HOUSE SALE PRICES PREDICTION

Nopasorn Kowathanakul

THE PROJECT

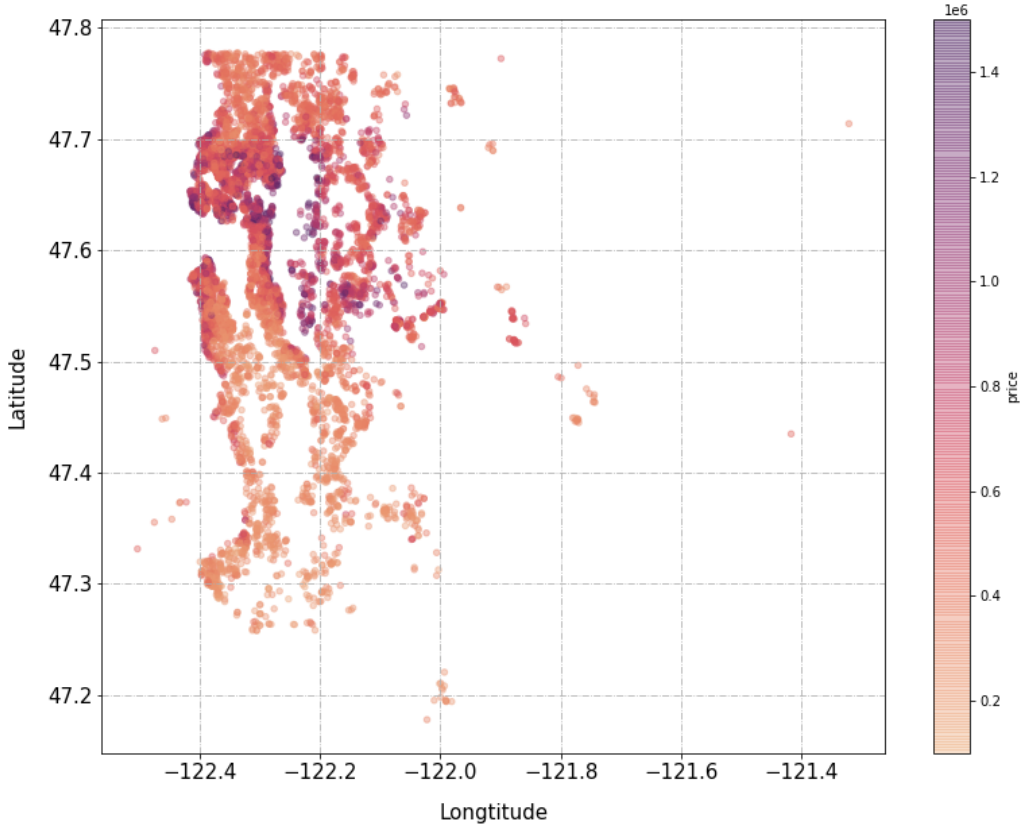
From the sales data in King County, I will do the data analysis and select influence factors for making a prediction model. Then, the following questions will be answered:

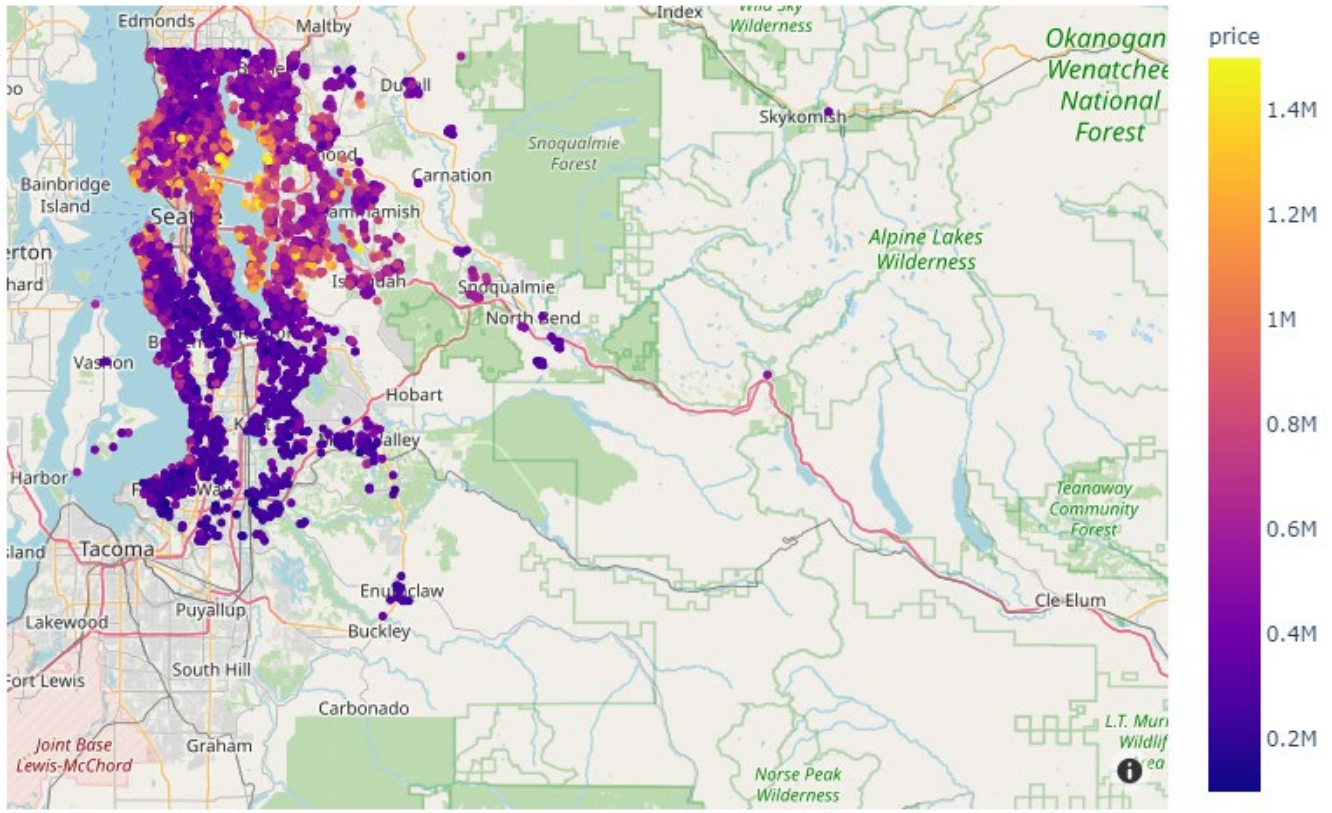
1. How does the location affect the price?
2. Do built and renovated times affect the price?
3. Does selling time affect the price?
4. Can we predict the price from a condition?

