HOUSE SALE PRICES PREDICTION

Nopasorn Kowathanakul

THE PROJECT

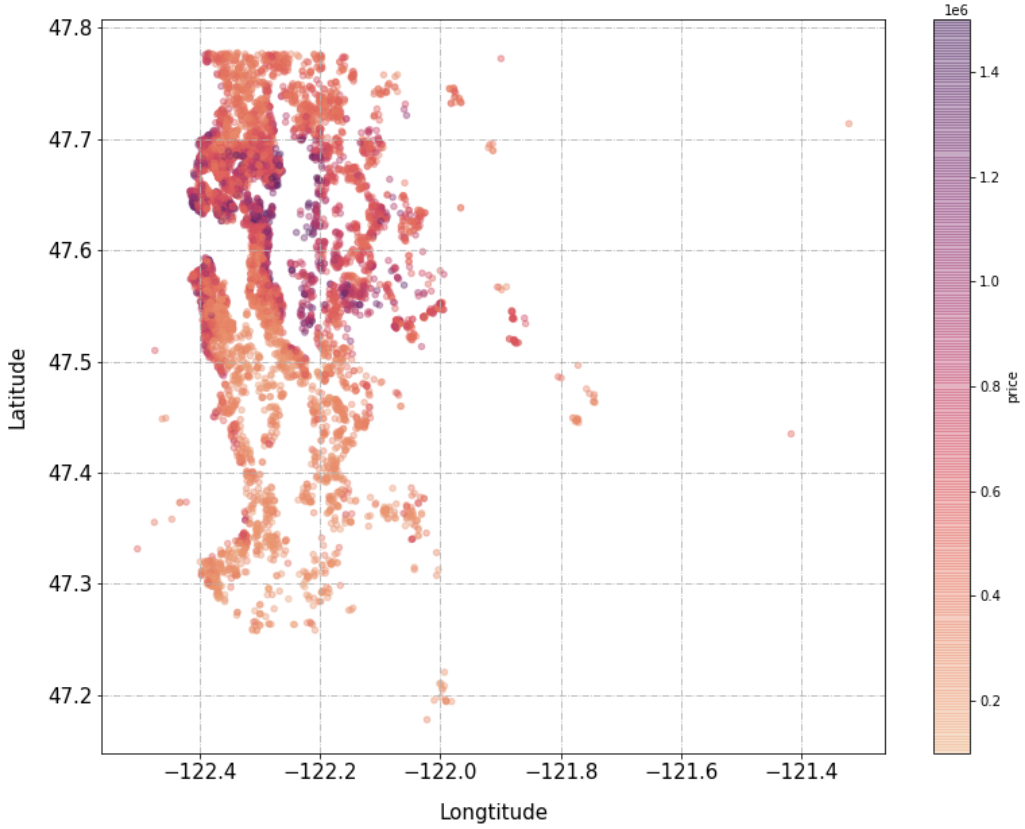
From the sales data in King County, I will do the data analysis and select influence factors for making a prediction model. Then, the following questions will be answered:

