The most challenging aspect of applying regression analysis to this dataset involved the 'CarName' column, which contained 147 unique text values that needed conversion to numeric values. Typically, I would use pd.get\_dummies for this task. However, adding 147 columns seemed excessive and likely to degrade the model's performance. Consequently, I opted to extract the unique company names from each record and assign ratings from 1 to 3. This required devising a method to isolate the company name from the 'CarName' column without the exact model specifications. I achieved this by using a lambda function on a new column, (a copy of 'CarName'), to split the text at the first space, retaining only the company name. Next, I introduced an additional column that utilized a map function to assign ratings to the car companies based on the new column's records.

Moreover, numerous car names were misspelled, with variations like 'toyota' and 'toyouta', 'porsche' and 'porschce', among others. This necessitated including both spellings for all errors in the dictionary used by the map function to rate the companies. Afterward, I removed the original 'CarName' column from the dataset, replacing it with the newly created ratings column.

This experience taught me that real-world data is often unclean and requires thorough cleaning; neglecting this step can significantly impair your model's performance. Additionally, it's crucial to manually review the data before executing any machine learning models, as I only noticed the spelling errors during this process.at.

I formed a subset of the dataset by including only those columns that showed a correlation with the 'price' column above a certain threshold. Naturally, the 'price' column had a perfect correlation score of 1 with itself, so it was initially included in the subset. However, it was necessary to remove this column from the dataframe to separate the 'X' and 'Y' data. Issues arose when I attempted to execute the cell that adjusted the correlation cutoff value (which was above the one that removed the 'price' column) with all the subsequent cells that used the correlation cutoff. Since the 'price' column was no longer present in the dataframe, it resulted in an error of looking for a column name that was already removed. To address this issue, I enclosed the code that removed the 'price' column within a try-except block to catch the Exception and used a ‘pass’ statement to ignore the error.

***Reflect on the interpretation of the regression model results. How did you derive actionable insights from the model outcomes to make data-driven decisions for the real-world problem?***

For my ensemble models which included Random Forest and Gradient Boost I used all columns for the ‘X’ values. For linear regression I used both the full dataset and a trimmed subset with only closely correlated columns, and for the polynomial model I was only able to use the trimmed subset to get meaningful results.

What I realized from the results of all my models was that without exception when using a very small amount of testing data the model performed much better. Although this seems not so impressive since we are only testing a small amount of records in the testing data, but since I was using cross validation and ran the models with eighty and ninety folds, that shows that the model was consistently performing well. I was also able to see which features were more closely correlated. The highest correlations to price were engine size, cylinder number, wheelbase, horsepower, curb weight, car length, and car width. This can be useful when it is costly or time-consuming to gather all the necessary data. All of these were generally about a .7 correlation score and working with these ‘X’ values my linear and polynomial models were able to achieve an R squared score of .82 (with 93 cross validation folds) and .83 (with 83 cross validation folds) respectively. The ensemble methods performed even better both getting an R squared of approximately .92 and an RMSE score of approximately .17% of the mean price, and an MAE of approximately .12% of the mean price.

In the process of interpreting the regression model results, I utilized a variety of models, including ensemble models like Random Forest and Gradient Boost, linear regression, and polynomial models. These models were trained using different subsets of data, with the ensemble models using all available features, the linear model using both the full dataset and a subset of closely correlated features, and the polynomial model only achieved meaningful results with the subset of the data.

The insights derived from these models were quite revealing. It was observed that the models performed significantly better when a smaller amount of testing data was used. This might initially seem unimpressive due to the smaller sample size, but the consistent performance across multiple cross-validation folds (80 and 90 folds) indicated the robustness of the models.

Further analysis of the feature importance revealed that engine size, cylinder number, wheelbase, horsepower, curb weight, car length, and car width had the highest correlation with the price. These features had a correlation score of about 0.7. This insight is particularly valuable when data collection is costly or time-consuming, as it allows for a focus on the most relevant features.

The linear and polynomial models, when trained with these highly correlated features, achieved an R-squared score of 0.82 and 0.83 respectively, with a high number of cross-validation folds. However, the ensemble methods outperformed these models, achieving an R-squared score of approximately 0.92, and an RMSE and MAE score of approximately 17% and 12% of the mean price respectively.

These results provided actionable insights for making data-driven decisions. For instance, knowing which features are most closely correlated with the price can guide data collection efforts, and understanding the performance of different models on different subsets of data can inform the choice of model and data preprocessing steps for future predictions. The consistent performance of the models across multiple cross-validation folds also provides confidence in their robustness and generalizability to new data. Thus, these insights can guide the development and deployment of effective predictive models for real-world problems.

How did the application of regression analysis in the case study scenario enhance your critical thinking and problem-solving skills? Provide specific examples of how regression analysis aided you in making informed decisions [Practice this question as if you were in an interview! ]

The application of regression analysis in the case study scenario made good use of my critical thinking and problem-solving skills. Here are some specific examples:

1. **Feature Selection**: The first step in my analysis involved deciding if all features were useful in predicting the price. This required an evaluation of the data and an understanding of how different features might influence the outcome.
2. **Correlation Analysis**: I used trial and error to find a correlation value that accurately identified which features were predictive of the price.
3. **Understanding Error Metrics**: When dealing with machine learning algorithms, I had to critically understand the error metrics, how to interpret them, and which ones to use. This involved somewhat of a dive into the mathematics behind these metrics and a careful consideration of their implications for my model.
4. **Data Presentation**: It is crucial to present the error metrics (RMSE and MAE values) in a way that is understandable. In this case it meant presenting these metrics as a percentage of the mean price.
5. **Nonlinear Relationships**: The regression analysis made me think outside the box by considering that the relationship between some features and the price might not be linear. This led me to explore other models, such as polynomial regression.
6. **Polynomial Regression**: When running the polynomial model, I had to account for multiple variables, such as the number of folds for cross-validation, the degrees for the polynomial, and a cutoff point for my correlation value. This required a systematic approach to model selection and optimization.
7. **Extrapolation of Knowledge**: The knowledge and skills I gained from this analysis can be applied to many other areas of inquiry and research. For example, the understanding that not all data and features are equal and need to be weighted accordingly, and that relationships between causes and effects can be nonlinear and interconnected in complex ways, can guide my approach to future data analysis tasks.

The most valuable lesson I learned from completing the case study was that one must employ both a high-level broad approach when trying to decipher trends in a dataset and granular calculations. Furthermore, a trial and error approach is also necessary to reach usable insights.

Ultimately, the goal of data analysis is to make informed decisions based on data. I plan to apply the skills and insights I have gained from this case study to real-world problems, using regression analysis and other data analysis techniques to derive actionable insights from data. This could involve anything from predicting sales for a business, to identifying risk factors for a health condition, to informing policy decisions in government.