Analysis on Used Car Market Based on Prices

By Juan Torres

Research Question

- What are the main characteristics which have the most impact on the car price? Describe the relationship between the price and characteristics (positive/negative/weak relationship).
- What is the correlation between the price of the used vehicle and selected characteristics (e.g. body-style, horsepower, mileage) of the car?

Methodology

Data points:

- make
- num-of-doors
- body-style
- drive-wheels
- engine-location
- wheel-base
- length
- width
- height
- curb-weight
- engine-type
- num-of-cylinders
- engine-size
- fuel-system
- bore
- stroke
- compression-ratio
- horsepower
- peak-rpm
- city-mpg
- highway-mpg
- price
- city-L/100km
- horsepower-binned
- fuel-type-diesel
- fuel-type-gas

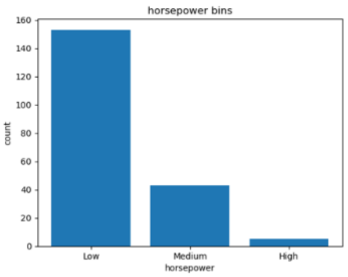
Data Cleaning

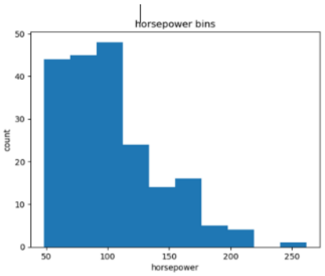
Familiarized myself with the data.

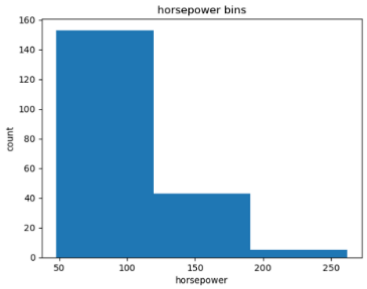
Displayed descriptive statistical metrics.

Data Wrangling

- Replaced null values with useful ones, such as mean values.
- Calculated means of selected columns to understand trends in data.
- Changed data types.
- Standardization of data metrics.
- Normalized data.
- Created histogram to understand the central tendency of in horsepower and pricing.







- Created indicator variables to create categorical binary variables (e.g. whether car use gas or diesel)

Exploratory Data Analysis

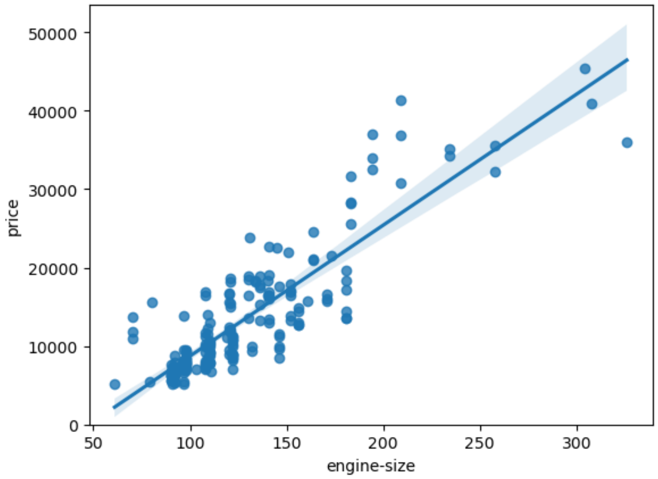
- Linear regression.
- ANOVA hypothesis testing and F-testing.
- P-value and Pearson Correlation Coefficient assessment.
- Descriptive statistical analysis.
- Value counts and statistics by grouping

