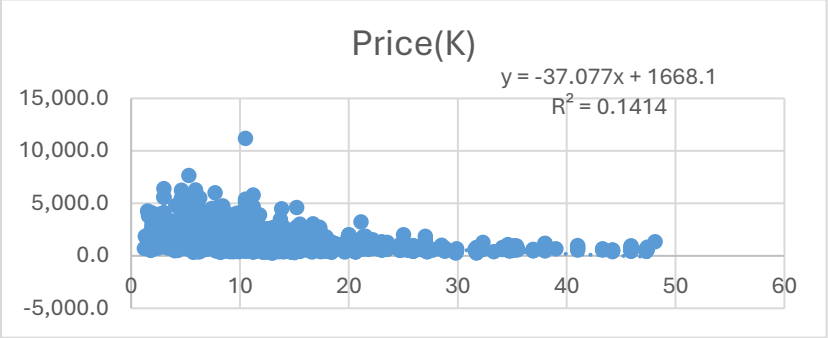
TASK 2.2 (C)

Model 2.2 (A):



Model 2.2(a) is a linear regression model where,

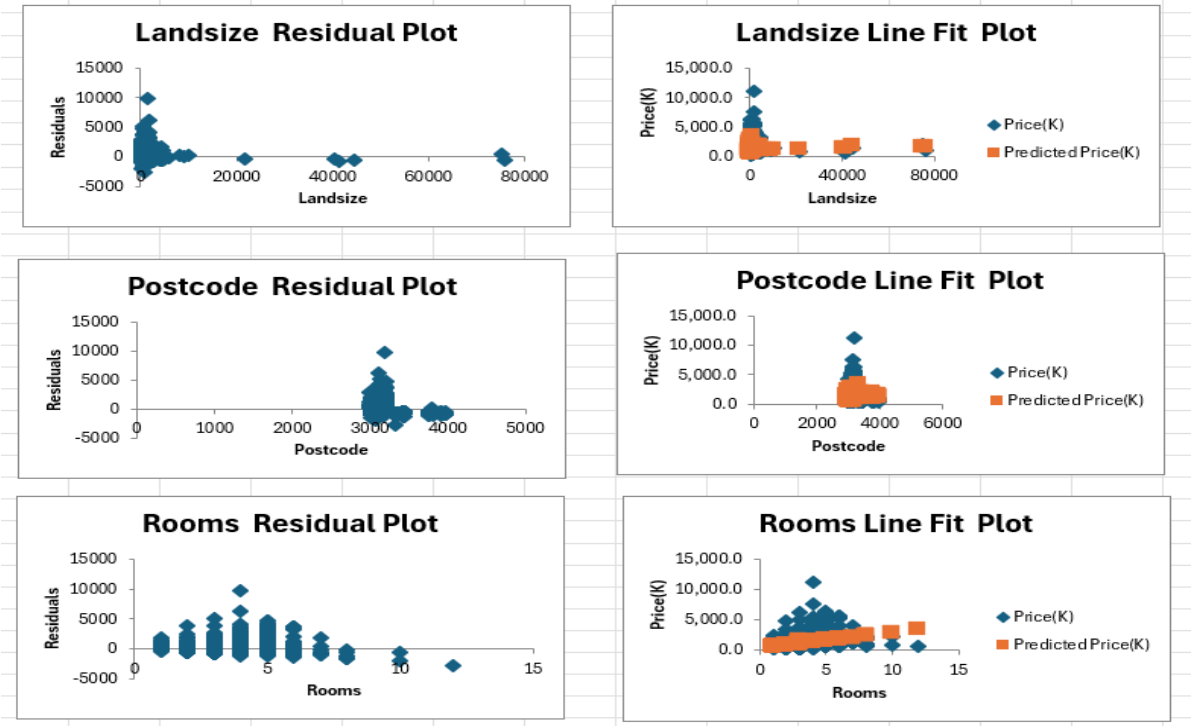
- Independent Variable: Distance
- Dependent Variable: Price(K)
- Equation: y=-37.077x + 1668.1
- R^2 Value = 0.1414
    - Which shows only 14.14% variance can be seen in price due to distance.
- The scatterplot shows a weak negative relationship, with significant dispersion around the regression line.
- correlation -0.37599 Correlation between distance and price comes in negative.
    - Which shows distance from CBD is strongly related to house Price.

