1. I will perform the analysis of this problem myself.

2. Problem Statement: as a bike lover, I have no idea which car I should buy after graduation. Many people advised me to buy a car that has a good Return on Investment i.e resell-able in the future with the best price. But what resale price should I make for my used car? What elements should I look for to maximize the price? Will it be the brand of the car, or the historic usage of the car, or some specific elements of the car, or the listing location, or even the ZIP code of the dealer, or a combination of some of them?

3. Data Source: https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset/data

4. About the data
    a. 3 million rows, 66 columns
    b. 10GB
    c. I want to predict the price of a used car if it is being re-listed, indicated by the column 'price', which is a numerical value.
    d. Other columns have a mix of numerical and text values.
    e. There exists null values in the data.
    f. Limitation: the metadata from the data source only gives descriptions of half of the 66 columns, but since I don't have much domain knowledge for this dataset, I will analyze the features based on their data types and traditional statistical analysis.
5. My plan for the data:
    a. First, I will perform the traditional statistical analysis.
        i. Collect data
        ii. List all possible features
        iii. Investigate possible new features using existing features
        iv. Investigate Data
            1. Check massive Missing Data
            2. Check duplicate columns / rows
            3. Transform data type of Categorical / DateTime data (still careful encoding NaN)
            4. Check small Missing Data
            5. (optional) Encoding
        v. While True:
            1. Visualize the data -- Correlation Matrix
            2. (Optional) Transform the data
            3. Regression Full Model
            4. Check Assumptions and Metric
            5. If we're happy: collect the model
            6. Else:
                a. Use (8) to create new model
                b. Back to step 5
        vi. Improvement:
            1. Special Points
                a. Outliers
                b. Influential Points
                c. Leverage Points
            2. This line is valid!
                a. Hypothesis Testing for coefficients
                b. Confidence Interval for coefficients: includes 0 or not
                c. Plot of forecast and residuals
                d. Exaggregated T-Stats
                e. Intuitively wrong sign of correlation coefficient:
                    => possibly multicollinearity or suppressor variable
            3. This line can be better!
                a. Adjusted R^2
                b. Omitted variable bias
                    i. Fixed effect
                    ii. Random effect
                c. Outliers
                d. Influential Points
                e. LOOP:
                    i. Bias-Variance Tradeoff:
                        1. Underfitting / High Bias
                        2. Overfitting / High Variance
    b. Then, I will optimize the runtime by purely using mathematical manipulation and libraries we learned in class.
        i. Gradient Descent
        ii. QR Factorization
        iii. and more.