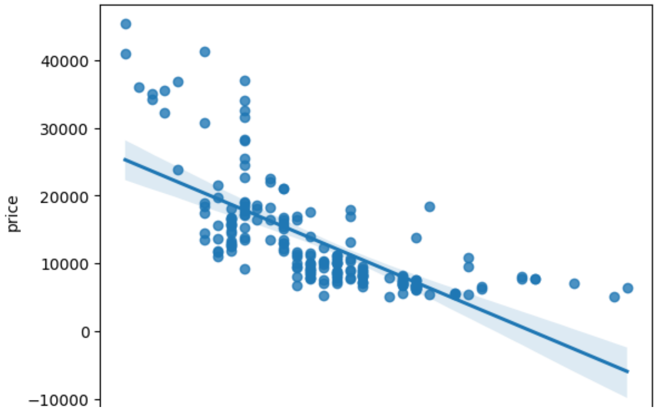
Hypothesis

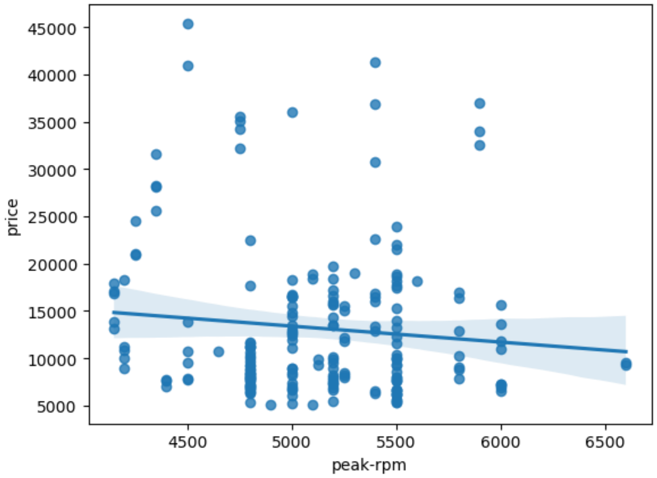
- Quantitative variables like highway-mpg, engine size, and peak-rpm tend to increase the market prices of used vehicles.
- ‘engine-location’ has a significant effect on the pricing of used vehicles.
- ‘drive-wheel’ property has a significant effect on pricing of used vehicles.
- ‘body-style’ property has a significant effect on the pricing of used vehicles.

Linear Regression for Quantitative Features of Interest: highway-mpg, engine size, and peak-rpm

We can see a good positive correlation between these engine-size and price since the regression line is almost a perfect diagonal 45-degree angle line. This is because the more diagonal the regression line the stronger the correlation between the two variables.







Mathematically, can model the relationship between the engine size and the price of the used vehicle with the following functions:

f(x)=166.86x-7,963.34

Meaning that for each cubic inch that the engine increases, the price of the used vehicle increases by $166.86. The correlation coefficient also indicates a strong correlation between the engine size and the price of the vehicle:

P=0.87…

Conversely, we see negative correlation between highway mileage per gallon (mpg) and the price:

f(x)=-821.73x+38423.31

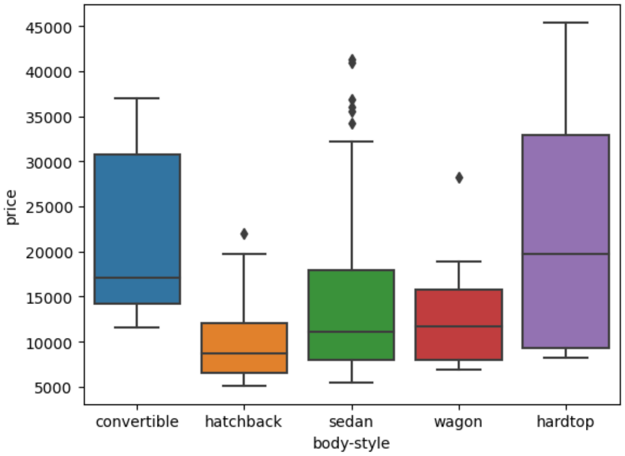
That is, we have a decrease in the overall price of the used car by $821.73 per highway mpg increase. Although the correlation coefficient is not a big as that of the relationship between engine size and price, we can confidently see a strong negative correlation between highway mileage and car price:

P=-0.70

However, where we don't see a strong correlation is between the price of the used car and the peak revolutions per minute (acceleration power). The relationship is technically negative, but the rate of change is a decrease of $1.69 per increase of peak-rpm and its correlation coefficient is close to zero that we might not take it into account:

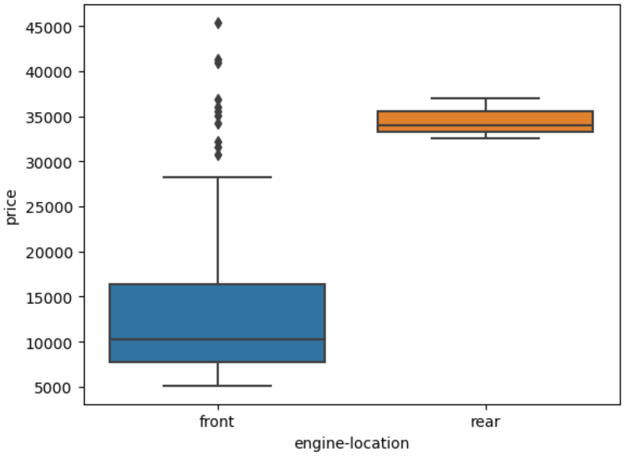
f(x)=-1.69x+21,851.01   P=-0.10

Categorical groups: body-style, engine location and drive-wheels



ANOVA results for the body-style and the price indicate that there is no real difference between the tested groups. However, observe that convertible and hardtop options have a higher median in price and have a higher variability on pricing than the hatchback, sedan, and wagon groups.