With the help of Correlation

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Jpper 95%*	*ower 95.0%*	*pper 9*
Intercept	1668.0759	15.446697	107.9892	0	1637.795514	1698.36	1637.8	1698
Distance	-37.076899	1.1152379	-33.2457	2E-224	-39.26311922	-34.891	-39.263	-34.

    - The p value of 0 shows that distance is statistically significant with Price.

Model 2.2(b)



Model 2.2(b) is a multilinear regression model. Which has,

- Independent Variables:
  - Rooms
  - Postcode
  - Landsize
- Dependent Variable: Price (K)
- R^2 Value: 0.31458
  - It indicates that 31.46% of the variance in the price is explained by the three variables combined.
- The residual plots indicate a slight difference in predictions for some variables like Rooms show more dispersion.
- VIF 1.10983 VIF less than 5 shows strong collinearity among variables.

	Coefficient	tandard Er	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	338.224	207.865	1.62713	0.10376	-69.2578	745.707	-69.2578	745.707
Rooms	251.905	9.37244	26.8772	3E-151	233.532	270.278	233.532	270.278
Postcode	0.01418	0.06696	0.21172	0.83233	-0.11708	0.14544	-0.11708	0.14544
Landsize	0.00574	0.00488	1.17523	0.23994	-0.00383	0.01531	-0.00383	0.01531

- P-value: Rooms have the lowest p-value than Post code and landsize. Which suggests that room and price has more statistically significant than postcode-price or landsize-prize.
- Co-efficient: Rooms have higher coefficient whereas postcode and landsize has less than 0 number of coefficient.

Comparison:

1) Introspect of R^2 Value:

The multiple regression model (Task 2.2(b)) has a significantly higher R^2 (0.31458) compared to the single-variable regression (0.1414). This suggests that the variation in price is not solely depends on distance but more than other variables such as Rooms.

2) In 2.2(b), Room has the highest significance on variation of price than landsize or postal code.
3) The linear regression model (Task 2.2(a)) is simpler and easier to interpret but lacks predictive due to the limited scope of information from a single variable.
   The multiple regression model (Task 2.2(b)) captures more relationships and offers better predictions but adds complexity.
4) The scatter plot for model 2.2(a) shows slightly non-linear relationship which shows Linear Regression Model might not be good fit for the prediction as much as Multilinear is.

