Used Car Data Analysis Project

by Sam Buwalda | Portfolio Project, 2025

Table of Contents

Project Description

This project analyzes a large-scale dataset of 426,880 used car listings scraped from Craigslist, with a focus on identifying key factors that influence car pricing, value retention, and common traits of higher-priced vehicles. Using Python and common data analytics libraries, the analysis progresses from descriptive insights to diagnostic evaluations and ends with predictive modeling, reflecting a structured and professional data analysis process.

The main purpose of this project is to showcase my data analytics capabilities through a real-world case study that demonstrates not only technical skill but also analytical thinking. This project serves as a core part of my professional portfolio, aimed at helping hiring managers, recruiters, or employers evaluate my ability to extract business-relevant insights from messy, real-world data. I specifically chose a well-known and respected dataset to ensure familiarity for the viewer while allowing me to display my data cleaning, analysis, and visualization skills in a meaningful way.

Who Might Find This Valuable?

- **Hiring managers and recruiters** seeking a data analyst with practical experience and a business-oriented mindset.
- **Marketplaces that offer used cars for sale**, as well as **sellers and buyers of used cars**, who are interested in understanding pricing patterns, value retention, and which features influence a car's resale value.
- **Fellow aspiring data analysts** looking for inspiration or structure for their own portfolio projects.

Business Questions Overview

This project is built around five business questions that reflect realistic challenges a used car marketplace might face. The questions are designed to mirror the type of insights that would help platforms like Carvana, Vroom, or CarMax—and by extension, individual sellers and buyers—better understand how used car prices are influenced. Together, these questions follow a natural progression from **descriptive** to **diagnostic** and **predictive** analysis, providing a comprehensive view of the data.

1. What are the most common traits of cars priced above $20,000?

This **descriptive** question explores which categorical features (such as fuel type, transmission, or manufacturer) are most frequently associated with higher-priced vehicles.

2. How does car age affect price?

This **diagnostic** question examines the relationship between a car's age and its listed price. It aims to answer how much value a car typically loses as it gets older—a key concern for both buyers and sellers.

3. Which car brands retain their value best over time?

This combined **descriptive and diagnostic** question compares brand performance in terms of value retention. It looks at how the average price of cars from different manufacturers changes with age, helping to identify which brands depreciate slower and thus offer better long-term value.

4. How do fuel type and transmission affect car price?

This **diagnostic** question investigates how two specific technical characteristics—fuel type (e.g., gas, diesel, electric) and transmission (automatic or manual)—influence a car's resale price.

5. What factors most influence the price of a used car?

This is the most advanced question in the project, since it combines both **diagnostic and predictive** analysis. To answer it, I built a model that estimates which variables—such as mileage, year, condition, or brand—have the strongest impact on price.

Dataset Description

This project uses a large, publicly available dataset containing 426,880 used car listings scraped from Craigslist across the United States. The dataset was sourced from Kaggle, and is commonly used in the data analytics community due to its size, realism, and relevance. It provides a reliable starting point for showcasing analytical skills within a practical business context.

- **Link to raw data source**: Kaggle - Used Car Dataset
- **Link to data files used**:
  - **Raw dataset**: https://drive.google.com/file/d/1TObFqXA3n7xa0W-Ml7C9dF7JRIE1OCTy/view?usp=sharing
  - **Cleaned dataset**: https://drive.google.com/file/d/1cRSfon5fKlvNKiB6BffyfoM-im7p92me/view?usp=sharing
- **Size**: 426,880 rows and 26 columns (before cleaning)

The dataset includes a wide range of real-world variables, making it well-suited for analyzing used car pricing, brand value retention, and vehicle traits. Key features used in this project include:

- **price** – Listed sale price
- **year** – Year of manufacture
- **manufacturer** – Car brand
- **fuel** – Type of fuel (e.g., gas, diesel, electric)
- **transmission** – Transmission type (automatic or manual)
- **odometer** – Mileage reading
- **type** – Vehicle type (e.g., sedan, SUV, truck)
- **state** – U.S. state where the car was listed

The dataset's volume, complexity, and diverse set of features make it ideal for applying core data analytics techniques—ranging from descriptive statistics to predictive modeling—in a realistic and business-oriented setting.

