Practical Exercise 2 | Statistics for Premasters DSS/CSAI

Hilal Caliskan Egilli u927185

The objective of this practical exercise is to get you accustomed to some data wrangling and visualisation methods within R. You may use your own style of visualisation (as long as you use R). If you want to make visualisations with ggplot, you can make use of the cheat sheet and learning resources here: https://ggplot2.tidyverse.org/.

Task 1. Load the data set from the following file: `cars_df.csv`. You can import the file using read.csv(), or by using the "Import Dataset" tab. The file is included in the PE 2 zip folder. The dataset contains a random sample drawn from a large set of cars sold in Spain. Go ahead and do a bit of data exploration on the dataset, like creating a summary.

```
library(lsr)
library(psych)
thedata <- read.csv("/Users/hilalcaliskan/Documents/PM_DSS_ Assignments/CSAI/Practical Exercise 2/cars_
colnames(thedata)
```

```
##  [1] "x"          "ID"         "make"       "model"      "version"
##  [6] "months_old" "power"      "sale_type"  "gear_type"  "fuel_type"
## [11] "kms"        "price"
```

```
head(thedata)
```

```
##       x     ID       make   model                       version months_old
## 1 57226  48363    Renault  Laguna               1.5Dci Emotion         47
## 2 84549  48076    Renault  Laguna       G.T. 2.0Dci Expression        140
## 3 31902  55048 Volkswagen GolfGti                       2.0 Tsi         89
## 4 84981 105416      Smart  Fortwo    Cabrio 52 Mhd Passion Aut.         62
## 5 93720  48523    Renault  Megane       M?gane Classic 1.9Dt Rn        230
## 6 32408  50050    Renault  Megane Dci 110 Zen Energy 81 Kw (110         13
##   power sale_type      gear_type fuel_type   kms price
## 1    81      used         manual    diesel 23714 11150
## 2   110      used         manual    diesel  5448  4900
## 3   155      used         manual  gasoline  4460 17000
## 4    52      used semi-automatic  gasoline 19443  4800
## 5    70      used         manual    diesel 10053   995
## 6    81      used         manual    diesel  6429 16990
```

