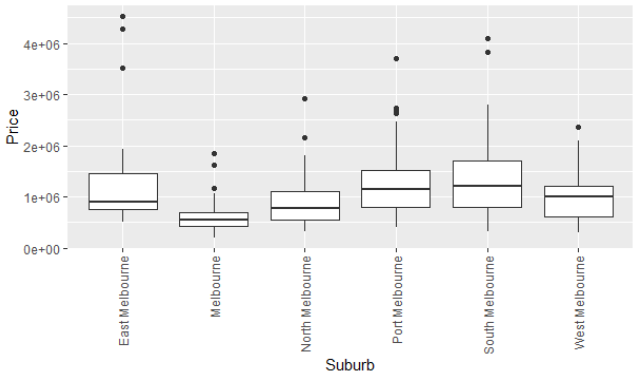
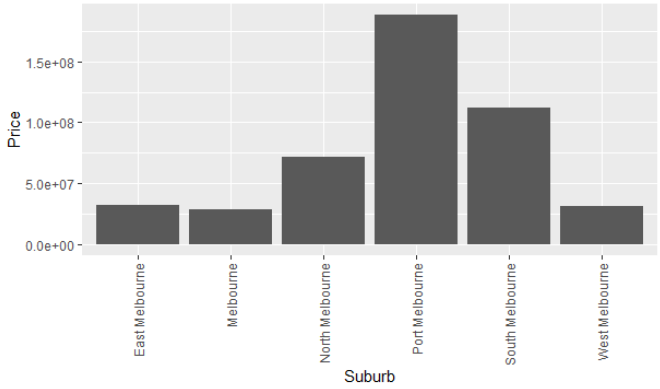
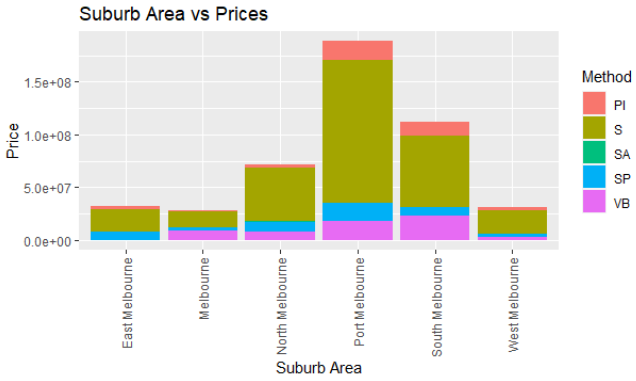
INDIVIDUAL ASSIGNMENT-4

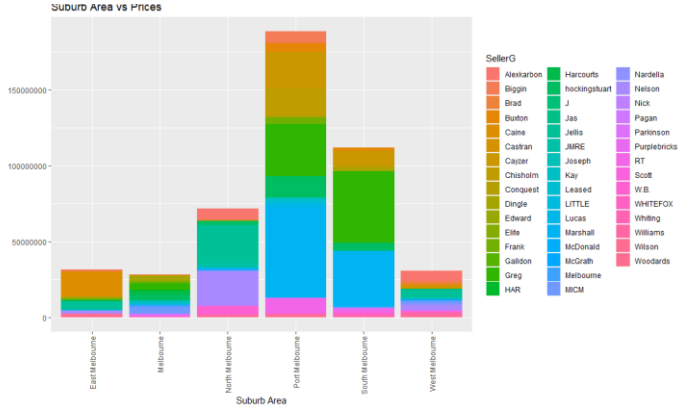
Problem Statement and Data:

The growing housing market in Australia is seeing a fluctuation in prices which can be assessed by analyzing the parameters that affect the prices of houses. A public dataset containing Melbourne's housing market data would be analyzed to find significant factors affecting housing prices and build a linear predictive model to accurately predict a home's price in Melbourne.









The preliminary analysis indicate that the 'Suburb', 'Seller' and 'Method' are few of the important factors affecting the prices. The other important factors will be assessed for their significance through inclusion as predictor variables in the model and comparing the different model accuracy.

Planning:

A combination of tibbles (8 nos.) were created using variables Price, Suburb, Seller, Type, Method, Bathroom, Bedroom, Car, Room and Landsize. Each tibble was wrangled to remove any missing data, filtered by Suburb area containing 'Melbourne' name, and finally the data were grouped by Suburb area and other category. The wrangled tibbles were fed into different combinations of linear regression models and their accuracies were compared using coefficient of determination values to arrive at the most accurate model. The selected model will be assessed for assumptions and inferences.

