Text

Description automatically generated

4.3 a)

A categorical variable is the one that makes it possible to classify a series of data by using fixed values, so in our dataset all the variables are categorical except price, km, Age\_08\_04, Quarterly\_tax, HP, CC and Weight.

The categorical variables are the ones that takes fixed values to classify the series and they cannot be ordered.

4.3 b)

A binary dummy variable is a logical numeric variable to know if the data of a series meet or not a condition.

Binary dummy variables that are derived from another categorical variable show a 1 for observations that belong to that category and a 0 for observations that do not.

Table

Description automatically generated

4.3 c)

Exactly N - 1 binary dummy variables are needed to explain the information of a variable with N categories because each of them will show a 1 for observations that belong to the category and a 0 for observations that do not belong, the only problem is perfect correlation, so we simply remove one variable or category.

We can check it by executing the code

```

data<-dummy\_cols(Toyota.df$Color)

```

which creates binary dummy variables for the categorical variable Color, as we can see it creates N + 1 variables but the first of them shows the actual value of the observations instead of 1 and 0.

4.3 d)

The result of running the code

```

data<-dummy\_cols(Toyota.df$Color)

```

A screenshot of a computer

Description automatically generated with medium confidence

will be a matrix with N + 1 variables, the first one, "Data" will refer to the actual value of each observation and the rest of the variables will refer to each one of the categories of the variable Color. The variables will have the value 1 in case the observation belongs to the category and 0 otherwise.

4.3 e)

The result of running the code

```

cor(na.omit(Toyota.df[,-c(1,2,5,6,8,10:12,14:16,19:39)]))

```

Table

Description automatically generated with medium confidence

Creates a matrix of variances and covariances between the variables in the database except for 1,2,5,5,6,8,10:12,14:16,19:39, which shows the relationship between the selected variables, with values closer to 1 being the strongest direct relationships and values closer to -1 being the strongest indirect relationships.

Graphical user interface, text

Description automatically generated

114

a.)

As PC1 is the line that best fits the data and therefore best explains it, its values should be the highest. Any additional lines will only give a less satisfactory analysis of the data. This near-perfect fit may be due to the fact that the data are essentially on the same line that PC1 was able to identify and transform. This makes any further attempt to interpret the data completely inadequate.

b.)

In order to make the PC results similar to each other, a standardization approach was used. Since the scales and units of different variables vary widely, alternative interpretations may arise. In order to compare several unique results, standardization is necessary. This is the reason why it is often used.