AUTOMOBILE DATA ANALYSIS

Akhil A. Naik

Contents

1 Introduction **3**

2 Exploratory Data Analysis **3**

2.1 Attribute Information 3

2.2 Data Manipulation And Summary Statistics 4

2.3 Visualisations 5

3 Conclusion **11**

4 Code Snippets **11**

References **13**

1 Introduction

Automobiles when launched in the market come with a fixed price with some additional costs in the form of Government taxes. The fixed price is assigned by the manufacturer by considering several features associated with the car like its engine, body-style, fuel type etc. Also manufacturer/brand value is also associated with the price of a car. However, these vehicles are built with keeping budget in mind with higher the budget, better will be the features, stronger the customer appeal and these factors contribute in deciding the price.

Also, every car and light commercial vehicle such as a small van, new or old, falls into a particular insurance group, which dictates to an extent how much it will be to insure (Hayes 2022). The risks associated with the vehicles are quantified using these groups by the insurers. Higher cost contributions are associated with higher group numbers. Cars are initially assigned these risk factor groups/symbols associated with its price which is adjusted accordingly based on the risk with higher the risk higher the factor. This process is termed as *"symboling"* by the actuarians.

With symboling arises the questions of what factor influences the risk associated with a car and if these factors affect the price of a car in any way and if is there any way of predicting the price of the car by planning ahead on the features to be included. Answering these key questions is of utmost importance from the point of view of companies. With the motivation of answering these question, the objective is to analyze automobile data present in the UCI Machine Repository (Schlimmer 1987), which was collected from the 1985 Ward's Automotive Yearbook.

2 Exploratory Data Analysis

This data set consists of three types of entities mainly the specification of an auto in terms of various characteristics, its assigned insurance risk rating and its normalized losses in use as compared to other cars (Schlimmer 1987). The second rating corresponds to the degree to which the auto is more risky than its price indicates (symboling) with a value of +3 indicating that the auto is risky while -3 indicating that it is probably pretty safe. The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year (Schlimmer 1987). These entities amount to a total of 26 attributes with 205 observations/instances.

2.1 Attribute Information

1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.

5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

2.2 Data Manipulation And Summary Statistics

Our first step is to import the data set and prepare it for further analysis. The data is available in the *imports-85.data* file, with comma seperated values and some null values as *(?)*. Note that columns/attribute names are not present in the same file and must be imported/added from a different *imports-85.names* file. There are a total of 11 categorical values with rest being continuous. Note that categorical variables are present as characters, so must be converted to factors for getting levels. There are also missing values for few features as shown below.

Table 1: Automobile Data Features With Missing Values

	Attributes	Missing_Values
2	normalized.losses	41
6	num_of_doors	2
19	bore	4
20	stroke	4
22	horsepower	2
23	peak_rpm	2
26	price	4

Since enough observations are present and observations with missing values are few, replacing the missing values with mean for continuous variables and mode for categorical variables is

considered. After analysing the statistical summary for the features, mean and median for all features are found to be close values hence mean is used for replacing the missing values. Also symboling feature which is categorical variable of 6 levels ranging from (-3,3) can be presented as different variable indicating if vehicle is risky or not risky (safe), with risky mapping to values 1,2,3 and not risky to 0,-1,-2,-3. Following tables show the count automobiles w.r.t the risk categories:

Table 2: Count of Vehicles w.r.t symboling feature

symboling	Automobile_Count
-2	3
-1	22
0	67
1	54
2	32
3	27

Table 3: Count of Vehicles w.r.t new isRisky feature

isRisky	Automobile_Count
no	92
yes	113

The the above tables we can confirm that majority of observation have vehicles which are risky or not safe. Now we can look ahead the correlation heatmap to find if features are related significantly and analyse the data via visualisations.

2.3 Visualisations

As we can see from the below correlation heatmap (Fig 1), the symboling feature or the risk factor is strongly correlated with the body style (- *0.69*) and number of doors (*0.65*) on either ends. Even though these features are compared for correlation after converting them to numeric values, it still suggests that trying to compare the two features for variations in the risk factor may result is positive results. We can also see how various features like engine size, horse power show strong positive correlation with the price of the vehicle. Overall, we notice majority of the features show either strong positive or negative correlation with each other as expected.

