Mandeep Singh Narang | mxn170019@utdallas.edu

1. Categorial variables in "ToyotaCorolla.xlsx" data file are
   "Model","Mfg_Month", "Fuel_Type", "Cylinder","Colour", "Met_Color", "Automatic", "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_Rim", "Radio_cassette", "Parking_Assistant", "Tow_Bar"

   Note:
   "Cylinder" has only One value so that's why we can take them as categorical variables or we can drop it.
   "Model" here is a Entity but we can

   "Gears" is a numeric variable but it has very few values and by plotting it vs Price we can see that it does not has any direct relation with price we can take it as categorical variable. But It is an attribute of cars and at some point it is affecting the price. And for this analysis we do not need the categories for gears. So I am taking it numerical

   "Mfg_Month" is a date variable and here it is providing some information about "Age" of car .
   So I It can also be taken in Numeric but in the data set we have "Year" column the combination of both Month and Year will provide the relation about Age.
   Month is not independent it ill only have 12 values but Year is Free, That's why I am taking "Mfg_Months" as Categorial but "Mfg_Year" as numeric.

   Binary Variable are a type of categorical variable.

   For "CC" , "HP" we can observe that they have small set of values but they are not categories for this Analysis, Because their values have relation with price.

2. The Relationship Between a categorical variable and its binary dummy variables is :
   When we need to incorporate a categorical variable in a predictive analysis or data minings methods we need to convert the categorical variables into some numeric form because algorithms only take numeric value as intake and to solve this problem we make Dummy binary variables. There will be one specific dummy variable for each different category and the value of dummy variable will be 1 for that specific category(1 for TRUE CASES) and in other cases it will be zero.

3. In General a variable with N categories will be transformed into N or N-1 dummy variables. In the Dummy variables 1 will be assigned to only True cases of that specific category and in all other cases we assign 0. Generally while taking N for N categories we face multicollinearity error, So we use N-1 for Analysis . For e.g.

ID	MODEL	X-Types
1	ABC 1	X1
2	ABC 2	X2
3	ABC 3	X3

   In the Above table "X-Types" is a categorical variable with three categories. To make dummy variables we need to make 3 OR 3-1 new binary column one for each true case.

N Dummy variable for N categories: In this case there will be same number of variables as of categories in the categorical variable. Every category will have one specific column for its TRUE cases. (assuming 1 as TRUE and 0 as FALSE, see below table)

ID	MODEL	X1	X2	X3
1	ABC 1	1	0	0
2	ABC 2	0	1	0
3	ABC 3	0	0	1

N-1 Dummy variables for N categories: In this case there will be one less dummy column as of number of categories. 1 will be assigned to TRUE case of that perticular category and all zeros will be showing the missing category's TRUE case. (see red highlight for X1 TRUE cases)

ID	MODEL	X2	X3
1	ABC 1	0	0
2	ABC 2	1	0
3	ABC 3	0	1

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:

Code to convert categorical variable to Dummy binary variable :

```
>install.packages("fastDummies")
>library(fastDummies)
>ANSWER4<- dummy_cols(Toyota_data,select_columns = c("Mfg_Month", "Fuel_Type","Color")
remove_first_dummy = FALSE)
> names(ANSWER4)

 [1] "Id"              "Model"           "Price"
 [4] "Age_08_04"       "Mfg_Month"       "Mfg_Year"
 [7] "KM"              "Fuel_Type"       "HP"
[10] "Met_Color"       "Color"           "Automatic"
[13] "CC"              "Doors"           "Cylinders"
[16] "Gears"           "Quarterly_Tax"   "Weight"
[19] "Mfr_Guarantee"   "BOVAG_Guarantee" "Guarantee_Period"
[22] "ABS"             "Airbag_1"        "Airbag_2"
[25] "Airco"           "Automatic_airco" "Boardcomputer"
[28] "CD_Player"       "Central_Lock"    "Powered_Windows"
[31] "Power_Steering"  "Radio"           "Mistlamps"
[34] "Sport_Model"     "Backseat_Divider" "Metallic_Rim"
[37] "Radio_cassette"  "Parking_Assistant" "Tow_Bar"
[40] "Mfg_Month_10"    "Mfg_Month_9"     "Mfg_Month_7"
[43] "Mfg_Month_3"     "Mfg_Month_1"     "Mfg_Month_6"
[46] "Mfg_Month_8"     "Mfg_Month_11"    "Mfg_Month_2"
[49] "Mfg_Month_5"     "Mfg_Month_4"     "Mfg_Month_12"
[52]
"Fuel_Type_Diesel" "Fuel_Type_Petrol"
[61] "Fuel_Type_CNG"   "Color_Blue"      "Color_Silver"
[64] "Color_Black"     "Color_White"     "Color_Grey"
[67] "Color_Red"       "Color_Green"     "Color_Yellow"
[70] "Color_Violet"    "Color_Beige"
```

This is the code for creating the dummy binary variables and we are using - "remove_first_dummy = FALSE" to create N number of Dummy variables for N categories in one categorical variable. Red Highlighted column names are dummy variables, These are Sample of total Dummy Variable you can find total data in R file.

