Modeling the relationship between car sales price and different car features

Buying and selling cars is common experience especially among people leaving in rural areas with little or no transportation. It is thus interesting to study what factors do influence car sales price significantly and what can be improved to have better sales and fair car sales prices. Using differnt Machine Learning Regression methods we will develop models to predict car sales price using CarDekho Dataset .

Authors

- Mariana Khachatryan
- Adreja Mondol
- Amogh Parab
- Nasim Dehghan

KPI

Technical performance Key Performance Indicators (KPIs)

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)
4. R-squared- proportion of variance in target variable that is predictable from input features

Business impact KPIs

These KPIs measure how effectively the model adds value to the business, such as improving pricing strategies or increasing customer satisfaction.

1. Revenue Increase: Accurate price predictions could help optimize the selling price of cars, leading to better margins.
2. Cost Reduction: If the model is used to automate pricing, it can reduce the need for manual evaluation and pricing, saving labor costs.
3. Inventory Turnover Rate: How quickly cars are being sold after their prices are set by the model.

Key Stakeholders

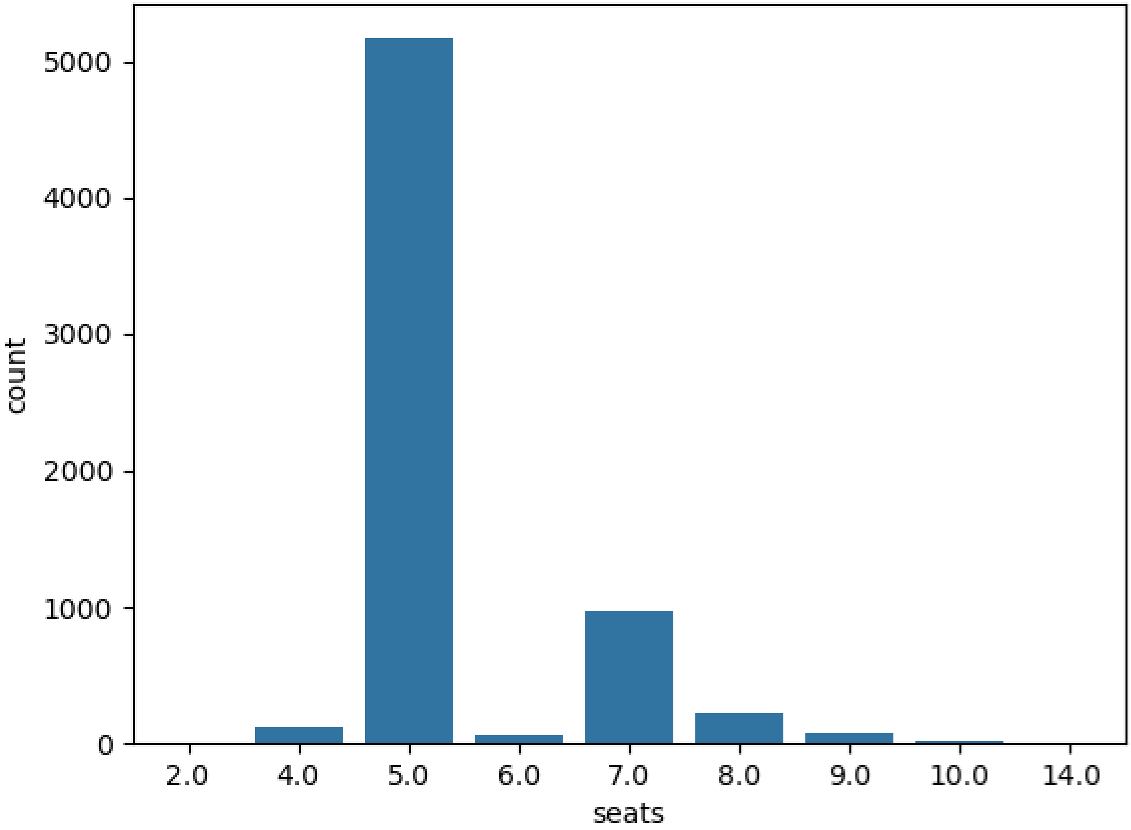
1. Car Dealerships need price prediction model to set competitive and accurate prices for cars. Dealerships want to maximize profit while ensuring quick car sales. Accurate price prediction results in competitive pricing and profitability.
2. Customers can use the model to estimate whether the set price is fare.