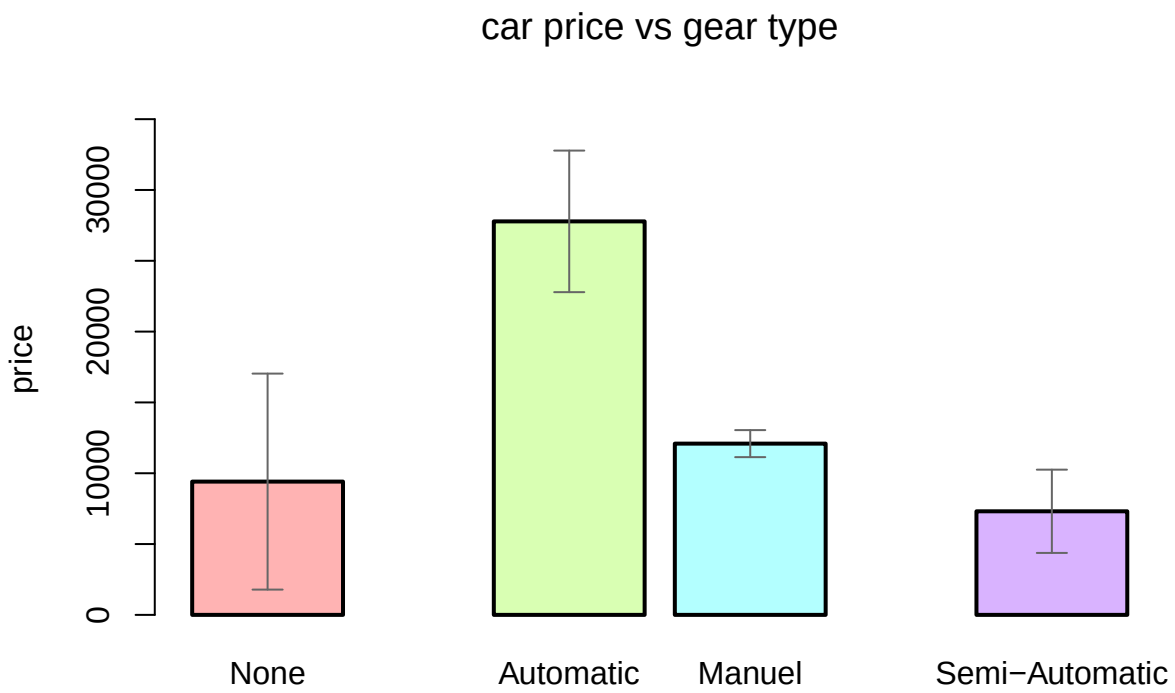
```
tail(thedata)
```

```
##         x    ID          make                 model
## 313   551 83732          Land RoverRangeRoverSport
## 314 32302 10624           Bmw                   116
```

```
## 315 62173 26388          Ford               Ka/Ka+
## 316 78631 66055       Citroen                   C4
## 317 83620 32163 Mercedes-Benz                 C220
## 318 92002 52469    Volkswagen                 Golf
##                                version months_old power sale_type gear_type
## 313     3.0Sdv6 Hse Dynamic 306Cv Aut.          7   225      demo automatic
## 314                   3P Pack M 116Cv         38    85      used    manual
## 315 Ka+ 1.2 Ti-Vct 63Kw Ultimate 85 5P          9    63      used    manual
## 316         1.4 16V Collection 5P 88Cv        136    65      used    manual
## 317                     Cdi Avantgarde        140   110      used    manual
## 318                       2.0 Highline        197    84      used    manual
##     fuel_type   kms price
## 313    diesel  2598 84950
## 314    diesel 18939 16990
## 315  gasoline     3 10000
## 316  gasoline 21148  6795
## 317    diesel 15505  5000
## 318  gasoline 14523  2000
```

Task 2a. Generate a bar graph that plots the mean value of car price and is organized by gear type. Be sure to include appropriate labels (meaningful title and label names). Include the plot in your answer.

```
library(lsr)
thedata$gear_type <- as.factor(thedata$gear_type)
bars(formula = price ~ gear_type,
     data = thedata,
     yLabel = "price",
     xLabels = c("None", "Automatic", "Manuel","Semi-Automatic"),
     main="car price vs gear type")
```



