



Phase 2 Project Presentation

NAME: PETER KIGOTHU WAITI

GROUP: DSFP COHORT 15

~Moringa School



HOUSE SALES PREDICTION IN KING COUNTY USING REGRESSION MODELLING

Business Problem

- A real estate Agency need to provide prospective home sellers with guidance on how to improve the value of their home prior to listing, including the predicted increase in value expected based on improvements to particular features(renovation).

Business Question?

- What features of their home can prospective home sellers change or improve to increase the value of their home, and by amount could this increase be specific to certain features?



OBJECTIVES



- Import the required libraries
- Load the given data
- Inspect the data
- Perform data cleaning
- Begin regression modelling
- Ask relevant questions that need to be answered in the form of visualizations.
- Derive conclusions and recommendations.

DATA USED

- This project uses the King County House Sales dataset.
 - The dataset contains information about the sale of each house as well as the number of predictor variables.
-
- | | |
|--|---|
| ❖ price - Sale price (prediction target) | ❖ condition - How good the overall condition of the house is. Related to maintenance of house |
| ❖ date - Date house was sold | ❖ grade - Overall grade of the house. Related to the construction and design of the house |
| ❖ bedrooms - Number of bedrooms | ❖ Sqft-above - Square footage of house apart from basement |
| ❖ bathrooms - Number of bathrooms | ❖ Sqft-basement - Square footage of the basement |
| ❖ Sqft-living - Square footage of living space in the home | ❖ Yr.-built - Year when house was built |
| ❖ Sqft-lot - Square footage of the lot | ❖ Yr.-renovated - Year when house was renovated |
| ❖ floors - Number of floors (levels) in house | ❖ zip code - ZIP Code used by the United States Postal Service |
| ❖ waterfront - Whether the house is on a waterfront | |
| ❖ view - Quality of view from house | |



Data Cleaning

- Ensuring that each column has the correct data type.
- Check the percentage of missing values
- Drop or replace null values in the affected columns or rows.
- The replace technique used was filling the null values using the median.
- Why median?

The choice was informed by the fact that the mean could be affected by outliers and thus could offset our values or rather may end up giving a false impression.

- Check for duplicates.

EXPLORATORY DATA ANALYSIS

Univariate Analysis

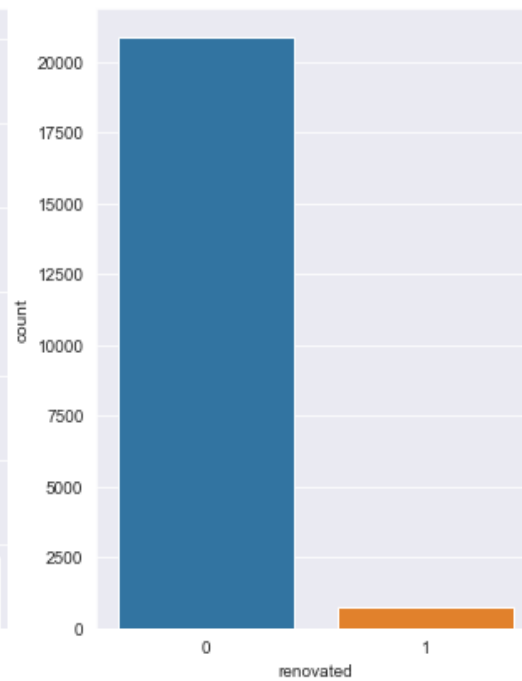
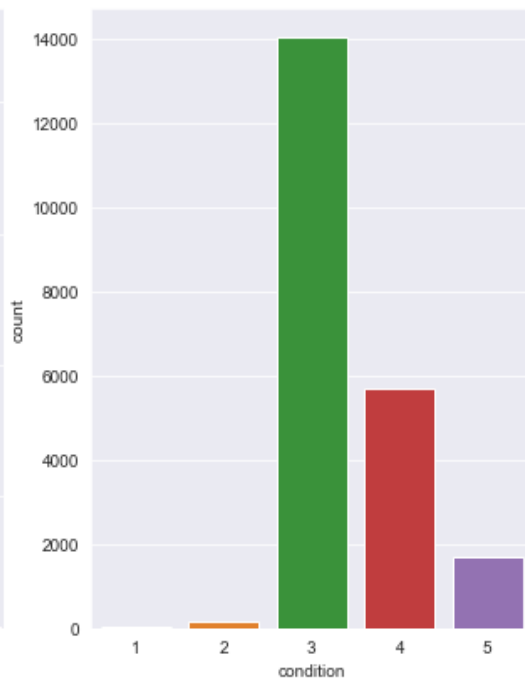
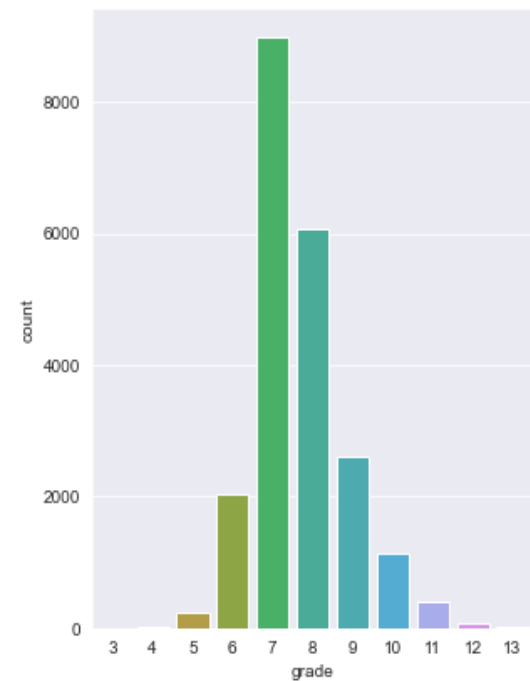
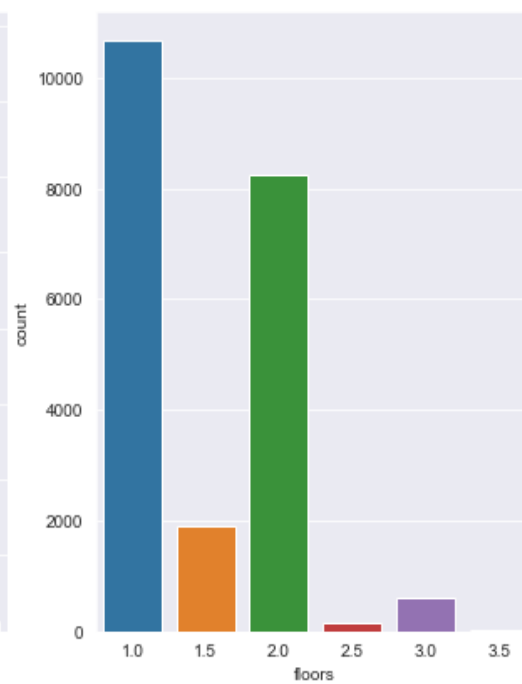
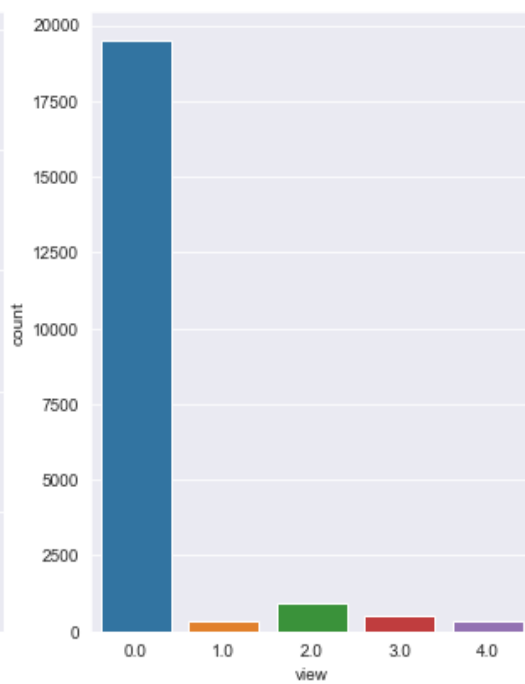
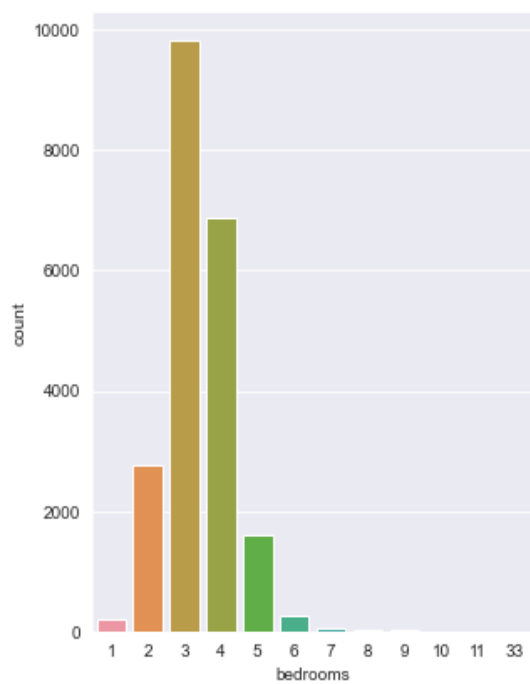


- From the distribution above we can see that most houses range from \$0 to around \$1.7. We can also see that there are some outliers in the distribution.

EXPLORATORY DATA ANALYSIS

Univariate Analysis (categorical variables)

- From the distribution we can see that 3 and 4 bedroom houses are with the most count.
- Most houses have low quality of view.
- 1 floor and 2 floor houses have the highest count.
- most houses range from \$0 to around \$1.7. We can also see that there are some outliers in the distribution.



EXPLORATORY DATA ANALYSIS

Univariate Analysis

Condition

- 1= Poor Many repairs needed.
- 2= Fair Some repairs needed immediately
- 3=Average Depending upon age of improvement
- 4= Good Condition above the norm for the age of the home.
- 5=Very Good Excellent maintenance and updating on home.