Dataset Cleaning

The cleaning strategy is designed to align with the business questions and ensure that the dataset is relevant, reliable, and focused. Before answering the question with the analysis, I will clean the dataset by:

- Dropping irrelevant columns that are not needed for answering the business questions.
- Dropping columns with a very high amount of missing values.
- Filtering out unrealistic prices and odometer readings.
- Removing duplicate rows.
- Dropping rows with missing values in key columns (like year, manufacturer, transmission, etc.).
- Keeping only features that are relevant in answering the business questions.

Tools and Technologies Used

Programming & Development Environment

- **Python** – Primary programming language used throughout the project
- **Jupyter Notebook** – Interactive development environment for combining code, visualizations, and documentation

Data Loading & Manipulation

- **pandas** – For data cleaning, manipulation, transformation, and aggregation
- **NumPy** – For numerical operations, especially when working with arrays or log transformations

Data Visualization

- **matplotlib** – Used for basic plotting and customizing figures
- **seaborn** – For creating statistical plots like bar plots, scatter plots, heatmaps, and boxplots
- **value_counts & groupby()** – Built-in pandas functions for summarizing categorical distributions and group analysis

Feature Engineering

- **Datetime parsing** (`pd.to_datetime`) – Used for calculating car age from posting date
- **Custom functions & binning** – For categorizing cars into age groups (e.g., 1–3, 4–6 years)
- **One-Hot Encoding** (`pd.get_dummies`) – For converting categorical variables into numeric format for modeling

Statistical Analysis

- **Correlation analysis** (`.corr()`) – To assess relationships between numerical variables
- **Descriptive statistics** (`.describe()`) – To understand distributions and identify outliers
- **Log transformation** (`np.log1p`) – To normalize price distribution before modeling

Machine Learning & Modeling

- **scikit-learn (sklearn)** – For predictive modeling and evaluation
  - **RandomForestRegressor** – Used to model feature importance for price prediction
  - **train_test_split** – For creating, training, and testing datasets
  - **mean_squared_error & r2_score** – For evaluating model performance

Analytical Approaches

Question 1: What are the most common categorical traits of cars priced above $20,000?

Objective:

This question is a Descriptive Analysis task. The goal was not to explain why certain traits lead to higher prices, but simply to identify which categorical features are most frequently found in the higher-end segment of the used car market (defined as listings priced above $20,000).

Analytical Reasoning and Approach:

To answer this, I applied a classic segmentation technique: filtering the dataset to isolate only vehicles priced above $20,000, creating a focused subgroup for trait profiling. This subset allowed me to examine categorical feature distributions specific to higher-priced listings, enabling straightforward comparison of frequency patterns.

I specifically chose the following categorical variables based on business relevance and data completeness:

- `manufacturer`

- `fuel`
- `transmission`
- `title_status`
- `type`

These features were selected because they are interpretable, available for the majority of listings, and provide actionable insights for businesses targeting the premium used-car segment (e.g., Carvana or Vroom).

Techniques and Tools Applied:

- **Filtering & Subsetting:** Filtered the dataset using pandas to create a high-price subset where `price > 20000`, narrowing the analysis scope to high-value listings.
- **Frequency Analysis:** Applied `.value_counts(normalize=True)` to compute percentage-based distributions of the selected categorical traits. This gave a clear picture of what dominates this segment without being skewed by dataset size.
- **Data Visualization:** Bar plots were created using Matplotlib and Seaborn to visualize the top 10 categories per trait. A horizontal layout was chosen for better readability and cleaner comparisons.

Question 2: How does car age affect price?

Objective:

This question falls under Diagnostic Analysis. The goal was to uncover whether a vehicle's age helps explain variation in used car prices—specifically, to determine if older cars consistently sell for less, and how sharply value declines over time. Unlike descriptive profiling, this analysis seeks causal patterns in how one variable (age) influences another (price).

Analytical Reasoning and Approach:

To answer this, I focused on examining the direct relationship between car age and its price. Since both variables are continuous, a scatter plot was the most appropriate visual tool—it allows us to inspect the trend across a broad range of values and identify non-linearities, clusters, or saturation points.

To support the visual evidence with quantifiable results, I calculated the Pearson correlation coefficient. This measures the strength and direction of the linear relationship between car age and price, confirming whether older cars reliably depreciate and how strongly age alone predicts value loss.

The broader goal was to diagnose whether car age is a primary driver behind lower prices in the used car market—and if so, whether that relationship is linear, exponential, or plateaus over time.

