**Dummy Variables**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Navya.P**

**Batch Id: WDEO 171220**

**Topic: Preliminaries for Data Analysis**

**Problem Statement:**

Data is one of the most important assets. It is often common that data is stored in distinct systems with different formats and forms. Non-numeric form of data makes it tricky while developing mathematical equations for prediction models. We have the preprocessing techniques to make the data convert to numeric form. Explore the various techniques to have reliable uniform standard data, you can go through this link:

<https://360digitmg.com/mindmap-data-science>

1. Prepare the dataset by performing the preprocessing techniques, to have the all the features in numeric format.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Animals | Gender | Homly | Types |
| 1 | Cat | Male | Yes | A |
| 2 | Dog | Male | Yes | B |
| 3 | Mouse | Male | Yes | C |
| 4 | Mouse | Male | Yes | C |
| 5 | Dog | Female | Yes | A |
| 6 | Cat | Female | Yes | B |
| 7 | Lion | Female | Yes | D |
| 8 | Goat | Female | Yes | E |
| 9 | Cat | Female | Yes | A |
| 10 | Dog | Male | Yes | B |

The data animal Category is analyzed to create the dummy variables from the original data. We have 3 types of dummy variable creations.

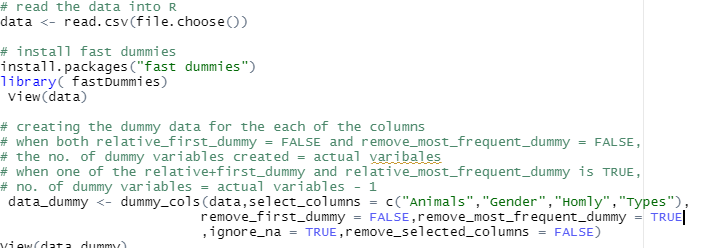
* + - 1. One hot encoding : makes use of only 0 or 1
      2. Label Encoding : Assigns alphabetical order to the data

One hot Encoding:

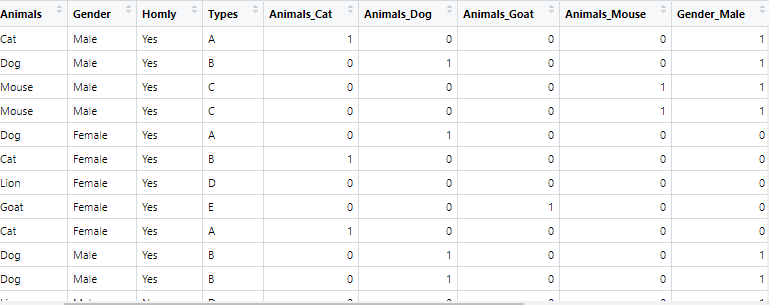
The package ‘ fast dummies’ is used to create dummy variables.

From the function dummy\_cols, when one of the columns in relative\_first\_dummy and relative\_most\_frequent\_dummy is TRUE, no. of dummy variables = actual variables - 1

The code is as follows.

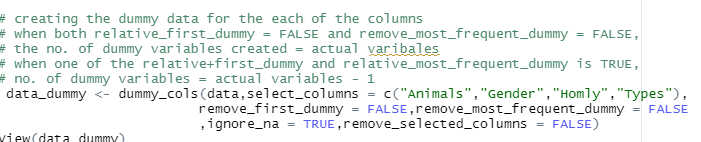


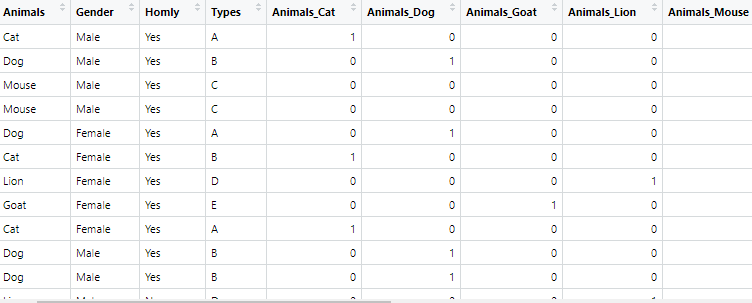
The Output is



Here, when both relative\_first\_dummy = FALSE and remove\_most\_frequent\_dummy = FALSE,