Exploratory Data Analysis and Feature Engineering

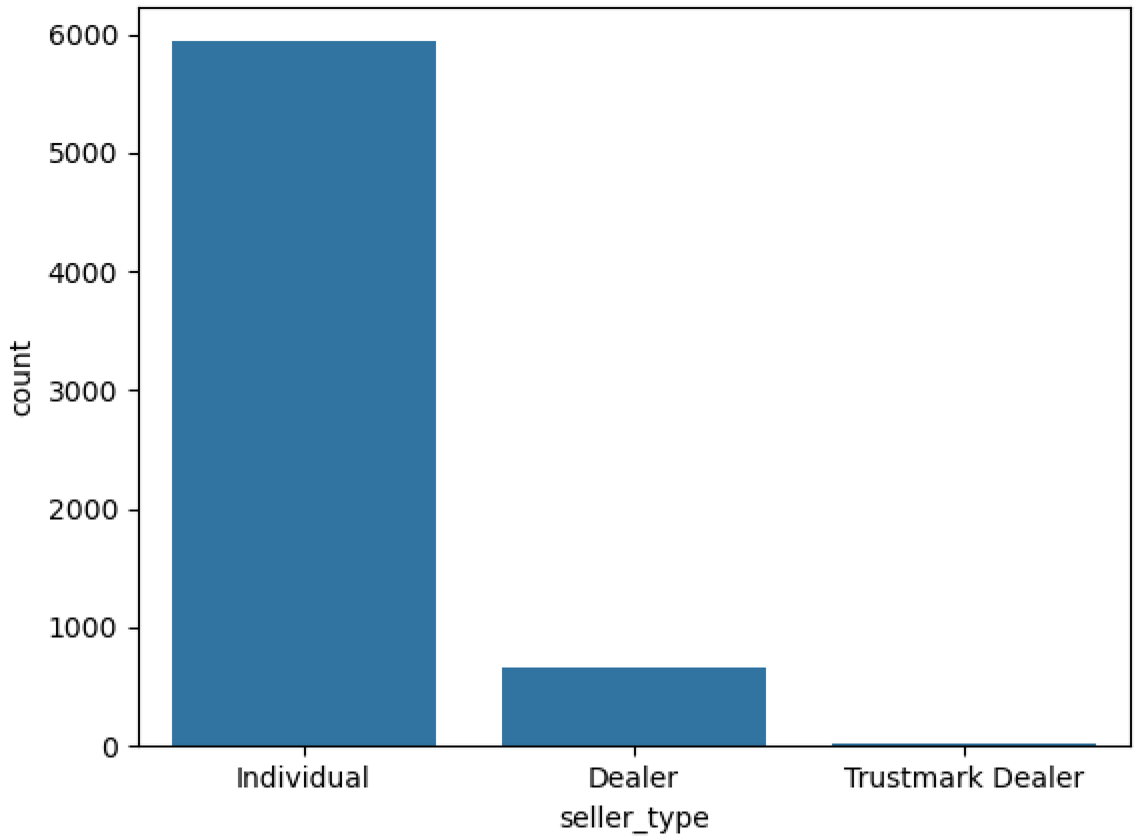
In this project we use data from CarDekho. Founded in 2008, it is India's leading online marketplace that helps individuals and dealers buy, sell, and manage their cars. The data contains 8128 entries and 12 features corresponding to car name, year bought, selling price, kilometers driven, fuel type, seller type, transmission type, owner type, mileage, engine in the units of cubic capacity (CC), max_power in the units of brake horsepower (bhp), torque and number of seats. We didn't use torque feature since the units used for different entries correspond to different physical quantities and some of them have fixed values while others show range of values. We also dropped car name feature, since it had too many unique categories.

- Data cleaning involved:
    - removal of duplicated rows, which dropped entries to 6926 (removed 14.79% of data)
    - removal of 209 rows with missing values for multiple columns, which reduced entries to 6717 (3.02% less statistics)
    - removal of 1.28% of remaining data corresponding to "LPG" and "CNG" underrepresented gas fuel types which left 6631 entries
    - removal of 0.08% data corresponding to "Test Drive Car" of owner category which left 6626 entries
    - removal of outliers (5.42% of remaining data)
- Final data set had 6533 data points with 9 features
- Feature engineering
    - we modified number of seats feature to have two bins for number of seats >5 and