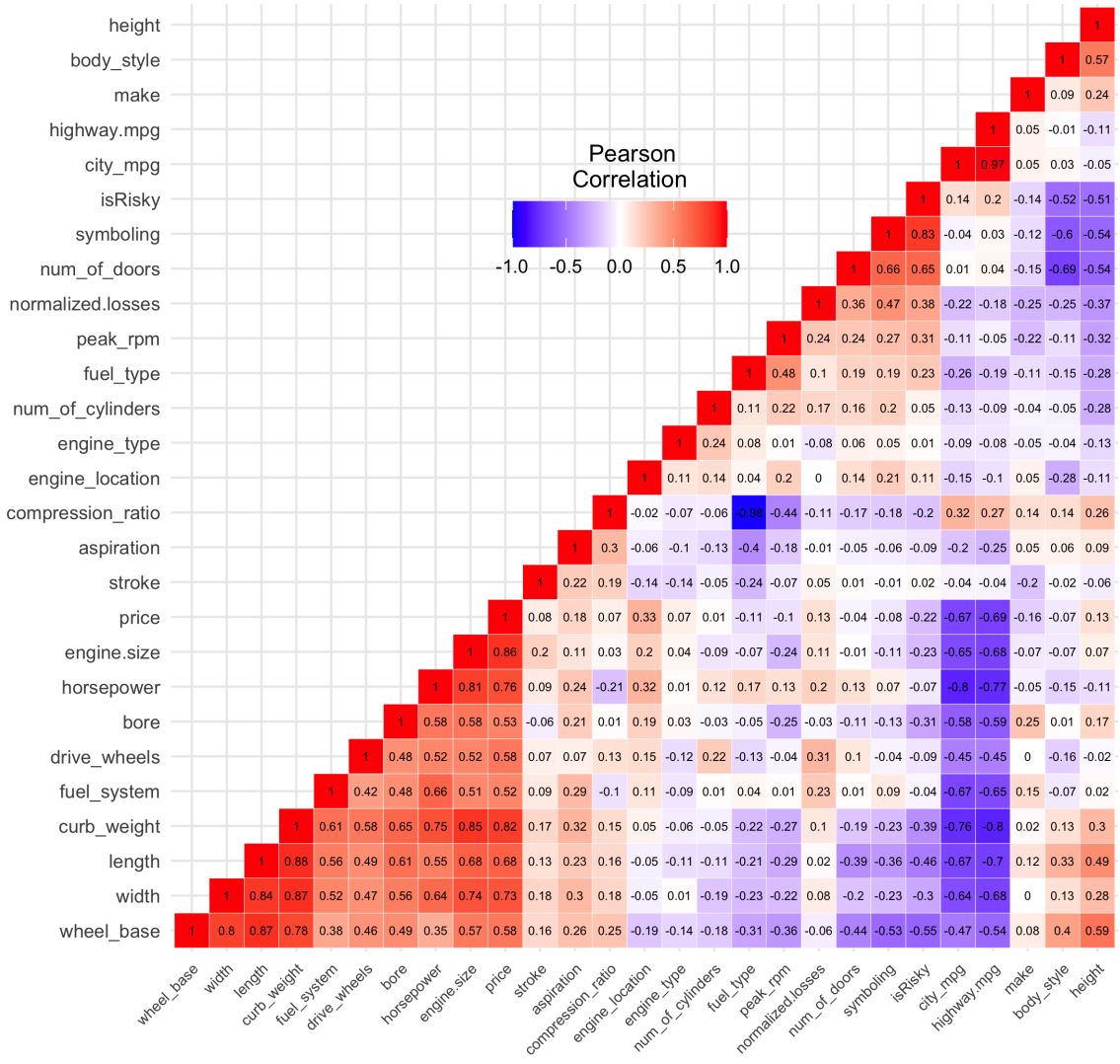

Figure 1: Correlation Heatmap

Another strong correlation we observed is between wheelbase which is the distance between the centers of the front and rear tires on a vehicle and the risk factor. Also it should be noted that wheelbase has a strong positive linear relationship between height of the vehicle which is expected as height increases the wheelbase will also increase. The following figure (Fig 2 (1)) shows variation of the risk factors associated with vehicles body style and number of doors. The symboling feature or the risk factor levels are also notably (Fig 2 (2)) related to the Wheel base and height of the car as seen ahead.