Techniques and Tools Applied:

- **Feature Engineering:** Created a new `car_age` column by subtracting each car's manufacturing year from the most recent posting year (`2021`) in the dataset. This transformation was necessary since price decay depends on current age, not production year.
- **Data Cleaning and Filtering:** Removed outliers and invalid entries (e.g., cars with `car_age < 0` or `> 80`, or `price <= 0`) to reduce distortion and ensure interpretability in the scatter plot. Antique vehicles and erroneous listings were excluded.

- **Data Visualization:** Used Seaborn's `regplot()` with LOWESS smoothing to display the relationship between car age and price. This approach reveals non-linear trends clearly—showing how prices drop sharply within the first 10 years, then stabilize around the 20-year mark.
- **Correlation Analysis:** Computed the Pearson correlation coefficient between `car_age` and `price`. The resulting value of `-0.43` indicated a moderate negative relationship—confirming that while age affects price, it's not the sole determinant.

Challenges and Solutions:

- **Non-linearity in Depreciation:** The relationship between age and price is not strictly linear. To account for this, I applied a LOWESS (Locally Weighted Scatterplot Smoothing) trend line to capture non-linear decay—especially the plateauing effect seen around 20 years.
- **Outliers and Saturation Effects:** Extremely old listings (80+ years) or suspicious price entries can distort visualizations. These were filtered based on realistic business thresholds (e.g., capping age at 80, excluding 0-priced listings).
- **Time Reference for Age Calculation:** The dataset contained various posting dates. I standardized the reference year to 2021, based on the most recent timestamp in the data. This ensured consistency across all calculated `car_age` values.

Question 3: Which car brands retain their value best over time?

Objective:

This question combines both Descriptive Analysis and Diagnostic Analysis. Descriptively, it summarizes how different car brands perform in terms of value retention over time. Diagnostically, it explores patterns in depreciation to understand which brands consistently hold their value—and how their price declines at various age intervals. The broader goal is to guide stakeholders toward brands with better long-term resale potential.

Analytical Reasoning and Approach:

To answer this, I examined brand-level price trends across distinct vehicle age brackets. This bucketed approach made it easier to compare how much value each brand loses over time, while reducing sensitivity to outliers or uneven distribution across exact age values.

I defined four custom `age_group` categories:

- `1–3 years` (newer cars)
- `4–6 years` (mid-age)
- `7–10 years` (older)
- `10+ years` (very old)

I then grouped the data by `manufacturer` and `age_group` to calculate the **mean**, **median**, and **count** of prices for each combination. This allowed for both trend visualization and statistical reliability filtering (i.e., removing brand-age combinations with low sample size).

Finally, I used two heatmaps:

1. A value heatmap showing average prices by brand and age group.

2. A depreciation heatmap showing percentage drops between age brackets and total long-term depreciation.

These tools allowed me to identify which brands consistently depreciate slowly—and which lose value sharply—helping answer the question with both numerical clarity and visual impact.

Techniques and Tools Applied:

- **Feature Engineering:** Created a custom `age_group` column based on car age (`car_age`), allowing for meaningful segmentation across consistent age ranges.
- **Group-Based Aggregation:** Used pandas `groupby()` and `.agg()` to compute the **mean**, **median**, and **listing count** for each brand-age group pair. This ensured robust trend summaries and enabled sample-size filtering.
- **Data Filtering:** Excluded combinations with fewer than 50 listings to reduce fluctuations in averages and avoid misleading results driven by small sample sizes.
- **Data Reshaping for Visualization:** Used `.pivot()` to reshape the aggregated data into a matrix format required for heatmaps—where rows represented brands and columns represented age groups.
- **Data Visualization:** Created two Seaborn heatmaps:
  - A **mean price heatmap** showing how each brand's resale value changes with age.
  - A **depreciation percentage heatmap** showing value loss between age groups and total depreciation from new to old, in percentages.
- **Custom Logic for Total Depreciation:** Wrote a custom function to calculate total depreciation from the `1–3 year` group to the latest available group (`10+`, `7–10`, or `4–6`), ensuring each brand had a fair total depreciation estimate. Without this, only manufacturers with listings for cars over 10 years old would be included in the total depreciation rate.

