

AKASH GUPTA(AXG170018)
Business Analytics with R(Homework)
Questions

The file ToyotaCorolla.xlsx contains data on used cars (Toyota Corollas) on sale during late summer of 2004 in the Netherlands.

It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications.

The goal will be to predict the price of a used Toyota Corolla based on its specifications.

1. Identify the categorical variables.
2. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.
3. How many dummy binary variables are required to capture the information in a categorical variable with N categories?
4. Use R to convert the categorical variables in this dataset into dummy variables, and explain in words, for one record, the values in the derived binary dummies.
5. Use R to produce a correlation matrix and matrix plot. Comment on the relationships among variables.

ANSWERS

ANSWER 1

Factors in the dataset which are categorical variables:

- 1.**Model**:This is a factor with 374 unique values.
- 2.**Fuel_Type**:Also a factor with 3 different values.
- 3.**Color**:This is a factor which has 10 values that is 10 different colors.

The numerical categorical variables are:

Mfg_Month,,Met_color,Automatic,Gears,Mfr_Guarantee,BOVAG_Guarantee,ABS,Airbag_1,Airbag_2,Airco,Automatic_Airco,BoardComputer,CD_Player,Central_Lock,Powered_Windows,Power_Steering,Radio,Mistlamps,Sport_model,Backseat_Divider,Metallic_Ring,Radio_Casette,Parking_Assistant, Tow_Bar.

The total number of categorical variables are 27:

Gears :This is taken as a categorical variable as it can have 3 values only.

I have neglected **doors** and **cylinders** as number of doors do not provide a significant correlation in price and cylinders are fixed in number throughout the dataset which as a result doesn't affect the data.

ANSWER 2

RELATIONSHIP BW CATEGORICAL AND BINARY DUMMY VARIABLE

A dummy variable is one which takes the value 0 or 1 resulting in indicating a presence or absence of some form of categorical effect which may be expected to shift the outcome. A categorical variable is one from which a dummy variable is derived. For every Categorical variable with N categories there are N-1 dummy variables created which in turn can answer the question like for example if a person is smoker or a non-smoker. When a Dummy variable is created for the same it either takes 0 which means "no" or 1 which means "yes". We can also take the example from this dataset considering the categorical variable **Fuel_Type** which has three possible values namely **CNG,DIESEL,PETROL**. Now since this categorical variable has 3 possible categories now (N-1) Dummy variables will be created which will show the result as 0 or 1.

ANSWER 3

In order to capture the information in a categorical variable with N categories we need (N-1) dummy binary variables.

For example if we have a Column as **Fuel_Type** which has three possible values **CNG,DIESEL AND PETROL**. R will create (3-1) dummy variables for this categorical variable.