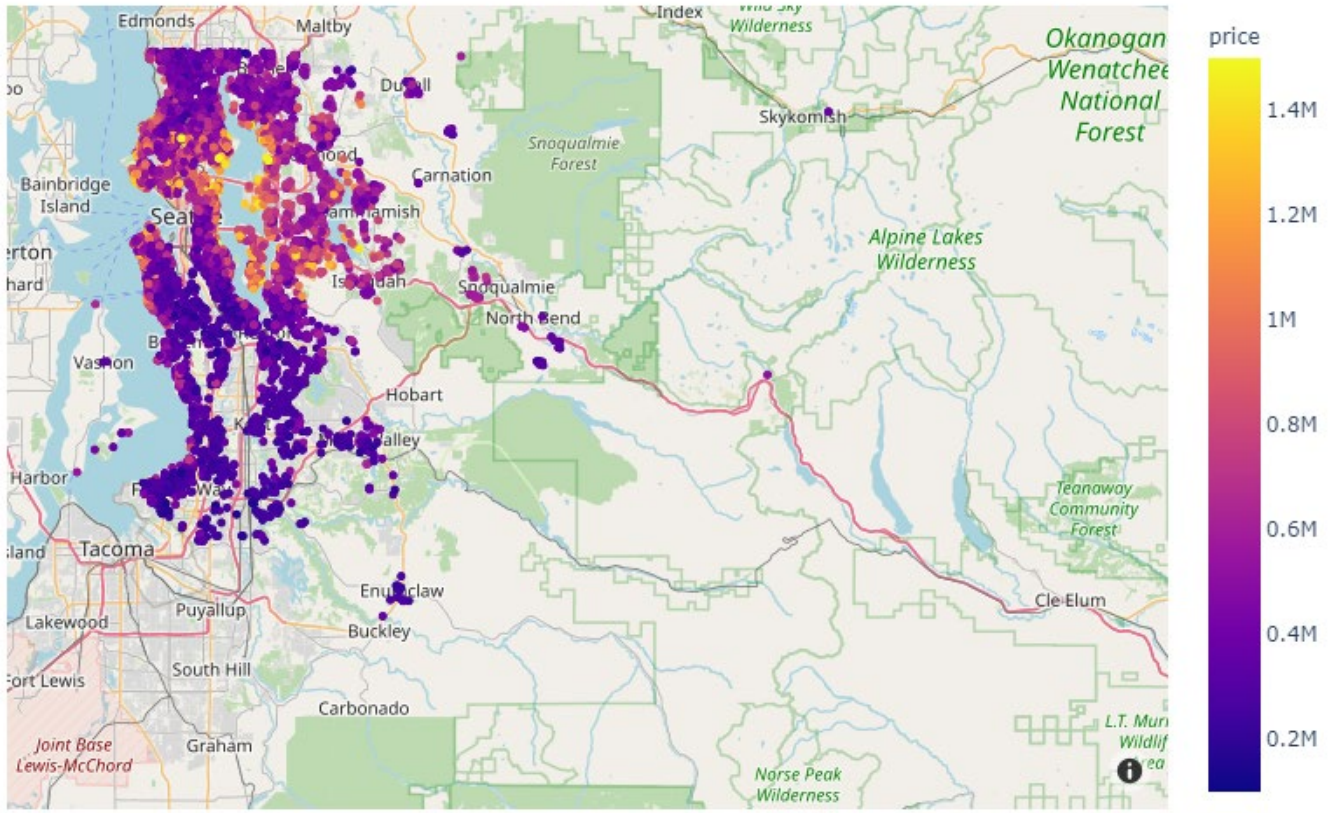
1. How does the location affect the price?
2. Do built and renovated times affect the price?
3. Does selling time affect the price?
4. Can we predict the price from a condition?

King County House Sales dataset

ID - UNIQUE IDENTIFIED FOR A HOUSE
DATE - DATE HOUSE WAS SOLD
PRICE - PRICE IS PREDICTION TARGET
BEDROOMS - NUMBER OF BEDROOMS/HOUSE
BATHROOMS - NUMBER OF BATHROOMS/BEDROOMS
SQFT_LIVING - SQUARE FOOTAGE OF THE HOME
SQFT_LOT - SQUARE FOOTAGE OF THE LOT
FLOORS - TOTAL FLOORS (LEVELS) IN HOUSE
WATERFRONT - HOUSE WHICH HAS A VIEW TO A WATERFRONT
VIEW - HAS BEEN VIEWED
CONDITION - HOW GOOD THE CONDITION IS (OVERALL)
GRADE - OVERALL GRADE GIVEN TO THE HOUSING UNIT, BASED ON KING COUNTY GRADING SYSTEM
SQFT_ABOVE - SQUARE FOOTAGE OF HOUSE APART FROM BASEMENT
SQFT_BASEMENT - SQUARE FOOTAGE OF THE BASEMENT
YR_BUILT - BUILT YEAR
YR_RENOVATED - YEAR WHEN HOUSE WAS RENOVATED
ZIPCODE - ZIP
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