Analysis:

The first model was built using Suburb, Seller and Method for predicting the price and the Rsq. Value came out to be 30%. The second model was built using Suburb, Type, Bathroom for predicting the price and the Rsq. Values came out to be 57.81%. The other variations of model with three predictor variables did not show any improvement as compared to second model.

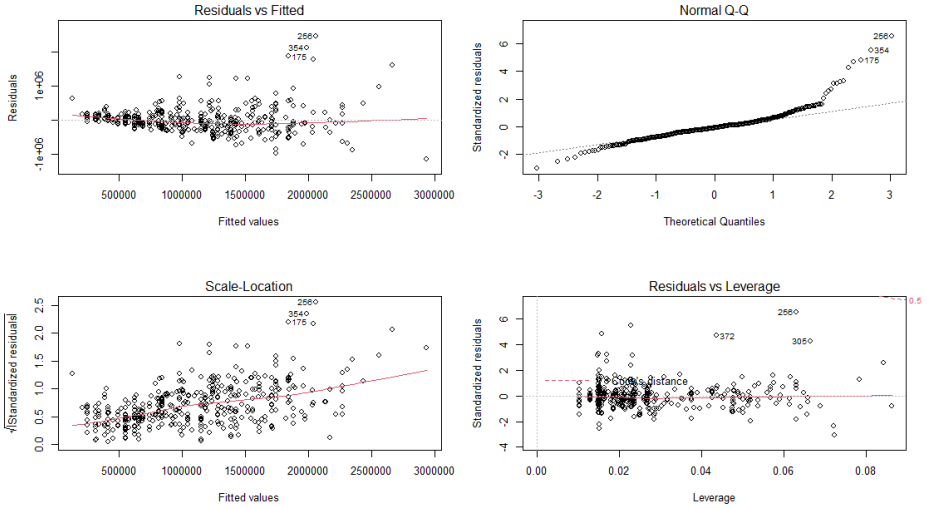
INDIVIDUAL ASSIGNMENT-4

The second model was chosen to add more predictor variables for increasing accuracy. The model with combination of Suburb, Room, Seller, Type, Bathroom, Car and Method as predictor variables for predicting price had an Rsq. Value of 68.63%. The other combinations of predictor variables could not match the Rsq. Value of this model. Hence, this particular model was chosen for predicting house prices.

After removing the insignificant variables Seller and Method from the chosen model, the Rsq. Value reduces to 65.5%. The model now has all significant variables which are Suburb, Rooms, Type, Bathroom and Car for predicting a house price. The final model equation is;

Y(Price) = 1072602 – 615305*Melbourne – 686558*NorthMelbourne – 583519*PortMelbourne – 548792*SouthMelbourne – 657781*WestMelbourne + 237936*Rooms – 409192*Typet – 626921*Typeu + 188632*Bathroom + 132230*Car + ϵ

Here, the categorical variable of Suburb and Type takes value 1 if the house of a particular type is in a particular suburb. ϵ is the error term for each predicted Y.



The Residual vs Fitted plot shows no definite pattern, hence it indicates that the model has no inherent problem. The Normal Q-Q plot indicates that most of the points (except outliers) do not significantly deviate from the reference line, so we can assume normality of residuals. The Scale-Location plot is not horizontal and indicates presence of heteroscedascity. The Residuals vs Leverage plot indicates that the outliers go beyond the +/-2 standardized residuals. The VIF values do not go beyond 2, which indicates presence of weak or almost no multicollinearity among predictor variables.

Conclusion:

The suburb of South Melbourne has the least negative coefficient and a house bought there would cost more. The suburb of North Melbourne has the most negative coefficient and a house bought there would cost less. Each new room added would increase the price by 237936$. A Type t house would cost more than a Type u house. Each new Bathroom would cost 188632$ and each new Car spot would cost 132230$. The predictive model accounts for 65.5% of the variation in prices, which indicates a good accuracy in the predictions. Although, the model violates the assumption of homoscedascity, the model's accuracy and reliability can be improved by excluding the outliers from the model data in future work.