seats ≤5



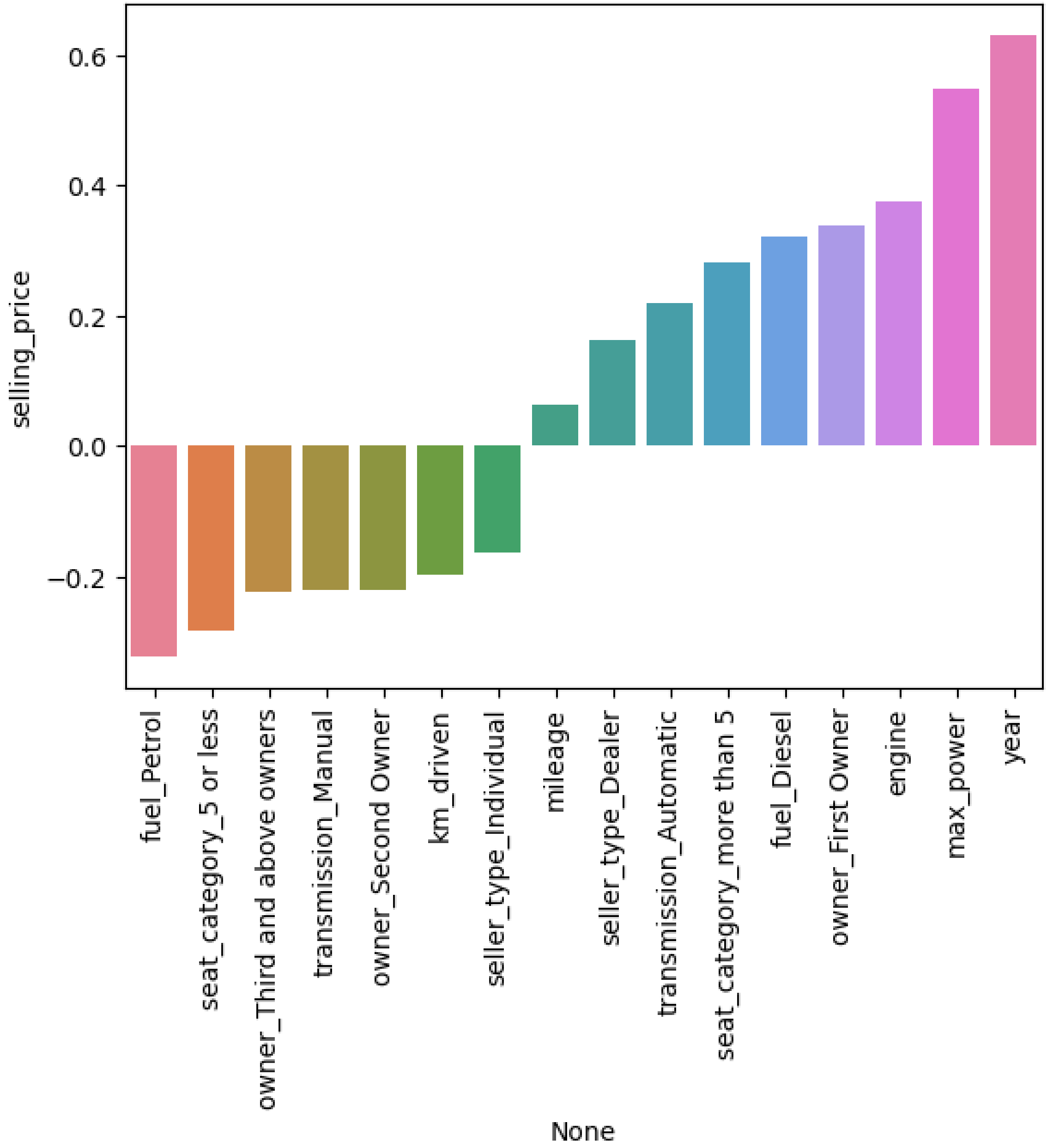
- we combined "Trustmark Dealer" with "Dealer" categories in seller type feature, to decrease imbalance between different categories



