The document is listed in a way in which anyone can understand the importance of dummy variables for creating a correlation matrix.

1. Identify the categorical variables.

|  |  |
| --- | --- |
| CATEGORICAL VARIABLE | DESCRIPTION |
| 1. Fuel\_Type | Fuel Type (Petrol, Diesel, CNG) |
| 1. Color | Color (Blue, Red, Grey, Silver, Black, etc) |
| 1. Met\_Color | Metallic Color? (Yes=1, No=0) |
| 1. Mfr\_Guarantee | Within Manufacturer's Guarantee period (Yes =1, No=0) |
| 1. BOVAG\_Guarantee | BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0) |
| 1. ABS | Anti-Lock Brake System (Yes=1, No=0) |
| 1. Airbag\_1 | Driver\_Airbag (Yes=1, No=0) |
| 1. Airbag\_2 | Passenger Airbag (Yes=1, No=0) |
| 1. Automatic | Automatic ( (Yes=1, No=0) |
| 1. Sport\_Model | Sport Model (Yes=1, No=0) |
| 1. Power\_Steering | Power Steering (Yes=1, No=0) |
| 1. Radio\_cassette | Radio Cassette (Yes=1, No=0) |
| 1. Powered\_Windows | Powered Windows (Yes=1, No=0) |
| 1. Radio | Radio (Yes=1, No=0) |
| 1. Parking\_Assistant | Parking assistance system (Yes=1, No=0) |
| 1. Metallic\_Rim | Metallic Rim (Yes=1, No=0) |
| 1. Backseat\_Divider | Backseat Divider (Yes=1, No=0) |
| 1. Mistlamps | Mistlamps (Yes=1, No=0) |
| 1. Tow\_Bar | Tow Bar (Yes=1, No=0) |
| 1. Central\_Lock | Central Lock (Yes=1, No=0) |
| 1. CD\_Player | CD Player (Yes=1, No=0) |
| 1. BoardComputer | Boardcomputer (Yes=1, No=0) |
| 1. Airco | Airconditioning (Yes=1, No=0) |
| 1. Automatic\_Airco | Automatic Airconditioning (Yes=1, No=0) |
| 1. Mfg\_month | Manufacturing month (1-12) |

1. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.

A categorical independent or predictor variable that contains values indicating membership in one of several possible categories. E.g., color (blue, red, grey, etc). The categories are often assigned numerical values used as labels, e.g. for Met \_Color, Yes = 1 and No = 0.

A dummy variable is derived from a categorical variable which have K number of categories.

A variable with K categories will be transformed into K dummy variables, with each dummy indicating whether a certain category is present or not. For example, Fuel\_type has 3 categories: Petrol, Diesel, and CNG. If we convert it to dummy variables, we get 3 dummy variables: Fuel\_Type\_Petrol (if the fuel type is Petrol then Fuel\_Type\_Petrol=1, otherwise Fuel\_Type\_Petrol=0), Fuel\_Type\_Diesel (if the fuel type is Diesel then Fuel\_Type\_Diesel=1, otherwise Fuel\_Type\_Diesel=0) and Fuel\_Type\_CNG (if the fuel type is CNG then Fuel\_Type\_CNG=1, otherwise Fuel\_Type\_CNG=0).

1. How many dummy binary variables are required to capture the information in a categorical variable with N categories?

N-1 binary variables are required to capture the information in a categorical variable with N categories. As stated in the previous example of Fuel\_Type, there are 3 categories (Petrol, Diesel, CNG). Therefore N=3. To capture the information in the categorical variables we need N-1 values i.e 3-1 = 2 values. If we know that Petrol = 0 and Diesel = 0, then the fuel\_type = CNG. Hence N-1 variables are required to capture the information in a categorical variable with N categories.

In some routines, e.g. linear regression and logistic regression, use of all N dummies will cause the routine to fail because the nth variable contains redundant information and can be expressed as a linear combination of the others. Only N-1 variables should be used (they contain all the available information about the variable from which they were derived.

1. Convert the categorical variables in this dataset into dummy variables, and explain in words, for one record, the values in the derived binary dummies.

R Script for creating dummy variables:

#Read the dataset

library(readxl)

ToyotaCorolla <- read\_excel("ToyotaCorolla.xlsx",

+ sheet = "data")

#Create Data Frame

Corolla <- as.data.frame(ToyotaCorolla)

#Create dummy variables using modelmatrix() function for Color and Fuel\_type categorical variables

dummy <- model.matrix(~ 0 + Fuel\_Type + Color, data = Corolla)

#View one record from the dataset

head(dummy, n=1)

Once the 1st record is viewed, the following data is obtained.

Fuel\_TypeCNG Fuel\_TypeDiesel Fuel\_TypePetrol ColorBlack

1 0 1 0 0

ColorBlue ColorGreen ColorGrey ColorRed ColorSilver

1 1 0 0 0 0

ColorViolet ColorWhite ColorYellow

1 0 0 0

The table shows that the selected record is a card which has fuel\_type = Diesel as the value of Fuel\_TypeDiesel = 1. The values in Fuel\_TypeCNG and Fuel\_TypePetrol = 0 which signifies that the car does not have a CNG or a Petrol tank.