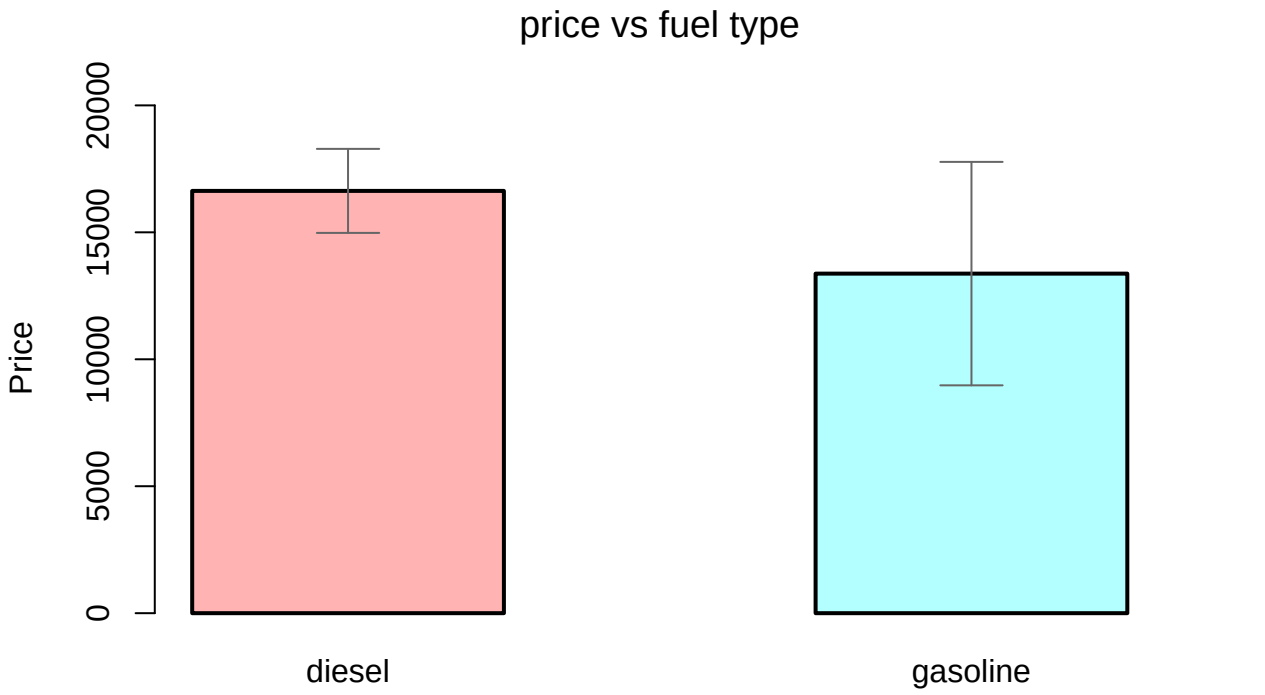
Task 2b. Which gear type has the highest mean price?

Automatic

Task 3a. Generate a bar graph that plots the mean value of price and is organized by fuel type. Make sure to include error bars, and some appropriate labels. Include the plot in your answer.

HINT: *Error bars are included in the bars() function within the lsr package by default.*

```
library(lsr)
thedata$fuel_type <- as.factor(thedata$fuel_type)
bars(formula = price ~ fuel_type, data = thedata,
     yLabel= "Price",
     xLabels = c("diesel", "gasoline"),
     main = "price vs fuel type")
```