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

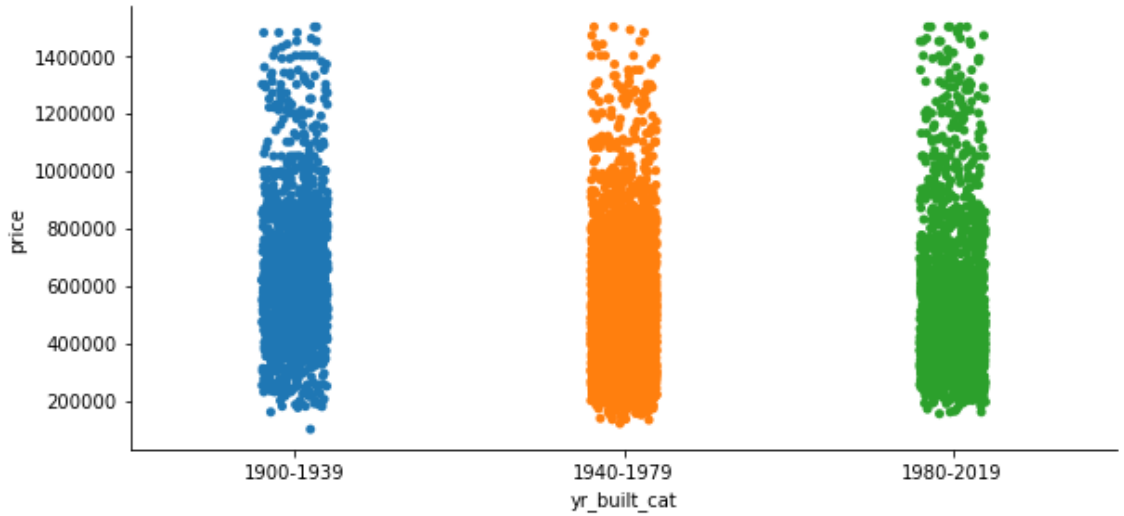
LOCATION EFFECT



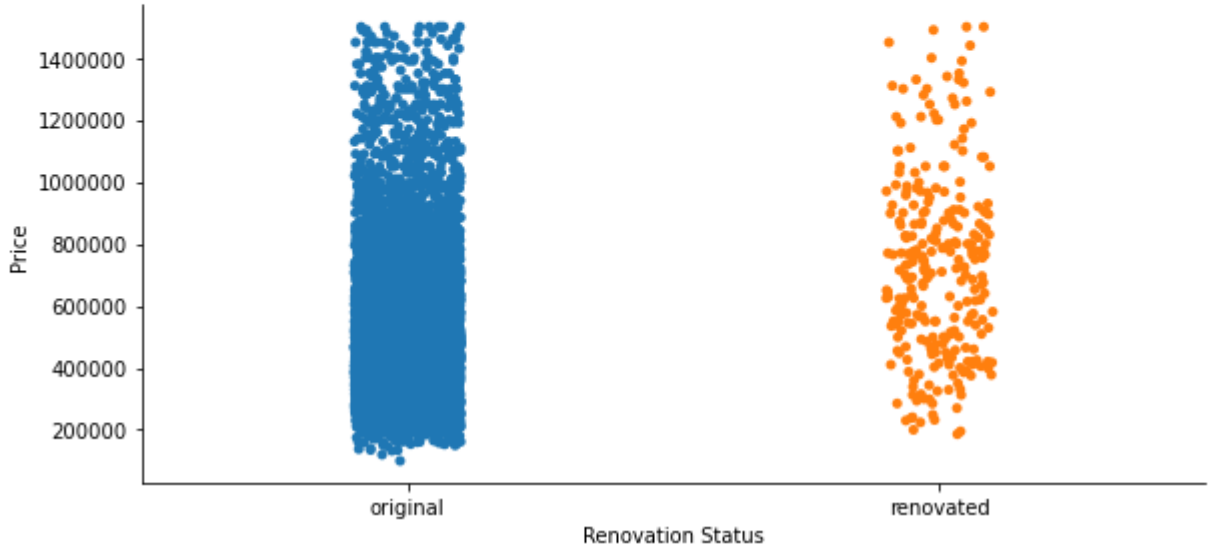


BUILT AND RENOVATED TIMES

```
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
```

