Data Analysis Project

By: Max Ooi Wei Xiang

Understanding the Problem

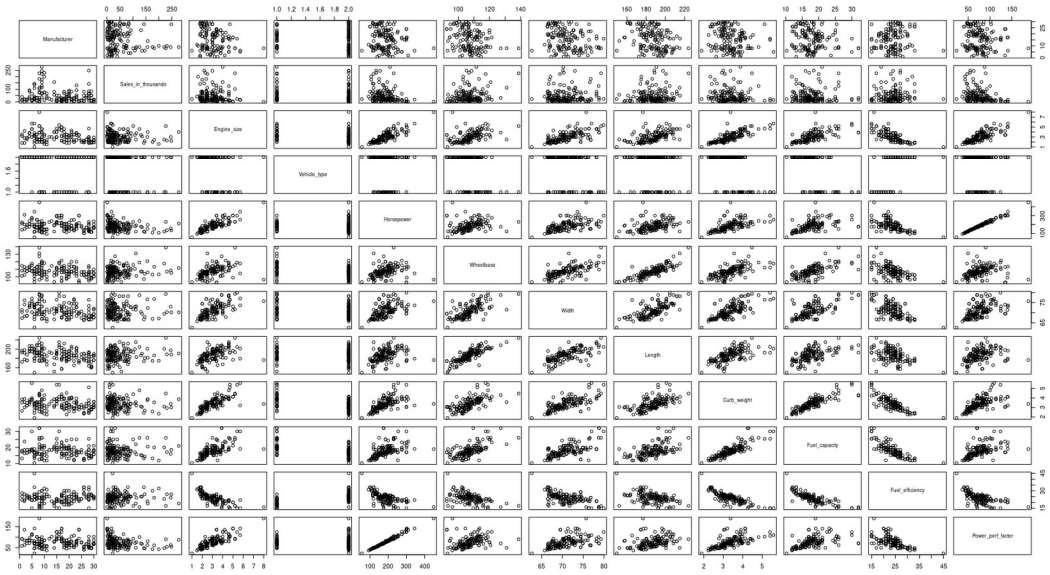
I want to understand what factors influence the sales of a car.

Plan and Properly Collect Relevant Data

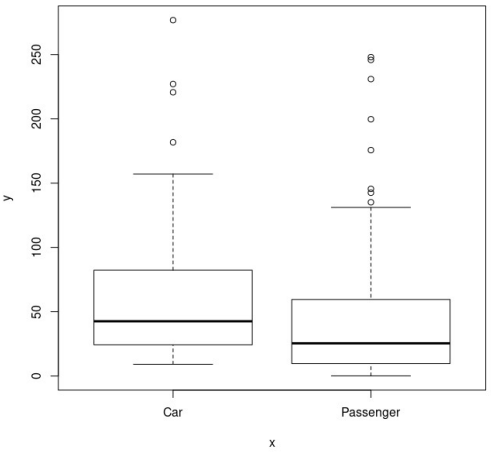
Data from Kaggle (https://www.kaggle.com/datasets/gagandeep16/car-sales) is used as a primary data source.