Grade

Grades 1 - 3 Falls short of minimum building standards. Normally cabin or inferior structure.

Grade 4 - Generally older low quality construction. Does not meet code.

Grade 5 - Lower construction costs and workmanship. Small, simple design.

Grade 6 -Lowest grade currently meeting building codes. Low quality materials, simple designs.

Grade 7 -Average grade of construction and design. Commonly seen in plats and older subdivisions.

Grade 8 -Just above average in construction and design. Usually better materials in both the exterior and interior finishes.

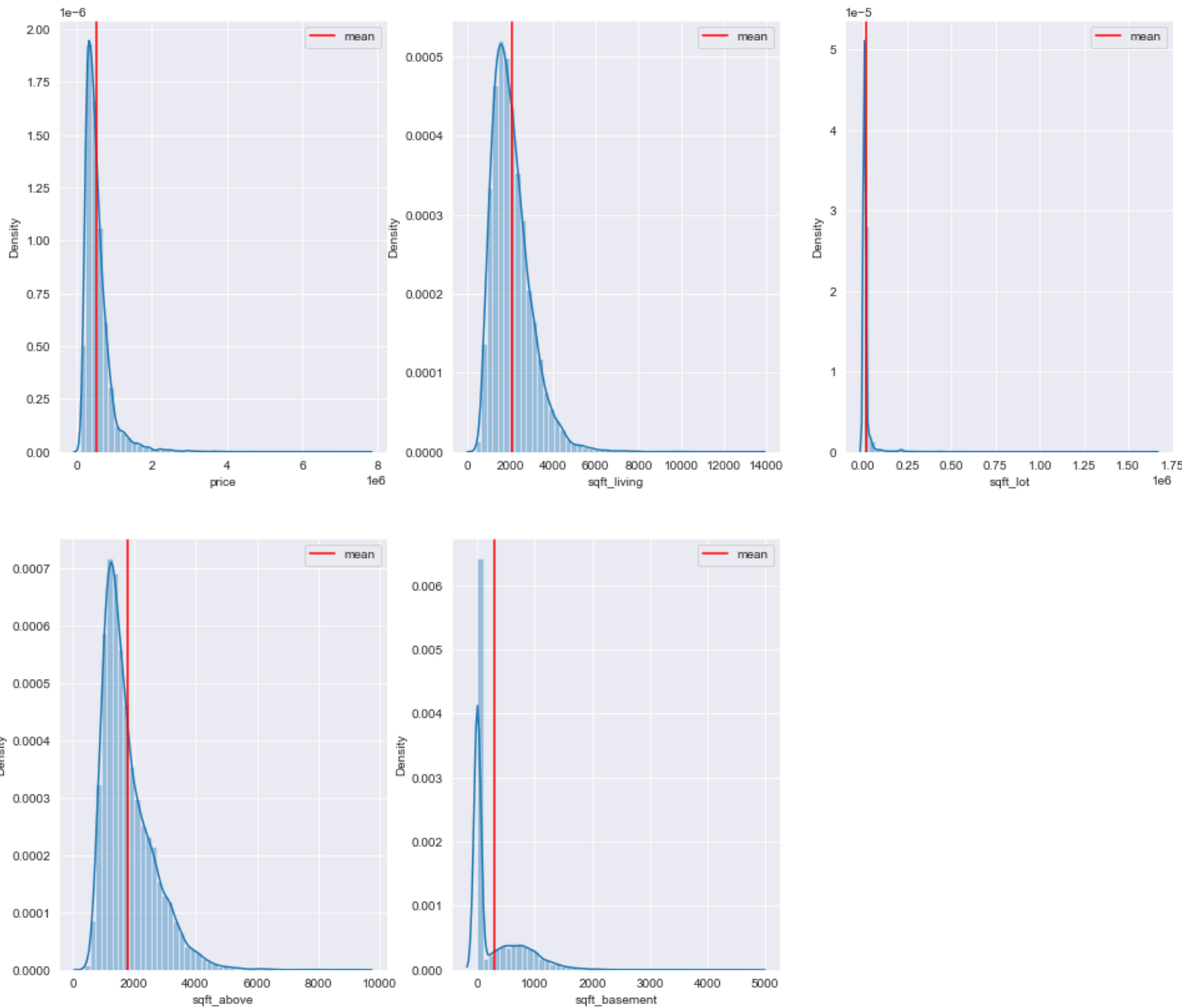
Grade 9 -Better architectural design, with extra exterior and interior design and quality.

Grade 10 -Homes of this quality generally have high quality features. Finish work is better, and more design quality is seen in the floor plans and larger square footage.

Grade 11 -Custom design and higher quality finish work, with added amenities of solid woods, bathroom fixtures and more luxurious options.

Grade 12 -Custom design and excellent builders. All materials are of the highest quality and all conveniences are present.

Grade 13 -Generally custom designed and built. Approaching the Mansion level. Large amount of highest quality cabinet work, wood trim and marble; large entries.

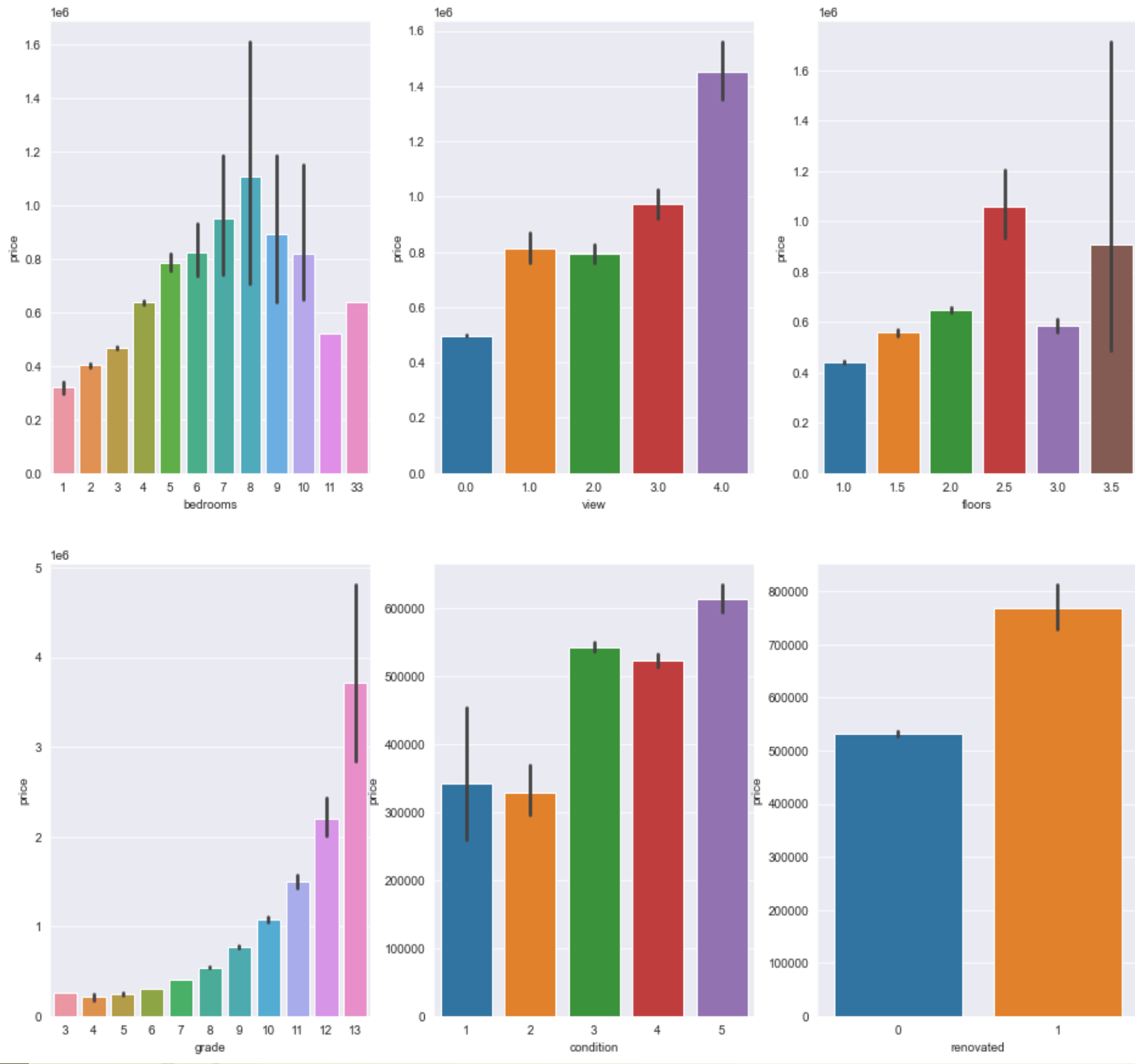


Q1: what is the distribution of all numerical features?