Challenges and Solutions:

- **Missing Age Groups for Some Brands:** Certain brands (e.g., Alfa Romeo) had price data only for newer vehicles. While their short-term depreciation looked impressive, they were excluded from final conclusions due to missing values in older age brackets—ensuring analytical fairness.
- **Low-Sample-Size Combinations:** Some brand-age group combinations had very few listings, which could lead to unreliable or misleading averages (e.g., a single expensive listing skewing the mean). To address this, I applied a strict filter: only combinations with at least 50 listings were included in the final analysis. This ensured statistical reliability and minimized volatility in computed values.
- **Logical Sorting of Age Groups:** To preserve chronological coherence in visualizations, I explicitly defined a `CategoricalDtype` to preserve the logical order (`1–3 < 4–6 < 7–10 < 10+`) in the heatmaps.

Question 4: How does fuel type and transmission affect car price?

Objective:

This question is a Diagnostic Analysis task. It investigates whether two categorical vehicle features—fuel type and transmission—help explain price variation in the used car market. The goal is to diagnose whether specific feature combinations (e.g., electric + automatic) are associated with higher or lower prices, and whether these traits can be considered pricing drivers within the dataset.

Analytical Reasoning and Approach:

To answer this, I first explored the distribution of `fuel` and `transmission` types in the dataset to identify dominant categories. To maintain interpretability and analytical clarity, I excluded ambiguous categories such as "other" from both variables—ensuring that only clean, well-defined feature values were retained.

Next, I examined how prices vary:

1. **Independently** across fuel types and transmission types using boxplots.
2. **Jointly** across fuel+transmission combinations using a grouped bar chart of median prices.

Boxplots were chosen to capture not only central tendencies (medians) but also variability and outliers within each category. A grouped bar chart was then used to identify how combined features interact—revealing nuanced price patterns that may not appear in univariate analysis.

Finally, I conducted a focused investigation on an unexpected result: the unusually high median price of hybrid+manual vehicles. Rather than dismiss it as noise, I traced the result back to a cluster of identical listings—all priced exactly at $19,990—which explained the statistical anomaly without skewing interpretation.

Techniques and Tools Applied:

- **Categorical Filtering:** Filtered the dataset to include only vehicles with clearly interpretable values in both `fuel` (`gas`, `diesel`, `hybrid`, `electric`) and `transmission` (`automatic`, `manual`). This removed vague categories that would otherwise reduce analytical precision.
- **Data Visualization – Boxplots:** Used Seaborn boxplots to compare the price distributions for each fuel type and transmission type individually. This provided insight into both median values and distribution spread, including outlier behavior across each category.
- **Data Aggregation – Median by Combination:** Grouped the filtered dataset by both `fuel` and `transmission` to compute the **median price** for each combination. This is more robust than the mean in datasets with extreme outliers or non-normal price distributions.
- **Data Visualization – Grouped Bar Chart:** Visualized the median price of each fuel+transmission combination using a grouped bar chart. This clearly illustrated which configurations tend to command higher prices and highlighted interactions between the two categorical variables.
- **Outlier Validation via Head Inspection:** Upon detecting an unusually high median price for hybrid/manual listings, I inspected the top entries in that group using sorting and `.head()`. This revealed a cluster of Hyundai Sonata Plug-in Hybrid listings, each priced exactly at $19,990—indicating a dealership-driven pricing pattern rather than an anomaly or data error.

Challenges and Solutions:

- **Ambiguous Categories (`other`)**: Both `fuel` and `transmission` contained "other" categories with unclear definitions. I excluded these entries to avoid misleading groupings and retained only well-defined categories for the analysis subset.
- **Unusual Outliers in Median Prices:** The high median price for hybrid/manual vehicles stood out as an anomaly. Instead of removing it arbitrarily, I investigated the listings directly and discovered that it was driven by bulk pricing patterns from a dealership (multiple entries at $19,990), not random error. This maintained analytical transparency.
- **Non-Normal Price Distribution:** Because used car prices are heavily skewed, I relied on **medians** instead of means to describe central tendencies. This prevented high-end listings from disproportionately influencing the analysis and made the results more representative of typical values.

Question 5: What factors most influence the price of a used car (based on the available data)?

Objective:

This question blends Diagnostic Analysis with Predictive Analysis. Diagnostically, it seeks to identify which features are most responsible for price variation in the dataset. Predictively, it uses machine learning to model these relationships—measuring each feature's contribution to price estimation. The broader goal is to move beyond visual trends and use a data-driven model to determine which variables have the strongest impact on price.

