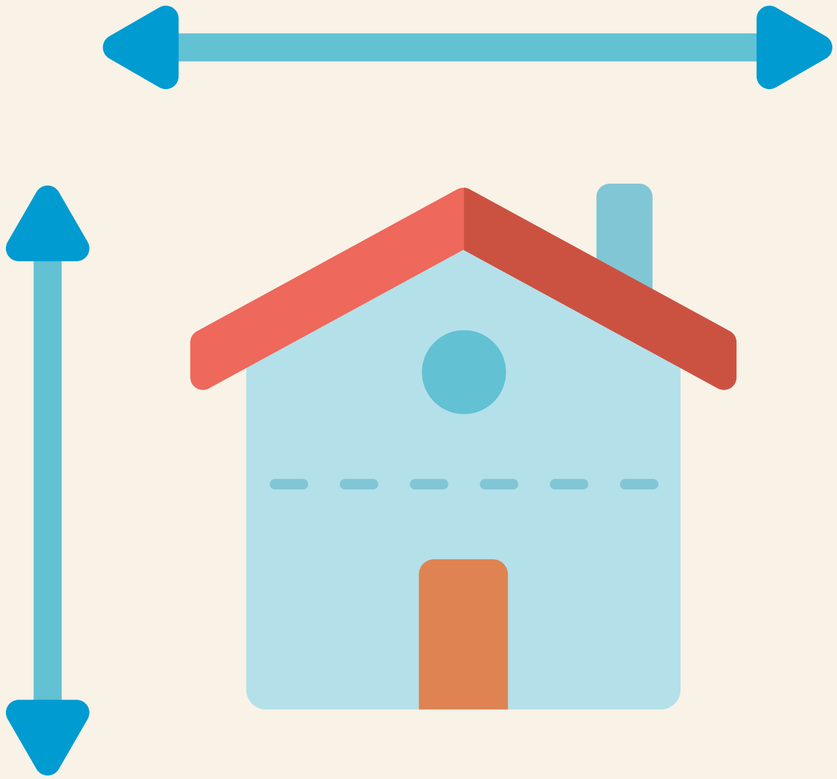

```
Feature 1: grade, Score: 0.3406
Feature 2: lat, Score: 0.3008
Feature 3: sqft_living, Score: 0.1613
Feature 4: long, Score: 0.0509
Feature 5: sqft_living15, Score: 0.0304
Feature 6: building_age, Score: 0.0298
Feature 7: sqft_above, Score: 0.0168
Feature 8: sqft_lot, Score: 0.0168
Feature 9: view, Score: 0.0155
Feature 10: sqft_lot15, Score: 0.0153
```

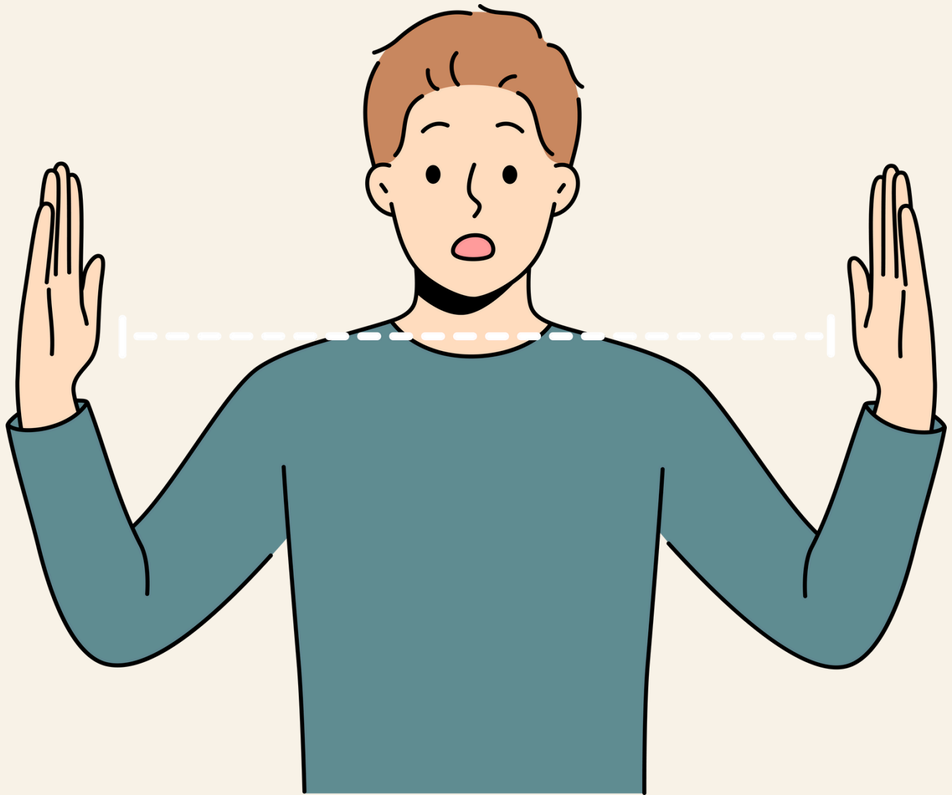

North
Seattle
East
South
Redmond
Bellevue
Renton
Kent
Tacoma
Leaky Lake
South Hill
Wild Sky Wilderness
Alpine Lakes Wilderness
Mount Si NRCA
Cle Elum Lake
Green River
White River
Snoqualmie National Forest
Snoqualmie River
Middle Fork Snoqualmie River
972 ft


N
W
E
S

373818.01

(Predicted by best model)






THANKS