Hi there, being studying the data Science form last 3 months, I have understood that the most of the time the data we get is in the form of qualitative data, and the very man-made intelligent thing which is called as “***COMPUTER***” only understands the ‘**Low level language’**. That is 0 and 1. Which is the sort of Quantitative data.

Now, as much as I have understood about pandas library, it is the “**RAM BAAN**” or one can say as ‘**Lamp in the darkest hour of Night’** to every critical and annoying dataset. Pandas has a function which can turn a categorical variable into a series of zeros and ones, which makes them a lot easier to quantify and compare.

Another nightmare in the Data Science industry is “***DATA CLEANING***”. At least 60 to 70 percent of the whole project time is spent on data cleaning in data science industries.

***So, the library “PANDAS”, comes with the solution. I can even say that everyone could conquer the world (or at least the data science world)*** 😊

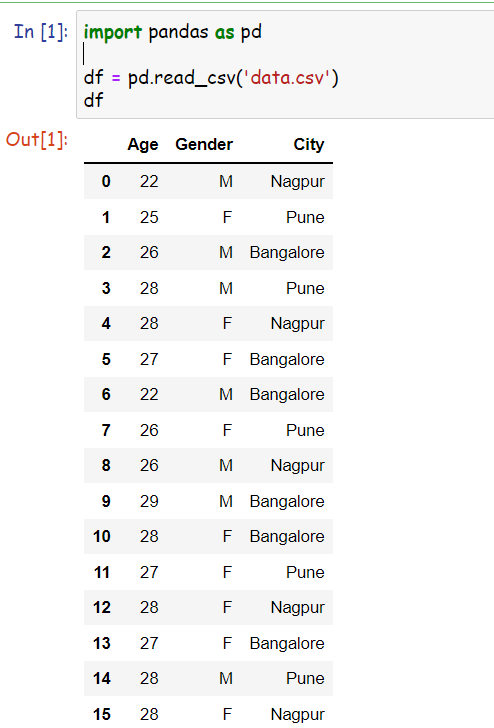
Among the most data cleaning methods, let me explain you with an example about the Dummy Variable and Dummy Variable trap method, most suitable for Nominal and Ordinal datasets.

Dummy Variable (as the name suggest) is the nothing but creating the dummy columns, by converting the categorical features of that column in to ‘***ZERO***’, and ‘***ONE***’ binary digit.

I started with importing the dataset which I prepared. It only contains the three columns age, gender and city of 21 people, (just a very small random data set).

In this below Image as you can see the glimpse of the dataset, where there are two columns ‘Gender’ and ‘City’, which are the categorical data.

Now, converting these data into numeric is our task, so that computer can understand the information.



To convert categorical features (Gender and City), we will create dummy variables using pandas libraries.



Above is the syntax for the dummy variable creation, where ***pd.get\_dummies*** will help in creating the 0 and 1 digits.

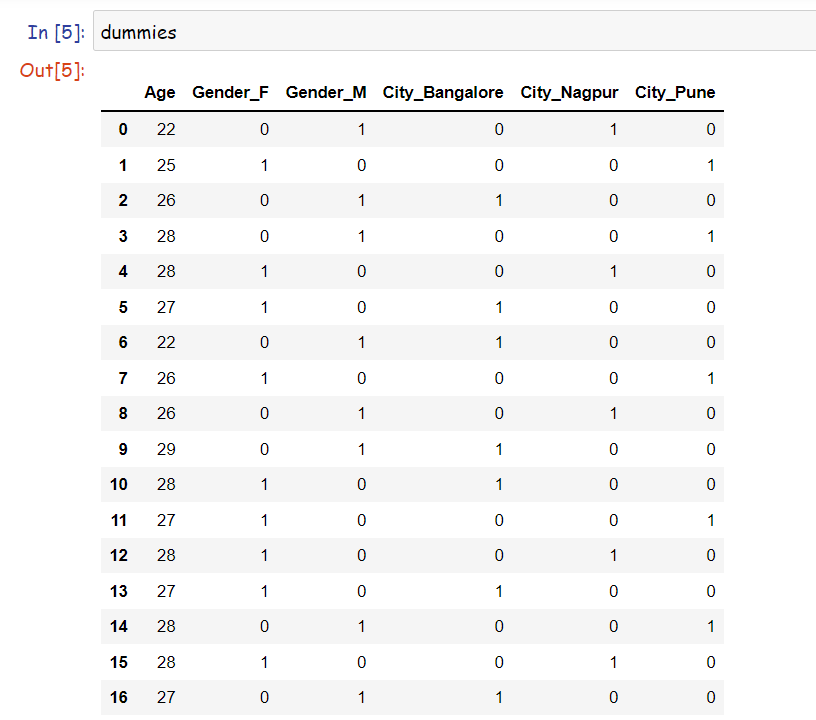
As, from the following image, you can see that the new columns has been automatically created by splitting the unique values in the column as the column variable name.

And now as we see in the original data set, the first person who was male has the value ‘***ONE***’ in Gender\_M column and he was from ‘***NAGPUR***’ so Nagpur column is also having ‘***ONE***’ in that row rest all the values are ‘***ZERO’***.

This signifies that importance of dummy variable creation.

Here, ‘***ONE***’ signifies the value and ‘***ZERO***’ signifies the null.

So now we can easily understand and have successfully converted the categorical data into numeric form, which computer can easily understand.



But, due to this, one more problem has been occurred, we have created three extra columns in our dataset.

Now, in the real-life condition, there can be dataset with hundreds of columns and top of that creating more columns will consume more resources and more time.

So, there is one more technique call as ***‘DUMMY TRAP’.***

To avoid Dummy Variable trap, we usually remove one column from these dummy variables. Because, when we create dummy variables, new entries will get added.

Suppose there are 3 categories in one feature column, so 3 different entries of dummy variables will get created.

Dummy variables trap is usually occurred while encountering Multilinear Regression Models which causes Multicollinearity. To avoid these issues, we simply drop any column from these dummies.

By dropping any one column the interpretation of categorical variables still remains the same.



In the above syntax, by just including the drop first entity, all the first columns of the dummy variable will be dropped and still we are able to interpret the results from that data.

As, in the following image we can see that the ***‘Gender\_F’*** variable and ***‘City\_Bangalore’*** is be dropped so for the first person we can understand that the person is male and belongs to Nagpur city. Whereas at the second index, the person is female, so the male column in zero, hence we can interpret that the person is female.

Hence, this is one of that ***‘Dummy Variable and Dummy Trap’*** affective method of data cleaning

***When to use Dummy Variable Trap Method:***

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups. In the simplest case, we would use a 0, 1 dummy variable where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don’t need to write out separate equation models for each subgroup. The dummy variables act like **‘switches’** that turn various parameters on and off in an equation. Another advantage of a 0,1 dummy-coded variable is that even though it is a nominal-level variable you can treat it statistically like an interval-level variable