ANSWER 4

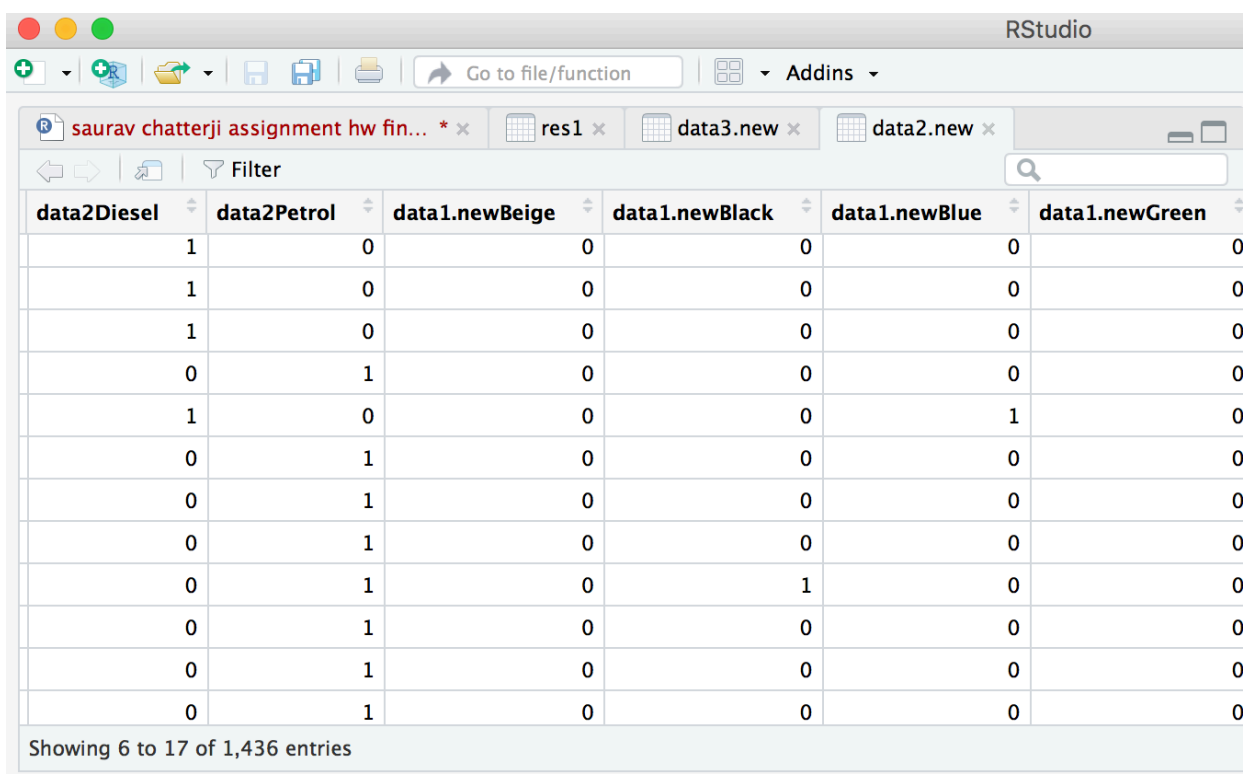
1. we convert the categorical variables in this dataset into dummy variables using `library(dummies)`
2. Now we convert the dataset to a dataframe and store it into data2
3. Now we make data1.new which creates dummy variables for data2\$Fuel_Type and column binds them to data2. Likewise we do this for Color as well and for the other categorical variables whose dummies we want to create.

Explanation for a record with visual:

We can explain this using the column **Fuel_Type** where the possible values from the dataset were:

- 1.PETROL
- 2.DIESEL
- 3.CNG

As we create the dummy variables now we have separate columns for these where each column reports values “0” and “1” which mean that if a “0” appears that the car is not that particular type and if “1” appears it means that the car belongs to that particular type.