EXPLORATORY DATA ANALYSIS

Bivariate Analysis

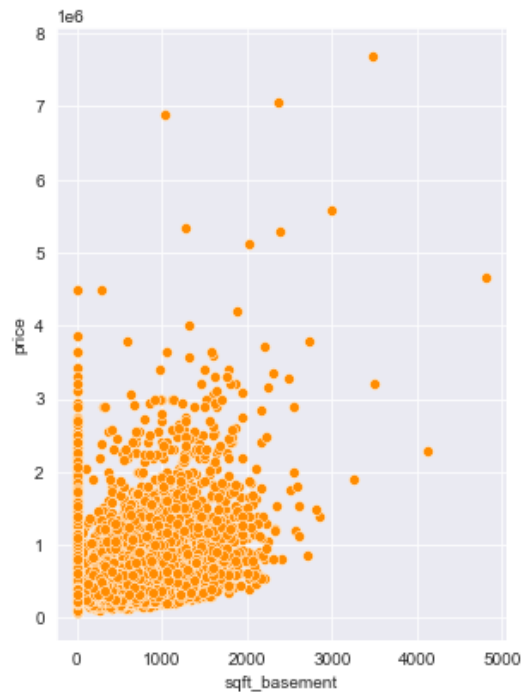
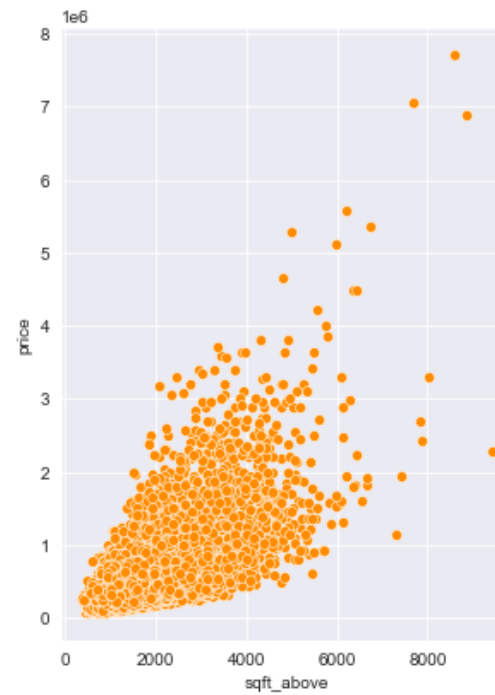
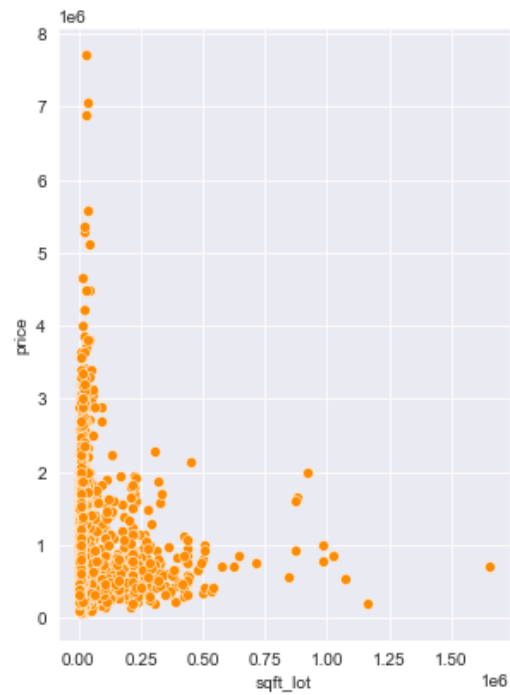
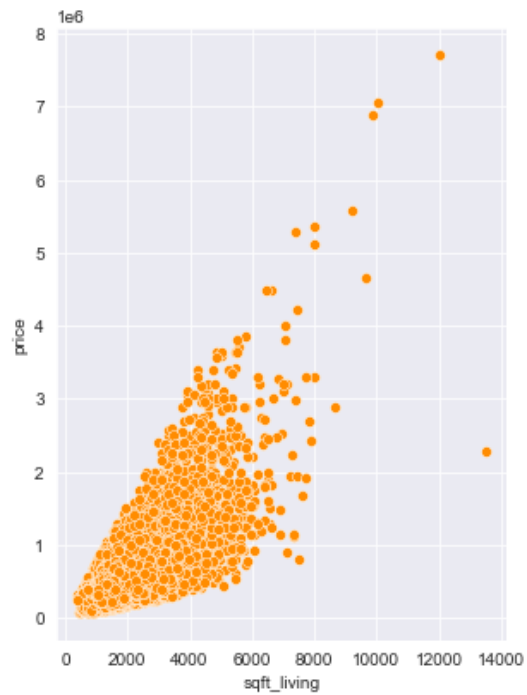
What is the Relationships between price and categorical variables?



EXPLORATORY DATA ANALYSIS

Bivariate Analysis

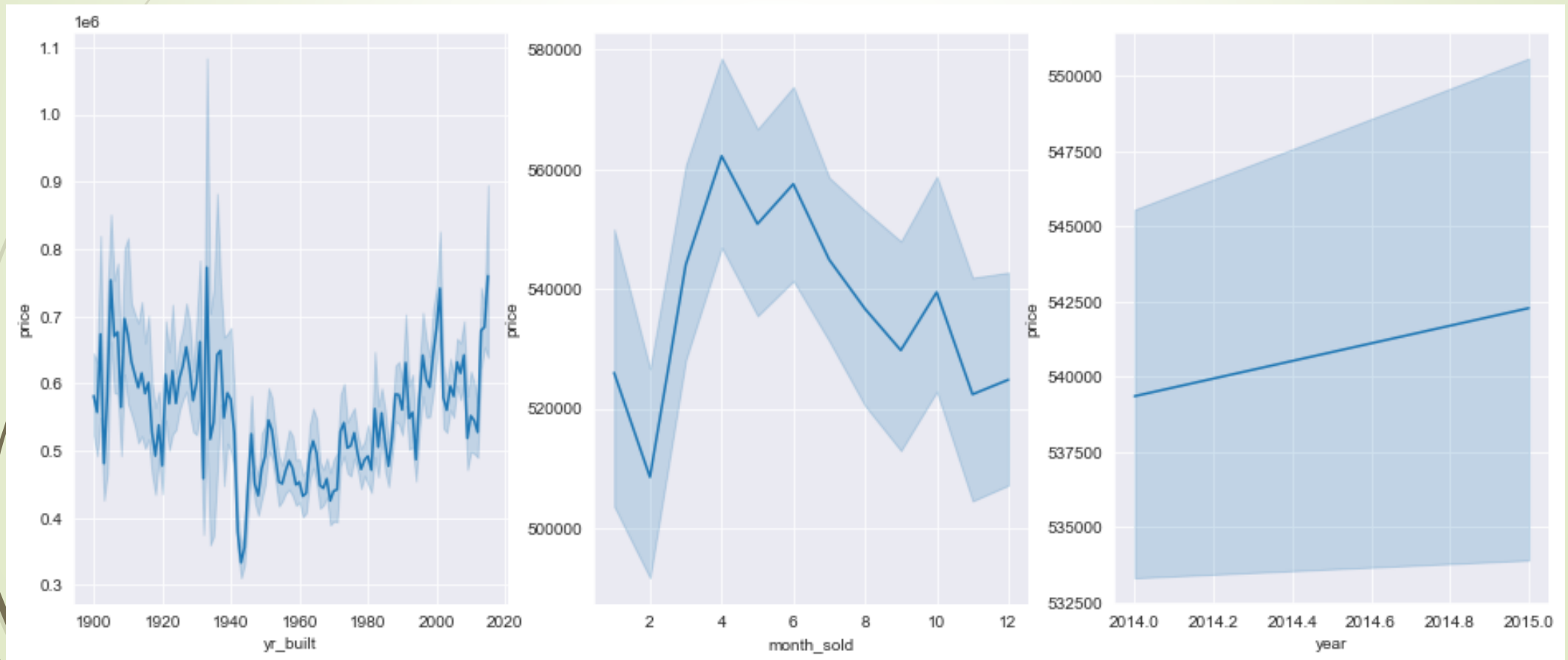
What Relationship between price and continuous variables?



EXPLORATORY DATA ANALYSIS

Bivariate Analysis

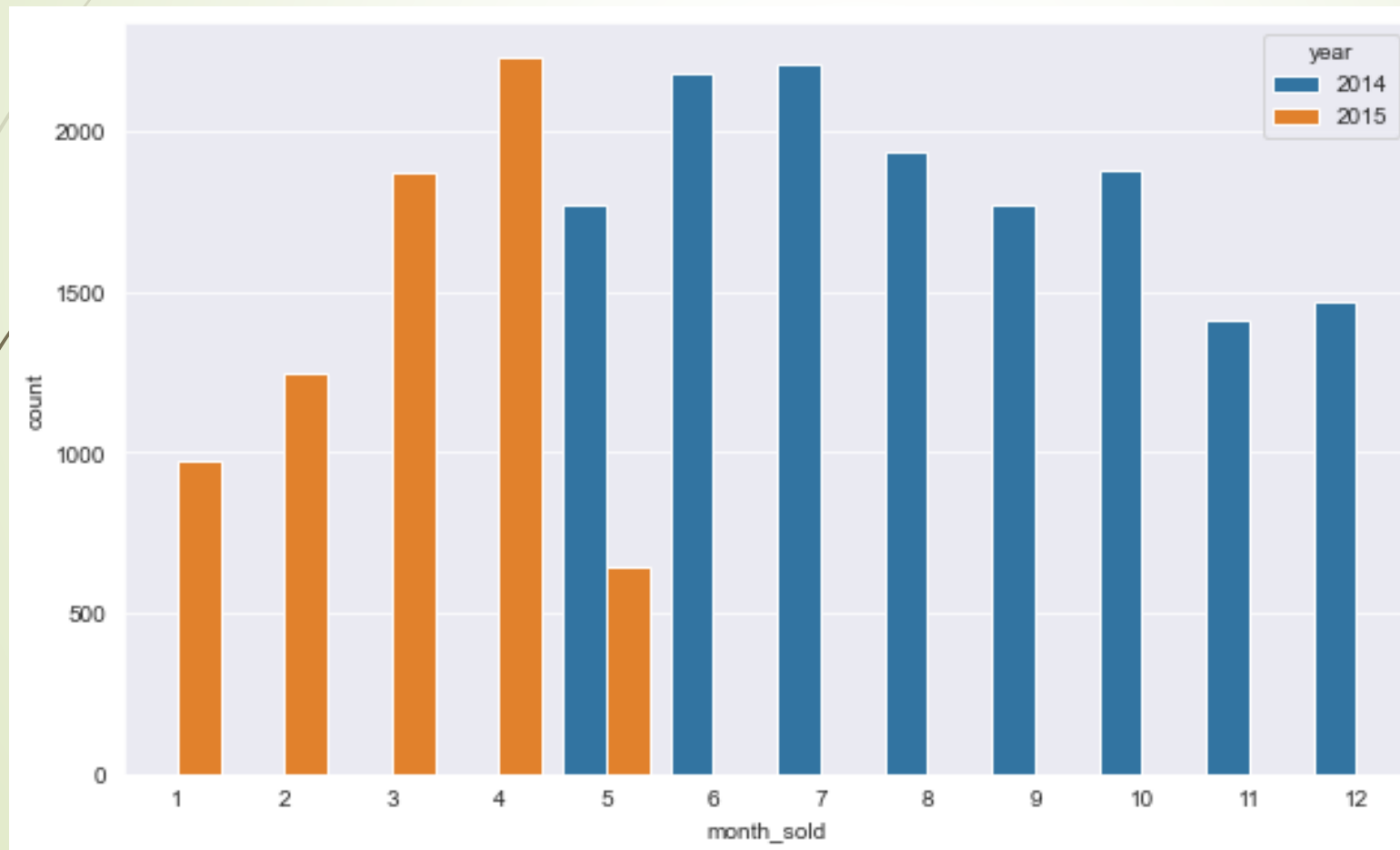
What Relationship between price and timeseries data?