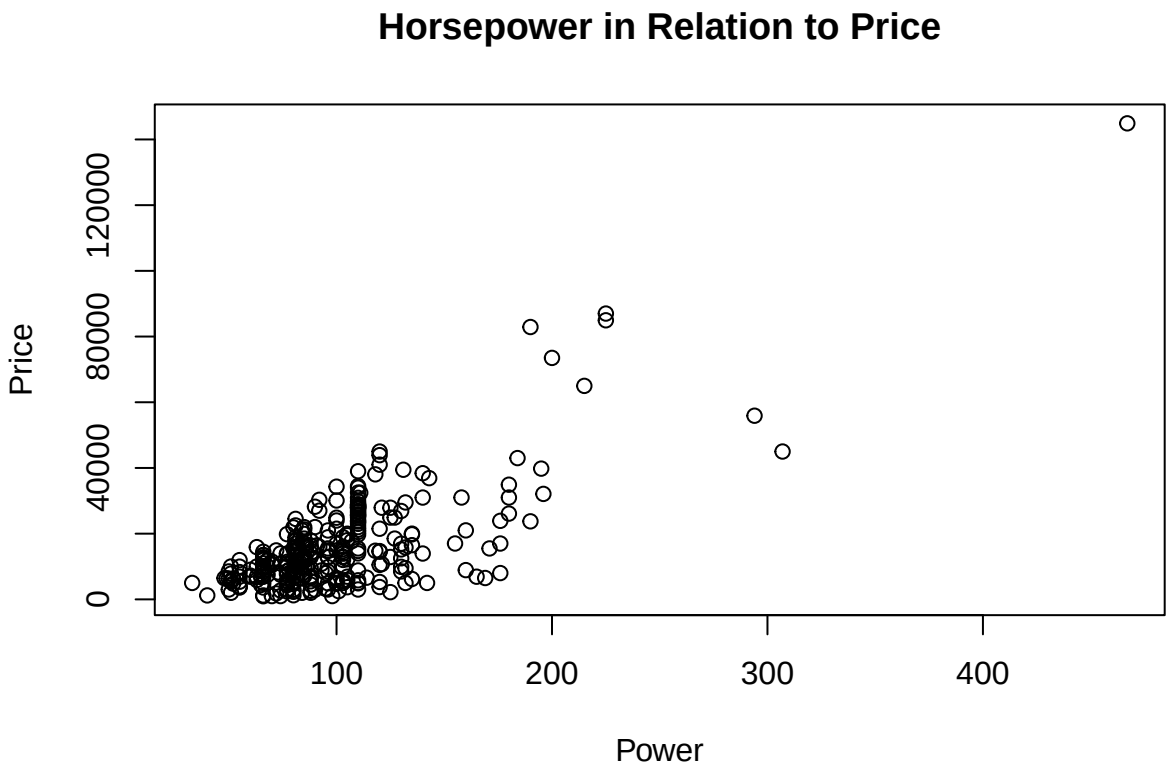


Task 3b. Which fuel type has the highest level of confidence over the mean price value? Why is this confidence level higher?

Diesel has the highest confidence as the confidence interval is more narrow.

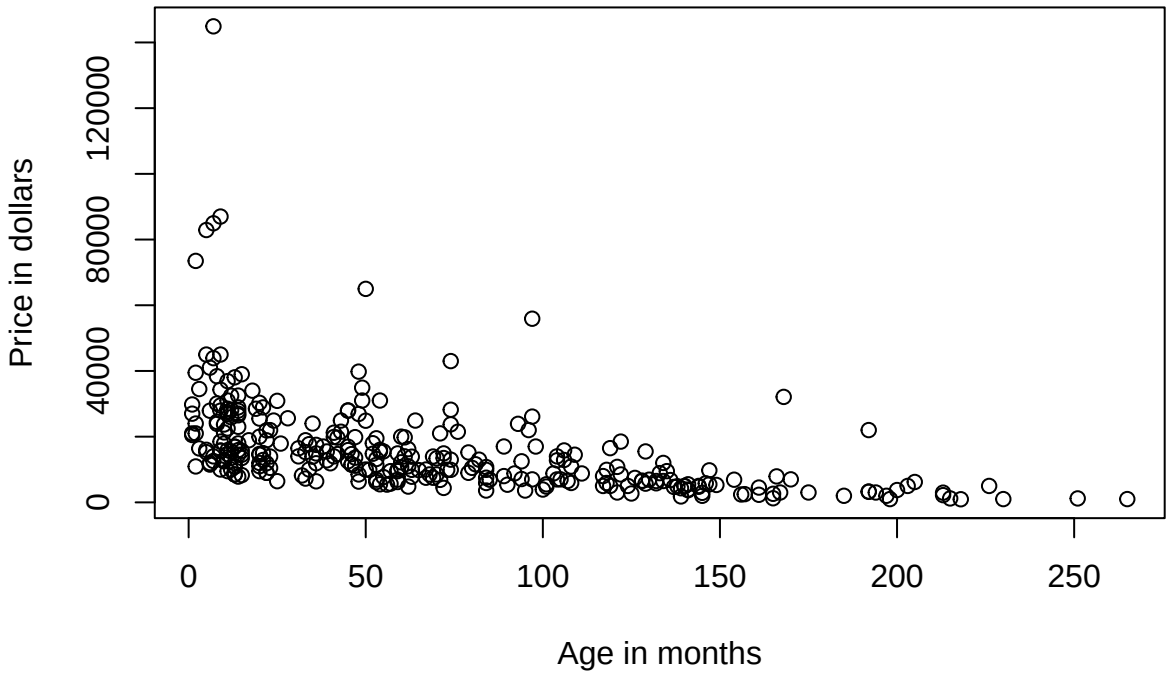
Task 4. Create a scatterplot with power (hp) on the x-axis and price on the y-axis. Be sure to include appropriate labels and add the plot in your answer.

```
plot(x = thedata$power, y = thedata$price,
     ylab = 'Price',
     xlab = 'Power',
     main = "Horsepower in Relation to Price")
```