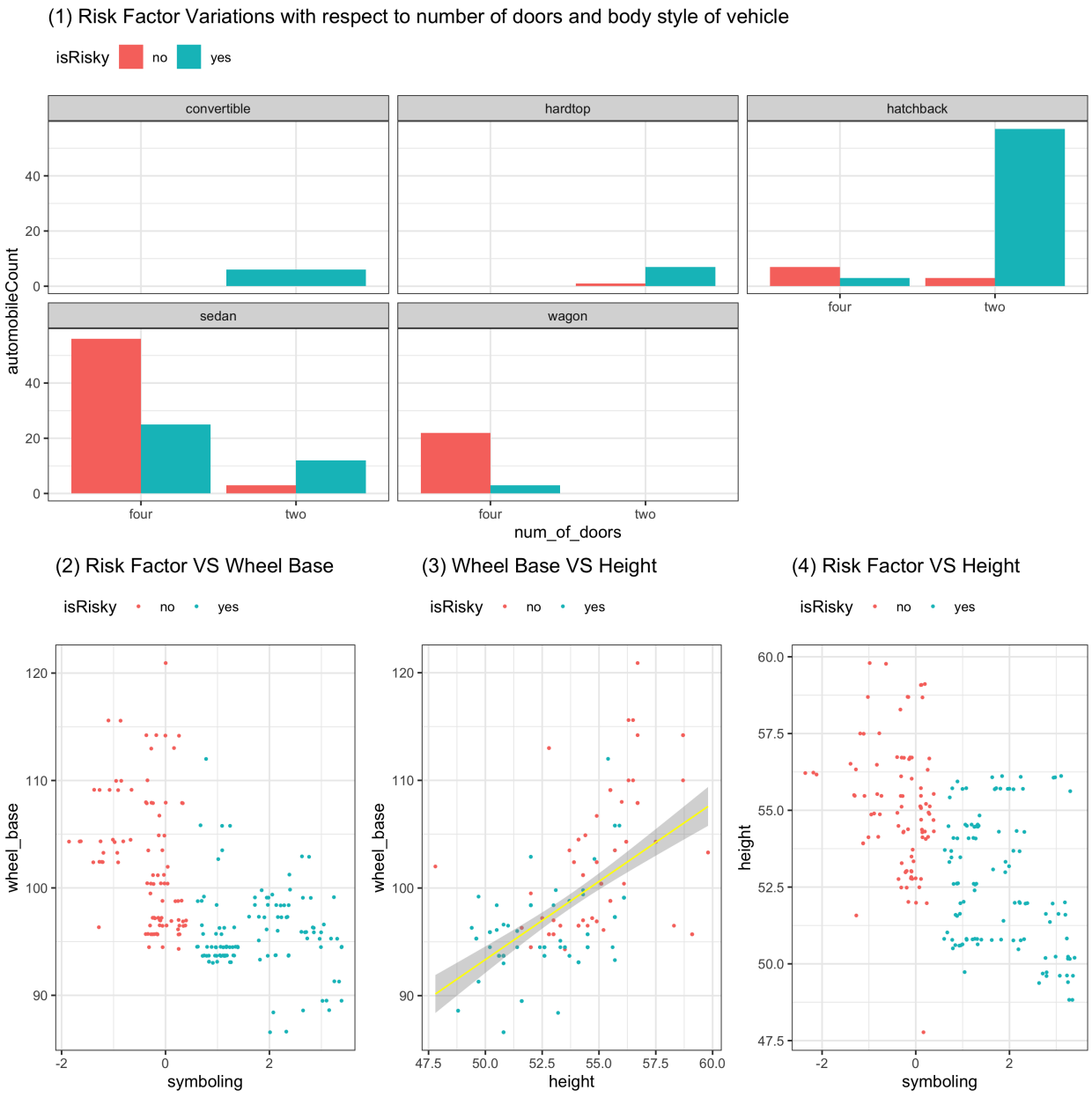

Figure 2: Risk Factor variations w.r.t (1) number of doors and body style, (2) wheel base, (4) Height and (3) Relation between wheel base and height

