In God we trust

Foundations of machine learning – 1st assignment

1- In this task you will carry out data analysis and build a regression a model for real state price prediction. The dataset for this task is available at https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction .
    a. Explore and clean the dataset. Explain your steps and reasons behind your actions (such as choosing a specific imputation technique).
    b. Visualize your data to gain better insights. Explain your findings. You can get some ideas from https://www.data-to-viz.com.
    c. Conduct the following tests. Include Null and Alternative hypothesis:
        i. Test if the average price per unit area of houses above the median age is significantly different from those below the median age. Use p-value method.
        ii. Investigate if the average price per unit area significantly differs by the number of convenience stores (categorized by "X4 number of convenience stores").
        iii. Choose two categorical variables and test if there's a significant association between them. (You may need to categorize some of the numerical variables.)
    d. Calculate the correlation between each variable and the target variable ("Y house price of unit area") to find which variable has the highest influence on the target.
    e. Use scikit learn to test-train split train simple linear regression model and extract the weights to confirm your findings from the last question.
2- In this we will analyze the life expectancy dataset provided by WHO. You can download it from this link: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who .
    a. Perform steps a and b similar to previous task.
    b. Conduct the following tests. Include Null and Alternative hypothesis:
        i. Investigate the relationship between Status (Developed vs. Developing) and the prevalence of a disease (for example, "Hepatitis B" vaccination rates).
        ii. Compare the life expectancy ("Life expectancy") between two groups defined by the "Status" column (Developed vs. Developing countries).

    iii. Compare the life expectancy ("Life expectancy") across multiple groups defined by the "Year" column.
    iv. Design 2 more hypothesis tests to further explore the dataset.
c. Calculate the correlation between each variable and the target variable ("Life expectancy")
d. Test-train split your dataset such that there are no countries in both test and train splits. Now train two linear regression models, one including the column "Country" and another one with this column removed. How did it affect your model performance on test split? Which approach is correct?
e. Now test-train split the dataset randomly and perform the previous steps. How did the results change? Compare these results to part d. Which approach is correct?
f. Train a simple linear regression model and another one with L1 regularization. Compare the model weights. Explain the results with the correlation values you calculated before.