**Machine Learning Kaggle Notebook Summary:**

[**Basic Time Series Analysis & Feature Selection**](https://www.kaggle.com/code/creatrol/basic-time-series-analysis-feature-selection)

**Methodology:**

1. **Data Preprocessing:** 
   1. Few functions for data transformation were created (log transformation).
   2. Columns that hold more 20% missing values and above were removed from the data set.
   3. Duplicated features were removed.
2. **Exploratory Data Analysis:** 
   1. The distribution of log price was examined.
   2. The trend of price by months was examined.
3. **Feature Engineering:**
   1. **Time frame columns:** time frames columns such as year, month, and date were created.
   2. **Log Price:** log transformation to the response variable (price\_doc)
4. **Feature Selection:**
   1. Correlation Test between every feature in the data set
   2. Correlation Test between every feature in the data set with log price
   3. Using XGBoost for feature selection.
   4. Choosing features according both correlation test and XGBoost importance.
5. **Data Preprocessing (Before deploying second XGBoost):**
   1. Missing values imputation using most frequent value.
   2. Normalizing the data using ‘l2’.
   3. Label Encoder was used to transform category features.
6. **XGBoost Model:**
   1. Training a model using K-Fold Cross Validation method with a rmse metric.
   2. Get the top 10 important features and examine their relationship with log price.

**Results:**

1. Some of the features are highly correlated with each other.
2. The highest correlation with the log price is less than 0.040.
3. Total area of the apartment is the most important feature to the first XGBoost model.
4. Time frame features created on the feature engineering phase have high importance.
5. Model’s RMSE is 0.46%

**Conclusions:**

1. Reducing the number of features is a crucial step because of the high correlation.
2. No strong linear relation with log price was detected.
3. Time features seem to contribute a lot to the model even though there’s no meaning to the order of the values (the values weren’t sorted).
4. After capping outliers for each of the top 10 features they seem to have a strong relationship with log price.

**Critics:**

1. Training XGBoost using cross validation is a smart decision, however when using a time series data, it’s important to consider the time which the event occurred while training the model. Thus, we suggest using a time series version of the cross-validation method.
2. Removing correlated features from the data set might overcome this problem however one can also use PCA to reduce the dimensionality in the data. Thus, we suggest also trying using other methods but removing.