As seen from above plots (Fig 2 (1)) which show variations in risk factor in association with body style and number of doors of the automobile, it is observed that four door vehicles are safe compared to two door vehicles. Also two door hatchback vehicles involve most risk compared to the safe 5 door sedan. In general we can say that two door vehicles are aimed at enthusiast who enjoy driving for fun rather than the four who vehicles which are aimed for families, which will be driven with more caution. Also as we can see (Fig 2 (2)) higher the wheel base, safer the vehicle. Cars with long wheelbases tend to have better ride quality than those with short wheelbases. This is simply because there's more time between the front and rear wheels hitting any bumps, so the car is less likely to become unsettled. Wheel

base and height are linearly related (Fig 2 (3)) as expected and hence height can also be said to impact the risk factors and levels directly (Fig 2 (4)). Similarly there are other feature like lenght and width of car which also have correlation with symboling and can impact the risk factor in a vehicle.

Our next step to is to check how price varies depending on different features of the automobile.

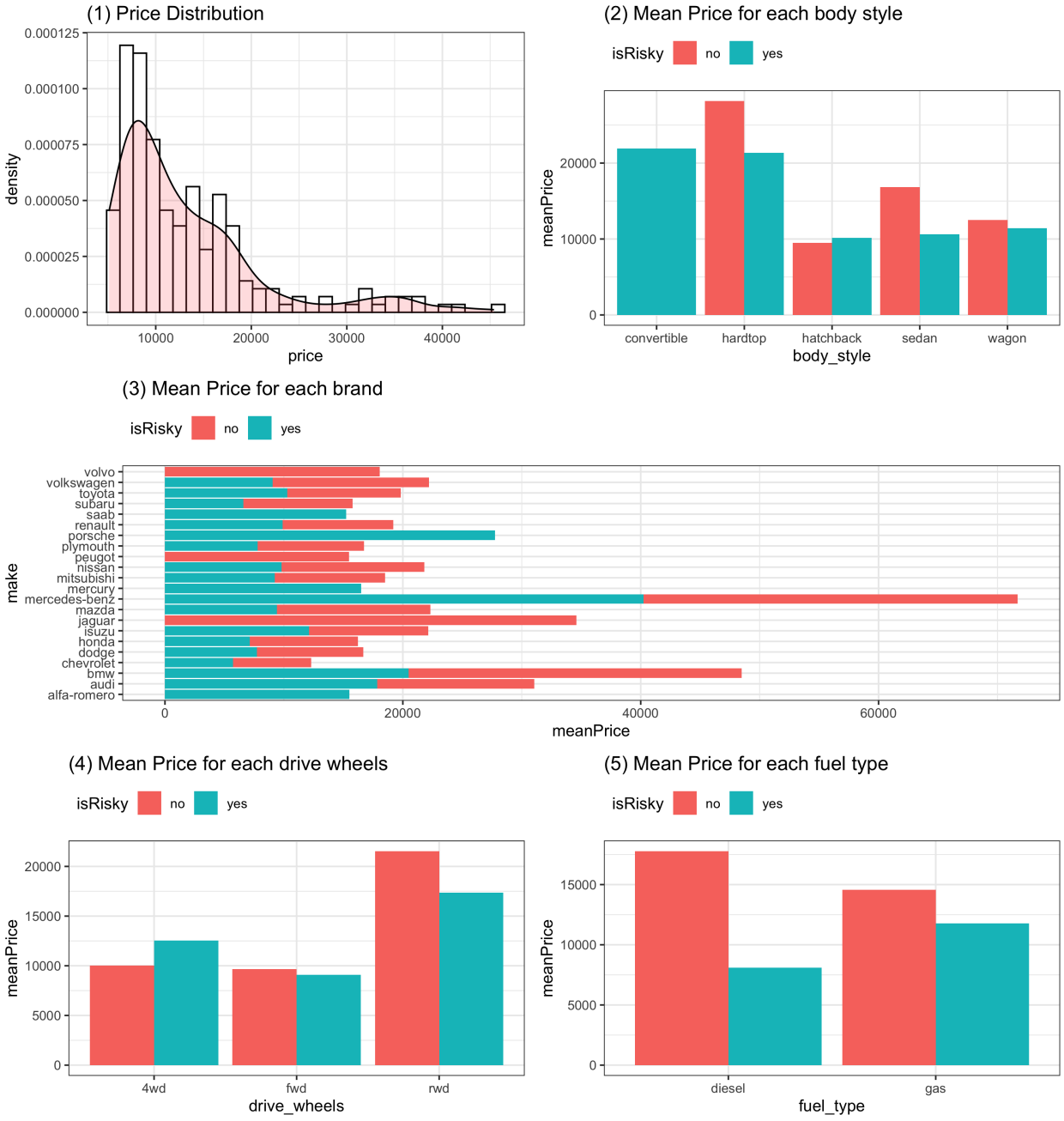

Figure 3: Price Distribution (1) and its average variation w.r.t body style (2), make (3), drive wheel types (4) and fuel type (5), corresponding to the risk factor

