A Road Trip Through Data: Predicting Used Car Prices

We're going to travel through our data, stopping at interesting landmarks along the way

Cameron Bell
Israel Kollie
Riccardo Valsecchi
Sosina Tefera

Why are we here?

Unpredictable used car prices

2021 Mercedes CLA Coupe

€34,000

2021 Mercedes CLA Coupe

€70,000

Many buyers don’t know **what truly affects a car’s price**

Our goal: Find **key patterns that influence pricing** and make **better buying decisions**

Let us take you on a journey

- A road trip through rows of data and columns of possibilities

The Map: Understanding Our Dataset

Numerical

- Price
- Km
- Year
- kW
- Horse
- Order

Categorical

- Country
- Manufacturer
- Model
- Automatic_Manual
- Fuel

Total Number of records: 100,000

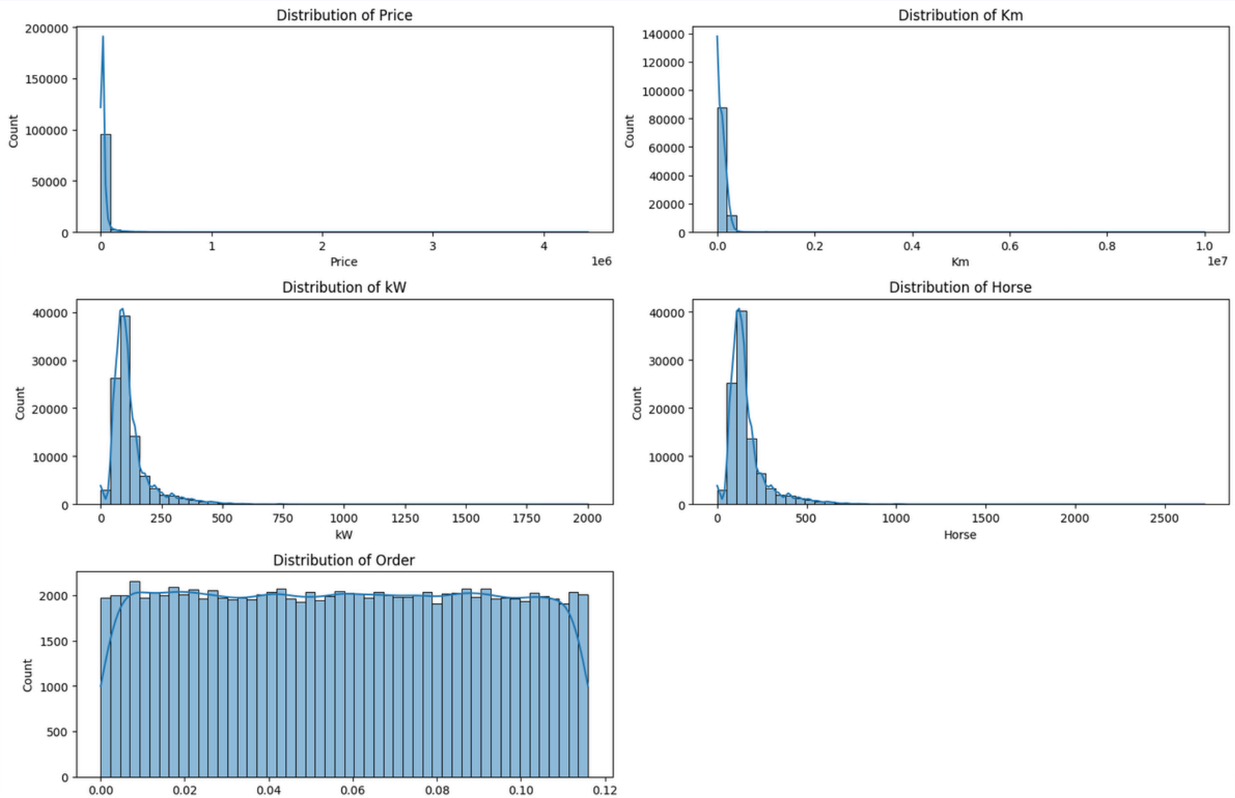
Total Number of variables: 11

Number of missing values: 0

Number of duplicates: 0

Data Types: object; int64; float64

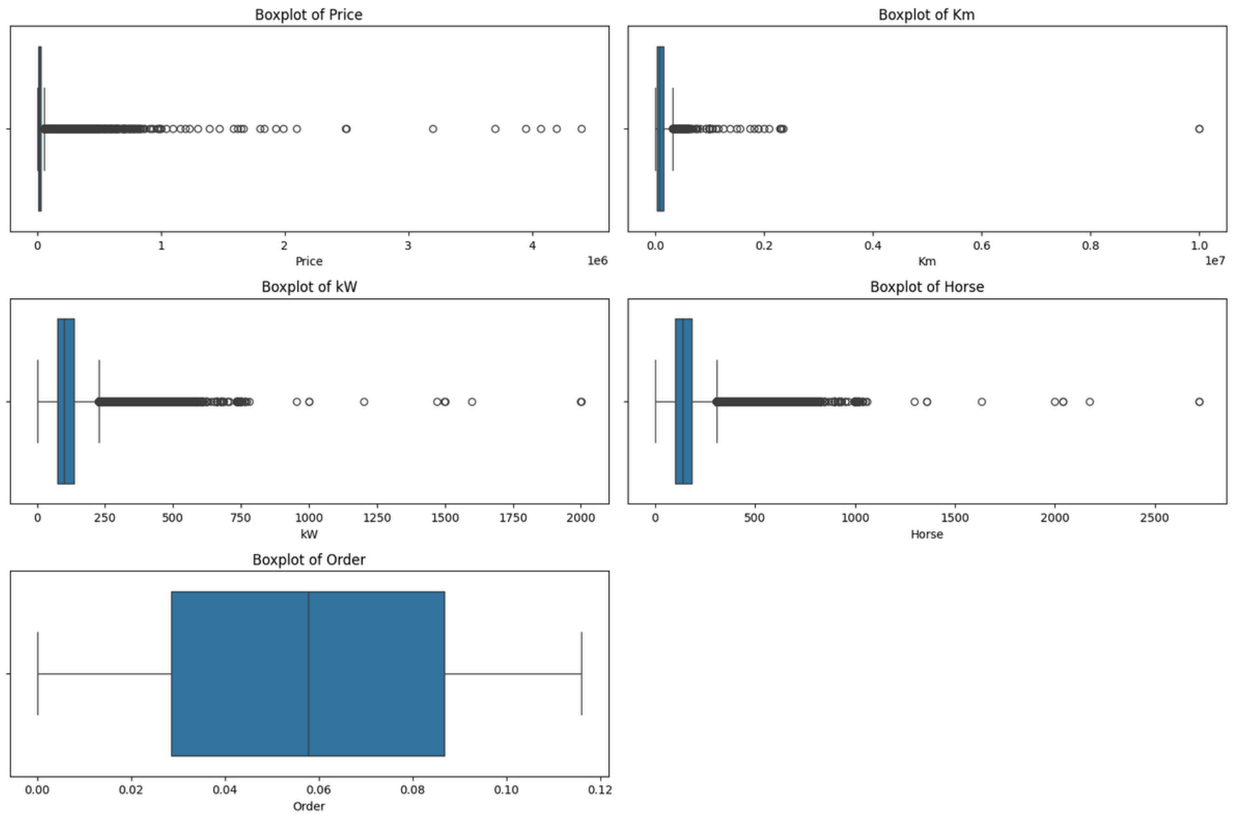
Investigating Our Clues: Preliminary Descriptive Statistics



Insights:

- **Price**: skewed to the right with high-priced outliers.
- **Km**: wide range; some vehicles extremely high mileage.
- **Kw & Horsepower**: Most cars have moderate power, but high-performance models exist.
- **Order**:Seems to be a ranking column rather than a meaningful variable.