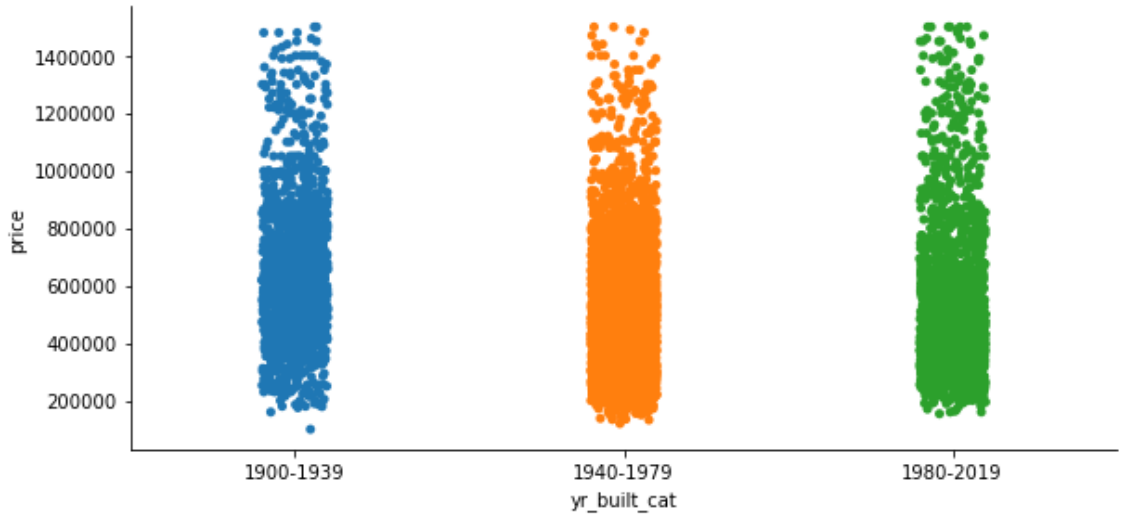
LOCATION EFFECT



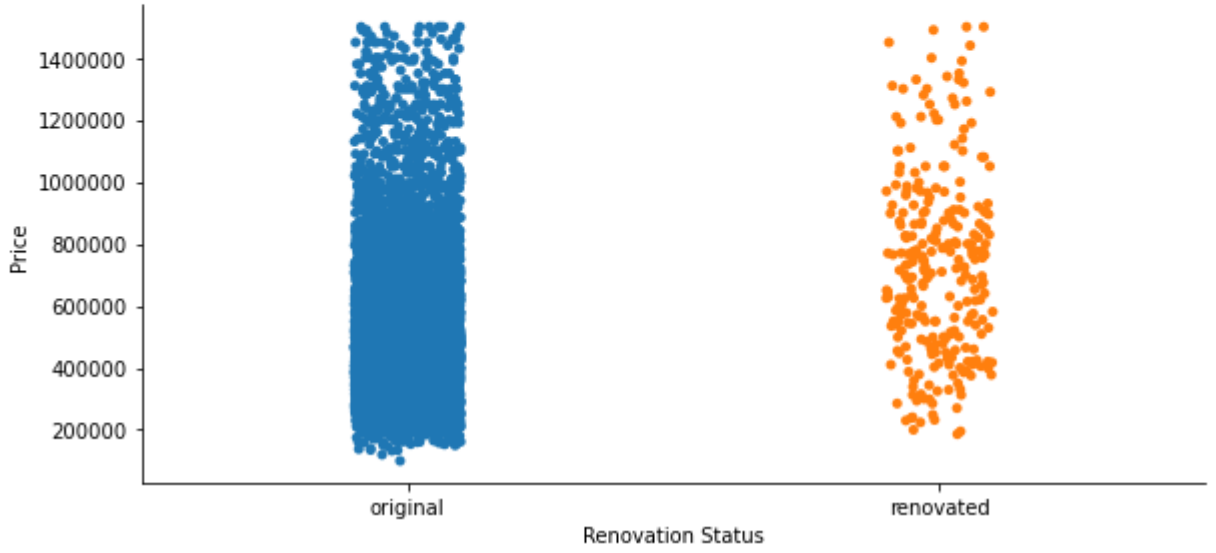


BUILT AND RENOVATED TIMES

```
Mean prices for yr_built_cat
1900-1939    639911.185238
1940-1979    503076.052348
1980-2019    538176.536572
Name: price, dtype: float64

Median prices for yr_built_cat
1900-1939    597000.0
1940-1979    459975.0
1980-2019    475000.0
Name: price, dtype: float64
```