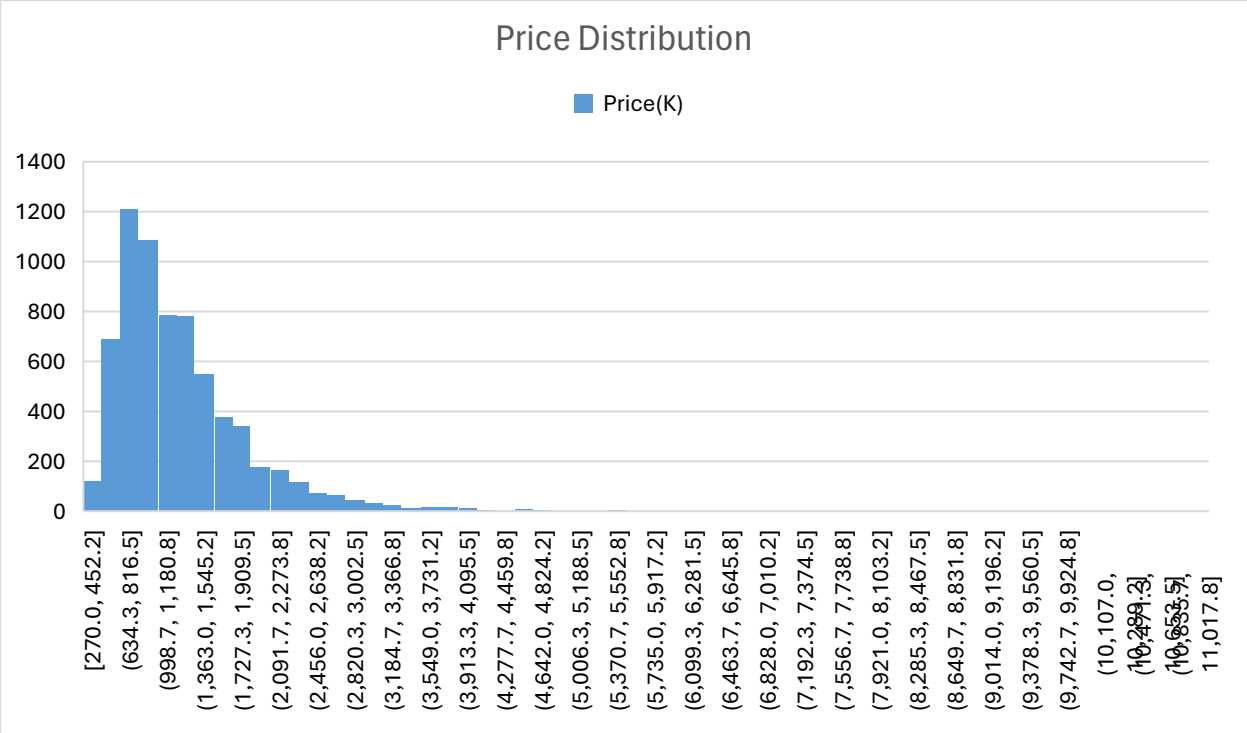
By incorporating multiple features, Task 2.2(b) reduces the omitted variable bias that affects Task 2.2(a). Due to high R^2 value and comparison it is safe to say that Multilinear Regression Model made in 2.2(b) is better than Linear Regression model 2.2(a).

TASK 2.3

Key findings:

1.) From TASK 2.1(a):

In this we perform initial distribution analysis on ‘Price’.



- The histogram shows a wide range of prices, suggesting a diverse market with properties ranging from relatively affordable to very expensive. Since it is mostly rightly skewed, the majority of prices fall in lower range.
- Most properties fall within the range of $270K to $1,363.7K, as indicated by the higher frequency in the initial bins which shows the base budget of housing for common people. A significant drop in frequency is observed as prices increase beyond $2,000K, with very few properties exceeding $6,000K which might be the property of upper middle-class people.
- The long tail on the right suggests the presence of some high-priced outliers. These outliers could be luxury properties or properties in highly desirable locations.

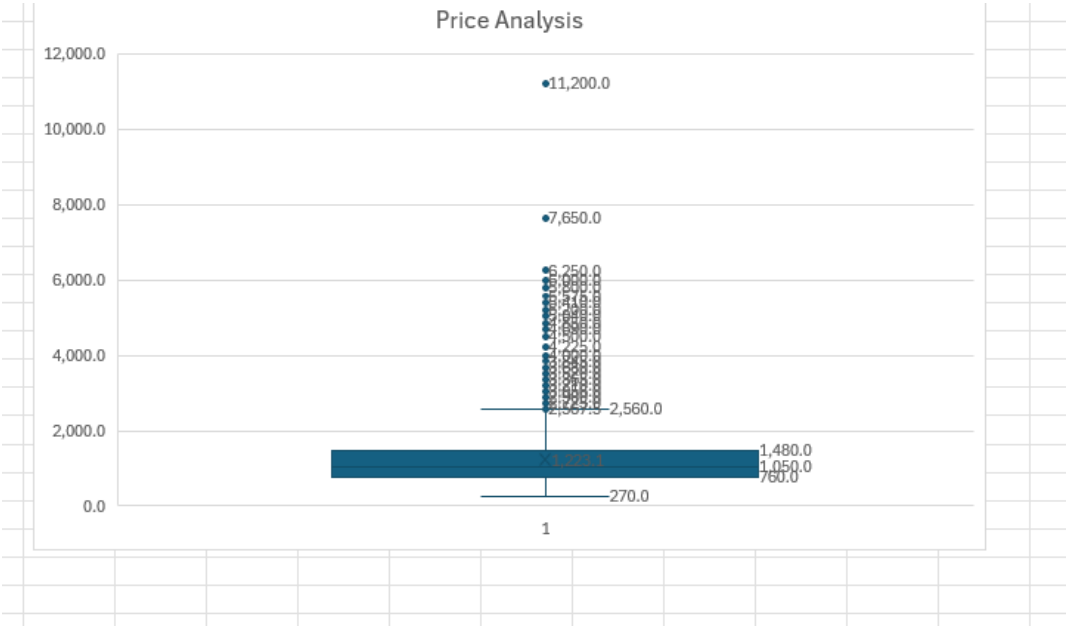
2.) From Task 2.1(b):

