The Assignment: Auditing the Used Car Data

The initial dataset, while comprehensive in scope, suffered from several significant data quality issues that made it unsuitable for direct use in a machine learning model. The most critical problems were the high volume of missing data in key predictive columns like *condition* and *cylinders*, and the presence of extreme, nonsensical outliers in the *price* and *odometer* fields. Values such as a price of $0 or a multi-million-dollar listing, combined with thousands of duplicate entries, would have severely compromised the performance and reliability of any predictive analysis.

Through a systematic process of cleaning, imputation, and outlier removal, we have curated a robust and reliable dataset. The exploratory analysis on this cleaned data revealed strong, logical relationships that confirm its predictive power. Key insights include the clear negative correlation between a car's price and its age or mileage, and the significant impact that features like condition, manufacturer, and transmission have on its value. For instance, vehicles in 'excellent' condition command a demonstrably higher price than those in 'good' condition, validating the quality of the underlying data.

Based on this comprehensive audit, the data is now in a suitable state to begin building a baseline price prediction model. The cleaned features provide a solid foundation, and the insights gained confirm that the necessary signals for accurate prediction are present. My final recommendation is to **proceed with the modelling phase** using this curated dataset.