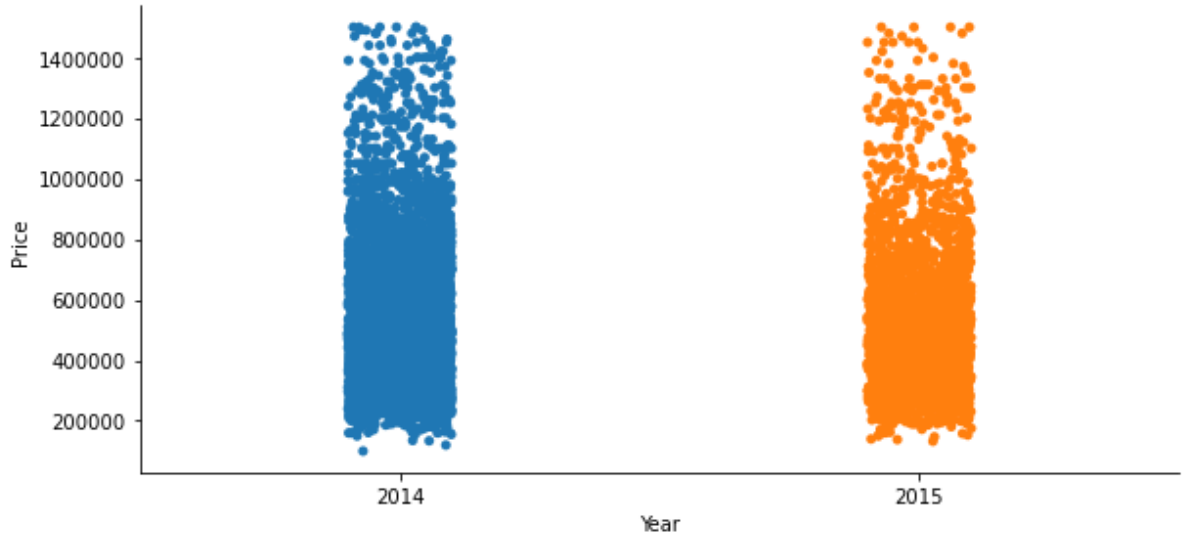



```
Mean prices for renovated
original     532330.073109
renovated    731288.941176
Name: price, dtype: float64

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

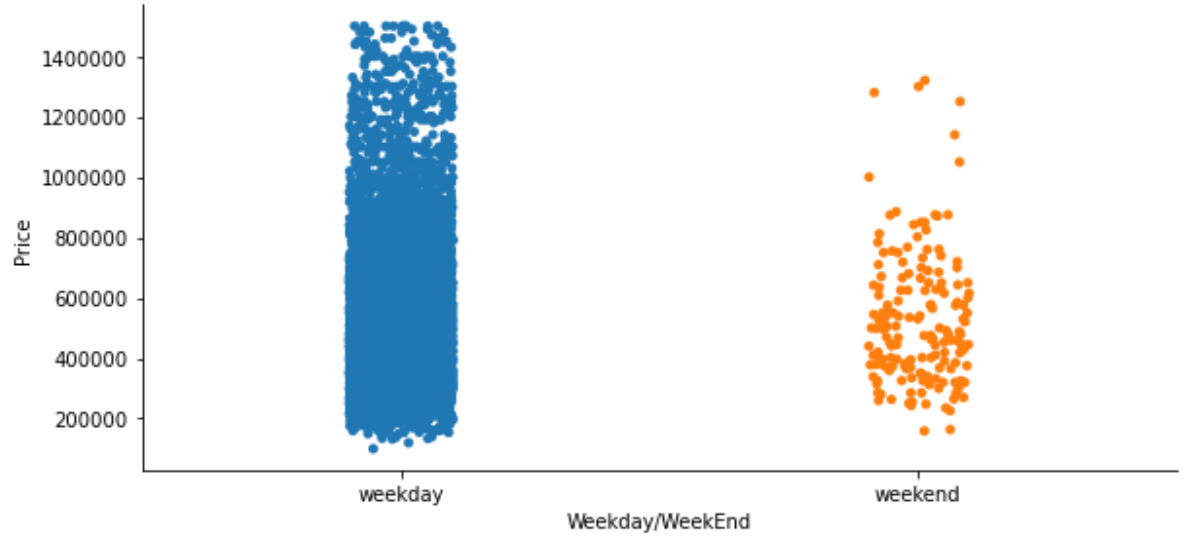


SELLING TIME

```
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 = 0.2729x sqft_living + 0.4687x grade + 0.2516x condition +0.0877x floors