In this we performed descriptive analysis on price. Where,

Price Analysis:

- Mean Price: 1223.094
- Median Price: 1050
- Mode Price: 1100
- Standard Deviation: 681.8202
- Skewness: 2.585034
- Minimum Price: 270
- Maximum Price: 11200
- Range of Prices: 10930

- Interquartile Range: 720

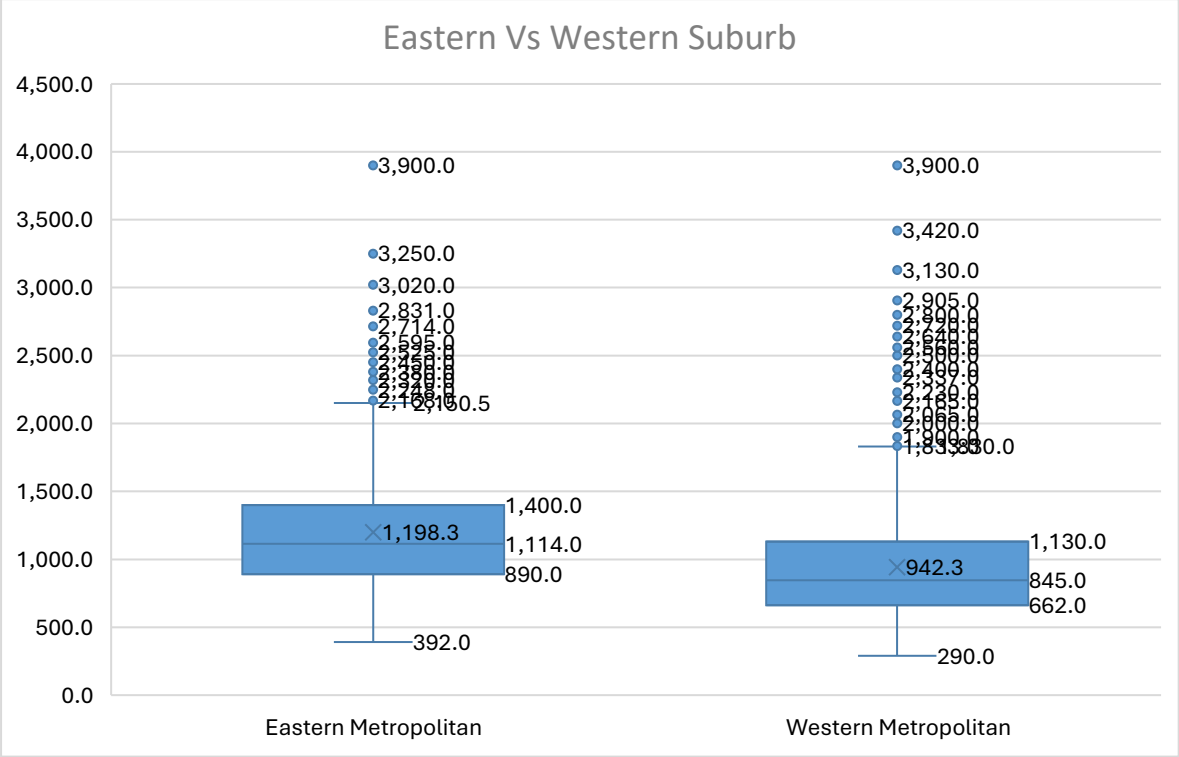


- Mean And Median: The mean price is higher than the median, which is again consistent with the right-skewed distribution. The mode price is also lower than the mean, indicating that the most frequent price point is lower than the average.
- Skewness: Skewness is 2.585034. It is rightly skewed, which represents more price distribution in lower ranges.
