Title (Topics): Data Analysis on Housing Price Prediction
Group Members: Raymond Pepper, Sophia Sieli, Stephanie Dong, Dylan Maray, Nancy Huang, Madeline Groth
Background of the topic & Importance of our topic: Several elements contribute to the variability in house prices. In this project, we plan to focus solely on how the location of houses and their size (number of rooms to be exact) affect their future market price. This is important for the future because as housing markets become more competitive and complex, data-driven insights will be crucial for managing affordability, planning sustainable communities, and making smart investment decisions, while preventing market bubbles.
Objective (Goal of the Project):
In this project, we will apply ElasticNet regression to predict housing prices. After loading and cleaning the dataset, including handling missing values and scaling numerical features, we will split the data into training and testing sets. Since regularization models are sensitive to feature scale, we will standardize all predictors. We will choose ElasticNet based on the analysis goals. Cross-validation will be used to tune the regularization parameter. We will evaluate model performance using RMSE and R^2 on both training and test sets, and interpret model coefficients to identify key predictors and assess the impact of regularization on model accuracy.
Plans:

- Data Preprocessing: We will load the dataset into a Jupyter Notebook, check for missing data, and handle it appropriately to ensure the data is ready for modeling.
- Exploratory Data Analysis (EDA): We will generate basic plots (scatterplots and heatmaps/correlation plots) to better understand the distribution and relationship among the variables.
- Testing the 5 Assumptions of Linear Regression before finalizing the model:
    - Linearity: Residuals V.S Fitted plot
    - Homoscedasticity: Constant variance of residuals across predicted values
    - Independence: Durbin-Watson Test for residual autocorrelation
    - Normality: Shapiro-Wilk Test and QQ-plot of residuals
    - Multicollinearity: Variance Inflation Factor
- Model Fitting: We will build a multiple linear regression model to predict the sale price using selected housing features by using RMSE and R-squared score.
- Model Diagnostics and Improvement: We will check for outliers using standardized residuals, identify leverage points, and detect influential points using Cook's Distance for each observation in the linear model and sum them. We might also apply transformation if needed and consider using ElasticNet Regression to improve the model stability.
- Model Selection: Choose models through AIC or BIC
- Hypothesis Testing:
    - Hypothesis 1: Test whether the area of a house significantly influences its sale prices
        - H_0: The area of a house does not affect the price **(β_area=0)**
        - H_1: The area of a house significantly affects the price **(β_area≠0)**
    - Hypothesis 2: Test the effect of area on price depending on how many bedrooms a house has. (effect of area)
        - H_0: There is no interaction between the area and the number of bedrooms.
        - H_1: The effect of area on price varies depending on the number of bedrooms.

Software:
We will use the Python language, specifically in a Jupyter Notebook environment with libraries including pandas, numpy, seaborn, matplotlib, and statsmodels.