The price distribution in above figure (Fig 3) shows most of the vehicles have budget (low) prices. The average price for each body style is high for convertible and hard top. Mercedez-benz shows the highest average price whereas chevrolet shows the lowest average price for their vehicles. Rear wheel drive vehicles cost more on average compared to the others. While diesel vehicles cost more on average for safe ones, the risky vehicles cost more on average for gas type compared to diesel type. However as expected, the safe vehicles when compared with each category cost more on average than the risky ones as we know additional cost in incurred in making them more safe.

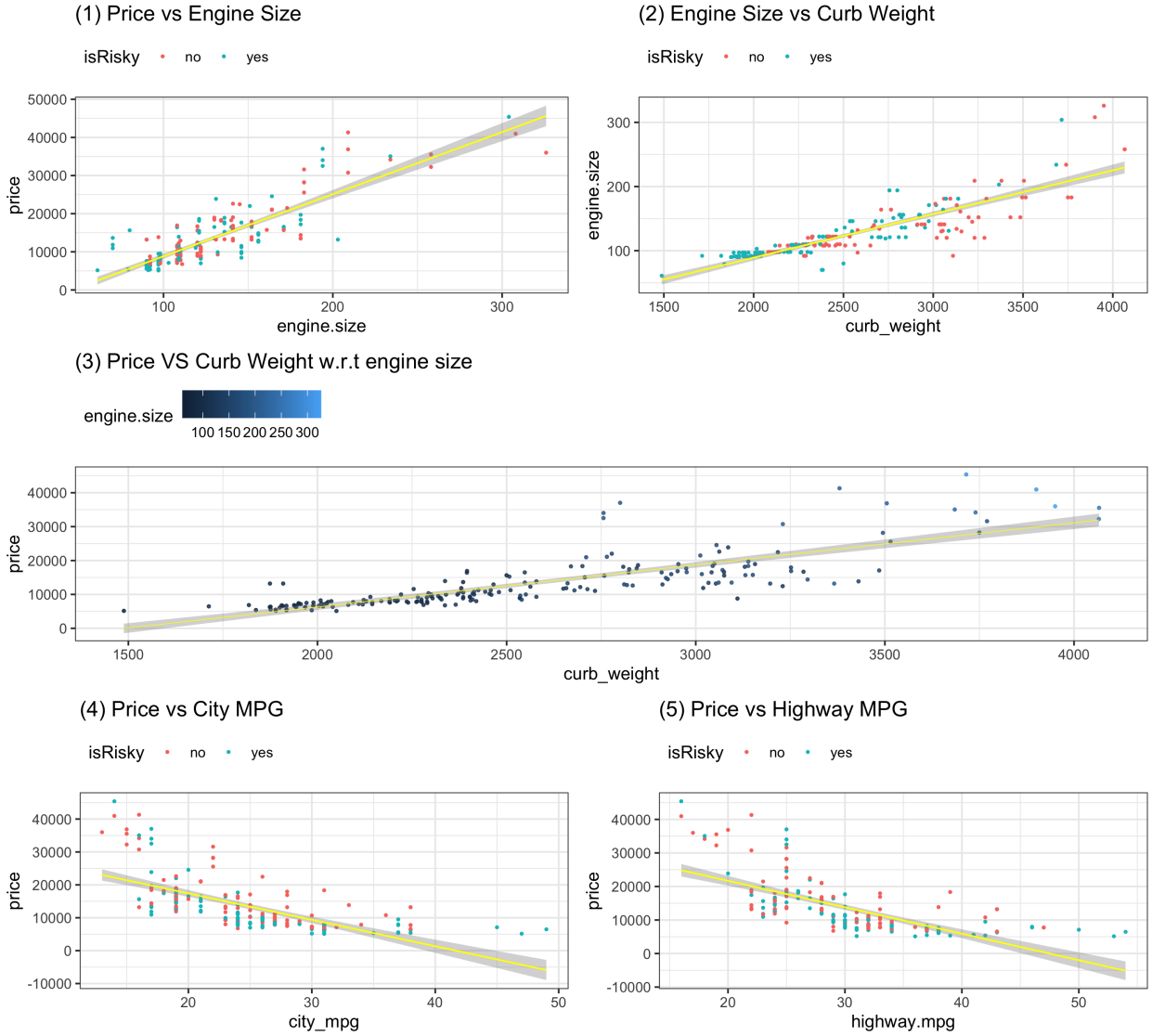

Figure 4: Price variations with several continuous features and w.r.t risk factor

The price shows strong positive correlation with engine size, horsepower, curb weight, length and width and shows strong negative correlation with fuel efficiency i.e. city_mpg and highway_mpg. Low the engine size, low is the price and maximum vehicles have engines with low size (Fig 4). Engine size and Curb weight show strong positive relation hence we check

price w.r.t curb weight of the vehicle associated with engine size and we confirm our initial reading of low price associates with lower engine size as well as lower curb weight. Both the fuel efficiency features show negative correlation. We can assume here that most consumers opt for low budget vehicles with high fuel efficiency.