Explore Data

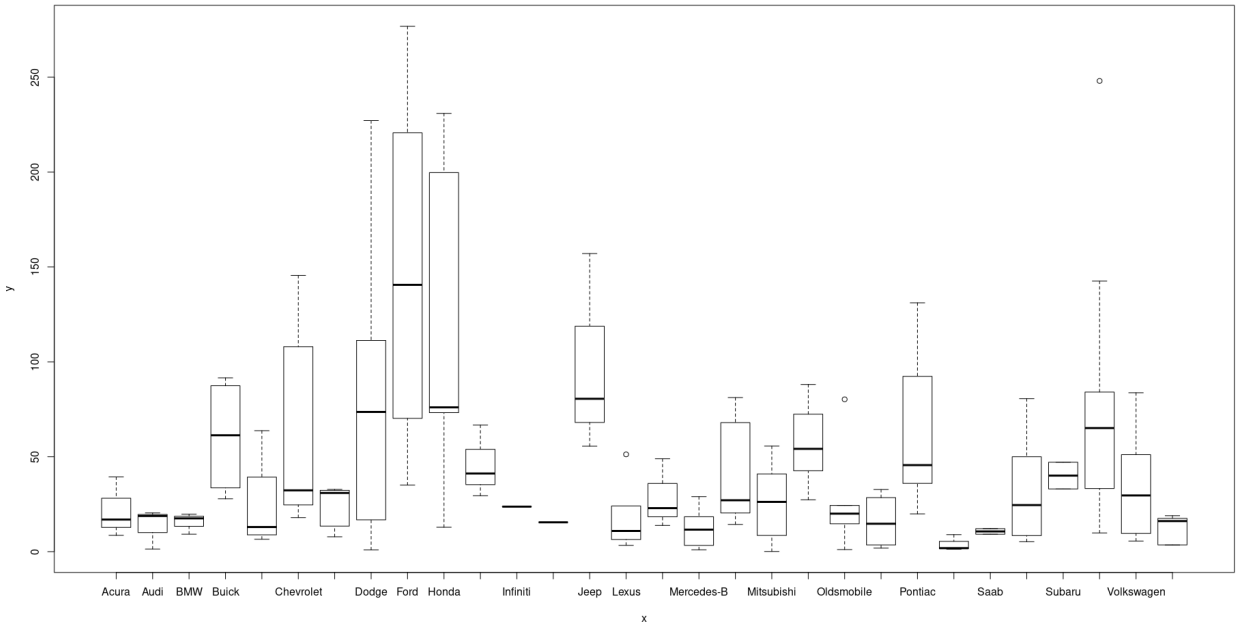
After taking the interesting features, omitting the NA values and removing 1 outlier, we are left with 151 rows. Columns with discrete levels are *Manufacturer* and *Vehicle_type*. Running a pairwise plot shows a high degree of correlation between *Horsepower* and *Power_perf_factor*, hence the latter was ignored.



Plotting the sales against *Vehicle_type*, the difference between Car and Passenger type is relatively small, hence *Vehicle_type* is ignored in the model to reduce complexity.



Plotting the sales against *Manufacturer*, it is apparent that the Manufacturer has a big influence over the mean and variance of the sales.



Postulate a Model

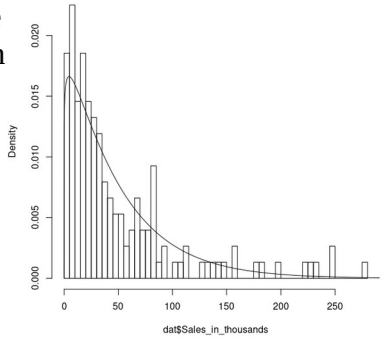
To reduce the complexity of the model, Manufacturers are first grouped together based on their similar distribution. The table below shows the grouping. This produces 9 levels for Manufacturer.

Group Name	Manufacturers
Ford	Ford
Honda	Honda
Jeep	Jeep
Porsche	Porsche
Subaru	Subaru
CPT	Chevrolet, Pontiac, Toyota
Dodge	Dodge
BMNV	Buick, Mercury, Nissan, Volkswagen
Other	Other

```
Call:
lm(formula = Sales_in_thousands ~ Manufacturer + Engine_size +
    Horsepower + Wheelbase + Width + Length + Curb_weight + Fuel_capacity +
    Fuel_efficiency, data = dat3)

Residuals:
    Min      1Q  Median      3Q     Max
-110.00  -20.59   -1.88   13.06  176.34

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -104.8460   114.6366  -0.915 0.362047
ManufacturerCPT       23.4086    12.6418   1.852 0.066273 .
ManufacturerDodge     32.9852    17.8709   1.846 0.067136 .
ManufacturerFord      93.4018    15.9097   5.871 3.24e-08 ***
ManufacturerHonda     71.2098    20.4371   3.484 0.000667 ***
ManufacturerJeep      77.2719    27.6903   2.791 0.006030 **
ManufacturerOther    -18.4417    10.6729  -1.728 0.086310 .
ManufacturerPorsche   17.9020    29.8895   0.599 0.550226
ManufacturerSubaru     4.1960    30.6942   0.137 0.891472
Engine_size            8.1358     9.0523   0.899 0.370396
Horsepower            -0.2537     0.1398  -1.814 0.071886 .
Wheelbase              2.4559     1.0165   2.416 0.017041 *
Width                 -2.1014     1.7368  -1.210 0.228445
Length                 0.5102     0.5280   0.966 0.335626
Curb_weight           -4.2503    15.4647  -0.275 0.783865
Fuel_capacity         -1.9689     2.0135  -0.978 0.329897
Fuel_efficiency        0.4028     1.7173   0.235 0.814906
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.84 on 134 degrees of freedom
Multiple R-squared:  0.5332,    Adjusted R-squared:  0.4774
F-statistic: 9.565 on 16 and 134 DF,  p-value: 1.651e-15
```

