Regression Techniques with the Melbourne Dataset

Pair Programming Task in your groups:
Data Science Teams

1. **Data Preparation**
   a. Load the Melbourne Housing dataset "Dataset_Melbourne.csv" from Moodle into a Pandas DataFrame
   b. Analyze the missing data and the data types
   c. Define X (features) and y (target variable)
      i. Y should be "Price"
   d. Conduct a train-test split
   e. If missing values: instantiate necessary imputers including various strategies
      i. Use a Column Transformer to apply the imputers to the correct columns
2. **Model Building**
   a. Create a Linear Regression-Classifier
   b. Create a Regression Tree-Classifier
   c. Create a Random Forest regressor
   d. For each of these models:
      i. Create a Pipeline with two steps: preprocessing (Column Transformer) and classifier
      ii. Train the Pipeline on the train set
      iii. Predict the house prices for the test set
      iv. Calculate MAE and RSME
3. **Model Optimization**
   a. Perform GridsearchCV with random forest regressor
   b. Potential parameters: n_estimators, max_depth, min_samples_split, min_samples_leaf
   c. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
4. **Report your scores**
   a. Report (only) your best scores for MAE and RSME here:
   b. https://docs.google.com/spreadsheets/d/1_dSok_0_nhMCf1tGpbldL4HDAmBCr08ryUtxiii-_c4/edit#gid=32989063
5. **Upload your notebook to Moodle**
   a. One notebook per group
   b. Notebooks don't have to be polished