Analytical Reasoning and Approach:

To answer this, I approached the problem as a supervised regression task, where the target variable (`price`) was to be predicted based on a variety of numerical and categorical features.

I began by preprocessing the dataset to ensure all variables were clean, interpretable, and model-ready. This included dropping irrelevant columns (like `id`, `description`, and `posting_date`), removing columns with too many unique values that could clutter the model (`model`, `region`), and replacing the raw `year` feature with the engineered `car_age`.

Since used car prices were heavily skewed—with most listings priced below $20,000 and a few rare listings priced far higher—I applied a log transformation (`log1p(price)`) to compress extreme high values and reduce the influence of outliers. This transformation reshaped the target variable into a more balanced form without changing the order of the data. It helped the model learn more evenly across the price range and focus on general patterns rather than being biased toward a small number of expensive listings.

I then trained a **Random Forest Regressor**, chosen for its ability to handle non-linear relationships and rank feature importance. After training the model, I extracted and visualized the top 20 most influential features—revealing which variables (e.g., age, odometer, manufacturer) had the largest impact on used car prices.

Techniques and Tools Applied:

- **Feature Engineering:** Replaced `year` with a `car_age` column (based on the 2021 dataset context), making the relationship with price more intuitive and interpretable.
- **Feature Selection and Cleaning:** Removed columns that were either:
  - Irrelevant (`id`, `description`, `posting_date`, `lat`, `long`)
  - Redundant (`year`)
  - Too many unique values for modeling (`region`, `model`)
- **One-Hot Encoding:** Applied one-hot encoding to convert categorical variables (`manufacturer`, `fuel`, `title_status`, `transmission`, `type`, `state`) into binary features suitable for the model.
- **Target Transformation:** Used `np.log1p(price)` to reduce skew in the target variable, which improves prediction quality in tree-based models and stabilizes variance across price ranges.
- **Model Training – Random Forest:** Trained a Random Forest Regressor on the full dataset (after encoding), using 80/20 train-test split. This model was chosen for its robustness, handling of nonlinear relationships, and built-in feature importance ranking.
- **Model Evaluation:** Evaluated performance using Root Mean Squared Error (RMSE = 0.39) and R^2 score (R^2 = 0.81), confirming that the model explained a strong proportion of price variance.
- **Feature Importance Visualization:** Extracted feature importance scores from the trained model and visualized the top 20 predictors using a horizontal bar chart. This provided a clear, ranked view of which factors most influenced price predictions.

Challenges and Solutions:

- **Too many variants in Categorical Features:** Features like `model` and `region` had hundreds of unique values, which could introduce noise or inflate dimensionality. I excluded them to focus on broader, more generalizable patterns (e.g., `manufacturer`, `state`).
- **Skewed Price Distribution:** Most cars were priced under $20,000, but a few were extremely expensive, creating a long right tail. I applied a log transformation (`log1p`) to compress these outliers and balance the distribution. This helped the model learn more evenly across all price levels without being biased toward rare high-priced listings.
- **Different Types of Features (Numbers vs Categories):** The dataset included both numeric columns (like mileage) and text-based categories (like fuel type and transmission). Since machine learning models can't use text directly, I converted all categorical features into numbers using one-hot encoding. This made the data usable for the Random Forest model without needing any additional scaling.
- **Feature Correlation vs. Importance:** While correlation analysis (e.g., Pearson r) can suggest linear relationships, it doesn't capture feature *importance* in multivariate models. Random Forest allowed me to overcome this by capturing nonlinear and interaction effects—offering a more realistic importance ranking.

Key Findings and Insights

Question 1: What are the most common categorical traits of cars priced above $20,000?