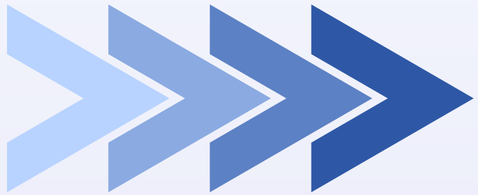
Investigating Our Clues: Preliminary Descriptive Statistics



Insights:

- Price, Km, kW, and Horsepower all contain outliers.
- These **extreme values** will need further investigation, as they could be errors or rare high-value vehicles.

Paving the Road: Data Cleaning



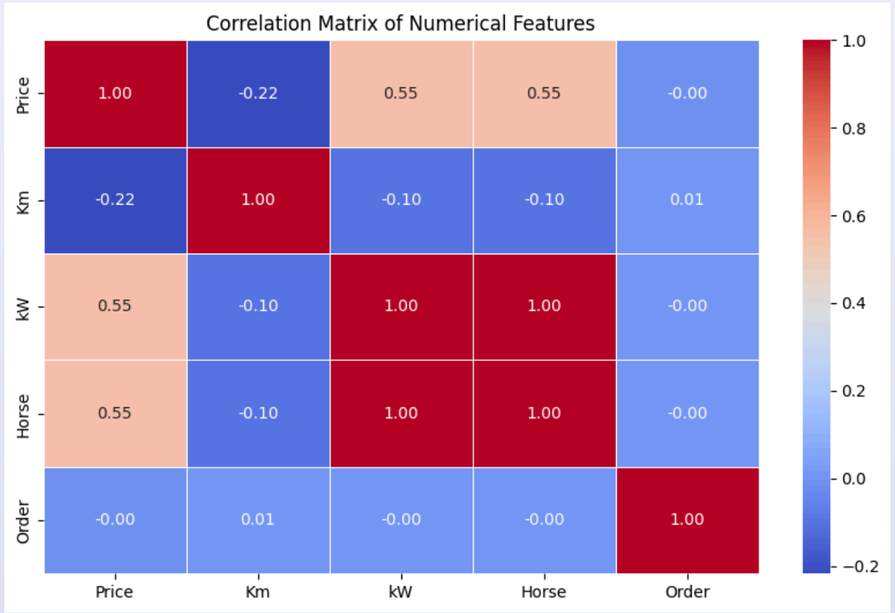
Horse

Order

Outliers

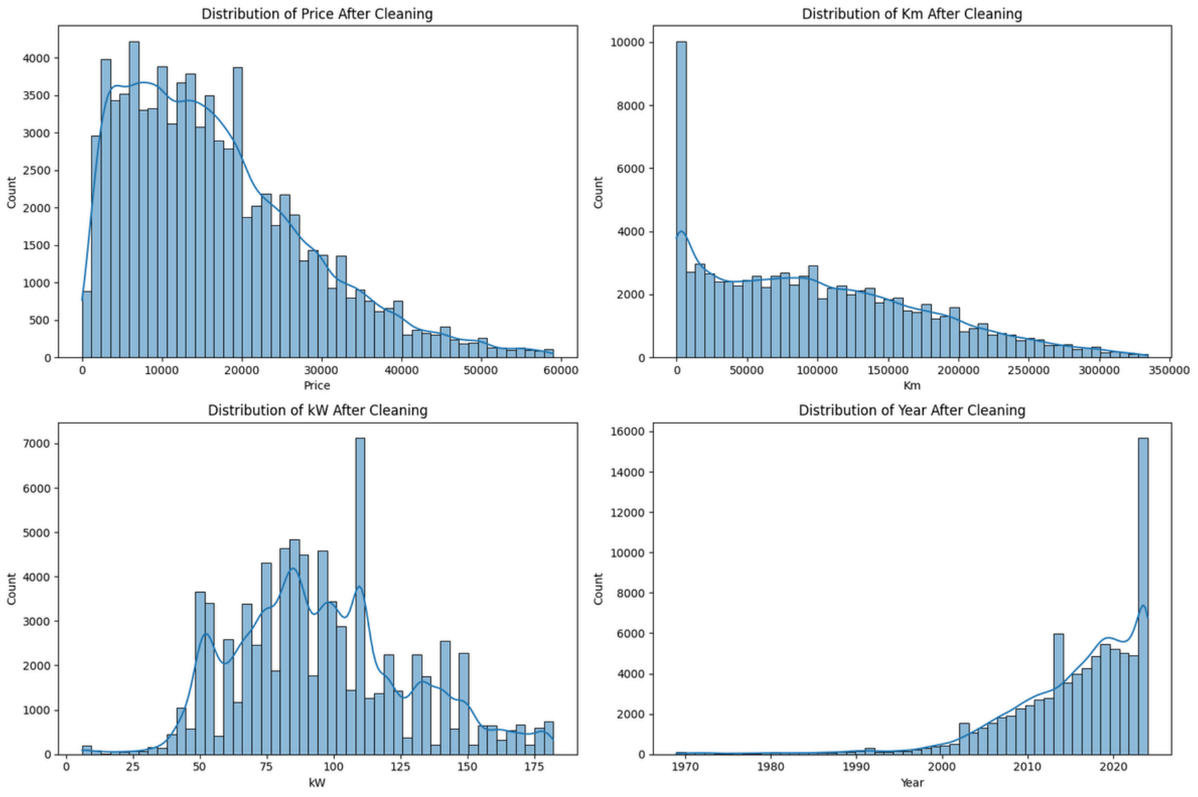
Year

Standardize text data



"Quality data for accurate predictions"

Investigating Our Clues: Outlier Detection & Removal



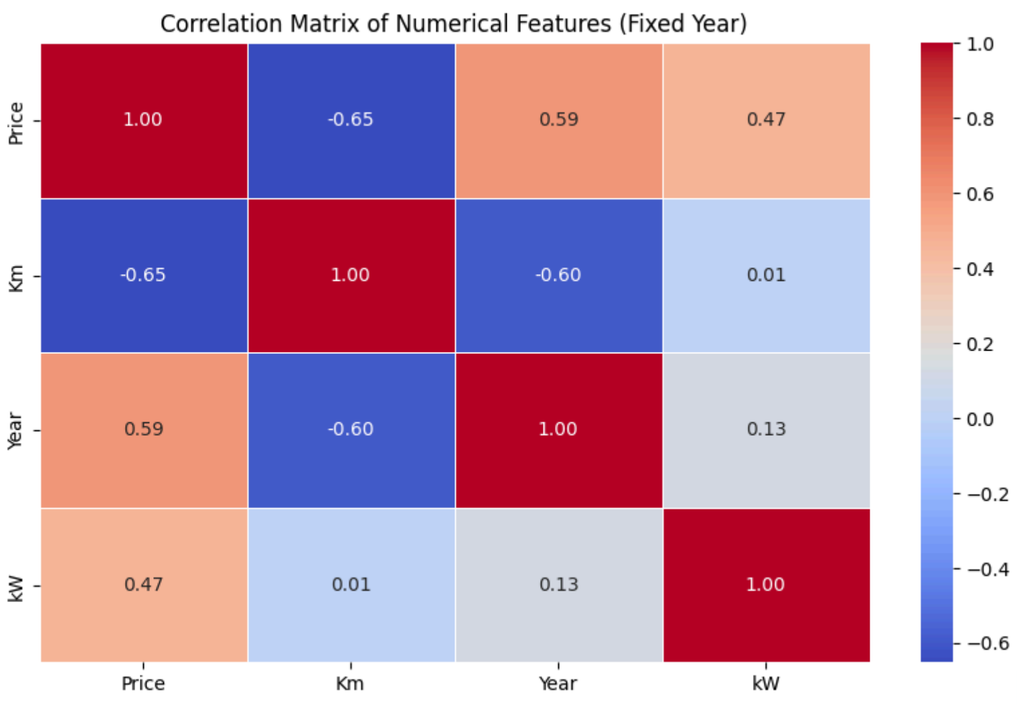
Methods:

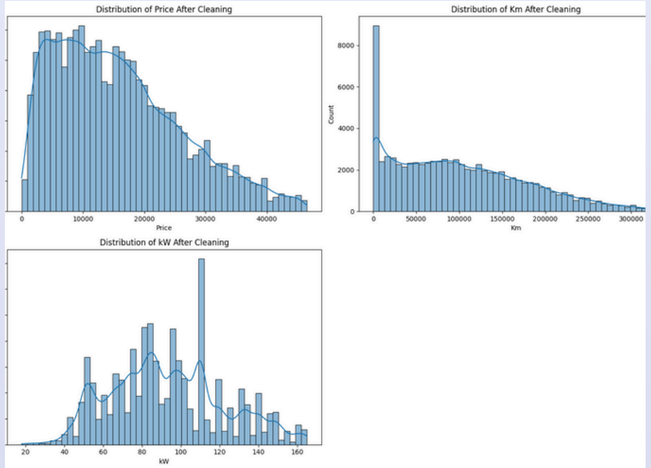
- Used **IQR (Interquartile Range)** to detect outliers in Price, Km, kW.
- Removed **extreme values** to reduce skew.
- **IsolationForest** also demonstrated for anomaly detection (though not the final approach).

Outcome:

- Significantly **improved** data distributions
- Final record count dropped from **100k** to **~82k** after outlier removal.

Exploratory Data Analysis (Post-Cleaning)





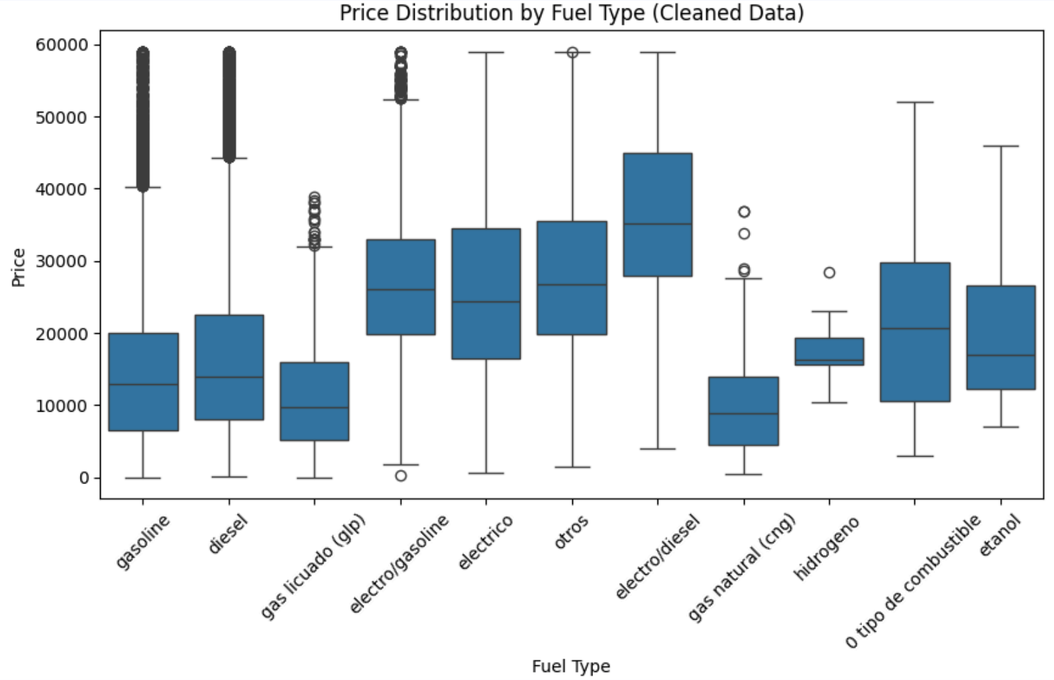
Key Plots:

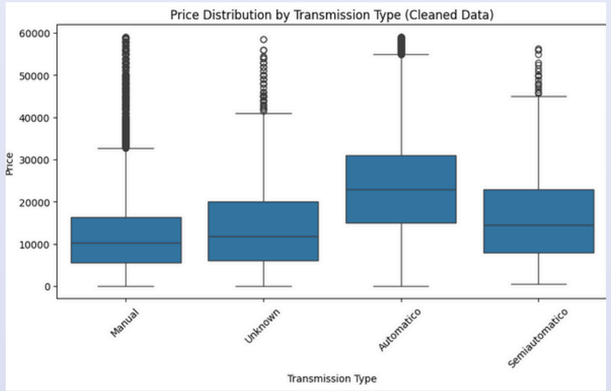
- Updated **Histograms & Boxplots** (Price, Km, kW).
- **Correlation Heatmap** (including encoded categorical features).

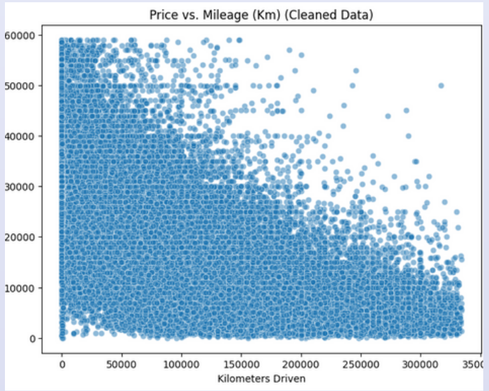
Findings:

- **Negative correlation** between Price & Km (*older, higher-mileage cars cost less*)
- **Positive correlation** between Price & kW (*more powerful cars command higher prices*)

Pit Stops: Hypothesis Testing







Understand which features drive used car prices

Hypotheses

Price vs Transmission Type: T-test results showed a statistically significant difference ($p \approx 0.0$).

Price vs Fuel Type: Gasoline vs. Diesel also significant difference in mean price ($p \approx 2.19e\text{-}141$).

Price vs Mileage (Km): Pearson correlation indicated an inverse relationship.

Conclusions

- **Transmission** (manual vs. automatic) and **Fuel Type** are associated with price variations.
- **Mileage** strongly impacts depreciation.

Refining the Clues: Feature Engineering

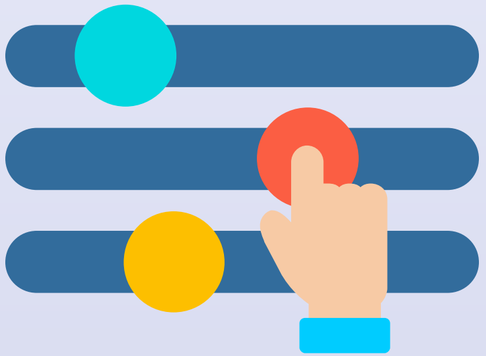
Raw Data	Label Encoding	Model Input
Manufacturer	182 Unique Values	Mercedes-benz -> 109
Automatic_Manual	4 Unique Values	Automatic -> 0; Manual -> 1
Fuel	11 Unique Values	Diesel -> 1; Electrico -> 2

Why not one-hot encoding?

High-dimensional data: feature has many unique categories

Tree-based model: Random Forest

These transformations help the model "read the map" more effectively



Finding the Right Vehicle: Model Selection

```python
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Convert "Year" column to Car Age
df["Car_Age"] = 2025 - df["Year"]


# Select Features (Keep kW, Remove Horsepower)
features = ["Km", "Car_Age", "kW", "Fuel", "Automatic_Manual", "Manufacturer"]
target = "Price"

X = df[features]
y = df[target]


X = df[features]
y = df[target]

# Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train Random Forest Regressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predictions
y_pred = rf_model.predict(X_test)

# Evaluate Model Performance
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
```

Random Forest (Baseline):

- R^2 ~ 0.86
- MAE ~ 2724.32
- RMSE ~ 4210.97

Split data into **training** (80%) and **testing** (20%)

Takeaways:

- Model explains ~86% of price variability.
- Car_Age, kW (power), Km, and Manufacturer rank as the most important features

We needed a model that handles both **numerical** and **categorical** data effectively.