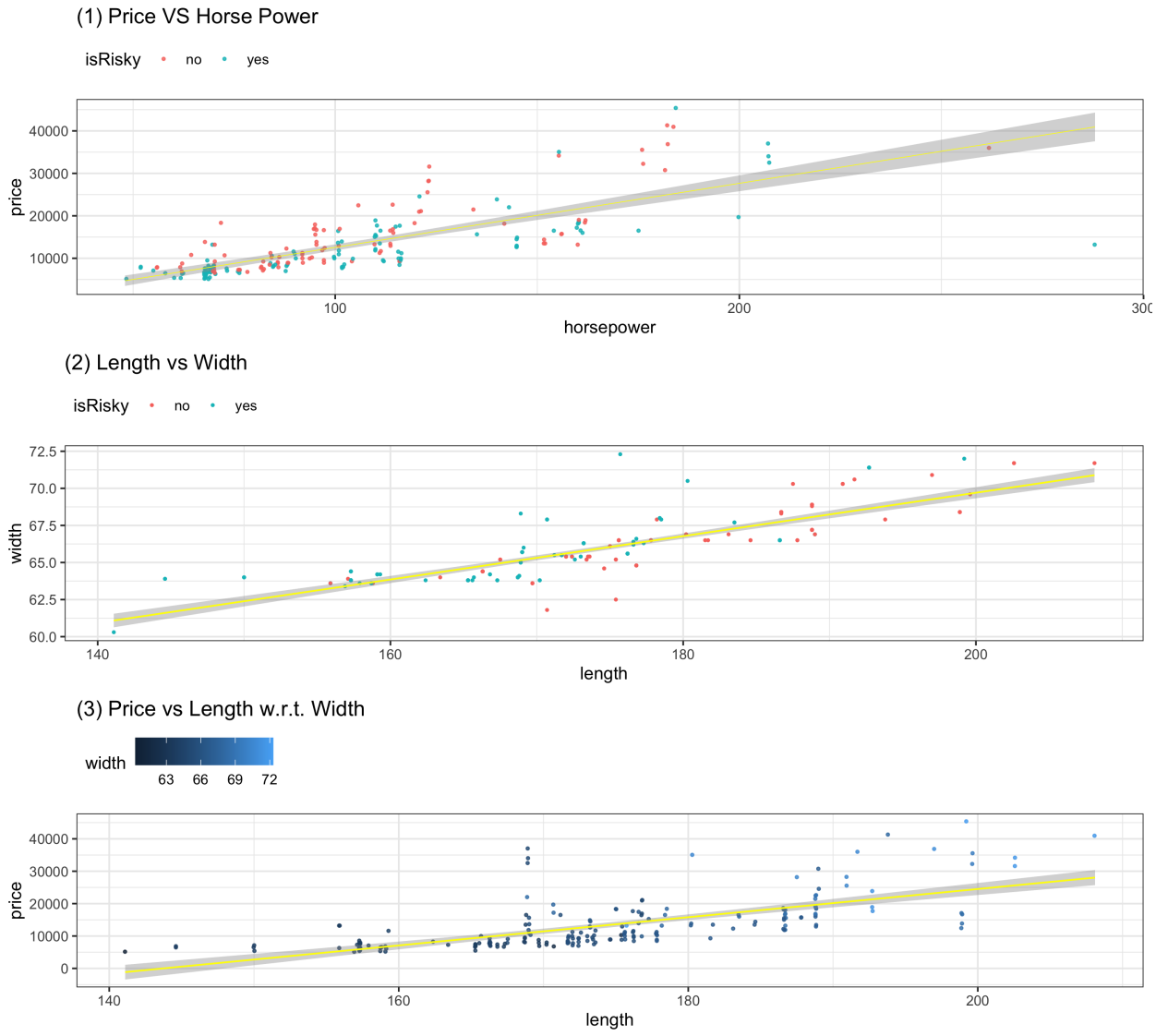

Figure 5: Price variations with horse power, length and width w.r.t risk factor

As observed (Fig 5), price has a linear relationship with horsepower and increases linearly. Length and width also have linear association with each other as well as the price. Lower the length and width (size in terms of vehicle classification (*Vehicle Size Class*, n.d.)) of the vehicle, less is the price.

3 Conclusion

The vehicle risk factor or levels were observed to be closely dependent on the body style, number of doors, height and wheel base. The price was observed to be dependant on related to engine size, curb weight, horsepower and many such continuous features. Although all the features were not analysed extensively, the correlation map can give us a rough idea as to how all the feature are co-related with each other either negatively or positively. We can use a GLM or random forest predictor with all the features included to get the best price predictor for better analysis.

4 Code Snippets

```
# Importing the data
columns = c('symboling','normalized-losses','make','fuel_type','aspiration',
            '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')
auto_data = read.table('imports-85.data',sep=',',na.strings = '?',
                       stringsAsFactors = TRUE,col.names = columns)
na_df = setNames(nm=c('Attributes','Missing_Values'),
                 stack(colSums(is.na(auto_data)))[2:1])
na_df = na_df[na_df['Missing_Values']>0,]
tprint(na_df,"Automobile Data Features With Missing Values")
```

```
# Function to find mode
getmode = function(v) {
   uniqv = unique(v)
   return(uniqv[which.max(tabulate(match(v, uniqv)))])
}
