**Machine Learning Kaggle Notebooks Summary:**

[A Very Extensive Exploratory Analysis in Python](https://www.kaggle.com/code/agzamovr/a-very-extensive-exploratory-analysis-in-python)

**Methodology:**

1. **Data Preprocessing:** Corrects data quality issues, ensuring the dataset’s integrity and reliability.
2. **Exploratory Data Analysis:**
   1. **Missing Values:** Visualize the number of missing values in each column.
   2. **Housing Internal Characteristics:** Investigates features like floor area, number of rooms and building material to understand their influence on apartment prices.
   3. **Time Series Analysis:** The author Investigated how does price changes in several time frames.
   4. **School Characteristics:** Explores variables related to school facilities and their association with housing prices.
   5. **Cultural Characteristics:** Analyzes the impact of proximity to cultural landmarks and recreational facilities on property prices.
   6. **Infrastructure Features:** Examines proximity to infrastructure such as public transport, parks, and utilities in relation to housing prices.
   7. **Variable Importance:** Built a Random Forest Regressor model to understand which features are most important (categorical features were encoded using label encoder).
   8. **Train –** **Test Comparison:** Compared between the train and test data to understand if they’re different or not.

**Results:**

1. **Housing Internal Characteristics:**
   1. Number of rooms in the apartment exhibits a high correlation with the total area of the apartment (makes sense)
   2. None of the features exhibit a strong linear relation with the price though number of rooms has a correlation score of 0.48 with the price.
   3. Housing internal features are the most important to the model.
2. **School , Cultural and Infrastructure Features:**
   1. Some of the features exhibits high correlation with each other.
   2. There is no strong linear relation between the features and the price.

**Conclusions:**

1. Most important variables are housing internal features.
2. There might be redundant information in the data (high correlation between predictors).
3. The relation between price and other features is probably not linear. However, a statistical test needs to be done to support this assumption.
4. The train and test datasets are different from each other and might hold different values / number of values in each column.

**Critic:**

1. Categorical features were encoded using label encoder. However, in some cases it’s better to use one hot encoding. Since label encoder might imply ordinality to a predictor which the ordinality is meaningless (such as sub area). Thus, we suggest using one hot encoding (dummies) instead.
2. Random Forest Regressor was used, and its performance wasn’t tested. This could’ve been addressed by Cross Validation to provide a more robust assessment of its predictive capabilities and help validate the results of the model.
3. Instead of removing missing values, considering imputation or interpolation method could retain valuable information and prevent potential bias in the analysis.
4. While using Random Forest’s feature importance is a smart decision, feature selection based on more than one method might lead to a better result. Thus, we suggest combining few feature selection methods to better understand which of the features contribute the most to the model.