1. Problem Statement and Data Summarization:

- **Problem Statement**: To understand the real estate market of the city Melbourne and to get the best deal of price with respect to all the important features to purchase a house.
- **Dataset summary:** This dataset is taken from source https://www.kaggle.com/dansbecker/melbourne-housing-snapshot which reflects the total number of houses in real estate market of Melbourne, their prices (in dollars), type of house, distance from CBD, region and many more features. The dataset has 13580 observations and 21 variables in total.

There are 11887 null values in the dataset. The null values are in variables Car, BuildingArea, YearBuilt and CouncilArea. It is not feasible to replace these null values with mean or median values as it may add complication in the dataset and may affect the model negatively. So, here all the null values are removed. The new dataset has 6830 observations and 21 variables.

2. Planning:

The variable "Date" is further split into day and year to find the age of house by calculating the difference between variable "Year" and "YearBuilt" and the new variable is named as "houseage". While exploring the dataset, it is observed that the variable houseage as negative values and variables Landsize and BuildingArea have zero values which is not reasonable. Thus, 1023 such observations are removed from the dataset that may affect the model negatively. Thus, the final dataset contains 5791 observations and 24 columns. This dataset can now be considered as a tidy dataset and further Exploratory data analysis can be carried out.

Comparing the influence of location vs features of house, the visual interpretation is as follows:

Figure 1: Price vs Regionname with respect to number of rooms.

- From Figure 1, as the number of rooms increases, the price of houses increases. This is most likely seen in region Southern Metropolitan. Considering the house size with number of rooms between 2 to 4, Southern Metropolitan region has houses with highest prices.
- Most of the houses in the dataset has rooms between 2 to 4. Also, regions like Western Metropolitan, Southern Metropolitan and Northern Metropolitan have more houses with 2 to 4 rooms.
- On the other hand, the price of house with 6 rooms in **Northern Metropolitan region** is less than price of house with 3-6 rooms in Southern Metropolitan.
- Secondly, regions like Eastern Metropolitan, Northern Metropolitan and Western Metropolitan have houses with nearly 8 rooms and still the price of houses are less compared to that in Southern Metropolitan.

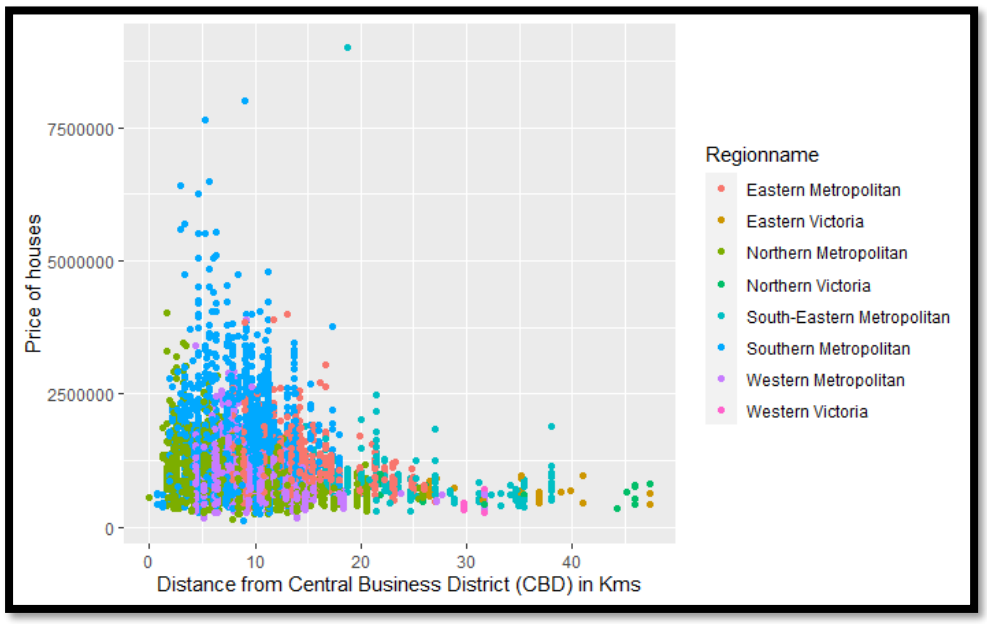
Figure 2: Price vs Regionname with respect to number of Bathrooms.

- From Figure 2, most of the houses in the dataset have bathrooms between 1 to 3. Also, regions like Western Metropolitan, Southern Metropolitan, Northern Metropolitan and Eastern Metropolitan have more houses with bathrooms from 1 to 3.
- Secondly, region like **Northern Metropolitan** has maximum number of bathrooms but still the price of house is less compared to prices in Southern Metropolitan region.
- As the number of bathrooms increases, the price of houses increases. This is most likely seen in region Southern Metropolitan. Considering the house size with number of bathrooms between 2 to 4, Southern Metropolitan region has houses with highest prices.

Figure 3: Price vs Regionname with respect to number of Car Garage.