# Replacing NA's
auto_data$normalized.losses[is.na(auto_data$normalized.losses)] =
  mean(auto_data$normalized.losses,na.rm=TRUE)
auto_data$num_of_doors[is.na(auto_data$num_of_doors)] =
  getmode(auto_data$num_of_doors[!is.na(auto_data$num_of_doors)])
```

```
# Creating the isRisky variable
auto_data = auto_data %>%
  mutate(
     isRisky = case_when(
         symboling==3 | symboling==2 | symboling == 1 ~ 'yes',
         symboling==-3 | symboling==-2 | symboling == -1 | symboling == 0 ~ 'no'
     ))
auto_data$isRisky = as.factor(auto_data$isRisky)
```

```
# Getting average price w.r.t body_style
d2 = auto_data %>%
  group_by(body_style,isRisky) %>%
  summarise(meanPrice = mean(price))
p6 = ggplot(data=d2, aes(x = body_style, y = meanPrice, fill = isRisky)) +
geom_bar(stat="identity", position=position_dodge())+
theme_bw()+ggtitle("(2) Mean Price for each body style")+
theme(legend.position='top', legend.justification='left',legend.direction='horizontal')
```

```
# Plotting Price vs Engine Size
p13 = ggplot(data=auto_data, aes(x = engine.size, y = price))+
    geom_point(size = 0.5,aes(col = isRisky))+
  stat_smooth(method = "lm", col = "yellow", size=0.5) +
  ggtitle("(1) Price vs Engine Size")+
    theme_bw()+
  theme(legend.position='top',
        legend.justification='left',
        legend.direction='horizontal')
#...
```

References

Hayes, Russell. 2022. *How Do Insurance Ratings for New Cars Work? The Car Expert*. https://www.thecarexpert.co.uk/how-do-insurance-ratings-for-new-cars-work/.

Schlimmer, Jeffrey C. 1987. *Automobile Data Set. UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/automobile.

Vehicle Size Class. n.d. *Wikipedia*. https://en.wikipedia.org/wiki/Vehicle_size_class.