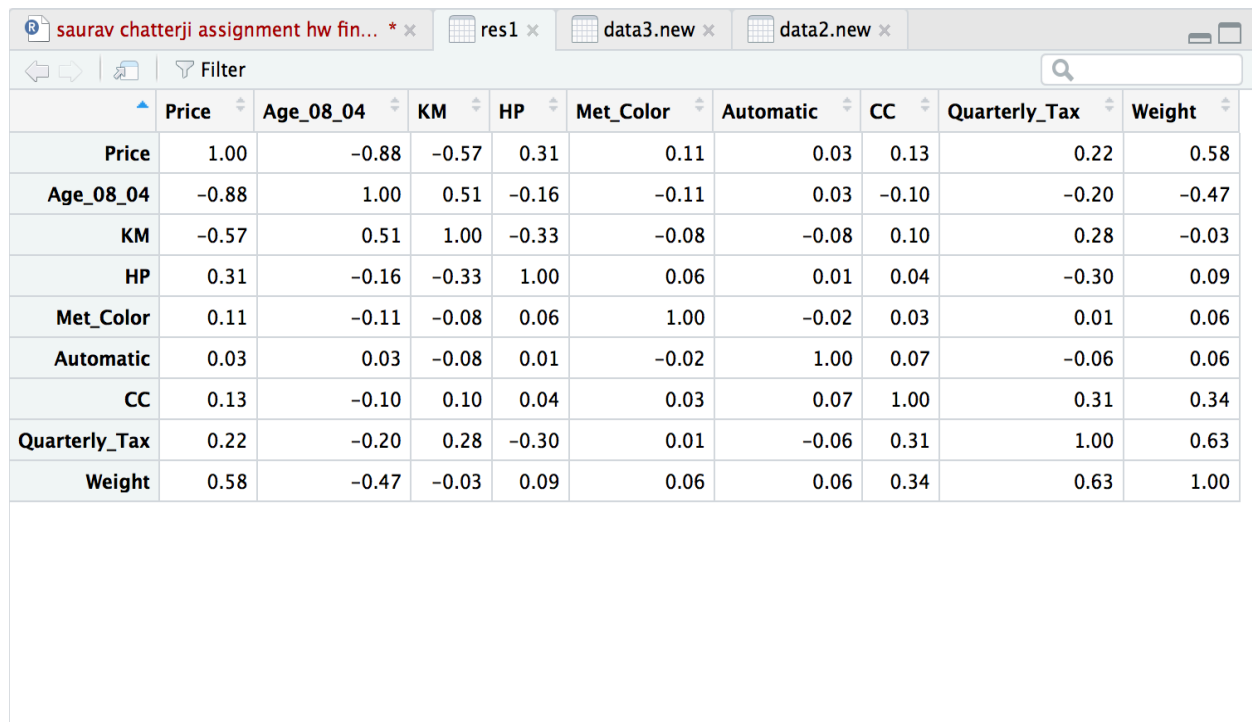


The screenshot shows the RStudio interface with a data table displayed. The table has six columns: data2Diesel, data2Petrol, data1.newBeige, data1.newBlack, data1.newBlue, and data1.newGreen. The data is organized into rows, with the first three rows showing data2Diesel = 1 and data2Petrol = 0, and the remaining rows showing data2Diesel = 0 and data2Petrol = 1. The color columns (data1.newBeige, data1.newBlack, data1.newBlue, data1.newGreen) show binary values (0 or 1) indicating the presence of a specific color. The table is filtered to show entries 6 to 17 of 1,436 total entries.

data2Diesel	data2Petrol	data1.newBeige	data1.newBlack	data1.newBlue	data1.newGreen
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	0
1	0	0	0	1	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	1	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0

Showing 6 to 17 of 1,436 entries

ANSWER 5



The screenshot shows a software window with multiple tabs: 'saurav chatterji assignment hw fin...', 'res1', 'data3.new', and 'data2.new'. The active tab displays a correlation matrix for ten car features: Price, Age_08_04, KM, HP, Met_Color, Automatic, CC, Quarterly_Tax, and Weight. The matrix is a 10x10 grid where each row and column represents a feature, and the cells contain the correlation coefficients between them. The diagonal elements are all 1.00, indicating perfect self-correlation. The off-diagonal elements show the relationships between different features, such as the negative correlation between Price and Age_08_04 (-0.88).

	Price	Age_08_04	KM	HP	Met_Color	Automatic	CC	Quarterly_Tax	Weight
Price	1.00	-0.88	-0.57	0.31	0.11	0.03	0.13	0.22	0.58
Age_08_04	-0.88	1.00	0.51	-0.16	-0.11	0.03	-0.10	-0.20	-0.47
KM	-0.57	0.51	1.00	-0.33	-0.08	-0.08	0.10	0.28	-0.03
HP	0.31	-0.16	-0.33	1.00	0.06	0.01	0.04	-0.30	0.09
Met_Color	0.11	-0.11	-0.08	0.06	1.00	-0.02	0.03	0.01	0.06
Automatic	0.03	0.03	-0.08	0.01	-0.02	1.00	0.07	-0.06	0.06
CC	0.13	-0.10	0.10	0.04	0.03	0.07	1.00	0.31	0.34
Quarterly_Tax	0.22	-0.20	0.28	-0.30	0.01	-0.06	0.31	1.00	0.63
Weight	0.58	-0.47	-0.03	0.09	0.06	0.06	0.34	0.63	1.00

From the Correlation matrix we can interpret the following results:

Price is inversely related to Age:

This means that with the Increase in the Age of the car the price will start to decrease.