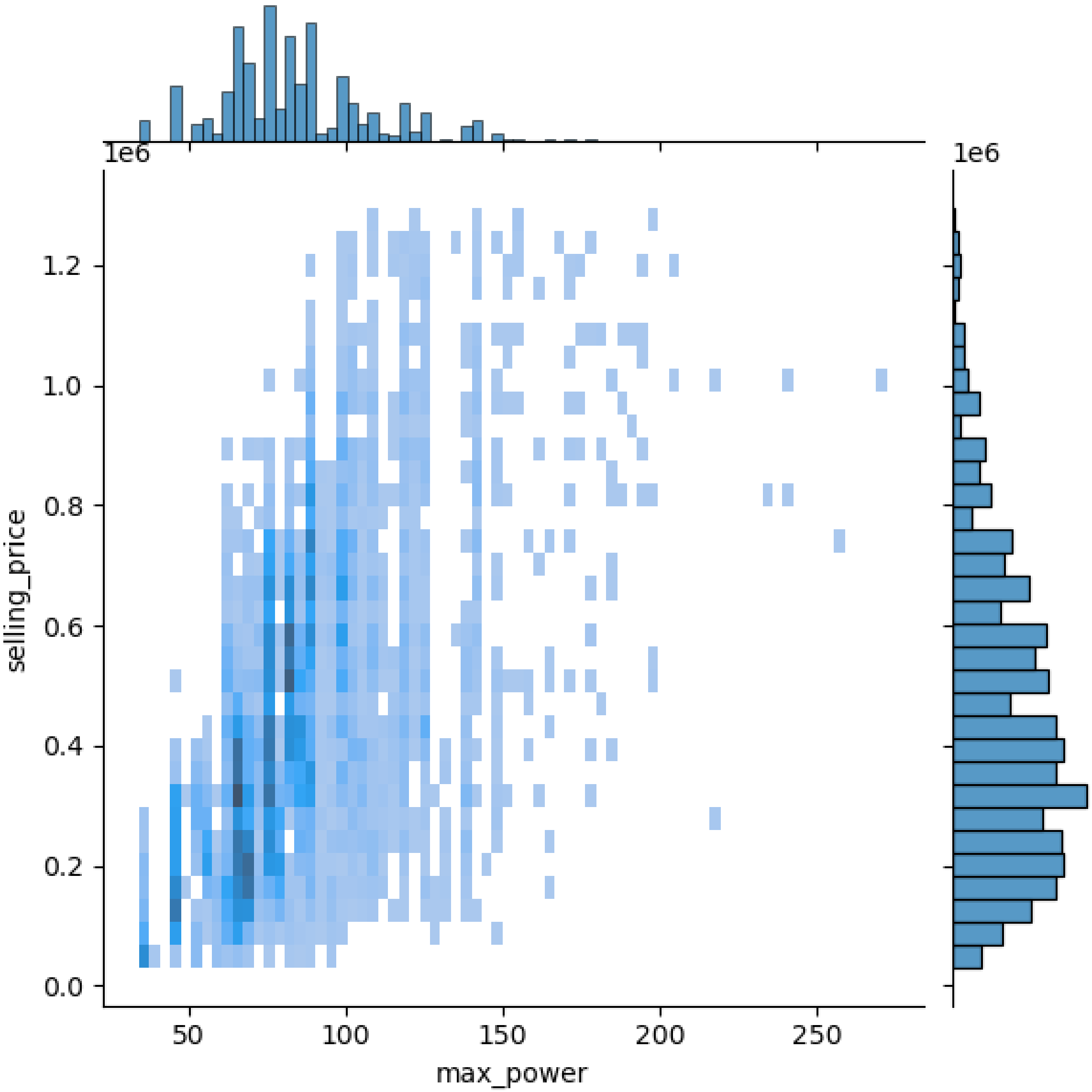
We then used one hot encoding for categorical features (fuel type, seller type, transmission type, owner type, number of seats bin category).