Next, it shows that the ColorBlue = 1 which signifies that the color of this particular car is blue. The values of all other dummy color variables = 0.

1. Producing a correlation matrix and matrix plot.

#Read the dataset

library(readxl)

ToyotaCorolla <- read\_excel("ToyotaCorolla.xlsx",

+ sheet = "data")

library(ggplot2)

library(gplots)

library(corrplot)

#Create data frame

Corolla <- as.data.frame(ToyotaCorolla)

#Create Correaltion matrix (remove columns which are not required)

cor.mat <- cor(Corolla[,-c(1,2,8,10,11,12,13,15,19,20,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39)])

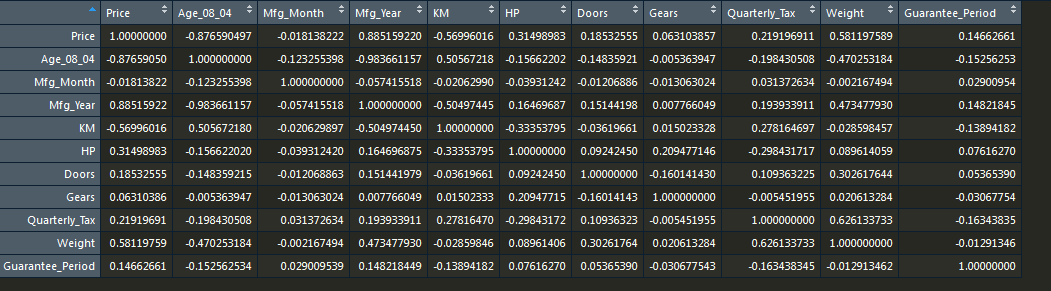
View(cor.mat)

#Plot the matrix using a heatmap

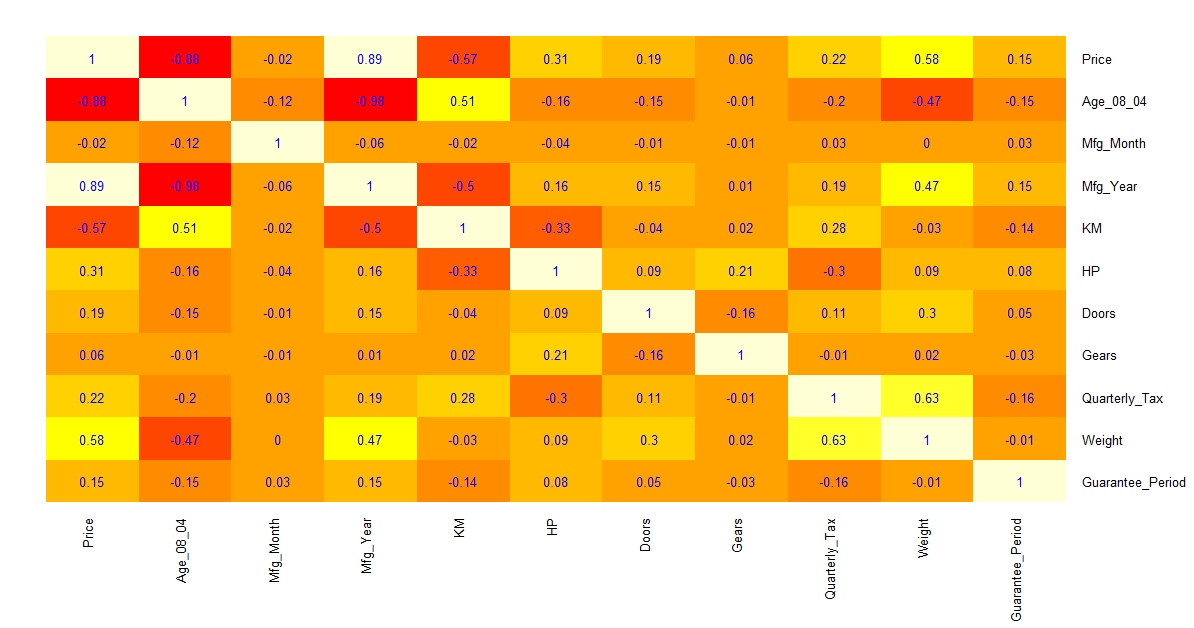
heatmap.2(cor(Corolla[,-c(1,2,8,10,11,12,13,15,19,20,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39)]), Colv = FALSE, Rowv = FALSE , dendrogram = "none",

cellnote = round(cor(Corolla[, -c(1,2,8,10,11,12,13,15,19,20,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39)]), 2),

notecol = "blue", key = FALSE, trace = 'none' , margins=c(9,9), cexRow=1, cexCol = 1)



Correlation matrix



Heatmap of correlation matrix.

**Comments:**

1. It can be seen that as the age\_08\_04 increases the price of the car decreases significantly. As the correlation between the two variables is -0.88.
2. Similarly we can see that when the mfg\_year increases the age\_08\_04 decreases (correlation of -0.98). Hence they have a large negative relationship.
3. Weight is positively correlated with price.
4. KM is negative correlated with price.
5. Age\_08\_04 is negatively correlated with price.