# Model Refresh Review Guidelines –

In this document, we lay out various steps to review the refreshed models. This guide is based on the three primary scripts used for EDA, Pricing Model and Win Loss Model construction (sample scripts attached). We dive into different sections of the script and layout any minor and/or major changes which may be required while refreshing the model. The steps are to serve as basic guidelines to help beginners towards this process and in no way are these to be taken as a comprehensive list. The review of each market may require additional steps and checks based on factors like market feedback and model upgrades which needs to be accounted separately.

Exploratory Data Analysis Script: (add a link to a sample script and comment in the script with each section)

1. Data Selection –
   1. Check the file and directory paths of master data, upcoming leases data, inflation data and yardi files (lease\_data.csv, Cleaned\_imputed\_property\_data.csv, new\_units\_data.csv)
   2. Check the dates of the data.
   3. Check the market name and market codes. Are there multiple market codes and market names for a single market? Are we including all of them while selecting the data?
2. Check on Filters –
   1. Check on the defined filters and if something doesn’t make sense. (Example selection of time range, amendment type, property type, amendment term, inactive properties etc)
3. Inflation (rent growth) Adjustments – Is data adjusted for inflation (rent growth)? Inflation adjustments should reflect in Starting Cash, NER, TI, Broker Commission amount, Brokerage, etc.
4. Feature Transformation and Feature Selection –
   1. What criteria/method is used to select the features (eg PCA, ICA, Decision Trees)? Are all the features testable with this criteria (check if assumption of the methods are satisfied by the choice of features)? Is the current method applicable for the market under review or alternate feature selection strategy is required?
   2. Are all the relevant features included in the model? What metrics and thresholds are used to select the feature? Do the features require any transformation?
   3. Are there any nans or incorrect values which require imputation?
   4. Class balance – Is the class balanced for win loss model?
5. EDA -
   1. Mixed Modelling - Which market variables are showing significant changes in their original relationship with NER? Are their contributions significant?
   2. DV vs IDV relationship – Is the linear relationship consistent between the DV vs IDV for all features cases?
   3. Cluster Analysis – Have any new properties been added? Do the clusters remain same or change? – Expected to remain mostly same.
   4. Outliers – Are there any new outliers appearing in the historical data and upcoming data? Do any of the previous datapoints now appear to be outlier with additional data? Alternatively, do any of the previous outliers appear as a relevant datapoint? Is current method of outlier detection leading to significant loss of data or biased datasets?
   5. Market Segmentation – Are all the relevant variables chosen for market segmentation? Does market segmentation change in comparison to the original distribution? Are there significant number of points corresponding to each submarket? If not, should the modelling strategy be changed?

Linear Regression Price Prediction –

1. Data Selection –
   1. Check the file and directory paths of master data, upcoming leases data, inflation data, yardi files (lease\_data.csv, Cleaned\_imputed\_property\_data.csv, new\_units\_data.csv) and distance data.
   2. Check the dates of the data.
   3. Check annual pricing, units of area, currency
   4. Check the market name and market codes. Are there multiple market codes and market names for a single market? Are we including all of them while selecting the data?
2. Inflation Adjustments – Is data adjusted for inflation? Inflation adjustments should reflect in Starting Case, NER, TI, Broker Commission amount, Brokerage, etc.
3. Check on Filters –
   1. Ensure the filters remove all the outliers identified in EDA.
   2. Check for any additional data removed for creating the model and reasoning behind the removal. Can removal of this data lead to incorrect predictions or increase bias?
   3. Do the filters incorporate the market feedback? Have the data points provided in the market feedback analysed in the EDA?
   4. Check on additional filters – like filters on DV, NRA, submarket names, amendment ID etc
   5. Check on filters for upcoming leases – Example cut off set on MLA
4. Handling missing data –
   1. Check the number of missing datapoints before imputation.
   2. Check which columns can be imputed and the method of imputation. Can imputation have any adverse effects? Is it better to remove the points of impute the data?
5. Model Construction –
   1. Feature engineering and transformation –
      1. Do any features require transformation or new features need to be created based on EDA?
      2. Is the transformation of features chosen for historical data also applied to upcoming data?
   2. Variable Selection – Which variables are originally selected? Do any new variables need to be added? How is the variable importance defined – what variable selection method is used?
   3. Model Assumption - Are the assumptions of linear regression satisfied?
      1. Weak exogenity – IDV can be assumed to be fix values with constant variance captured by error term.
      2. Linearity – There exist a linear relationship between IDV and DV.
      3. Constant Variance/Homoscedasticity – Variance does not depend on IDV.
      4. Independence of errors – Errors/Residuals are uncorrelated with normal distribution.
      5. Lack of multicollinearity – Predictor variables are uncorrelated with each other.

To what degree are the above assumptions satisfied? In case of significant violations, what possible strategy can be adopted?

1. Model Performance and Accuracy –
   1. How do the previous and new models perform against the training, test and validation datasets? Has the performance improved in comparison with previous version of the model?
   2. Are there any indication of overfitting in the new model?
   3. Are there any indications of failure of linear model based on residual analysis and regression metrics?
   4. Error Buckets – Check if errors are higher for a
      * 1. specific NRA bucket
        2. New vs Renewed lease
        3. Specific year
        4. Driving distance buckets
        5. Any particular region of other categories