- Central Tendency: Standard Derivation The standard deviation is relatively high, indicating that prices are spread out over a wide range. This is also confirmed by the large range and interquartile range.
- The presence of a few high-priced items influences the mean price, while the median and mode provide a more representative measure of central tendency.
- Minimum and Maximum Range: Minimum range suggest no price is less than 270k and maximum range suggest no house price is more than11200k.
- IQR: IQR suggest measure of spread. Here it shows that price is distributed between 720 ranges.

3.) From Task 2.1(c):

In this we compared prices of houses in eastern vs western metropolitan area.

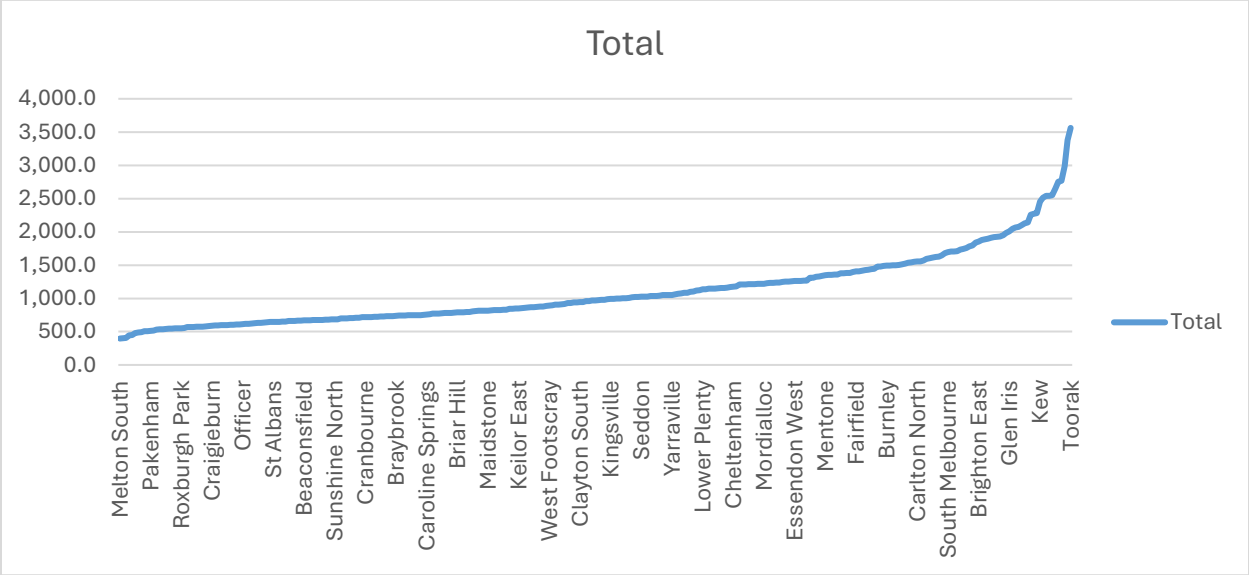
IQReast	510		
IQRwest	468		
Eastern		Western	
Lower boui	125		428
Upper boui	698		1364