OLS Regression Results

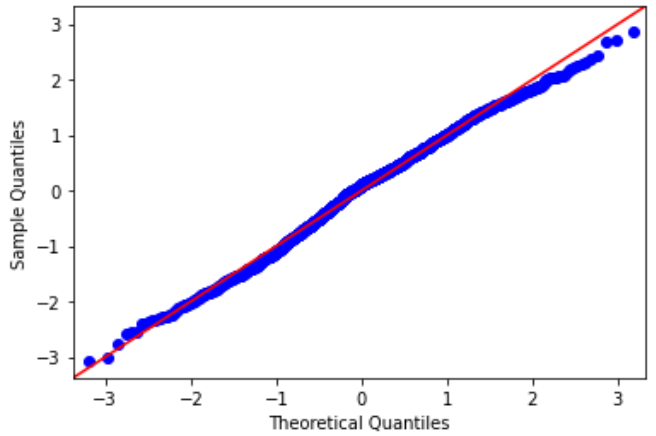
Dep. Variable:	price	R-squared (uncentered):	0.961
Model:	OLS	Adj. R-squared (uncentered):	0.961
Method:	Least Squares	F-statistic:	3.402e+04
Date:	Thu, 27 May 2021	Prob (F-statistic):	0.00
Time:	12:51:06	Log-Likelihood:	3864.7
No. Observations:	5590	AIC:	-7721.
Df Residuals:	5586	BIC:	-7695.
Df Model:	4		
Covariance Type:	nonrobust		

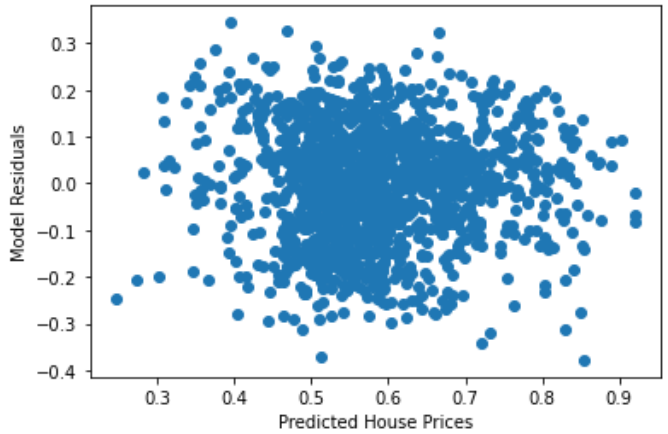
	coef	std err	t	P>|t|	[0.025	0.975]
sqft_living	0.2729	0.012	23.421	0.000	0.250	0.296
grade	0.4687	0.015	30.641	0.000	0.439	0.499
condition	0.2516	0.008	33.200	0.000	0.237	0.266
floors	0.0877	0.006	13.693	0.000	0.075	0.100

Omnibus:	37.982	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36.015
Skew:	-0.167	Prob(JB):	1.51e-08
Kurtosis:	2.791	Cond. No.	12.0

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

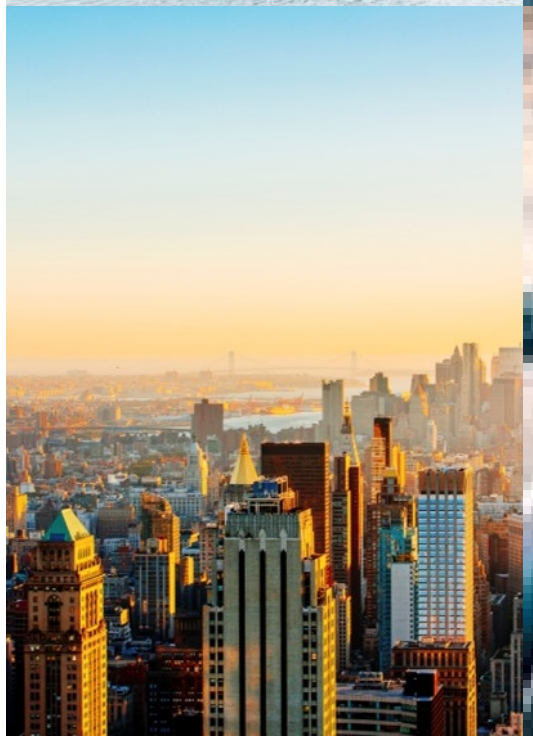
Train Mean Squarred Error: 0.014690038703237523
Test Mean Squarred Error: 0.014832862439659537