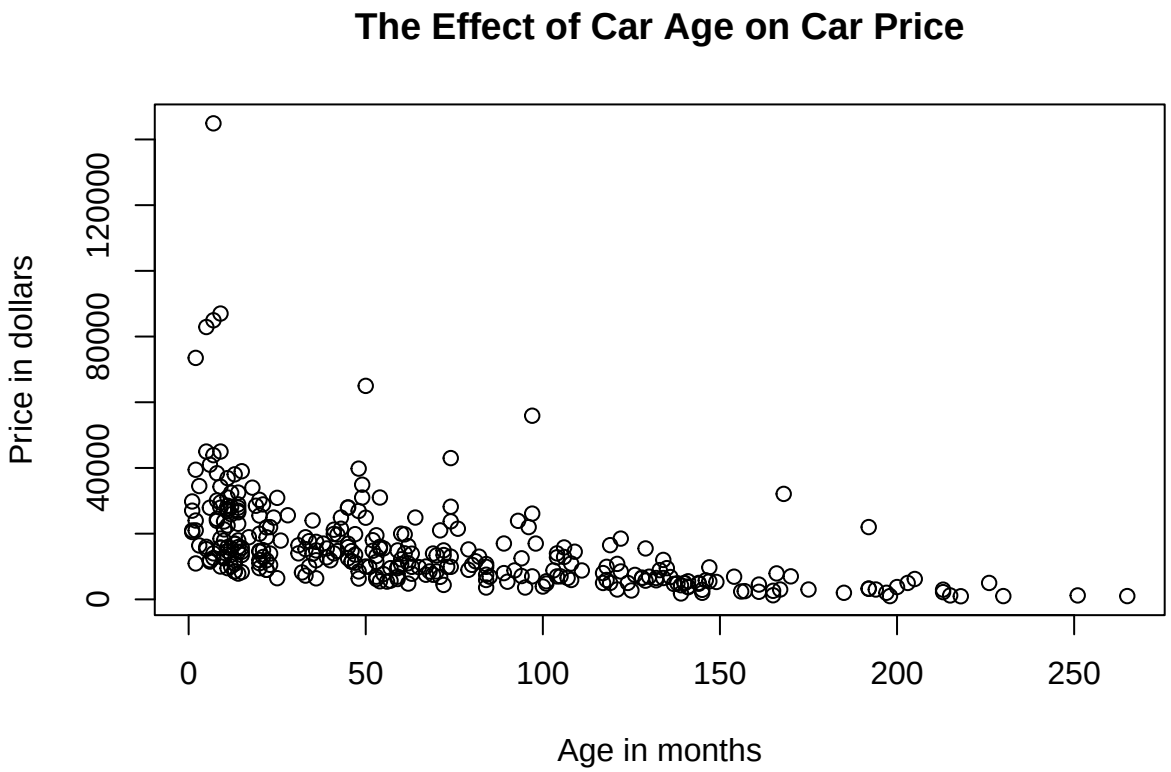
Task 5. Create a scatterplot of months_old on the x-axis and price on the y-axis. Be sure to add appropriate labels and include the plot in your answer.

```
plot(price ~ months_old, data = thedata, # The dependent variable goes first
     main = "The Effect of Car Age on Car Price",
     xlab = "Age in months",
     ylab = "Price in dollars")
```




```
## Alternatively, we can specify the x and y axis like this instead of the formula:

plot(x = thedata$months_old, y = thedata$price, # The dependent variable is the y variable
     main = "The Effect of Car Age on Car Price",
     xlab = "Age in months",
     ylab = "Price in dollars")
```