- **Median Values:** The median value for Eastern Metropolitan Suburbs is around **1400.0**, while for Western Metropolitan Suburbs it is around **1130.0**. This suggests that the Eastern Metropolitan Suburbs generally have higher values compared to the Western Metropolitan Suburbs.
- Eastern Suburbs have a few higher values that pull the distribution to the right.
- The Eastern Metropolitan Suburbs have a wider range of values, from longer boxes and whiskers to the box plot. This suggests greater variability in the data for the Eastern Suburbs compared to the Western Suburbs.
- Since data points show prices of houses it is safe to say that house prices are higher in eastern metropolitan areas than in western.
- IQR: IQR for eastern metropolitan 510 and IQR western metropolitan is 468. Which price is more distributed in eastern metropolitan areas.
- Lower and Upper range: For Eastern Metropolitan they are 125 to 68 and for Western Metropolitan they are 428 to 1364.

4.) From task 2.1(d):

We perform analysis and find out which suburb has expensive house.

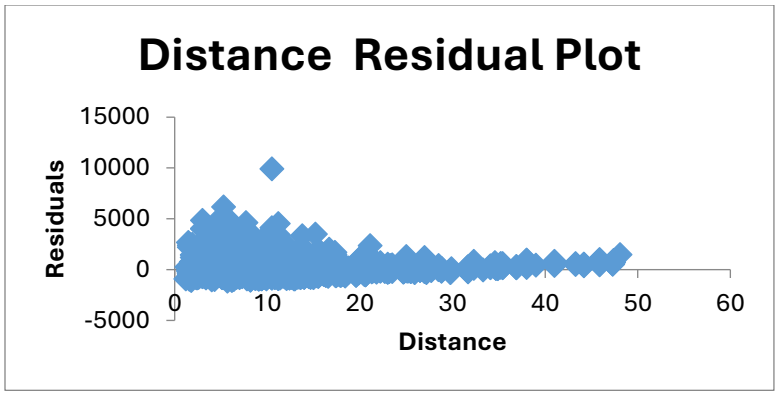


- The trend line suggests that the values generally increase as we progress through the list of suburbs.
- In between suburb prices of houses vary.
- There are a few points where the line rises sharply, indicating significantly higher values compared to surrounding suburbs. These points could be considered outliers.
- Here we can see Toorak has the highest values houses whereas Melton South suburb has less valued houses.

5.) From 2.2(a) and 2.2(b):

Task 2.2(a) is linear model where distance is independent variable and price is dependent variable. Which shows how much value of price vary solely based on distance. Here distance is variable which means distance from CBD.

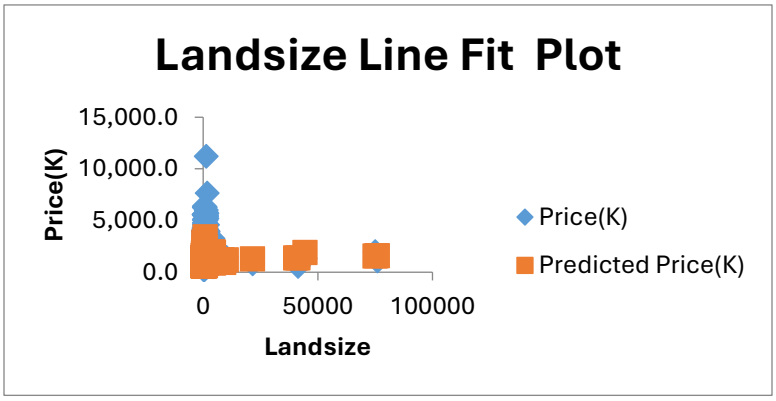
Here it is distance to price residual plot.



- From discussion in TASK 2.2(C) and the graph we can say that only distance might not be effective to predict accurate prices of houses. Hence, we add more independent variable in task 2.2(b).