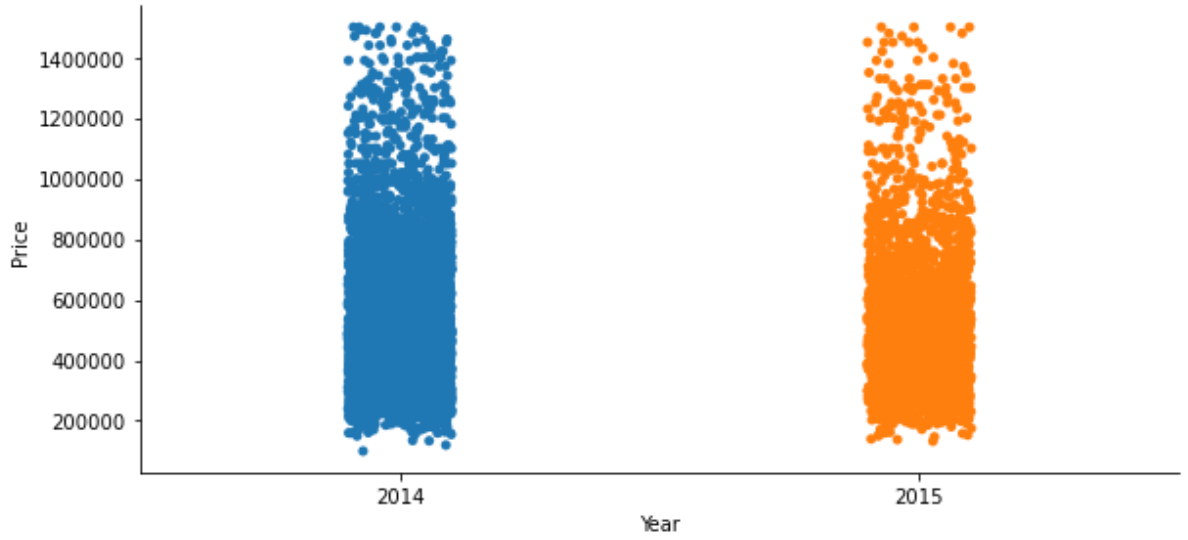



```
Mean prices for renovated
original     532330.073109
renovated    731288.941176
Name: price, dtype: float64

Median prices for renovated
original     484000.0
renovated    721000.0
Name: price, dtype: float64
```

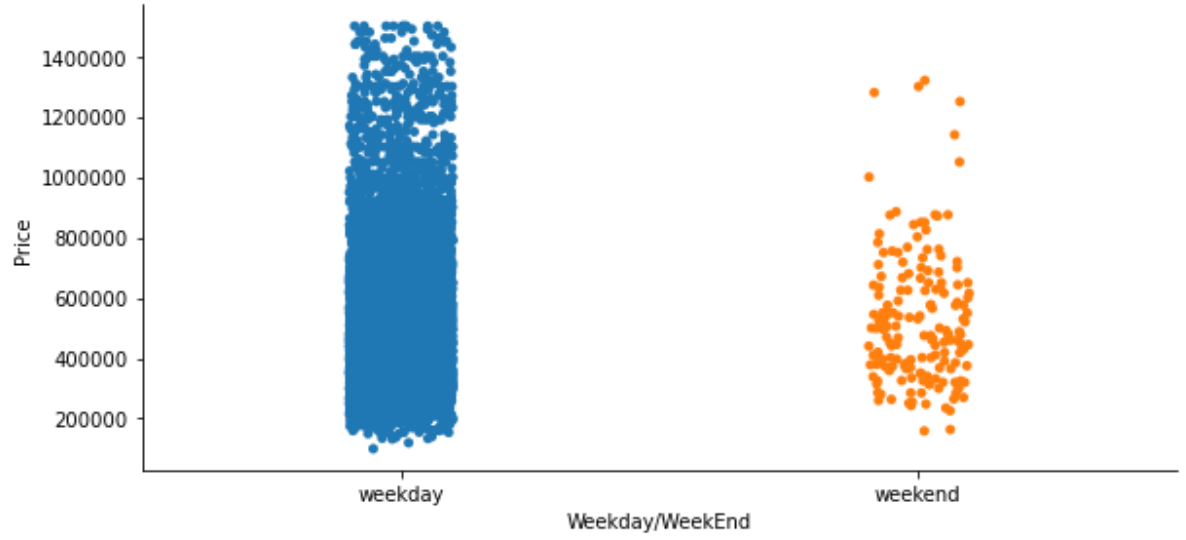


SELLING TIME

```
Mean prices for year
2014    538130.161706
2015    544165.609507
Name: price, dtype: float64

Median prices for year
2014    489000.0
2015    499000.0
Name: price, dtype: float64
```




```
Mean prices for day_cat
weekday    540430.406543
weekend    525963.441860
Name: price, dtype: float64

Median prices for day_cat
weekday    490000.0
weekend    484000.0
Name: price, dtype: float64
```