- Exploratory data analysis:

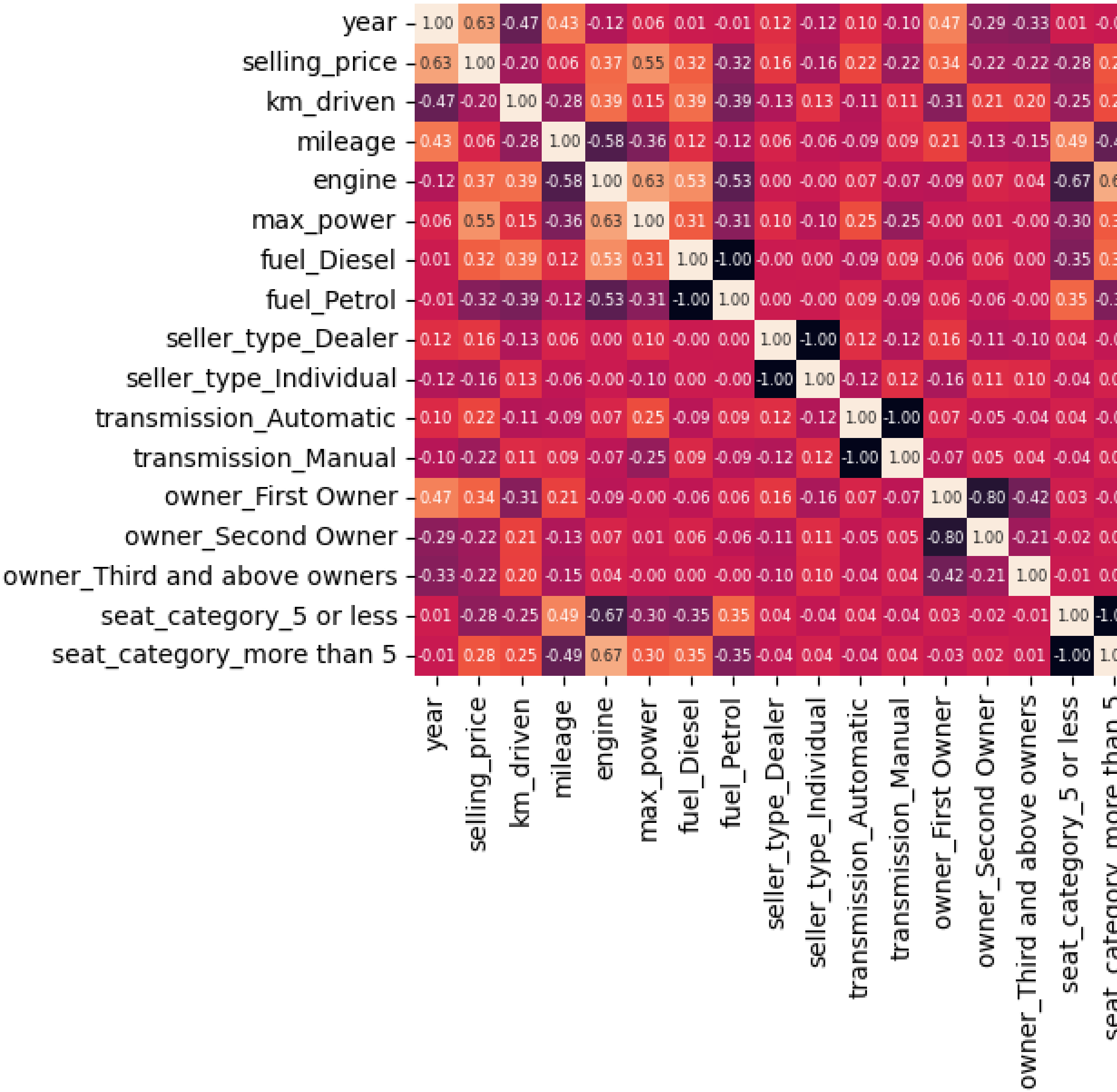
- we have looked at the correlation of sales price with different feature