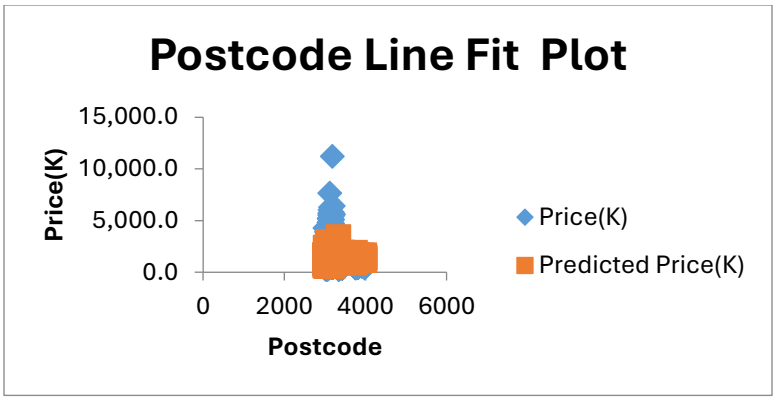
Task 2.2(b) we made Multilinear Regression Model which has three independent variables: Rooms, Landsize and Postcode.

➢ Landsize



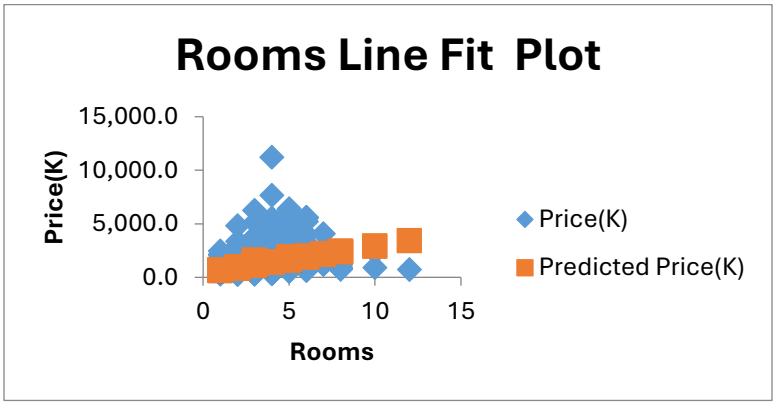
- For Landsize, predicted price and actual price are different.
- It shows upword trend means with the increases in landsize theres increased in price of house.
- Data points suggest that landsize might not only be the factor who affects prices.
- The orange bars follow a straight line, suggesting a linear relationship between land size and price. This line likely represents a regression model that attempts to predict prices based on land size.

➢ Postcode:



- This suggests the relationship between Price and Post code. It is not perfectly linear one.
- The blue scattered around the predicted prices, showing that postcode is not the sole determinant of price. Other factors like location within the postcode, property size, age, condition, etc., also play a significant role.

➢ Rooms:



- This presents the relationship between number of Rooms and house prices.
- The trend shows upwards which means with the increase in the number of rooms prices get high.

From 2.2(b) we can see the variation in house price is not solely based on one variable distance but more of combination of variable.

- With the help of correlation of independent variables, we can find which variable affects the house prices more.

	Rooms	*Postcode*	*Landsize*
Rooms	1		
Postcode	0.09513	1	
Landsize	0.05745	0.08412	1

This is correlation between Rooms, Postcode and Landsize.

- Rooms and Postcode: There's a positive correlation 0.09513 between them. This suggests that properties with more rooms tend to be in areas with higher postcodes, although they can't say that for certain.

- Rooms and Landsize: There's a weak positive correlation, 0.05745 between them. This indicates that properties with more rooms tend to have slightly larger land sizes, but the relationship is not very strong.

- Postcode and Landsize: There's a weak positive correlation, 0.08412 between them. This suggests that properties in areas with higher postcodes tend to have slightly larger land sizes, but the relationship is not very strong.

It shows that its heavily based on Rooms other than any individual variable.

Room as independent variable and price as depended variable and build linear regression model:

- The regression statistics for that is below.

Regression Statistics	
Multiple R	0.314261
R Square	0.09876
Adjusted R Square	0.098626
Standard Error	647.325
Observations	6715

- R^2 value of 0.09876 means that only 9.876% of the variation in property prices is explained by the variables included in the model. This suggests that the model has limited explanatory power.
- A standard error of 647.325 indicates that there is a considerable amount of variability in the predictions, meaning the model may not be very accurate in predicting individual property prices.