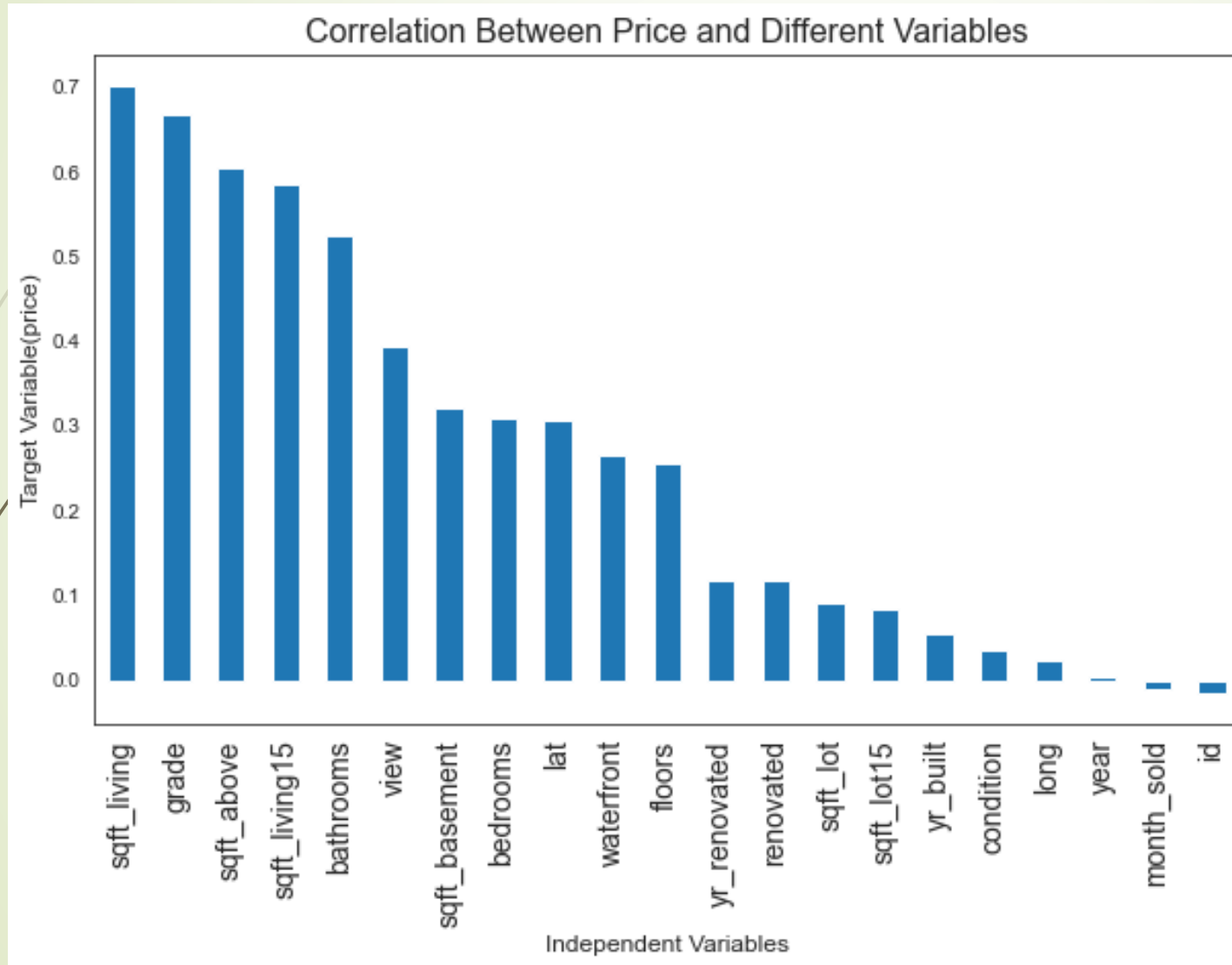
EXPLORATORY DATA ANALYSIS

Bivariate Analysis

check which month and year the houses sold the most?



Q2: What is the correlation between price(target variable) and independent variables?



	price
price	1.000000
sqft_living	0.701917
grade	0.667951
sqft_above	0.605368
sqft_living15	0.585241
bathrooms	0.525906
view	0.393497
sqft_basement	0.321108
bedrooms	0.308787
lat	0.306692
waterfront	0.264306
floors	0.256804
yr_renovated	0.117855
renovated	0.117543
sqft_lot	0.089876
sqft_lot15	0.082845
yr_built	0.053953
condition	0.036056
long	0.022036
id	0.016772
month_sold	0.009928
year	0.003727

Checking for Multicollinearity

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	...	sqft_basement	yr_built	yr_renovated	lat	lon
id	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False
price	False	True	False	False	True	False	False	False	False	False	...	False	False	False	False	False
bedrooms	False	False	True	False	False	False	False	False	False	False	...	False	False	False	False	False
bathrooms	False	False	False	True	True	False	False	False	False	False	...	False	False	False	False	False
sqft_living	False	True	False	True	True	False	False	False	False	False	...	False	False	False	False	False
sqft_lot	False	False	False	False	False	True	False	False	False	False	...	False	False	False	False	False
floors	False	False	False	False	False	False	True	False	False	False	...	False	False	False	False	False
waterfront	False	False	False	False	False	False	False	True	False	False	...	False	False	False	False	False
view	False	False	False	False	False	False	False	False	True	False	...	False	False	False	False	False
condition	False	False	False	False	False	False	False	False	False	True	...	False	False	False	False	False
grade	False	False	False	False	True	False	False	False	False	False	...	False	False	False	False	False
sqft_above	False	False	False	False	True	False	False	False	False	False	...	False	False	False	False	False
sqft_basement	False	False	False	False	False	False	False	False	False	False	...	True	False	False	False	False
yr_built	False	False	False	False	False	False	False	False	False	False	...	False	True	False	False	False
yr_renovated	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	False
lat	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	False
long	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	True
sqft_living15	False	False	False	False	True	False	False	False	False	False	...	False	False	False	False	False
sqft_lot15	False	False	False	False	False	True	False	False	False	False	...	False	False	False	False	False
year	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False
month_sold	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False
renovated	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	False

cc

pairs

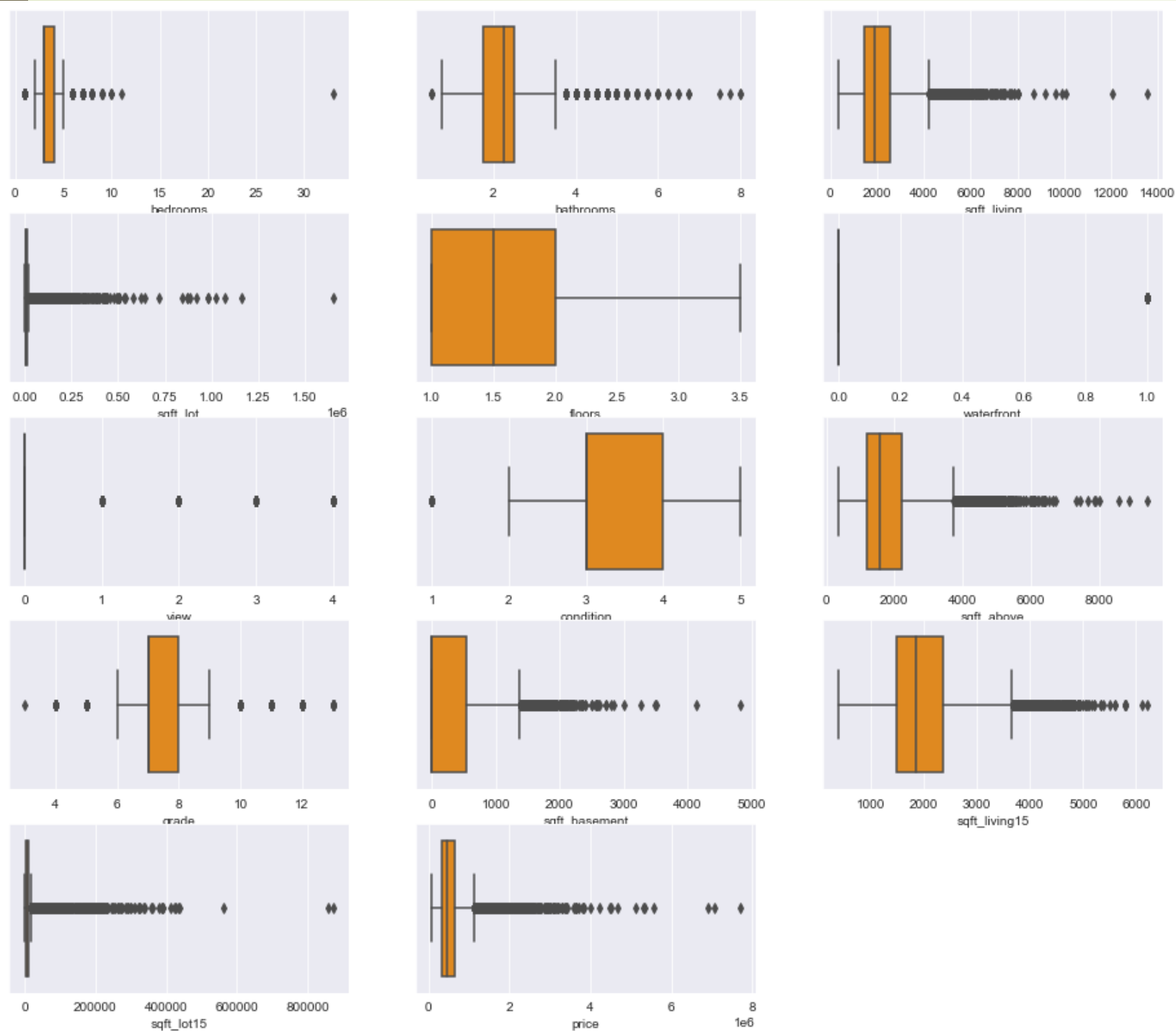
(yr_renovated, renovated)	0.999968
(sqft_above, sqft_living)	0.876448
(year, month_sold)	0.782325
(grade, sqft_living)	0.762779
(sqft_living15, sqft_living)	0.756402
(grade, sqft_above)	0.756073
(sqft_living, bathrooms)	0.755758

Q4: Which columns to drop?