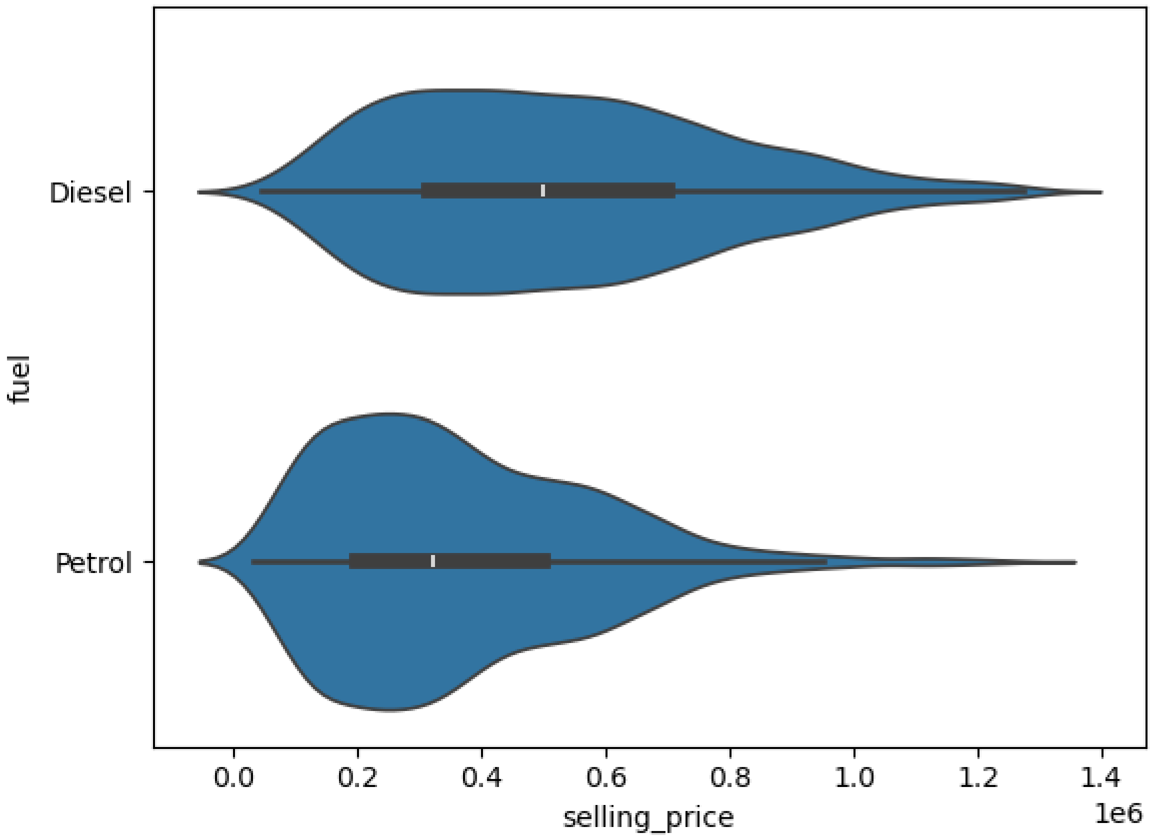
- sales price has highest correlation with max power

