Melbourne Housing Data

1. Data used and Problem statement

This data set includes details on 13,580 property sales in Melbourne, Australia, with 21 variables such as the suburb, address, number of rooms, price, method of sale, type of property, real estate agent, date of sale, distance from the city centre, postcode, number of bedrooms and bathrooms, car spaces, land and building size, year of construction, council area, latitude, longitude, region name (Appendix -1).

Distance from city centre, region name, number of rooms, type of property, bathrooms, and secondary bedrooms are used to predict property prices. These features were chosen because they have a major influence on house prices and are important factors for purchasers to consider when buying a property. Distance and region name are two location-based variables that might influence property pricing, while the others are indicative of property size and quality. My objective is to develop a regression model that can accurately predict the property prices. To do this, I will evaluate several models and analyse the relationship between dependent variable 'Price' and multiple independent variables.

2. Planning

We start with looking at the distribution of prices, correlation of the numerical variables with 'Price' and their significance (Appendix-2) for identifying our initial predictor variables.

For the baseline model, type and region name were parsed as factors, all the categorical and numerical variables thrown into the regression model which showed that type-h is not significant while Eastern Victoria is the most significant among the regions. I decided to use these insights for building Model 1. Then I iteratively added other features by checking AIC / Multiple & Adjusted R -Squared to build Model 2. I followed the stepwise procedure for model 3 but instead of repeatedly using add1(), I automated the selection of the variables that provide the lowest AIC value and incorporated them to the model until no other variables could improve the model performance or it has checked for all the potential variables.

For having a reliable regression model which generalizes well, certain **assumptions** should be checked.

- Quantitative or categorical predictor variables, and continuous and unbounded outcome (The variables in the model satisfied these requirements.)
- Non-zero variance (The predictor variables had non-zero variance.)
- No perfect multicollinearity (Verified through visual inspection and the VIF test)
- Predictors uncorrelated with external variables (Assumed to be true since it is hard to confirm.)
- Homoscedastic, independent, and Normal residuals (Analysed below)
- Linearity (Verified through inspection of scatterplots.)

3. Analysis

```
Model 1: Price ~ Rooms + Bathroom + Distance
+ Type_t + Type_u + RN_EV

Model 2: Price ~ Rooms + Bathroom + Distance
+ Type_t + Type_u + RN_EV + RN_NM + RN_NV +
RN_SEM + RN_SM + RN_WM
```

Model	Multiple R^2	Adjusted R^2
Model 1	0.4437	0.4435
Model 2	0.5792	0.5788
Model 3	0.5792	0.5788

```
Model 3: Price ~ Rooms + RN_SM + Distance + Type_h + Bathroom + RN_WM + RN_NM
+ Type_u + RN_SEM + RN_EV + RN_NV
```

Variance explained and multicollinearity: Models 2 & 3 explain the highest variance of 57.88%. They have very close adjusted R^2 and multiple R^2 which means that most of the predictors used are significant. It was observed that bedroom2 and rooms are highly correlated which inflates the VIFs. Since rooms is used the three models have no perfect multicollinearity.

Residuals: The Durbin Watson not significant for the 3 models at 5% significance but the test statistic was very close to 2, so I continued with assumption of independence of residuals. The Q-Q plots of the residuals did not look normal because of some high prices which deviated the plot on one tail, however due to the large same size we proceed with assumption of normality citing the CLT.

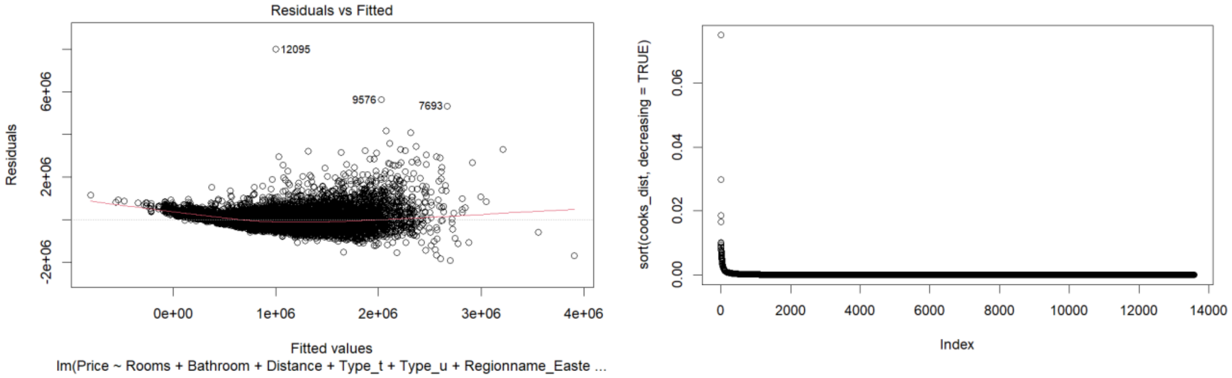
D-stat	p-value
1.364	0
1.647	0
1.647	0

Outlier and Influential points: To check for the possible outliers I tried to find out the % of standardized residuals that lie outside [+1.96, -1.96]. All models had less than 5%. The maximum cook's distance was calculated to check if points exert undue influence on the models, they were found to be less than 1. So, there were no outliers and influential points.

Comparison of models:

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	13573	3.09E+15				
2	13568	2.34E+15	5	7.52E+14	873.59	< 2.2e-16 ***
3	13568	2.34E+15	0	-2.00E+00		

Model 2 has a much smaller residual sum of squares than Model 1, according to the F-statistic and the associated p-value (Pr(>F)), with a p-value of less than 2.2e-16, suggesting strong evidence against the null hypothesis that the two models are equal. Model 3, on the other hand, has no significant difference in the total of squares when compared to Model 2. Therefore, Model 2 is the better model compared to the baseline Model 1 and Model 3.



The residual vs fitted values plot and Cook's distance for best performing model i.e., Model 2

4. Conclusion

The analysis revealed that Model 2 is the best performing regression model for predicting property prices in Melbourne. We can say that on holding all other factors fixed, the Rooms coefficient shows that adding one more room to a house result in an increase in price of $188,053.3 on average. Similarly, the coefficient for Regionname Southern Metropolitan indicates that, on average, residences in this region are more expensive than those in the other 7 regions. Similar interpretations can be done for type of the property.