- **Manufacturer**: Ford and Chevrolet were the most common manufacturers among vehicles priced above $20,000 in this dataset. These were followed by RAM, Toyota, and GMC—brands typically associated with trucks and larger vehicles, which tend to occupy the higher end of the used car market.
- **Fuel Type:** Gasoline-powered vehicles dominate the $20,000+ price segment, accounting for 74% of listings. Diesel follows at a distance, while hybrids and electric vehicles represent a small minority. This distribution likely reflects the historical dominance of internal combustion engines in the used market and the relatively recent mainstream adoption of electrified vehicles.
- **Transmission:** Automatic transmissions dominate the $20,000+ segment, making up over 60% of listings. Manual options are rare within this higher price bracket, which likely reflects the market's preference for automatics in newer, more expensive vehicles.
- **Title Status:** Nearly all vehicles priced above $20,000 have a clean title, indicating that high-value listings are almost exclusively limited to vehicles with verified, damage-free histories. This reflects both buyer expectations and the reduced resale value of branded-title vehicles—cars whose titles have been officially marked due to serious past issues like accidents, major repairs, or legal/financial complications.
- **Vehicle Type:** In the $20,000+ price range, the most common vehicle types are pickups, SUVs, and sedans. Trucks and coupes are also well represented, though at slightly lower volumes. This suggests strong demand for both utility-focused and traditional passenger vehicles in the higher end of the used market.

Question 2: How does car age affect price?

- **Overall Relationship:** There is a clear negative relationship between car age and price—**older cars tend to sell for less**. However, the price drop is **non-linear**: most depreciation happens in the first 10 years, after which prices level off.
- **Early Depreciation:** Vehicles lose value most rapidly within their first 3–10 years. This steep decline reflects how newer cars depreciate quickly as they exit warranty, accumulate mileage, or are displaced by newer models.
- **Stabilization Plateau:** After roughly 20 years, the price curve flattens. At this point, cars are typically priced low regardless of further aging, suggesting a floor effect in used car pricing.
- **Correlation Strength:** The Pearson correlation coefficient between `car_age` and `price` was **-0.43**, indicating a moderate negative linear relationship. Age clearly influences price, but it's not the only factor.
- **Non-Linear Trend Capture:** A LOWESS-smoothed scatterplot revealed that the depreciation curve is steep early on, then flattens gradually—validating that linear models would miss important curvature in the price-age relationship.

Question 3: Which car brands retain their value best over time?

- **Tesla** retains the most value by far—maintaining **79.9%** of its original price over time, with only **20.1% total depreciation**. Uniquely, Tesla showed *price appreciation* from the 1–3 to the 4–6 year age group. This likely occurred because the few newer listings in the dataset were mostly lower-trim models (e.g., base Model 3s), while the older group had more premium variants—raising the older group's average

price. Combined with strong demand and recent new-car price increases, this temporarily made some older Teslas more expensive than newer ones.

- **Mitsubishi** ranks second, retaining **40.1%** of its original value (**59.9% total depreciation**). It was the **best-performing internal combustion engine (ICE) brand** in the dataset in terms of long-term value retention—outperforming many mass-market and even premium competitors.
- **Porsche** comes third, holding **38.4%** of its original value—equivalent to **61.6% total depreciation**. It was the best-performing luxury brand in the dataset in terms of percentage-based retention. Its performance likely reflects strong brand reputation and enthusiast demand.
- **Fiat**, **RAM**, **Chevrolet**, and **Ford** also performed well, with depreciation rates between **61% and 71%**—placing them just outside the top 3 but still indicating strong long-term value compared to the rest of the market.
- **Alfa Romeo** showed very low short-term depreciation (**8.5%** from 1–3 to 4–6 years), but was excluded from final rankings due to missing data in older age brackets, making its long-term value retention unreliable to assess.

Question 4: How does fuel type and transmission affect car price?

- **Diesel Automatics** have the highest median prices in the dataset. This is likely because they're often found in trucks, heavy-duty pickups, and commercial vehicles—segments that naturally cost more than standard sedans or compact cars.
- **Electric Automatics** also show high median prices, reflecting the newer model years of most EVs and their typically automatic-only configurations.
- **Manual Transmission Vehicles**, across most fuel types, tend to have lower median prices. This likely reflects both reduced consumer demand and the fact that manual vehicles are often older.
- **Hybrid + Manual Combinations** show an unexpectedly high median price. This turned out to be caused by a concentrated set of identical Hyundai Sonata Plug-in Hybrid listings, each priced at $19,990. This case illustrates how bulk dealership listings at fixed price points can inflate median values—something that can potentially affect any category pairing and should be checked for during analysis.

Question 5: What factors most influence the price of a used car (based on the available data)?