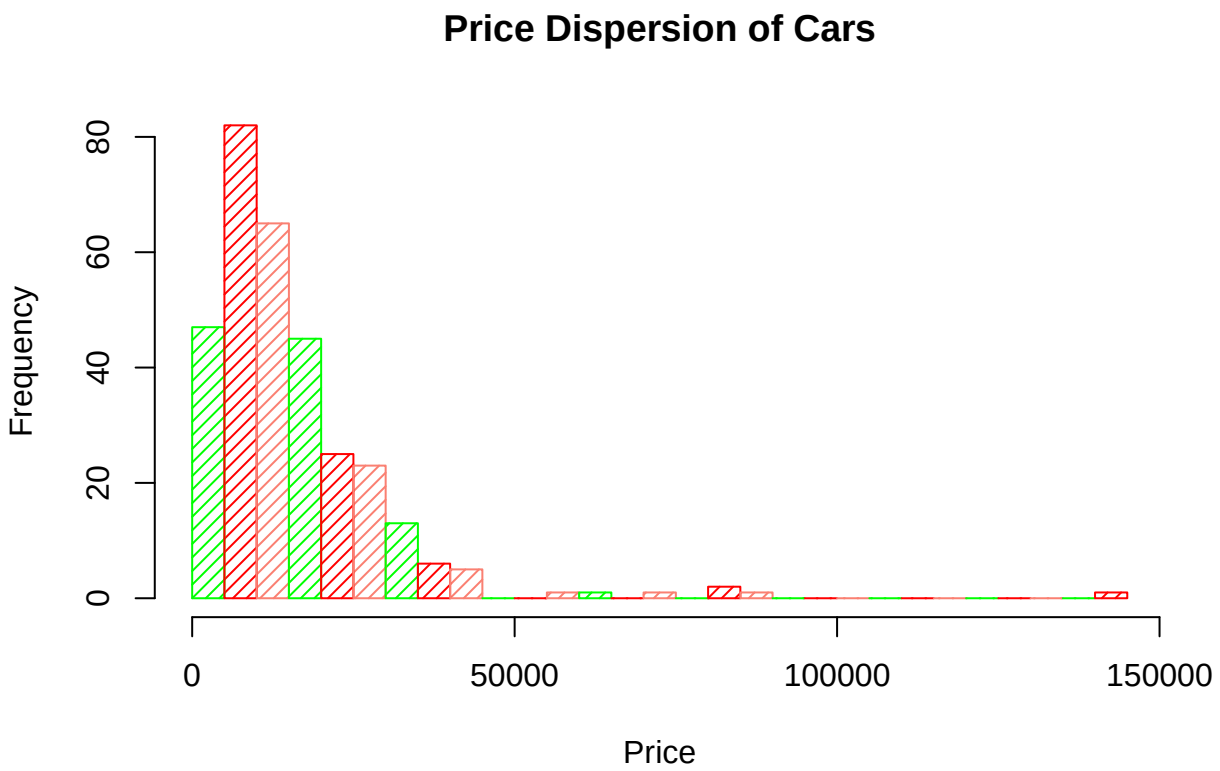
Task 6. Based on the plots, what can you say about the relationship between variables: power and price (task 4), and, months_old and price (task 5). Which relationship seems stronger?

We will talk about correlation in later lectures. For now our intuition is fine.

Visualisation 1: More power results in higher price. Visualisation 2: A car becomes cheaper over time.

Task 7a. Make a histogram of the price variable that features coloured bars and appropriate labels. Include the plot here.

```
hist(x = thedata$price,
     main = "Price Dispersion of Cars",
     xlab = "Price",
     breaks = 40,
     density = 20,
     col = c("green", "red", "salmon"))
```



Task 7b. Based on your histogram, what can you say about the distribution of the price values?

That the prices might be very skewed to the positive side. Most car prices are in the under 25k category.