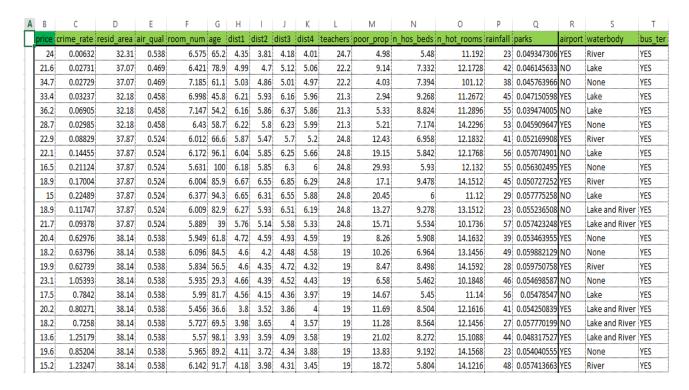
#### **DOCUMENTATION OF DATA PROCESSING STEPS**

Identify and separate the categorical from the numerical data.



# **Feature Engineering**

Refer to DATA PREPROCESSING sheet for follow up

### Missing values

From our EDD, we notice the counts of the *n\_hos\_beds* is not equal to 506 which is the number of observations. This shows that there are missing values in that feature, and to take care of this, we take the average of value (7.899767068) of the column feature and fill in the missing gaps with that value.

### **Outliers**

From our descriptive statistics shown on EDD sheet, there is a huge deviation shown in the mean and median, minimum and 25<sup>th</sup> percentile (smallest 25) value, maximum and 75<sup>th</sup> percentile (largest 125) value in the *crime\_rate* column feature. All these signs are indications of outliers. So we proceed to replacing all values of the feature greater than the 95<sup>th</sup> percentile (15.78915) with a value equals to 2 \* 95<sup>th</sup> percentile.

# Variable transformation

There is sort of redundancy in the dist1, dist2, dist3 and dist4 column features. To handle this, we create a new feature, **av\_dist**, by taking the average of these features and removing them afterwards.

### Feature encoding

Since the regression model cannot handle text or strings format, we have to encode the categorical features. We do this by using the IF() function, where set our value to 1 if true and 0 if false.

airport	Lake	River	Lake and River	bus_ter
=IF(\$N3="YES", 1, 0)	=IF(\$Q3="Lake", 1, 0)	=IF(\$Q3="River", 1, 0)	=IF(\$Q3="Lake and River", 1, 0)	=IF(\$S3="YES", 1, 0)
=IF(\$N4="YES", 1, 0)	=IF(\$Q4="Lake", 1, 0)	=IF(\$Q4="River", 1, 0)	=IF(\$Q4="Lake and River", 1, 0)	=IF(\$S4="YES", 1, 0)
=IF(\$N5="YES", 1, 0)	=IF(\$Q5="Lake", 1, 0)	=IF(\$Q5="River", 1, 0)	=IF(\$Q5="Lake and River", 1, 0)	=IF(\$S5="YES", 1, 0)
=IF(\$N6="YES", 1, 0)	=IF(\$Q6="Lake", 1, 0)	=IF(\$Q6="River", 1, 0)	=IF(\$Q6="Lake and River", 1, 0)	=IF(\$S6="YES", 1, 0)

### Feature Selection

After the correlation analysis was done on the CORR sheet, there seems to be a high positive correlation between *air\_qual* and *parks*. This high correlation can disrupt our model from learning properly, so we have to remove one out of the two features. Our preference goes to the feature with a higher correlation to the target variable and this is *air\_qual* so we remove the *parks* feature.

The data is now ready for use. (see PREPROCESSED DATA sheet)

#### **INFERENCES**

Upon correlation analysis, the two most correlated features to the target variable (price) are room\_num and poor\_prop. With room\_num having a positive correlation and prop\_prop, a strong negative correlation. Three regression models were therefore built with price as target variable, they are;

- 1. Simple linear regression on room\_num feature.
- 2. Simple linear regression on poor\_prop feature.
- 3. Multi regression on all features.
- (1) Simple linear regression on room\_num feature; The r-square value of this model found under the summary statistics of the Linear Reg 1 sheet is seen to be 0.484838974, signifying that about 48.48% of the changes in the data is explained by our model. The standard error is 6.597015858, signifying that our predicted values of price by the model has an error of +/- 6.597015858 on average. Also, on our Anova table, the F value is far greater than the F critical value meaning our regression model is significant.
- (2) Simple linear regression on poor\_prop; The r-square value of this model found under the summary statistics of the Linear Reg 2 sheet is seen to be 0.548837968, signifying that about 54.88% of the changes in the data is explained by our model. The standard error is 6.1736542, signifying that our predicted values of price by the model has an error of +/- 6.1736542 on

- average. The F value is greater than the F critical value, signifying that the regression model is significant.
- (3) Multi regression on all features; The r-square value of this model found under the summary statistics of the Multi Reg sheet is seen to be 0.720994242, signifying that about 72.10% of the changes in the data is explained by our model. The standard error is 4.923793596, signifying that our predicted values of price by the model has an error of +/- 4.923793596 on average. Also, on our Anova table, the F value is far greater than the F critical value meaning our regression model is significant.

The Multi Regression model seems to be the most effective since its R square value is larger and the standard error is the most minimal. This is accompanied by the Linear regression model on the poor\_prop feature and lastly the room\_num feature.