- **Car Age** was by far the most influential variable affecting price. As cars get older, their prices generally decline—which aligns with my expectations.
- **Odometer (Mileage)** was second most influential factor. Higher mileage was associated with lower prices, confirming that lower-mileage cars are typically priced higher, since they've usually been driven less and are in better condition.
- **Fuel Type** (especially diesel) and **Vehicle Type** (especially pick-ups and trucks) strongly influence price, relative to the majority of the features.
- **Manufacturer** (e.g., Ford, Toyota, Chevrolet) also plays a significant role in determining price. This likely reflects brand-related factors such as perceived quality, reliability, and how each brand is positioned in the market
- **State** had a smaller influence on pricing, though likely due to cost-of-living and tax differences, or regional market preferences.

- **Log Transformation on Price:** Before modeling, the `price` variable was log-transformed to reduce the effect of outliers and stabilize variance. This made the model more accurate by ensuring that very high-priced cars did not disproportionately influence the prediction of average listings.
- **Overall Model Performance:** The Random Forest Regressor achieved strong predictive accuracy, with an R^2 score of 0.81. This means the model explained 81% of the variation in used car prices using the selected features. The top 20 features were visualized to provide a (visual) ranking of their influence.

Business Insights and Recommendations

Used Car Platform Companies

Recommendation: Implement Advanced Search Filters for High-Value Cars to Optimize Listing Visibility

Used car platform companies can maximize revenue by enhancing their user experience (UX) to prioritize high-value cars, particularly those priced above $20,000. Vehicles with diesel fuel type**,** automatic transmissions, and clean titles tend to have higher median prices. By introducing smart filters that spotlight vehicles with these high-value traits, the platform can increase user engagement with premium listings. This strategy would drive more relevant traffic to these high-value cars, potentially increasing conversion rates, and leading to higher commissions or transaction fees for the platform.

Car Dealers on These Platforms

Recommendation: Focus on High-Value Diesel Automatics and Clean-Title Vehicles for Increased Profitability

Car dealers on used car platforms should focus on sourcing and listing high-value vehicles, especially diesel-powered vehicles with automatic transmissions, which show the highest median prices in the dataset. Diesel automatics, often found in premium trucks and commercial vehicles, consistently demonstrate higher prices. Additionally, prioritizing clean-title vehicles, which are more desirable to buyers and generally sell faster, can reduce negotiation time and improve inventory turnover. By focusing on these high-demand vehicles, dealers can potentially command higher prices and experience quicker sales.

Private Individuals (Buyers/Sellers)

Recommendation: Buy and Sell High-Value, Brand-Specific Vehicles to Maximize Resale Value

Private individuals can optimize their car buying or selling strategy by focusing on brands and vehicle types that retain value best over time. Brands like Tesla and Porsche, which exhibit strong value retention (79.9% and 38.4%, respectively), are ideal targets for both purchasing and resale. Sellers can capitalize on these high-demand brands by pricing them based on their market longevity, while buyers can ensure that their investment holds value, leading to better long-term financial returns. This insight helps both buyers and sellers make informed decisions that are likely to result in higher profits when selling or more value for money when buying.

Reflection

This project was my first complete, end-to-end data analysis using Python—and it taught me more than any course or tutorial ever has. I gained hands-on experience with essential tools like Jupyter Notebook, Anaconda, and pandas, and became comfortable using visualization libraries such as Seaborn and Matplotlib. I learned how to structure a professional data analytics project: from defining realistic business questions, to transforming messy raw data, selecting appropriate graphs, interpreting results, and delivering stakeholder-ready insights.

I developed a clearer understanding of key technical concepts, such as feature engineering, one-hot encoding, log transformations, and training a machine learning model. More importantly, I learned to stay skeptical—digging deeper into surprising results instead of accepting them at face value. For example, I learned to spot when a bulk dealership listing was distorting the data and validated it with direct inspection.

Building the README was also a learning experience in itself. I now know what sections are expected in a polished portfolio project and how to present complex analysis in a readable way for non-technical stakeholders. I also learned that memorizing syntax isn't as important as knowing how to look things up quickly and focus on solving the actual business problem. ChatGPT was a valuable assistant during this process—it helped me debug, clarify unfamiliar concepts, and iterate faster without taking shortcuts in learning.

One of the biggest takeaways was how much faster I learned through building something real. Compared to passive learning or courses, actively making a full project gave me the confidence and context I needed to internalize what a data analyst actually does.