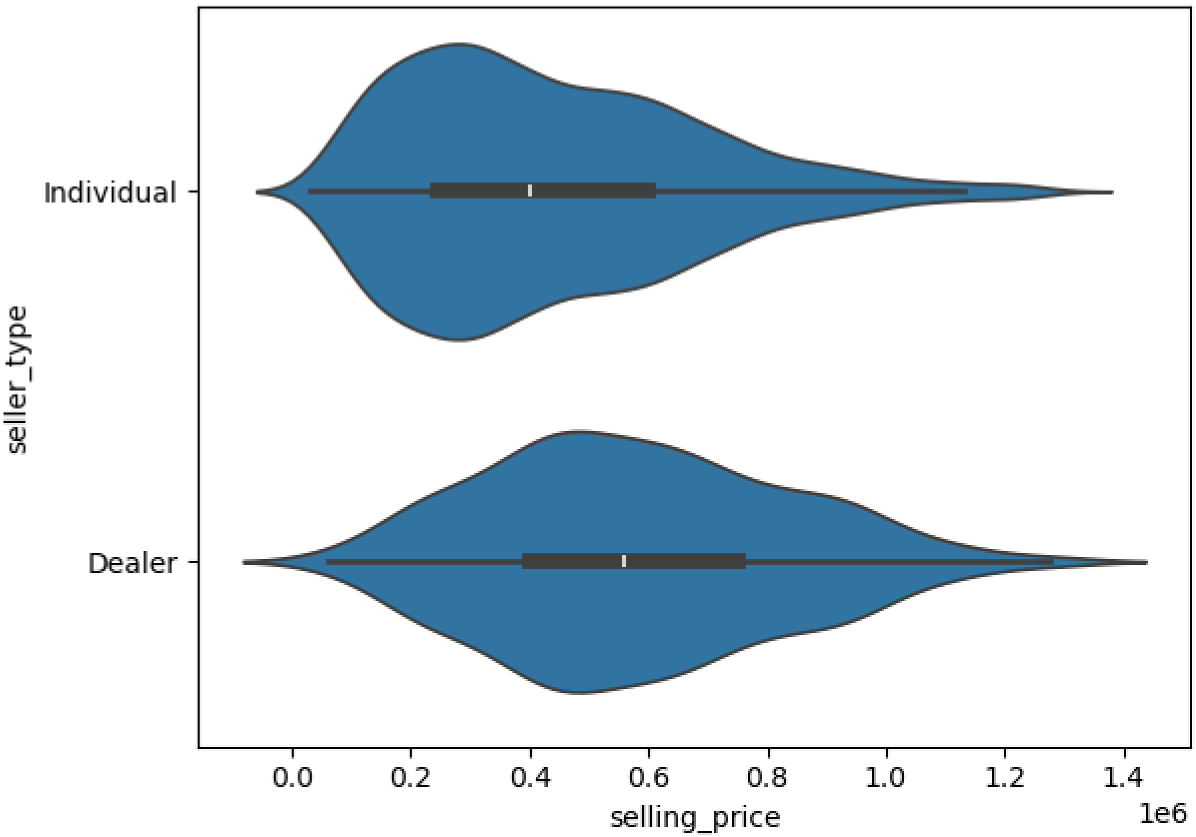


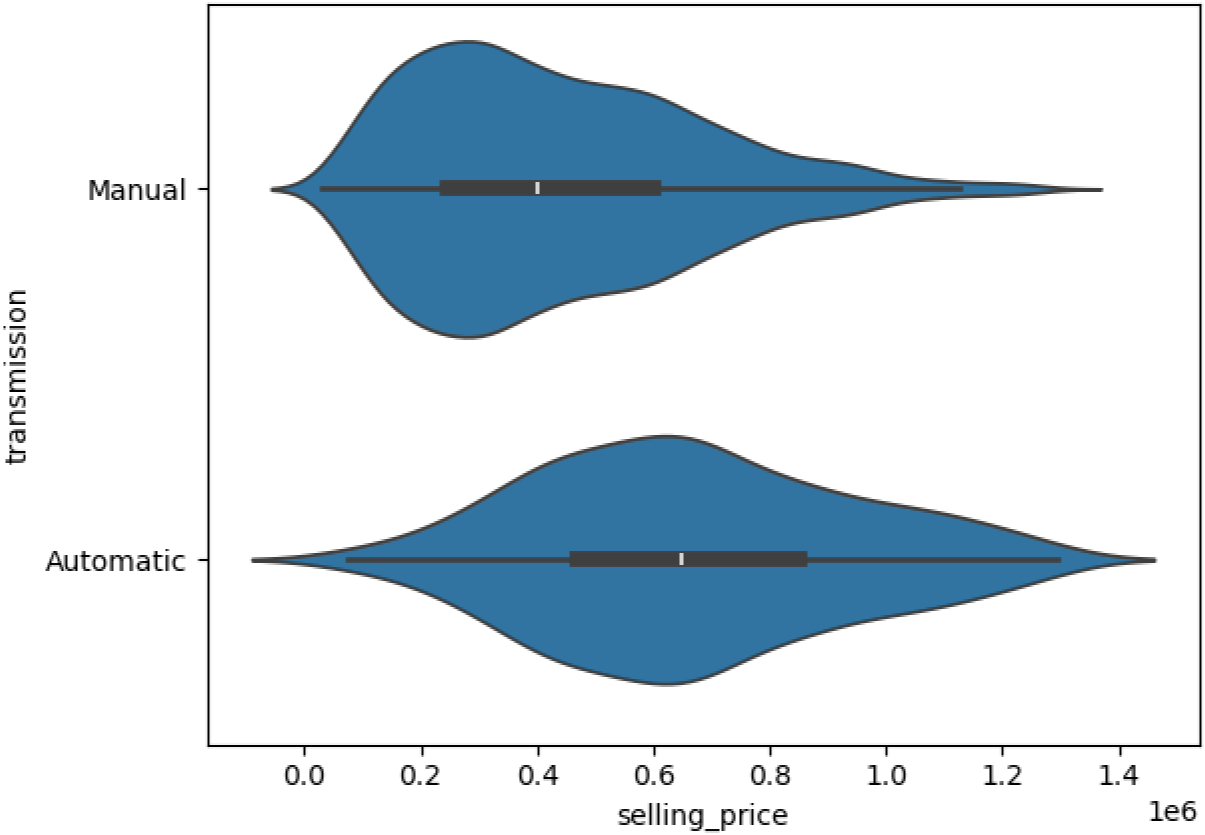
- Before model training, correlated features were removed, i.e. only one feature among features with correlation ≥8 was kept.