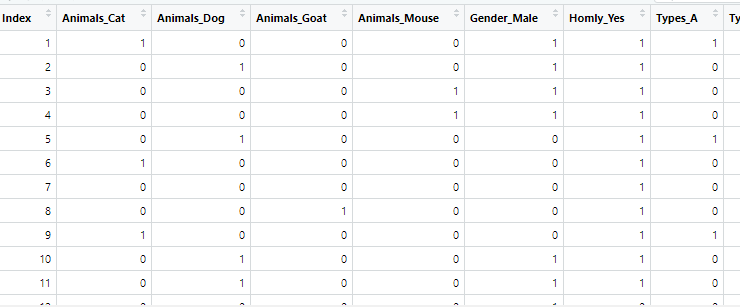
the no. of dummy variables created = actual variables.





When we set remove\_Selected\_comuns = TRUE, the output is as follows:

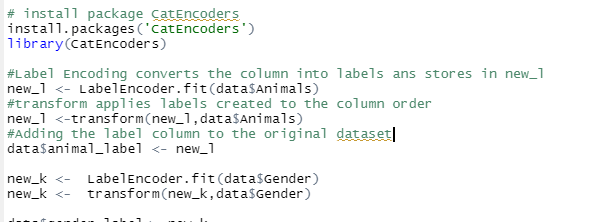




Label Encoding:

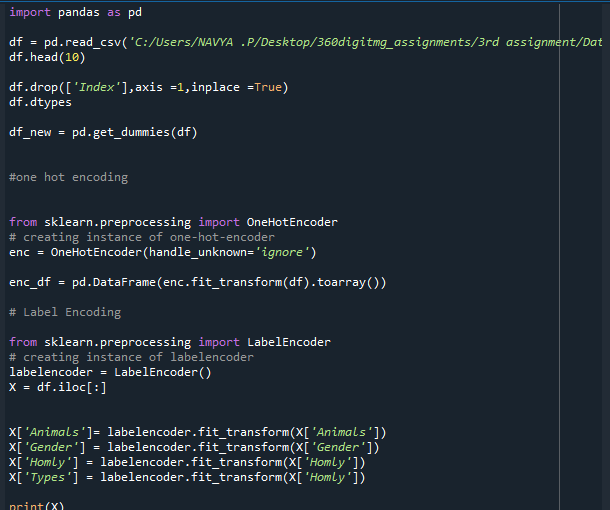
In Label Encoding we name the data based on the order. The Label encoding is done in R using a package CatEncoders.

The code is as follows.

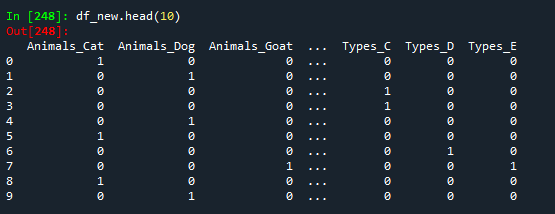


The output is

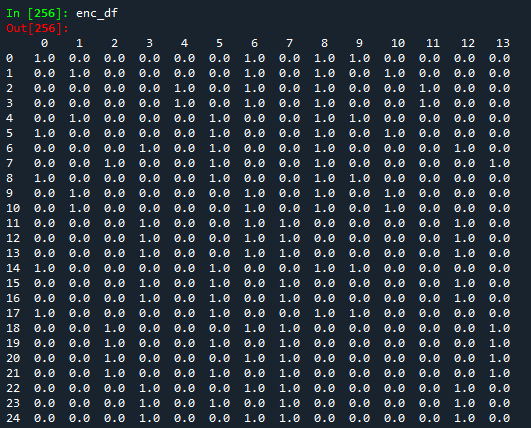




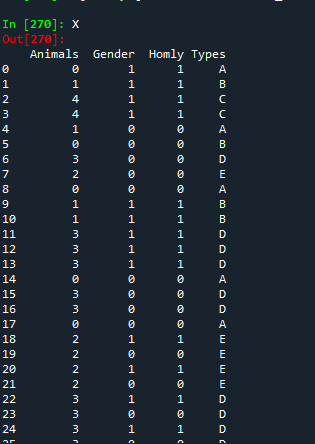
Creating dummy variables



One Hot Encoding



Label Encoding



**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Refer to animal\_category.csv data set.
3. Research and perform all possible steps for obtaining solution
4. All the codes (executable programs) should execute without errors
5. Code modularization should be followed
6. Each line of code should have comments explaining the logic and why you are using that function