Price is inversely related to KM:

This means that with the increase in number of kms the car is driven the price will decrease.

Price is directly related to HP:

This means that with the increase in number of HP of the car the price will also increase.

Price is directly related to Met_Color:

This means that addition of a Met_Color will result in increase of the car price as well.

Price is directly related to Automatic:

This means that if the car has a feature of automatic in it the price of that car will be more as compared to one without automatic.

Price is directly related to CC:

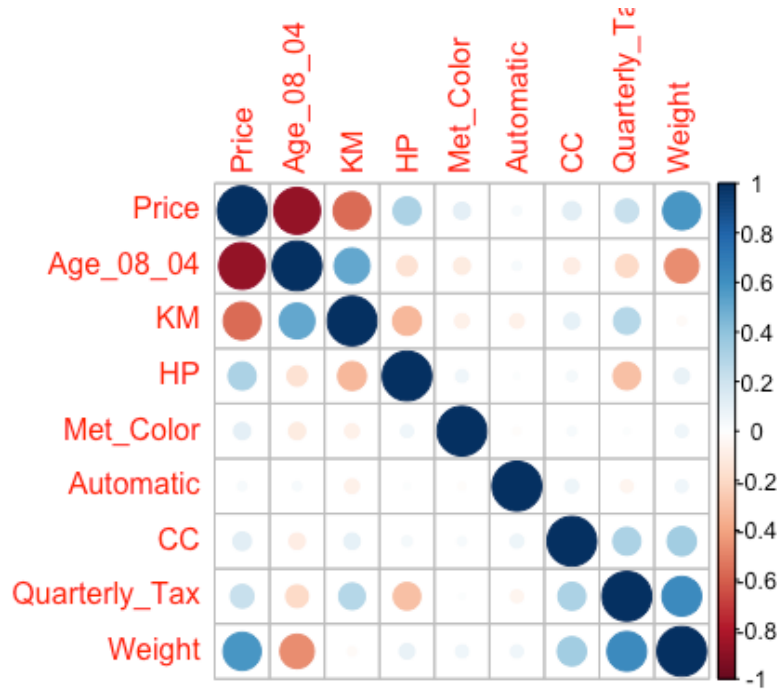
This means that more the CC value of the engine more will the price be.

Price is directly related to Quaterly_Tax:

This means that with the increase in Quarterly_Tax the price of the car will increase.

Price is directly related to Weight:

This means that with the increase in weight of the car price of the car will also increase.



TO PLOT THE CORRELATION PLOT WE USED THE VARIABLES BELOW:

🔍 saurav chatterji assignment hw fin... * x
📄 res1 x
📄 data3.new x
📄 data2.new x

🏠
📁
🔍 Filter
🔍

	Price	Age_08_04	KM	HP	Met_Color	Automatic	CC	Quarterly_Tax	Weight
1	13500	23	46986	90	1	0	2000	210	1165
2	13750	23	72937	90	1	0	2000	210	1165
3	13950	24	41711	90	1	0	2000	210	1165
4	14950	26	48000	90	0	0	2000	210	1165
5	13750	30	38500	90	0	0	2000	210	1170
6	12950	32	61000	90	0	0	2000	210	1170
7	16900	27	94612	90	1	0	2000	210	1245
8	18600	30	75889	90	1	0	2000	210	1245
9	21500	27	19700	192	0	0	1800	100	1185
10	12950	23	71138	69	0	0	1900	185	1105
11	20950	25	31461	192	0	0	1800	100	1185
12	19950	22	43610	192	0	0	1800	100	1185
13	19600	25	32189	192	0	0	1800	100	1185
14	21500	31	23000	192	1	0	1800	100	1185
15	22500	32	34131	192	1	0	1800	100	1185

Showing 1 to 15 of 1,436 entries