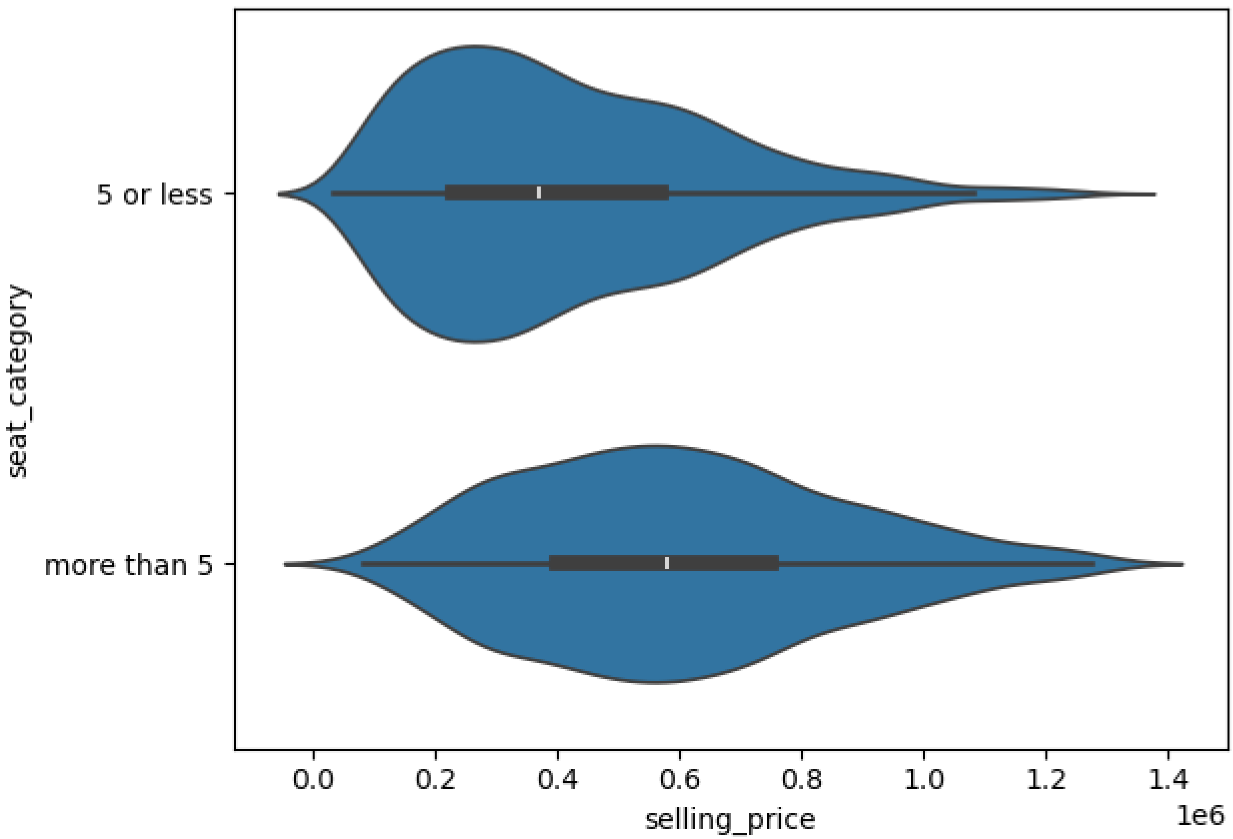


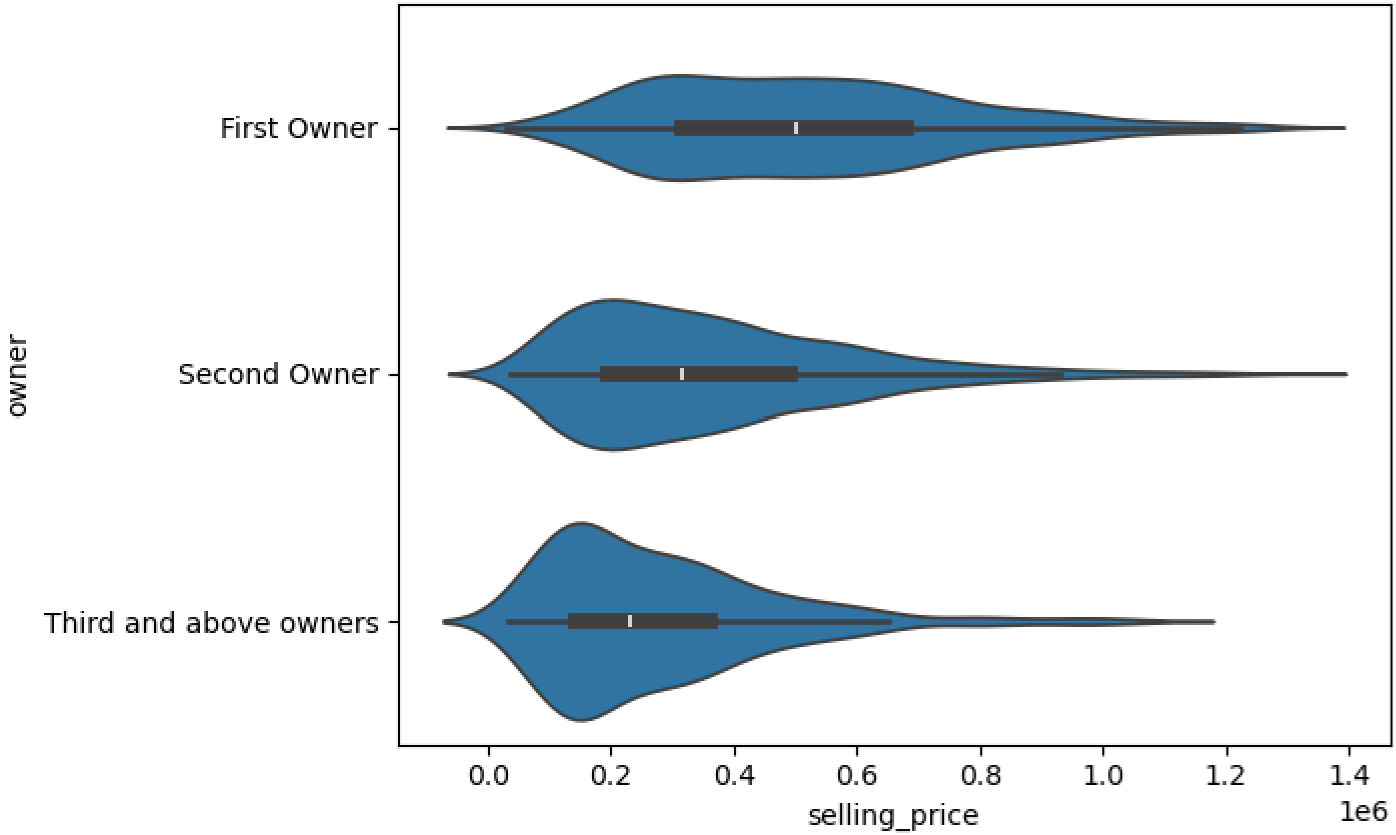
We can see the distributions of sales price for differnt categorical features in the violin plots below. the distributions look different for different categories of given categorical feature, which suggests that they should all be included in training of model.


Diesel
Petrol
fuel
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
selling_price
1e6


Individual
Dealer
seller_type
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e6
selling_price


Manual
Automatic
transmission
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
selling_price
1e6


5 or less
more than 5
seat_category
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
selling_price
1e6



Approach

We compare predictions of different Machine Learning models:

- Linear Regression
- Tree Methods
- Support Vector Machines We start with linear regression models, such as Elastic Net from python sklearn library. Elastic Net uses combination of Lasso and Ridge regularizations. We will then compare results from different regression models. We use grid search to find optimal model parameters for different regression m