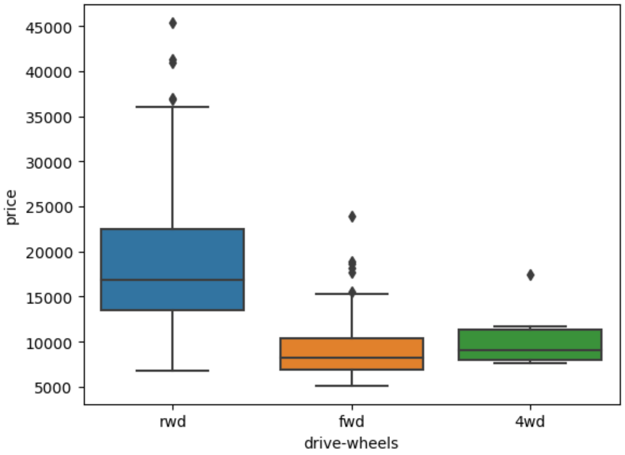
ANOVA Test for body style	
F	*9.13*
P	*8.78*



Where a good indicator of a factor having a significant change in the price of the used cars is in the engine location:

ANOVA Test for engine location	
F	*24.50*
P	*1.58*

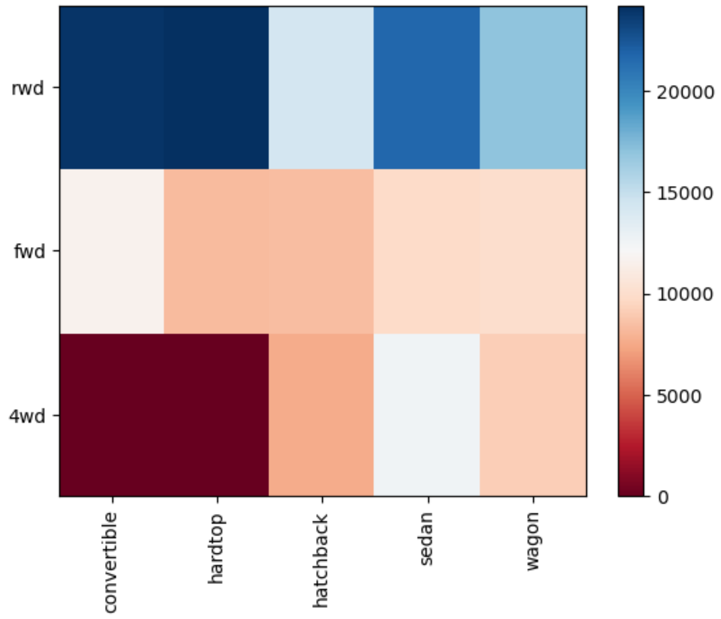
Cars with engines located in the rear have a tendency to be located in around the price of $35,000 with a narrower range price. This is while cars with front engines sit at the $10,000 median with some outliers overpassing the price of rear-engine cars.



ANOVA Test for drive wheels	
F	*67.95*
P	*3.39*

The distribution of price of price between the different drive-wheels categories do differentiate, with rear-wheel drive vehicles having a significantly higher median and variability in prices. The central tendency of front-wheel drive and 4-wheel drive are near the same price range of $1,000 and $1,100. However, front-wheel drive vehicles still have more variability than their 4-wheel drive counterparts.

By grouping the body-type and drive-wheel categories around a heatmap price gradient (dark blue being the most expensive and dark red being the least expensive) we can start to see that the most expensive vehicles are the convertible rear-wheel drive used cars.



Peason Correlations and P-Value Measurements of Selected Properties

Pearson Correlation and P-value Measures		
Variables	**Measure**	**Value**
Length and price	R	69.06%
	P	8.01
Wheel-base & price	R	58.46%
	P	8.08
Horsepower & price	R	80.96%
	P	6.37
Width & price	R	75.13%
	P	9.2
Curb-weight & price	R	83.44%
	P	2.19
Engine-size & price	R	87.23%
	P	9.26
Bore & price	R	54.32%
	P	8.05

City-mpg & price	R	-68.66%
	P	2.32
Highway-mpg & price	R	-70.46%
	P	1.75

Conclusion

For predicting price of used cars, we can narrow down the variables that have correlations. For numerical variables:

- Length
- Width
- Horsepower
- Curb-weight
- Engine-size
- City-mpg
- Highway-mpg

For categorical variables:

- Drive-wheels