PRICE PREDICTION

Est.Price = 0.2827x sqft_living + 0.4312x grade + 0.2161x condition + 0.0614x floors + 0.3554x lat + 0.2396x waterfront - 0.1997

OLS Regression Results

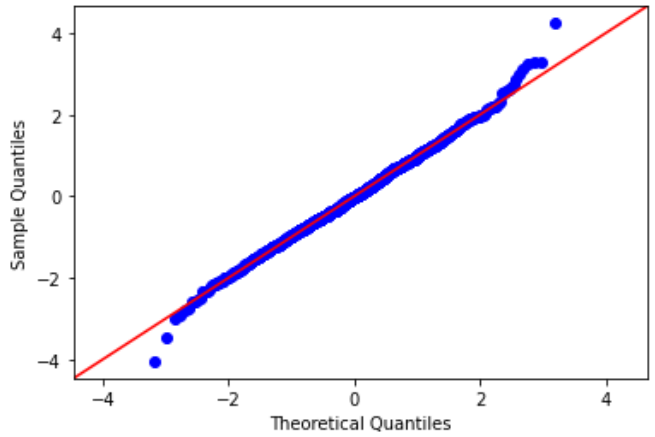
Dep. Variable:	price	R-squared:	0.634
Model:	OLS	Adj. R-squared:	0.634
Method:	Least Squares	F-statistic:	1615.
Date:	Fri, 28 May 2021	Prob (F-statistic):	0.00
Time:	11:14:10	Log-Likelihood:	5049.0
No. Observations:	5590	AIC:	-1.008e+04
Df Residuals:	5583	BIC:	-1.004e+04
Df Model:	6		
Covariance Type:	nonrobust		

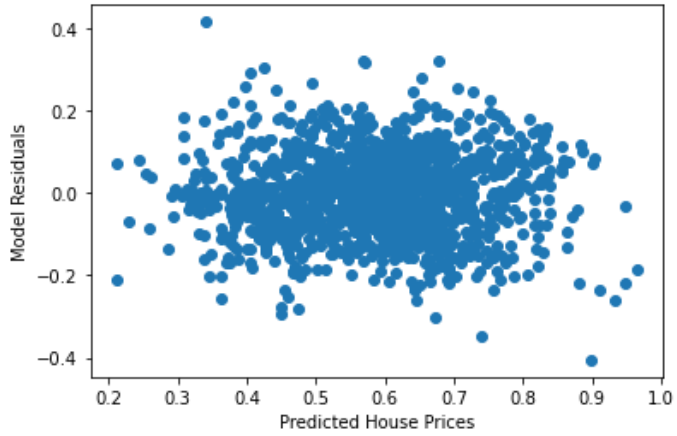
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1997	0.011	-17.959	0.000	-0.221	-0.178
sqft_living	0.2827	0.010	29.752	0.000	0.264	0.301
grade	0.4312	0.013	32.138	0.000	0.405	0.457
condition	0.2161	0.012	18.526	0.000	0.193	0.239
floors	0.0614	0.005	11.786	0.000	0.051	0.072
lat	0.3554	0.007	53.562	0.000	0.342	0.368
waterfront	0.2396	0.031	7.707	0.000	0.179	0.301

Omnibus:	6.964	Durbin-Watson:	1.974
Prob(Omnibus):	0.031	Jarque-Bera (JB):	6.923
Skew:	0.081	Prob(JB):	0.0314
Kurtosis:	3.060	Cond. No.	38.8

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

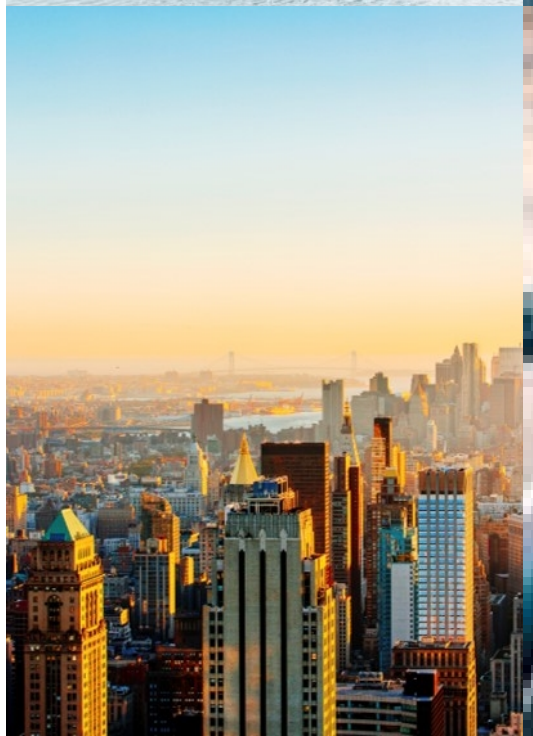
Train Mean Squarred Error: 0.009616167526054365
Test Mean Squarred Error: 0.009811975982764534