- From Figure 3, most of the houses in the dataset has Car parking between 1 to 4. Also, regions like Western Metropolitan, Southern Metropolitan, Northern Metropolitan and Eastern Metropolitan have more houses with car parking from 1 to 4.
- Secondly, regions like **Northern Metropolitan** and Eastern Metropolitan have maximum number of car parking but still the price of house is less compared to prices in Southern Metropolitan region.
- As the number of car parking increases, the price of houses increases. This is most likely seen in region Southern Metropolitan. Considering the house size with number of car parking between 1 to 4, Southern Metropolitan region has houses with highest prices.
- On the other hand, the price of house with nearly 8 car parking in Northern Metropolitan region is less than price of most of the houses with nearly 3-7 car parking in Southern Metropolitan.

Figure 4: Distance from CBD vs Price of houses



- It can be visualised from Figure 4 that in areas of "Southern Metropolitan", the prices of houses are higher when they are close to CBD.
- On the other hand, in areas of "Northern Metropolitan" the prices of houses are less compared to "Southern Metropolitan" even if the distances of these houses are same from CBD.

Thus, it can be concluded from the above interpretations that areas of "**Northern Metropolitan**" like Broadmeadows, Brunswick East, Pascoe Vale, etc. can give the biggest house for a particular price when compared to prices in **"Southern Metropolitan"** which has most expensive houses.

From the above graphs and analysis, it can be interpreted that variables Rooms, Bathroom, Car and Distance from CBD are more likely to have an impact on prices of houses.
To evaluate the most accurate model for predicting a house's price, the house type 't' is selected i.e., "**townhome**". Thus, the final dataset contains 553 observations and 24 variables.

3. Analysis of Multiple Linear Regression Models:

All the assumptions of multiple linear regression model are fulfilled. There are three models and their analysis is as follows:

Table 1: Analysis of 3 models

Functions	Model 1	Model 2	Model 3
Predictor Variable/s	Rooms	Rooms, Bathroom and Car	Distance
Outcome Variable	Price	Price	Price
B0	218464	107200	1007233
B1	238104	105322	-10298
B2	NA	187888	NA
B3	NA	97997	NA
p-value	< 2.2e-16	< 2.2e-16	0.00439 @ 99% CI
Multiple R-squared	0.1905	0.2763	0.01463
Adjusted R-squared	0.1891	0.2724	0.01284
95% Confidence Intervals	Intercept: 97001.16 & 339927.6 Rooms: 197034.85 & 279173.8	Intercept: -13327.26 & 227726.2 Rooms: 54716.85 & 155927.1 Bathroom: 133694.14 & 242081.4 Car: 45332.79 & 150661.0	Intercept: 929527.13 & 1084939.373 Distance: -17368.68 & -3226.278
Durbin-Watson test d-value	1.529764	1.554079	1.507124

From Model 2, the predicted price of a townhouse having 2 rooms and 2 bathrooms is 6,93,619 dollars. Thus, if one wants to sell this townhouse with 2 rooms and 2 bathrooms in a given location, then the best upgrade one can make is to increase the number of bathrooms to 3 and car garage to 2. The predicted price of this upgraded house will be 10,77,500 dollars. Thus, the price of house increases by **55.3%.**

The p-value for model 1 and model 2 are significant to say that the model is accurate. But, the p-value of model 3 is less in compared to model 1 and model 2. Thus, it can be interpreted that model 1 and model 2 are better models than model 3. Further, this can be clarified more by using ANOVA technique which compares the variances between the models and predicts the best model.

Table 2: ANOVA AIC value comparison

Model	Predictor Variable/s	Outcome Variable	AIC
Model 1	Rooms	Price	15650
Model 2	Rooms, Bathroom and Car	Price	15592.05
Model 3	Distance	Price	15758.74

It can be interpreted from Table 2 that Model 2 has the least AIC which means Model 2 is the best of all.

4. Conclusion:

- Areas of "**Northern Metropolitan**" like Broadmeadows, Brunswick East, Pascoe Vale, etc. can give the biggest house for a particular price when compared to prices in areas of **"Southern Metropolitan"** which has most expensive houses.
- From Model 2, the predicted price of a townhouse having 2 rooms and 2 bathrooms is 6,93,619 dollars. Thus, if one wants to sell this townhouse with 2 rooms and 2 bathrooms in a given location, then the best upgrade one can make is to increase the number of bathrooms to 3 and car garage to 2. The predicted price of this upgraded house will be 10,77,500 dollars. Thus, the price of house increases by **55.3%.**
- P-value of Model 3 is more compared to Model 1 and Model 2. Thus, it can be concluded that distance from Central Business District (CBD) does have much impact on price of the houses.
- The AIC of Model 2 is the minimum out of all 3 models. Thus, it can be concluded that Model 2 is the best model to predict the price of houses.
- Based on graph visualization, multiple regression models and ANOVA technique; it can be concluded that the three most important features to determine the price of houses in Melbourne city are Rooms, Bathroom and Car Garage.