- From the table above I realized that **sqft-living** has **high correlation** with **bathrooms**, **grade**, **sqft-above** and **sqft-living15**. Also I noticed **grade** has high correlation with **sqft-living**, **sqft-above**. From this observation I think I will drop both **sqft-living** and **grade** since they might cause problems of Multicollinearity to the model.
- since id doesn't have any correlation with house prices we drop the column.
- I'm also going to drop the "zip code" column, I don't think i will need it for this project. even if I fed it to our model it would be considered as a continues value although it isn't.
- I am going to drop date column since we already split the date into month and year sold

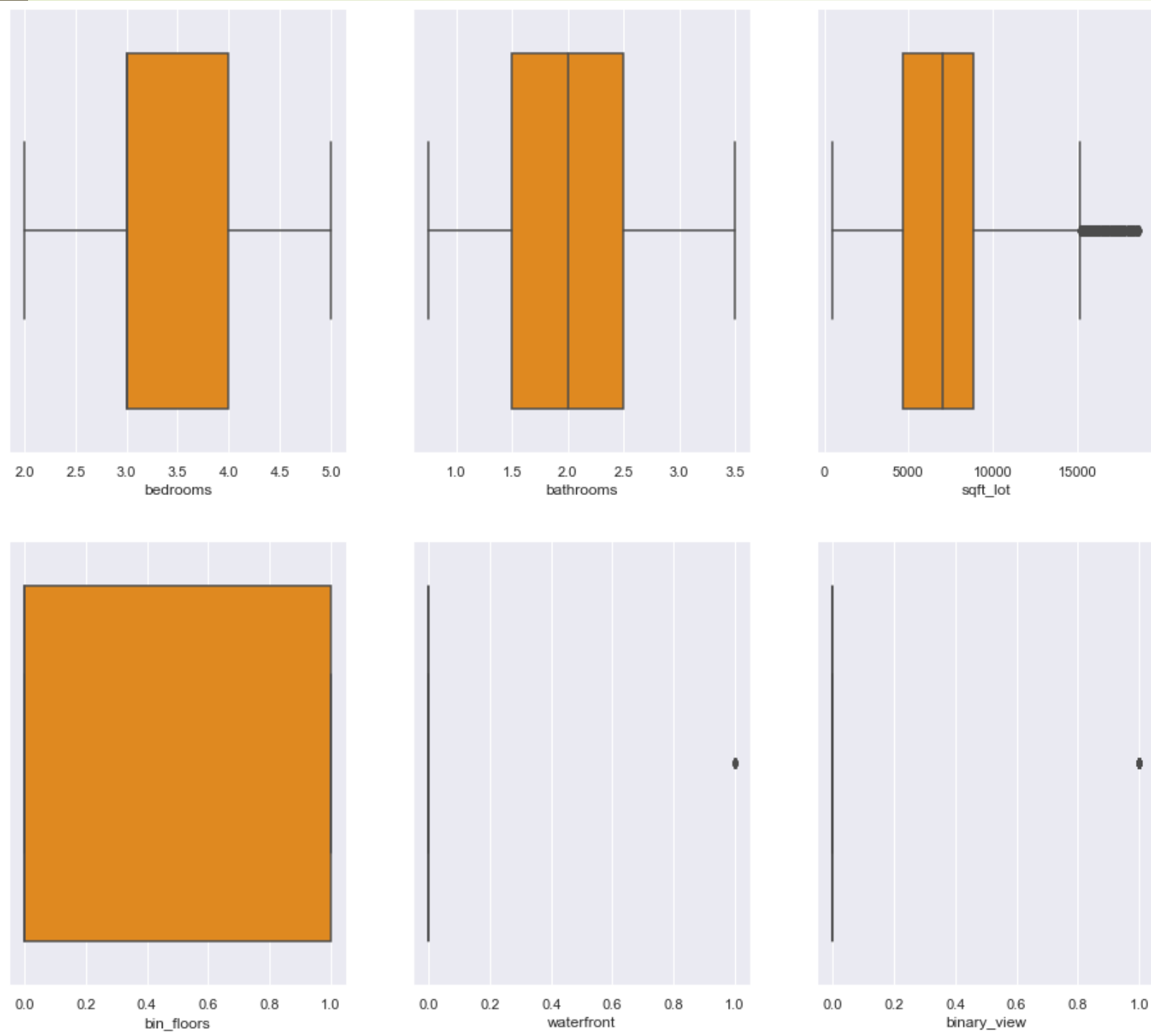
EXPLORATORY DATA ANALYSIS

check for Outliers



EXPLORATORY DATA ANALYSIS

check for Outliers after removal

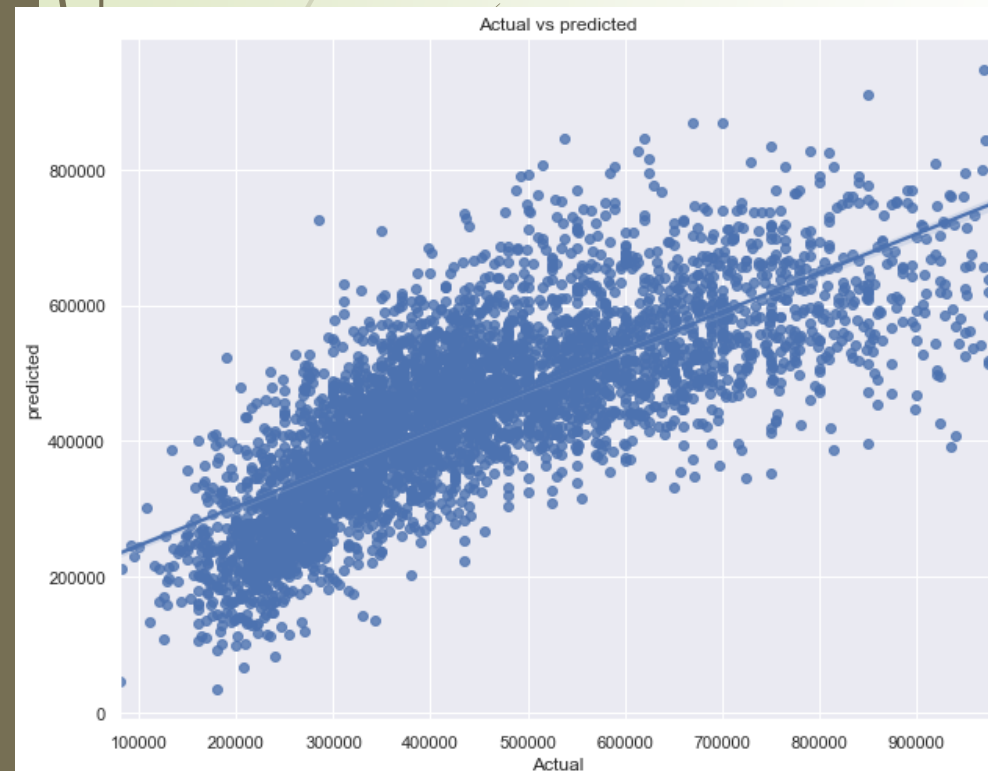


LINEAR REGRESSION MODELLING

Model 1

➤ The mean absolute error was found to be 90571.122 .

	waterfront	sqft_above	sqft_basement	lat	long	sqft_living15	sqft_lot15	renovated	binary_view	binary_condition	bin_floors
0	0.0	1180	0.0	47.5112	-122.257	1340	5650	0	0	1	0
1	0.0	2170	400.0	47.7210	-122.319	1690	7639	1	0	1	1
2	0.0	770	0.0	47.7379	-122.233	2720	8062	0	0	1	0
3	0.0	1050	910.0	47.5208	-122.393	1360	5000	0	0	1	0
4	0.0	1680	0.0	47.6168	-122.045	1800	7503	0	0	1	0



	PREDICTIONS	ACTUAL VALUES	error
8138	402299.393582	394000.0	-8299.393582
15491	504469.523846	700000.0	195530.476154
20451	661818.993613	870000.0	208181.006387
21179	420110.497760	445000.0	24889.502240
376	465656.695421	450000.0	-15656.695421
...
11671	503920.003532	500000.0	-3920.003532
13899	687132.261747	480000.0	-207132.261747
18020	646819.144646	585000.0	-61819.144646
6681	322604.849755	350000.0	27395.150245
5918	446327.086302	575000.0	128672.913698

