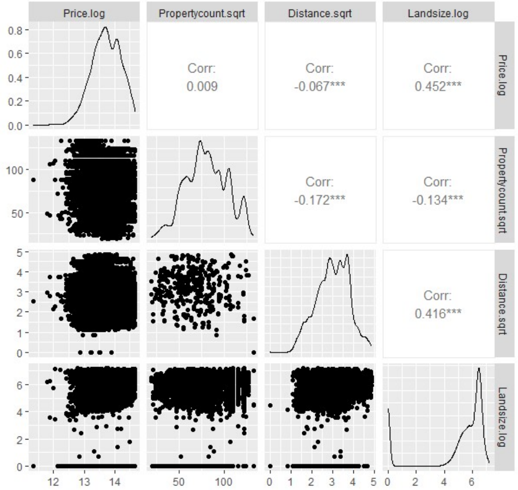
<u>Question</u>: What is the best possible combination of parameters out of Distance, Propertycount and Landsize, that could help determine the Price of housing units in Melbourne?

<u>Data Summary</u>: The Melbourne housing snapshot [1] is a snapshot of Tony Pino's Melbourne housing dataset [2]. The snapshot of the dataset is a tibble with 13,580 rows and 21 columns. As per our objective, there are 4 variables in our scope: Price (in Australian dollars), Distance (distance from central business district), Propertycount (number of properties that exist in the suburb) and Landsize (in metres), which are all numerical, continuous data.

<u>Data Cleaning</u>: There weren't any missing values for these 4 variables. I checked normality of data by visual inspection of histogram and Q-Q plots and understood that the data would need some transformations. First, I used the Interquartile range (IQR) method to filter the data into the acceptable range, using quantiles (Q1 and Q3). So, all data smaller than Q1-1.5*IQR and greater than Q3+1.5*IQR was filtered out, hence removing the outliers. After removing the outliers, there was still need for some transformations in order to better fit the data onto a Q-Q plot. Thus, post analysis, I applied log transformation on Price and Landsize, and square root transformation on Propertycount and Distance variables.



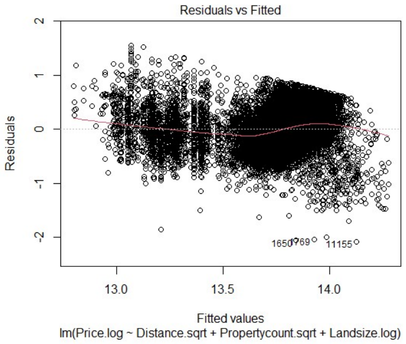
<u>Planning</u>: As per the problem objective I chose outcome variable as Price and predictor variables as Distance, Propertycount and Landsize.

<u>Assumptions</u>:

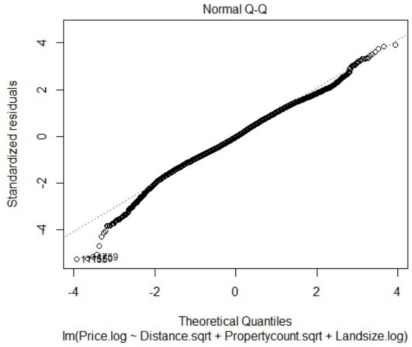
1. All predictor variables (Distance, Propertycount and Landsize) are quantitative, and outcome variable (Price) is also quantitative, continuous, and unbounded (because we cannot have negative price of a house).
2. non-zero variance: yes, the data clearly varies as per our plots.
3. no perfect multicollinearity: yes, visual inspection of scatterplots show that our 3 predictor variables do not correlate with each other.
4. predictors are uncorrelated with external variables: As per our objective, we've only selected these 4 variables (3 predictor and 1 outcome) from our dataset, so this condition satisfies as all are in scope.

We'll be validating rest of the assumptions once we have residuals for our analysis, which will be obtained once we obtain our model.

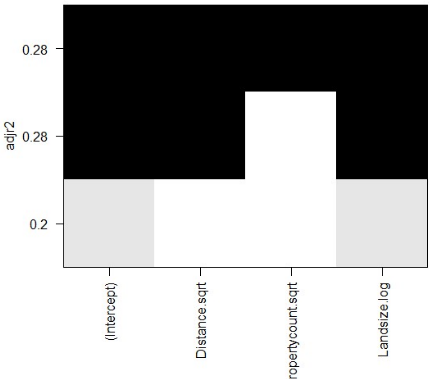
<u>Analysis</u>: The linear model(lm) function is used to compute a regression model. I regressed log(price) [outcome variable] on sqrt(distance), sqrt(propertycount) and log(landsize) [predictor variables]. On execution, the result was interpreted as: log(Price) = 13.6095 - 0.1832*[sqrt(distance)] + 0.0007*[sqrt(propertycount)] + 0.1201*[log(landsize)]. Upon summarizing the model, I found that the intercept, sqrt(distance), sqrt(propertycount) and log(landsize) are all significant. This signifies that we reject Null hypothesis and thus imply that these coefficients are all not equal to 0. Additionally, R^2=0.2844 denotes 28.44% variance is explained by this model.



<u>Assumptions continued</u>-
<u>Multicollinearity</u>: As an addition to visual inspection, we can test this using the VIF (Variance Inflation Factor): The largest VIF was 1.229, less than 10, and the average VIF was 1.159, close to 1. The lowest tolerance (1/VIF) was 0.813, much greater than 0.1(which would indicate a serious problem) and 0.2(which indicates a potential problem). We thus conclude that there is no collinearity amongst our predictor variables.
<u>Residuals</u>: On plotting the Residual v/s fitted plot of our model, some signs of constant variance(homoscedasticity) were visible; although the line wasn't perfectly linear which signifies there is room for improvement in our current model. Further, the Durbin-Watson test for independence was significant at 5% level of significance: D-W Statistic(d)=1.42, p-value=0. As 'd' is a little away from 2, it shows some degree of positive autocorrelation, which means the model could be fine tuned, but since D-W statistic is between 1 and 3, we fail to reject the null hypothesis that the errors are independent and continue with the assumption of independence met. Through the Normal Q-Q plot of residuals, we can see they are normally distributed.



<u>Choice of final model</u>: I used the all-subsets method for variable selection for our final model. This method constructs all possible combinations of predictors and compares their explanatory power. Here, I used the adjusted r^2 scale and the plot showed that either combination of all the 3 predictor variables or exclusion of only sqrt(propertycount) variable would show the equal, best result(0.28=28%).



<u>Comparison of 2 models</u>: I did comparison among 2 different models: Model1: Distance.sqrt + Propertycount.sqrt and Model2: Distance.sqrt + Propertycount.sqrt + Landsize.log. Then, I used the ANOVA (Analysis of Variance) test to check if they are significantly different. Results showed p-value<0.05 signifying that we reject H0 and conclude that there is significant difference among the 2 models. I also used the AIC (Akaike information criterion) to compare the best fit out of these 2 regression models. The aictab() function lists the model with the lowest AIC value first. From the output, we were able to see that Model2 had lowest AIC thus it was the best fitting model out of these.

<u>Conclusion</u>: We built a linear model predicting price based on the distance from central business district, number of properties in the suburb and the landsize of the housing units in Melbourne. Assumptions of the linear model were all met. As per the variables from our problem statement, we were only able to get 28% variance. Thus, we have room for improvement and possibly other variables from our dataset could be used to increase the prediction accuracy for price of housing units in Melbourne.

References:

[1] “Melbourne Housing Snapshot,” *Kaggle*, Jun. 05, 2018. https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot

[2] “Melbourne Housing Market,” *Kaggle*, Oct. 14, 2018. https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market