A preliminary linear model is fitted to act as feature selection, shown on the right. It can be seen that, among the existing features, *Fuel_efficiency* and *Curb_weight* have the lowest t values. Hence, they are removed from feature list, resulting in the final feature list of Engine_size, Horsepower, Wheelbase, Width, Length, and Fuel_capacity, on top of Manufacturer.

As we established that Manufacturers influence the mean and the variance of the distribution, a hierarchical model is proposed. A gamma distribution is used for the prediction of sales, with the alpha and beta calculated from the mean and the standard deviation. The mean is a linear model of the 6 built-in features. The intercept of the linear model, as well as the standard deviation, are functions of Manufacturer.

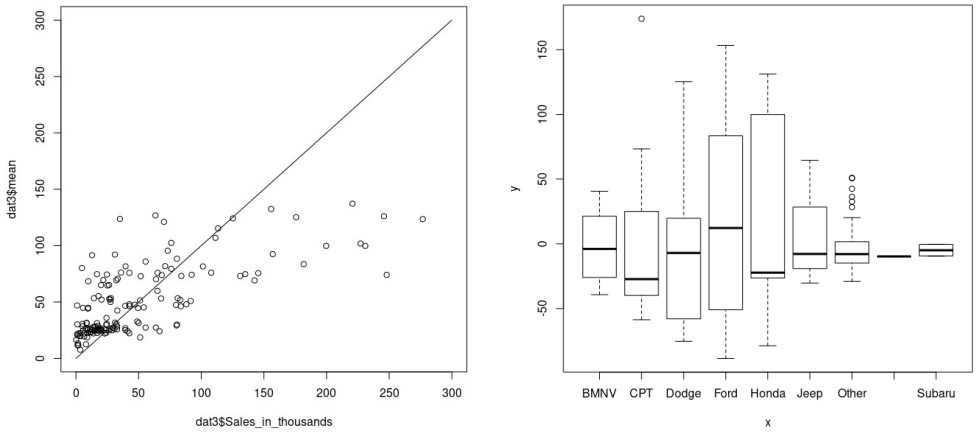


Fit the Model

The model was fitted with 10 000 points of burn-in, followed by 500 000 points of sampling. A scale reduction factors of at most 1.12 is achieved, with only 3 parameters being above 1.02 out of 24 parameters tracked (9 intercepts, 9 standard deviations, 6 coefficients). A mean deviance and penalty of 1377 and 69.7 was achieved.

Check the Model

On the left is a plot of the predicted expected value for each observation against the actual sales. On the right is (sales – expected value) against manufacturer. Although the model can fit the general trend, it is obvious that the variations are too high for this model to be useful. It can also be seen that variance increase with sales. Manufacturers with the highest variance (Ford, Honda and Dodge) are also the manufacturers with the highest sales.



Iterate if Necessary

For the next iteration, perhaps try higher-order terms and interactions could help to capture more complex behaviours. Making coefficients dependent on the manufacturer might also help to better fit the model.

Use the Model

Need to refine model before use.