4074 rows x 3 columns

OLS Regression Results

Dep. Variable:	price	R-squared:	0.579
Model:	OLS	Adj. R-squared:	0.578
Method:	Least Squares	F-statistic:	1524.
Date:	Tue, 05 Jul 2022	Prob (F-statistic):	0.00
Time:	19:32:47	Log-Likelihood:	-1.5996e+05
No. Observations:	12222	AIC:	3.199e+05
Df Residuals:	12210	BIC:	3.200e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	4.429e+05	1057.690	418.780	0.000	4.41e+05	4.45e+05
x1	7007.3102	1066.625	6.570	0.000	4916.557	9098.064
x2	6.875e+04	1871.327	36.740	0.000	6.51e+04	7.24e+04
x3	3.901e+04	1207.233	32.312	0.000	3.66e+04	4.14e+04
x4	8.466e+04	1094.219	77.373	0.000	8.25e+04	8.68e+04
x5	-1.056e+04	1220.632	-8.649	0.000	-1.29e+04	-8164.262
x6	3.818e+04	1625.570	23.489	0.000	3.5e+04	4.14e+04
x7	-2.176e+04	1289.117	-16.877	0.000	-2.43e+04	-1.92e+04
x8	1.29e+04	1065.481	12.105	0.000	1.08e+04	1.5e+04
x9	1.861e+04	1102.407	16.883	0.000	1.65e+04	2.08e+04
x10	3964.2001	1058.545	3.745	0.000	1889.284	6039.117
x11	4168.2373	1533.222	2.719	0.007	1162.879	7173.596

Omnibus:	758.130	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1010.575
Skew:	0.570	Prob(JB):	3.60e-220
Kurtosis:	3.827	Cond. No.	3.46

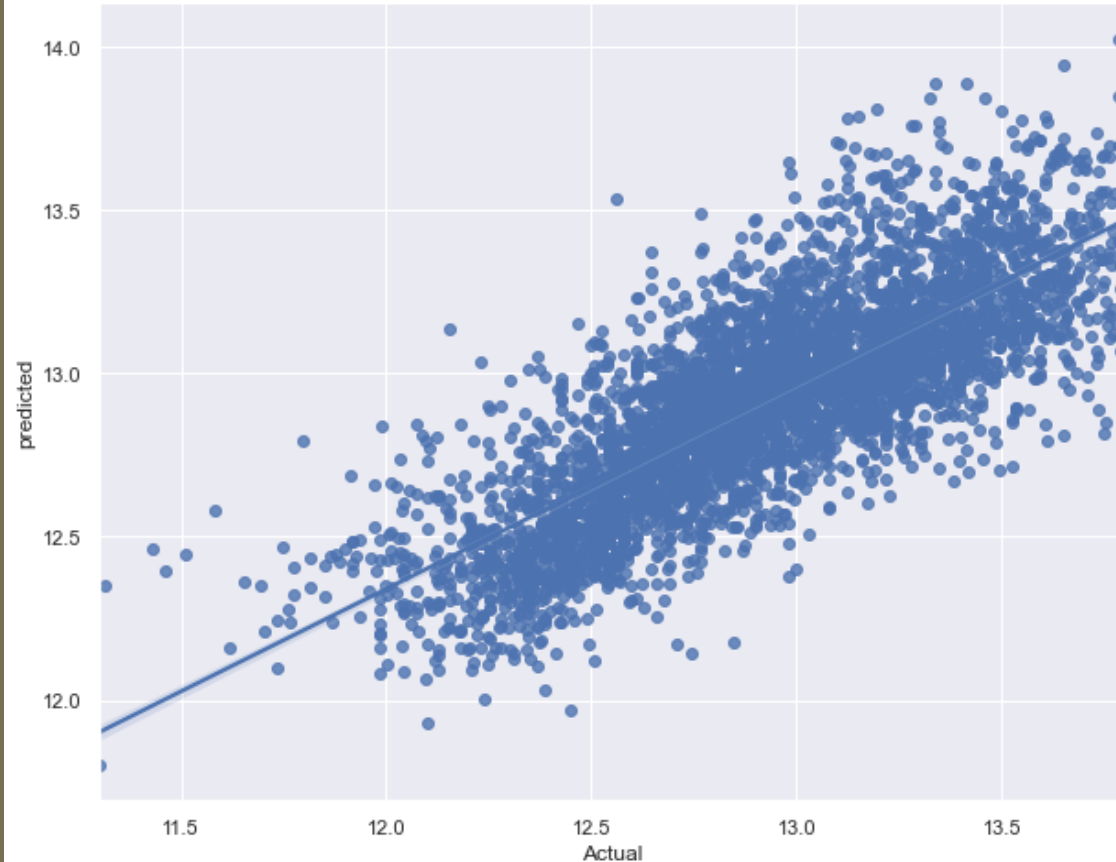
LINEAR REGRESSION MODELLING

Model 2

The mean absolute error was found to be 0.1993613898 .

	waterfront	sqft_above	sqft_basement	lat	long	sqft_living15	sqft_lot15	renovated	binary_view	binary_condition	bin_floors
0	0.0	1180	0.0	47.5112	-122.257	1340	5650	0	0	1	0
1	0.0	2170	400.0	47.7210	-122.319	1690	7639	1	0	1	1
2	0.0	770	0.0	47.7379	-122.233	2720	8062	0	0	1	0
3	0.0	1050	910.0	47.5208	-122.393	1360	5000	0	0	1	0
4	0.0	1680	0.0	47.6168	-122.045	1800	7503	0	0	1	0

Actual vs predicted model 2



	PREDICTIONS	ACTUAL VALUES	error
8138	12.857367	12.884106	0.026739
15491	13.056131	13.458836	0.402704
20451	13.443018	13.676248	0.233231
21179	12.903530	13.005830	0.102300
376	12.980163	13.017003	0.036840
...
11671	13.071649	13.122363	0.050714
13899	13.530583	13.081541	-0.449041
18020	13.395407	13.279367	-0.116040
6681	12.627553	12.765688	0.138135
5918	12.920988	13.262125	0.341137

OLS Regression Results

Dep. Variable:	price	R-squared:	0.627
Model:	OLS	Adj. R-squared:	0.627
Method:	Least Squares	F-statistic:	1870.
Date:	Tue, 05 Jul 2022	Prob (F-statistic):	0.00
Time:	19:32:52	Log-Likelihood:	-630.70
No. Observations:	12222	AIC:	1285.
Df Residuals:	12210	BIC:	1374.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	12.9171	0.002	5602.079	0.000	12.913	12.922
x1	0.0150	0.002	6.440	0.000	0.010	0.020
x2	0.1506	0.004	36.907	0.000	0.143	0.159
x3	0.0923	0.003	35.058	0.000	0.087	0.097
x4	0.2176	0.002	91.218	0.000	0.213	0.222
x5	-0.0140	0.003	-5.249	0.000	-0.019	-0.009
x6	0.0922	0.004	26.028	0.000	0.085	0.099
x7	-0.0570	0.003	-20.277	0.000	-0.062	-0.051
x8	0.0242	0.002	10.397	0.000	0.020	0.029
x9	0.0372	0.002	15.498	0.000	0.033	0.042
x10	0.0144	0.002	6.225	0.000	0.010	0.019
x11	0.0152	0.003	4.545	0.000	0.009	0.022

Omnibus:	159.567	Durbin-Watson:	1.971
Prob(Omnibus):	0.000	Jarque-Bera (JB):	223.240
Skew:	-0.167	Prob(JB):	3.34e-49
Kurtosis:	3.571	Cond. No.	3.46

LINEAR REGRESSION MODELLING

Model 3

➤ The mean absolute error was found to be 119748.75601170

	sqft_lot	sqft_above	sqft_basement	sqft_living15	sqft_lot15
0	5650	1180	0.0	1340	5650
1	7242	2170	400.0	1690	7639
2	10000	770	0.0	2720	8062
3	5000	1050	910.0	1360	5000
4	8080	1680	0.0	1800	7503

	PREDICTIONS	ACTUAL VALUES	error
8138	471225.564024	394000.0	-77225.564024
15491	452252.627490	700000.0	247747.372510
20451	606177.309314	870000.0	263822.690686
21179	483966.734862	445000.0	-38966.734862
376	352815.508551	450000.0	97184.491449

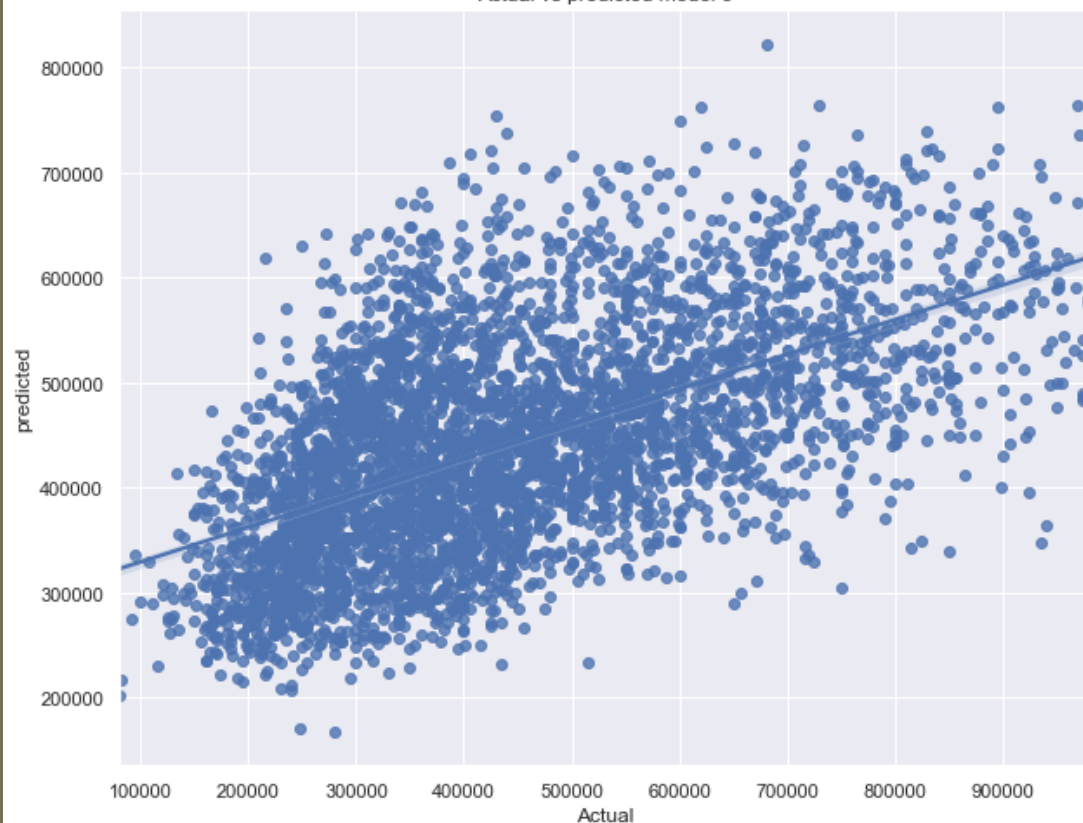
...
11671	536613.987924	500000.0	-36613.987924
13899	570085.026059	480000.0	-90085.026059
18020	505659.708124	585000.0	79340.291876
6681	422583.519371	350000.0	-72583.519371
5918	452866.337350	575000.0	122133.662650

	sqft_lot	sqft_above	sqft_basement	sqft_living15	sqft_lot15
0	5650	1180	0.0	1340	5650
1	7242	2170	400.0	1690	7639
2	10000	770	0.0	2720	8062
3	5000	1050	910.0	1360	5000
4	8080	1680	0.0	1800	7503