Please see data example below.

Sampledummy

Filter

	Id	Model	Color	Color_Blue	Color_Silver	Color_Black	Color_White	Color_Grey	Color_Red	Color_Green	Color_Yellow	Color_Violet	Color_Beige
1	1	TOYOTA Corolla 2.0 D4...	Blue	1	0	0	0	0	0	0	0	0	0
2	2	TOYOTA Corolla 2.0 D4...	Silver	0	1	0	0	0	0	0	0	0	0
3	3	TOYOTA Corolla 2.0 D4...	Blue	1	0	0	0	0	0	0	0	0	0
4	4	TOYOTA Corolla 2.0 D4...	Black	0	0	1	0	0	0	0	0	0	0
5	5	TOYOTA Corolla 2.0 D4...	Black	0	0	1	0	0	0	0	0	0	0
6	6	TOYOTA Corolla 2.0 D4...	White	0	0	0	1	0	0	0	0	0	0
7	7	TOYOTA Corolla 2.0 D4...	Grey	0	0	0	0	1	0	0	0	0	0
8	8	TOYOTA Corolla 2.0 D4...	Grey	0	0	0	0	1	0	0	0	0	0
9	9	TOYOTA Corolla 1800 ...	Red	0	0	0	0	0	1	0	0	0	0
10	10	TOYOTA Corolla 1.9 D ...	Blue	1	0	0	0	0	0	0	0	0	0
11	11	TOYOTA Corolla 1.8 VV...	Silver	0	1	0	0	0	0	0	0	0	0

Here in the first row we can see that the colour of the car is Blue and we have several Dummy columns for colour category and we only have 1 in the column named "Color_Black" and in the same column we can see in row 3 and 10 we have Blue color and 1 in the Color_Blue column all other entries are zero. Same for the other columns.

5. Correlation matrix:
In this correlation matrix I have excluded all the categorical and binary variables just to show is clear in this document. In the attached R code file and data files you can find total correlation matrix among all the binary, dummy binary and numeric variables.

:1_R_ShortIntro.R | Session2_DataViz.R | R_Code_DimensionReduction.R | PerfEval.R | Assignement 1 BA with R.R* | cor_mat

Filter

	Price	Age_08_04	KM	Gears	HP	CC	Mfg_Year	Quarterly_Tax	Weight	Guarantee_Period
Price	1.0000	-0.8766	-0.5700	0.0631	0.3150	0.1264	0.8852	0.2192	0.5812	0.1466
Age_08_04	-0.8766	1.0000	0.5057	-0.0054	-0.1566	-0.0981	-0.9837	-0.1984	-0.4703	-0.1526
KM	-0.5700	0.5057	1.0000	0.0150	-0.3335	0.1027	-0.5050	0.2782	-0.0286	-0.1389
Gears	0.0631	-0.0054	0.0150	1.0000	0.2095	0.0146	0.0078	-0.0055	0.0206	-0.0307
HP	0.3150	-0.1566	-0.3335	0.2095	1.0000	0.0359	0.1647	-0.2984	0.0896	0.0762
CC	0.1264	-0.0981	0.1027	0.0146	0.0359	1.0000	0.0919	0.3070	0.3356	-0.0177
Mfg_Year	0.8852	-0.9837	-0.5050	0.0078	0.1647	0.0919	1.0000	0.1939	0.4735	0.1482
Quarterly_Tax	0.2192	-0.1984	0.2782	-0.0055	-0.2984	0.3070	0.1939	1.0000	0.6261	-0.1634
Weight	0.5812	-0.4703	-0.0	~/Downloads/Week1_R_ShortIntro.R			0.4735	0.6261	1.0000	-0.0129
Guarantee_Period	0.1466	-0.1526	-0.1389	-0.0307	0.0762	-0.0177	0.1482	-0.1634	-0.0129	1.0000

- Price here is a dependent variable. And all others are independent variables.
- We can see from this "Price" and "Age_08_04" have very strong negative relation means if the age increases price decreases.
- "Age_08_04" and "KM" have a positive relation but the relation Is not that strong.
- "Mfg_Year" has a positive relation with price, it means newer is the car more wil be the price.

- There is a positive relation between “Quarterly_Tax” and “weight”. Weight also has positive relation with “Price” .

  Correlation heat map: we can see the darker is the colour more positive is the relationship.