Hitting the Road: Model Development

Hyperparameter Tuning (Gradient Boosting: GridSearchCV)

```python
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV, train_test_split

# Define parameter grid
param_grid = {
    "n_estimators": [100, 200, 300],  # Number of trees in the forest
    "max_depth": [10, 20, None],  # Maximum depth of trees
    "min_samples_split": [2, 5, 10],  # Minimum samples to split a node
    "min_samples_leaf": [1, 2, 4]  # Minimum samples per leaf node
}

# Perform Grid Search
grid_search = GridSearchCV(RandomForestRegressor(random_state=42), param_grid, cv=3, scoring="r2", n_jobs=-1)
grid_search.fit(X_train, y_train)
```

Model Improvement

Random Forest (Optimized):

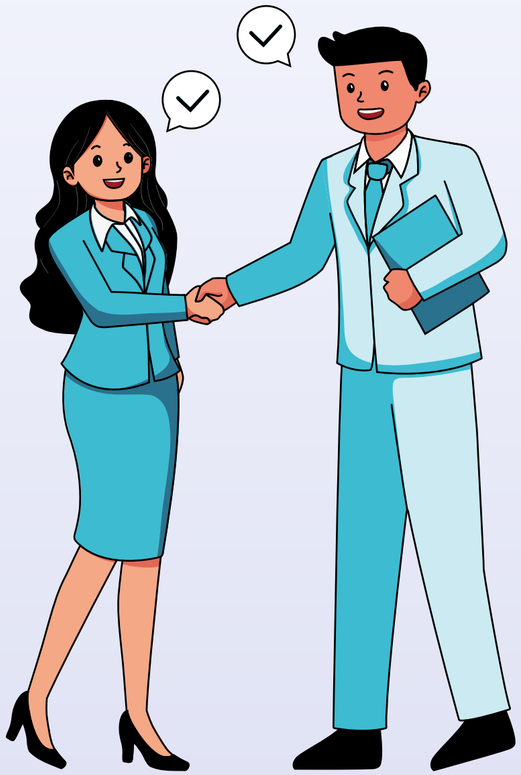
- R^2 ~ 0.8670
- MAE ~ 2660.12

Iterated to find the **best model** and how well it **fits** your data

Hitting the Road: Model Development

Determining whether it's a good deal

	Actual Price	Predicted Price	Price Difference	Percentage Difference	Deal Label
75721	5199	5053.88	-145.12	-2.791306	Fair Price
80184	67900	69942.41	2042.41	3.007968	Fair Price
19864	15395	15957.35	562.35	3.652809	Fair Price
76699	17001	25303.59	8302.59	48.835892	Good Deal
92991	18495	14555.25	-3939.75	-21.301703	Overpriced
...	...	...	...	...	...
32595	6200	6863.27	663.27	10.697903	Good Deal
29313	68900	67337.58	-1562.42	-2.267663	Fair Price
37862	6900	5280.06	-1619.94	-23.477391	Overpriced
53421	1490	3873.82	2383.82	159.987919	Good Deal
42410	34000	25303.10	-8696.90	-25.579118	Overpriced



Predict Price Specifications:

- Km, kW, Transmission, Fuel, Car_Age

Percentage Differences

Good Deal: >=10%

Fair Price: <= -10%

What We Learned from the Drive

Insights:

Feature Importance

which strongly influence a car's resale price

- **Depreciation**: Price drops as Km and Car_Age increase.
- **Performance**: Higher kW = higher price.
- **Fuel & Transmission**: Both matter; automatics & certain fuels often priced higher.

Recommendations:

- **Sellers** can highlight low-mileage or high-power aspects to maximize **price**.
- **Dealers** can use model predictions to set fair market prices.
- **Additional data** (maintenance history, accidents) could further refine the model.

Next Steps:

- Include **external market data** or brand reputation scores.
- **Resale Value Forecasting:** predict future resale price (need historical data)

Crossing the Finish Line: The Joy of Data Discovery

The Best Part?

This journey shows how numbers can tell stories, **helping dealers set fair prices**, **helping buyers find good deals**, and, ultimately, helping us **understand the hidden forces shaping the market**.