...
21592	1131	1530	0.0	1530	1509
21593	5813	2310	0.0	1830	7200
21594	1350	1020	0.0	1020	2007
21595	2388	1600	0.0	1410	1287
21596	1076	1020	0.0	1020	1357

OLS Regression Results

Dep. Variable:	price	R-squared:	0.339			
Model:	OLS	Adj. R-squared:	0.339			
Method:	Least Squares	F-statistic:	1253.			
Date:	Tue, 05 Jul 2022	Prob (F-statistic):	0.00			
Time:	19:32:55	Log-Likelihood:	-1.6271e+05			
No. Observations:	12222	AIC:	3.254e+05			
Df Residuals:	12216	BIC:	3.255e+05			
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	4.429e+05	1324.399	334.446	0.000	4.4e+05	4.46e+05
x1	-1.313e+04	2697.010	-4.867	0.000	-1.84e+04	-7840.141
x2	6.001e+04	2004.217	29.943	0.000	5.61e+04	6.39e+04
x3	5.291e+04	1457.696	36.296	0.000	5.01e+04	5.58e+04
x4	4.549e+04	1980.766	22.964	0.000	4.16e+04	4.94e+04
x5	-2.86e+04	2710.891	-10.551	0.000	-3.39e+04	-2.33e+04
Omnibus:	433.688	Durbin-Watson:	2.003			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	481.022			
Skew:	0.486	Prob(JB):	3.53e-105			
Kurtosis:	3.024	Cond. No.	4.09			

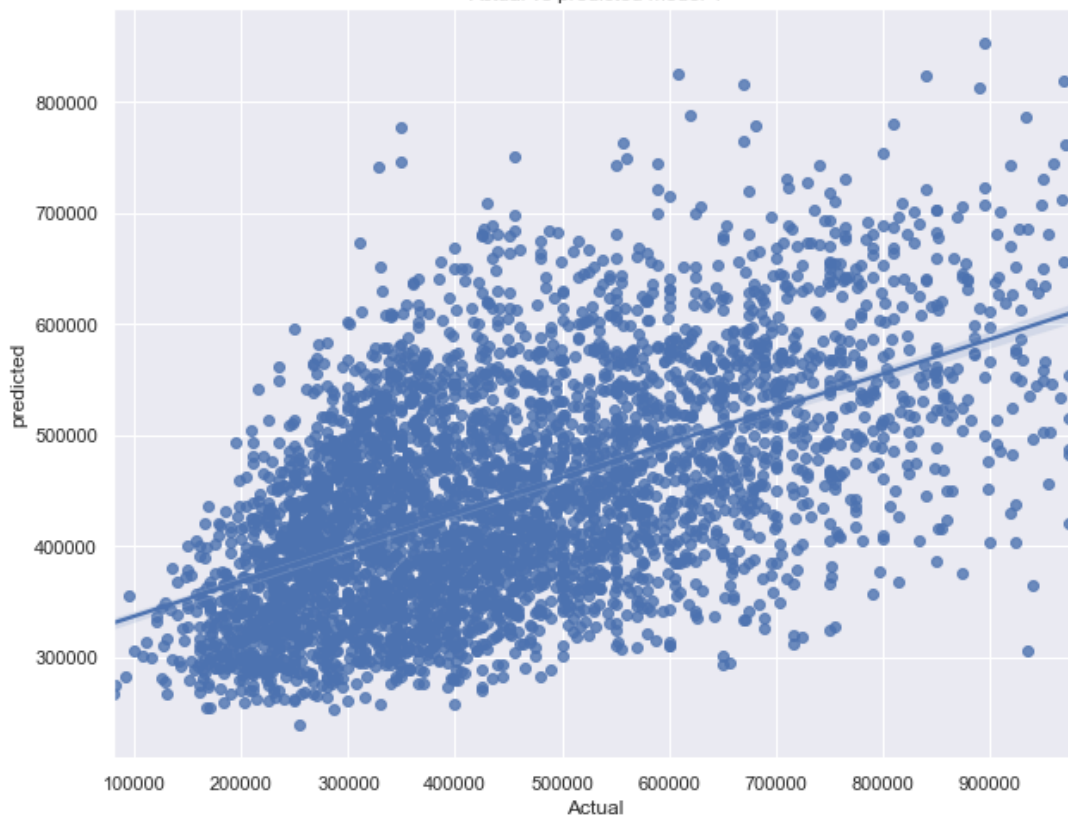


LINEAR REGRESSION MODELLING

Model 4

➤ The mean absolute error was found to be 122042.74460081532

Actual vs predicted model 4



	PREDICTIONS	ACTUAL VALUES	error
8138	448132.696386	394000.0	-54132.696386
15491	419102.112276	700000.0	280897.887724
20451	560376.086925	870000.0	309623.913075
21179	423232.018488	445000.0	21767.981512
376	430309.898241	450000.0	19690.101759

...
11671	530921.722788	500000.0	-30921.722788
13899	528624.163842	480000.0	-48624.163842
18020	551861.829617	585000.0	33138.170383
6681	450718.056231	350000.0	-100718.056231
5918	433973.905331	575000.0	141026.094669

4074 rows × 3 columns

```
performance4['error'].abs().mean()  
122042.74460081532
```

	waterfront	sqft_above	sqft_basement	lat	long	sqft_living15	sqft_lot15	binary_view	binary_condition	bin_floors
0	0.0	1180	0.0	47.5112	-122.257	1340	5650	0	1	0
1	0.0	2170	400.0	47.7210	-122.319	1690	7639	0	1	1
2	0.0	770	0.0	47.7379	-122.233	2720	8062	0	1	0
3	0.0	1050	910.0	47.5208	-122.393	1360	5000	0	1	0
4	0.0	1690	0.0	47.6168	-122.345	1890	7503	0	1	1

OLS Regression Results

Dep. Variable:	price	R-squared:	0.312
Model:	OLS	Adj. R-squared:	0.312
Method:	Least Squares	F-statistic:	925.3
Date:	Tue, 05 Jul 2022	Prob (F-statistic):	0.00
Time:	19:32:58	Log-Likelihood:	-1.6295e+05
No. Observations:	12222	AIC:	3.259e+05
Df Residuals:	12215	BIC:	3.260e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	4.429e+05	1350.807	327.907	0.000	4.4e+05	4.46e+05
x1	7.258e+04	2546.265	28.506	0.000	6.76e+04	7.76e+04
x2	5.457e+04	1736.291	31.428	0.000	5.12e+04	5.8e+04
x3	2.983e+04	2004.372	14.882	0.000	2.59e+04	3.38e+04
x4	4640.5333	1846.339	2.513	0.012	1021.416	8259.651
x5	2.048e+04	1388.569	14.751	0.000	1.78e+04	2.32e+04
x6	-2.003e+04	1701.733	-11.768	0.000	-2.34e+04	-1.67e+04

Omnibus:	494.390	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	556.495
Skew:	0.523	Prob(JB):	1.44e-121
Kurtosis:	2.993	Cond. No.	3.86

LINEAR REGRESSION MODELLING

Model 5

- The mean absolute error was found to be 90860.1851

Actual vs predicted model 5



	PREDICTIONS	ACTUAL VALUES	error
8138	402830.697976	394000.0	-8830.697976
15491	507284.436673	700000.0	192715.563327
20451	666792.849878	870000.0	203207.150122
21179	421900.789950	445000.0	23099.210050
376	468365.459426	450000.0	-18365.459426
...
11671	506545.153303	500000.0	-6545.153303
13899	688298.614154	480000.0	-208298.614154
18020	650857.213944	585000.0	-65857.213944
6681	323585.935826	350000.0	26414.064174
5918	449638.341424	575000.0	125361.658576

1074 rows × 3 columns

```
performance5['error'].abs().mean()
```

90860.18515098584

	waterfront	sqft_above	sqft_basement	lat	long	sqft_living15	sqft_lot15	binary_view	binary_condition	bin_floors
0	0.0	1180	0.0	47.5112	-122.257	1340	5650	0	1	0
1	0.0	2170	400.0	47.7210	-122.319	1690	7639	0	1	1
2	0.0	770	0.0	47.7379	-122.233	2720	8062	0	1	0
3	0.0	1050	910.0	47.5208	-122.393	1360	5000	0	1	0
4	0.0	1680	0.0	47.6168	-122.045	1800	7503	0	1	0

OLS Regression Results

Dep. Variable:	price	R-squared:	0.574
Model:	OLS	Adj. R-squared:	0.573
Method:	Least Squares	F-statistic:	1643.
Date:	Tue, 05 Jul 2022	Prob (F-statistic):	0.00
Time:	19:33:02	Log-Likelihood:	-1.6003e+05
No. Observations:	12222	AIC:	3.201e+05
Df Residuals:	12211	BIC:	3.202e+05
Df Model:	10		
Covariance Type:	nonrobust		