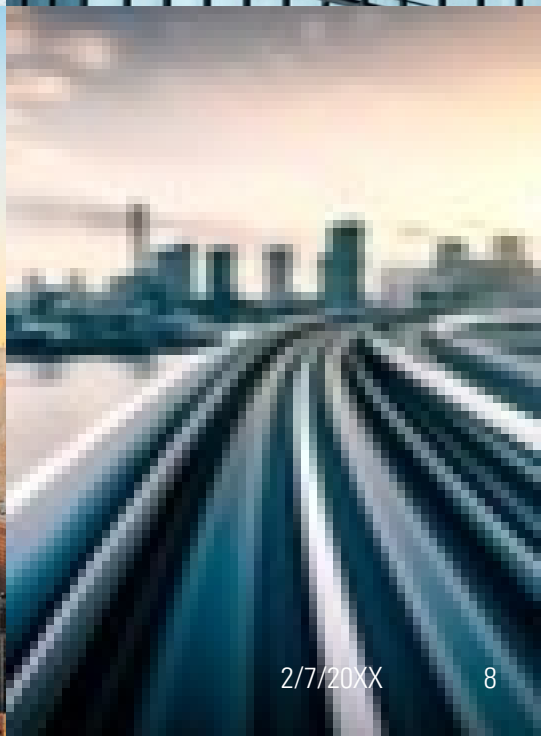


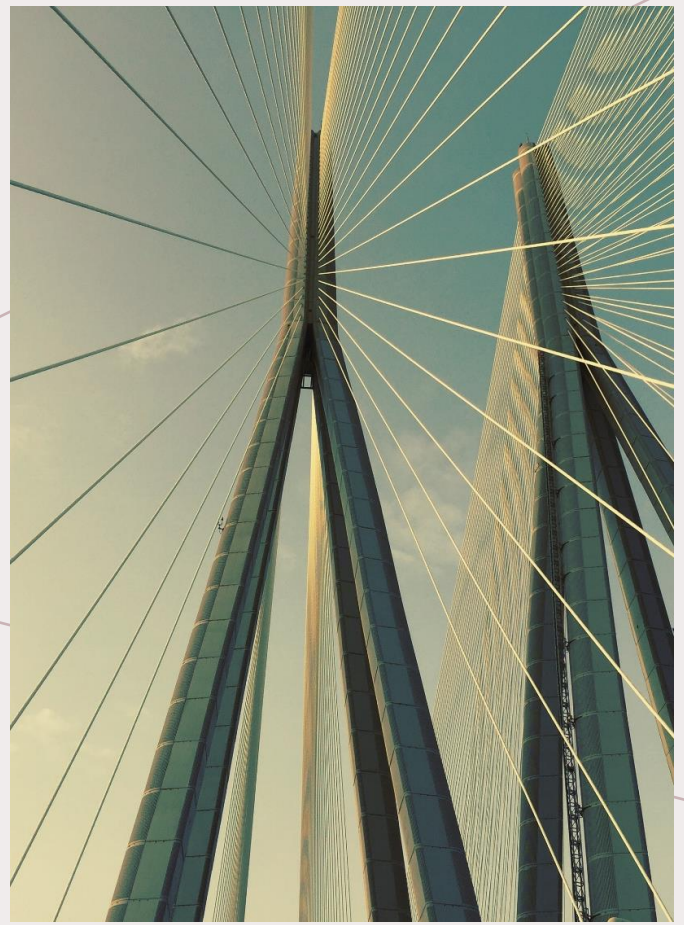


FUTURE PLAN

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