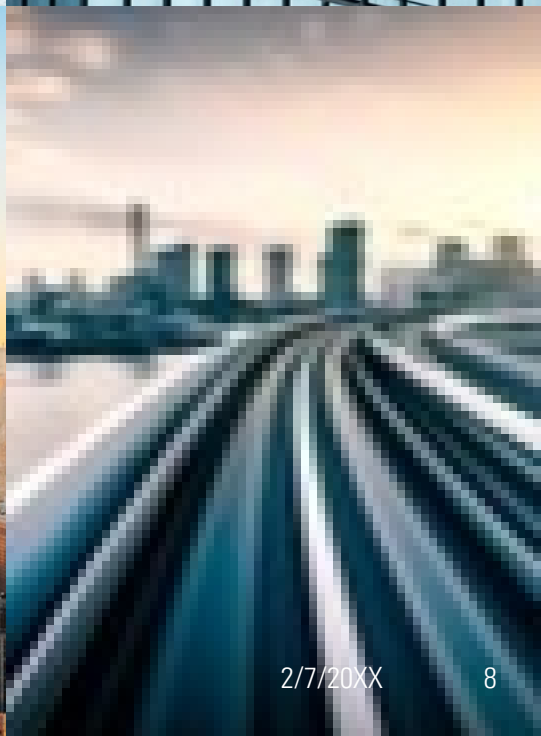


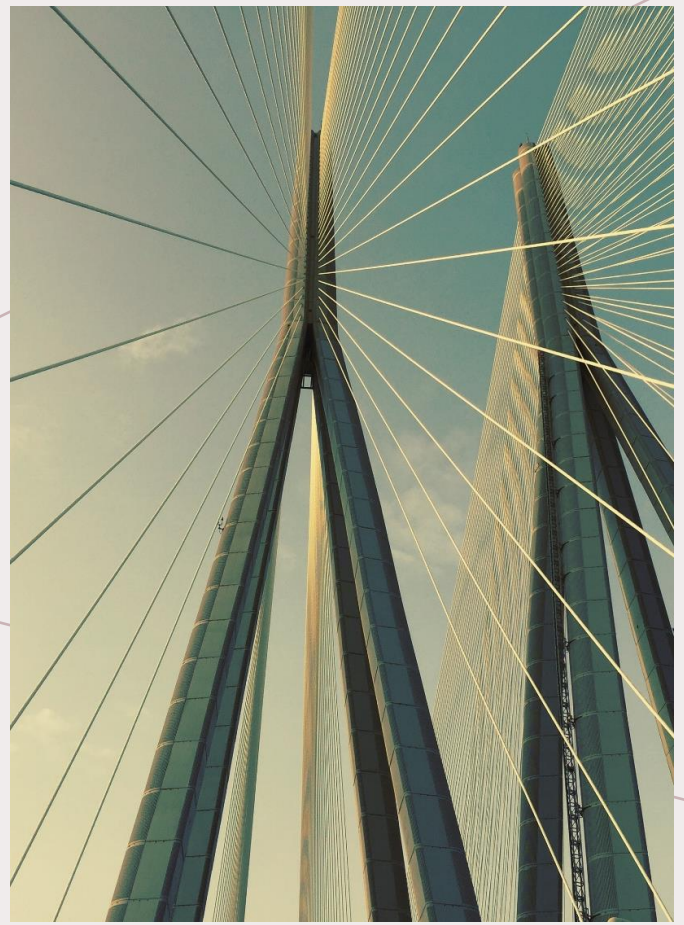


FUTURE PLAN

For the future, I should include more detail about location, such as downtown, shopping mall, supermarket or public transportation, to increase the accuracy of prediction.







THANK YOU