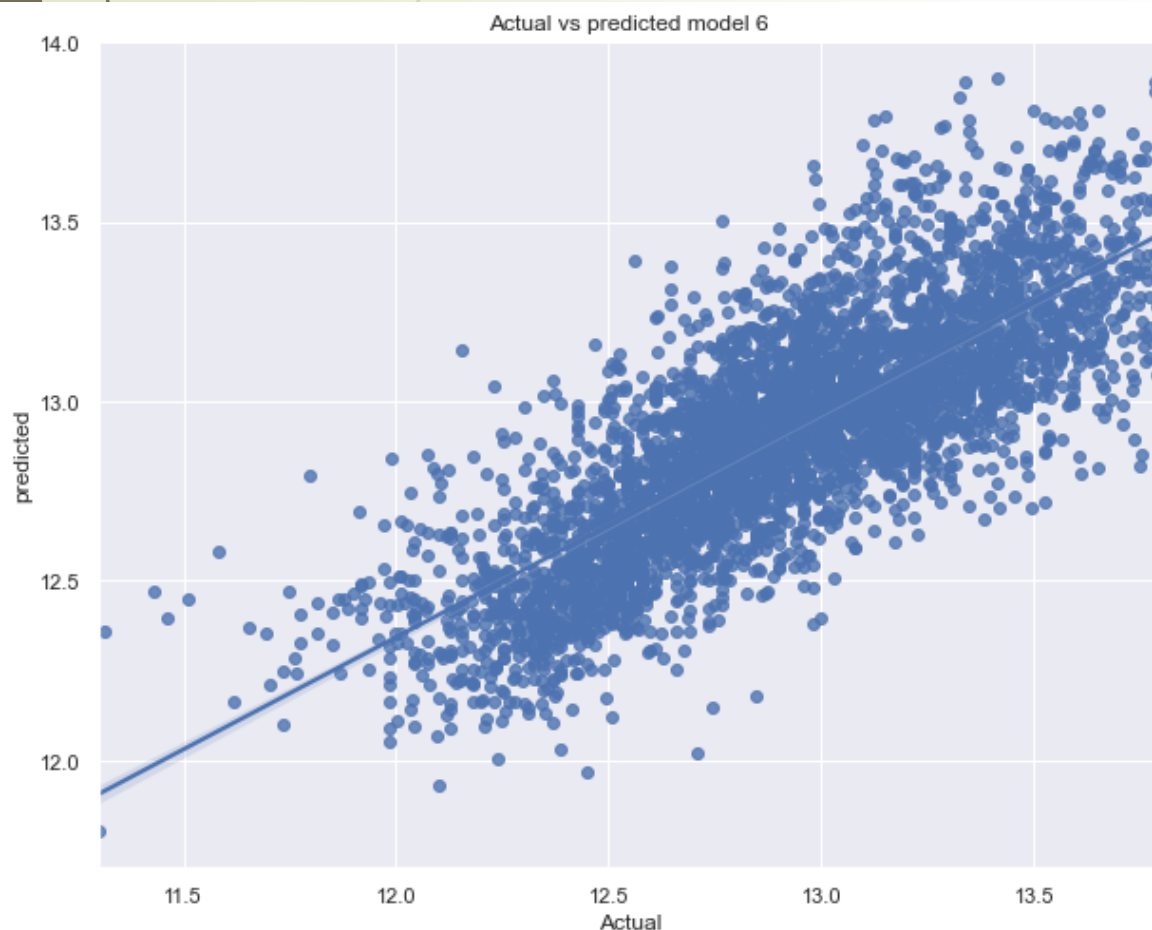
	coef	std err	t	P> t	[0.025	0.975]
const	4.429e+05	1063.974	416.307	0.000	4.41e+05	4.45e+05
x1	7607.5917	1071.802	7.098	0.000	5506.690	9708.493
x2	6.967e+04	1880.915	37.038	0.000	6.6e+04	7.34e+04
x3	3.979e+04	1212.661	32.813	0.000	3.74e+04	4.22e+04
x4	8.495e+04	1100.460	77.196	0.000	8.28e+04	8.71e+04
x5	-1.116e+04	1226.869	-9.095	0.000	-1.36e+04	-8752.923
x6	3.691e+04	1631.826	22.622	0.000	3.37e+04	4.01e+04
x7	-2.155e+04	1296.668	-16.623	0.000	-2.41e+04	-1.9e+04
x8	1.916e+04	1108.031	17.289	0.000	1.7e+04	2.13e+04
x9	4014.3685	1064.826	3.770	0.000	1927.140	6101.597
x10	4541.7673	1542.019	2.945	0.003	1519.165	7564.369

Omnibus:	812.584	Durbin-Watson:	1.971
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1092.690
Skew:	0.596	Prob(JB):	5.31e-238

LINEAR REGRESSION MODELLING

Model 6

- The mean absolute error was found to be 0.199754173



	PREDICTIONS	ACTUAL VALUES	error
8138	12.858362	12.884106	0.025744
15491	13.061402	13.458836	0.397433
20451	13.452331	13.676248	0.223917
21179	12.906882	13.005830	0.098947
376	12.985235	13.017003	0.031768
...
11671	13.076565	13.122363	0.045799
13899	13.532767	13.081541	-0.451225
18020	13.402968	13.279367	-0.123601
6681	12.629390	12.765688	0.136298
5918	12.927189	13.262125	0.334937

4074 rows × 3 columns

```
performance6['error'].abs().mean()
```

0.19975417315287974

waterfront	sqft_above	sqft_basement	lat	long	sqft_living15	sqft_lot15	binary_view	binary_condition	bin_floors
0.0	1180	0.0	47.5112	-122.257	1340	5650	0	1	0
0.0	2170	400.0	47.7210	-122.319	1690	7639	0	1	1
0.0	770	0.0	47.7379	-122.233	2720	8062	0	1	0
0.0	1050	910.0	47.5208	-122.393	1360	5000	0	1	0
0.0	1680	0.0	47.6168	-122.045	1800	7503	0	1	0

OLS Regression Results

Dep. Variable:	price	R-squared:	0.624
Model:	OLS	Adj. R-squared:	0.624
Method:	Least Squares	F-statistic:	2028.
Date:	Tue, 05 Jul 2022	Prob (F-statistic):	0.00
Time:	19:33:04	Log-Likelihood:	-684.56
No. Observations:	12222	AIC:	1391.
Df Residuals:	12211	BIC:	1473.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	12.9171	0.002	5577.671	0.000	12.913	12.922
x1	0.0161	0.002	6.900	0.000	0.012	0.021
x2	0.1523	0.004	37.194	0.000	0.144	0.160
x3	0.0937	0.003	35.511	0.000	0.089	0.099
x4	0.2181	0.002	91.067	0.000	0.213	0.223
x5	-0.0151	0.003	-5.652	0.000	-0.020	-0.010
x6	0.0899	0.004	25.300	0.000	0.083	0.097
x7	-0.0566	0.003	-20.057	0.000	-0.062	-0.051
x8	0.0383	0.002	15.866	0.000	0.034	0.043
x9	0.0145	0.002	6.238	0.000	0.010	0.019
x10	0.0159	0.003	4.734	0.000	0.009	0.022

Omnibus:	142.052	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	200.299
Skew:	-0.147	Prob(JB):	3.20e-44
Kurtosis:	3.554	Cond. No.	3.45




Conclusions

Based on Qn 1:

- The graph displays indicators of the revenue that the company may attain from the production of such kind of movies.
- It forms a guide that could lead the company down the right path if it releases similar kind of movies and even provides a benchmark to build on and work with moving forward.



Qn 2:

- The bar plot just gives a general idea of where to allocate most funds of the production budget based on the genres being produced.
 - It looks like horror films take up a major chunk of the money. This is important so as to allocate funds appropriately and plan accordingly in terms of the release date.
 - For example, horror films are majorly saturated during the Halloween season and it would be ideal to release them at such times.
- 

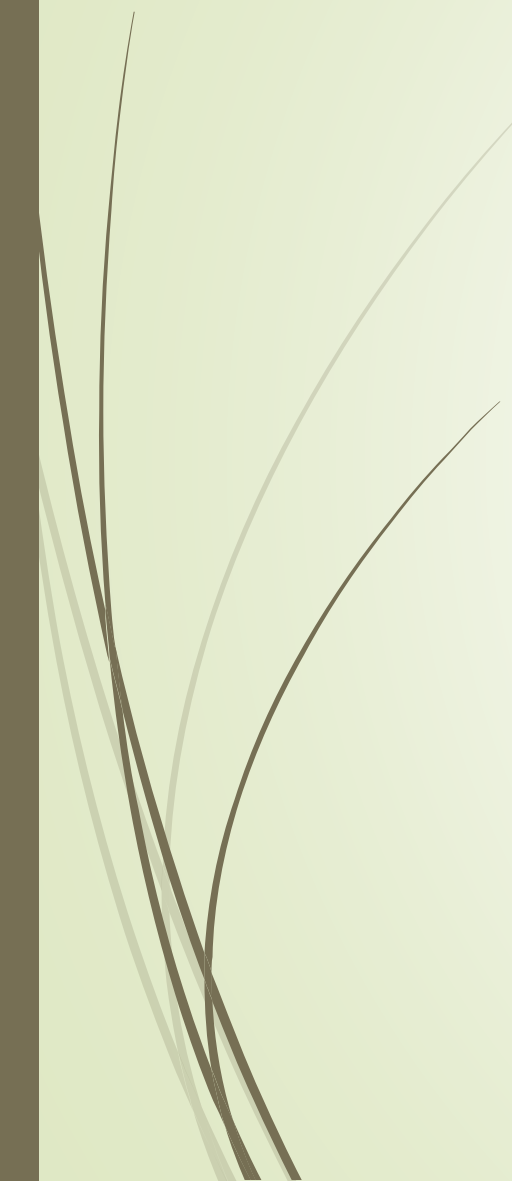


Qn 3:

- The bar plot displayed mainly provides insight on what the audience feels about the films. Higher ratings are obviously associated with good movies and vice versa. This is a plus because it is feedback from the audience. Both the fan base and critic base are useful gauges of the kind of movies you would want to create.



Qn 4:

- A majority of the movies have a runtime centered between 80 and 125 minutes with the average value being 104 minutes.
 - The length of the movie influences the audience's attention.
- 

➔ **THANK YOU FOR YOUR ATTENTION!**
QUESTIONS?

