

Dealing with Categorical Columns

Categorical data

- The categorical data consists of categorical variables which represent the characteristics such as a person's gender, hometown etc. Categorical measurements are expressed in terms of natural language descriptions, but not in terms of numbers. Sometimes categorical data can take numerical values, but those numbers do not have mathematical meaning.
- Categorical variables can take on only a limited, and usually fixed number of possible values. Besides the fixed length, categorical data might have an order but cannot perform numerical operation. Categorical are a Pandas data type.

Types of Categorical Data

In general, categorical data has values and observations which can be sorted into categories or groups. The best way to represent these data is bar graphs and pie charts. Categorical data are further classified into two types namely,

- **Nominal Data**-Nominal data is a type of data that is used to label the variables without providing any numerical value. Some of the few common examples of nominal data are letters, words, symbols, gender etc. These data are analyzed with the help of the grouping method.
- **Ordinal Data**-Ordinal data is a type of data that follows a natural order. The notable features of ordinal data are that the difference between data values cannot be determined. It is commonly encountered in surveys, questionnaires, finance and economics.

Categorical Variables

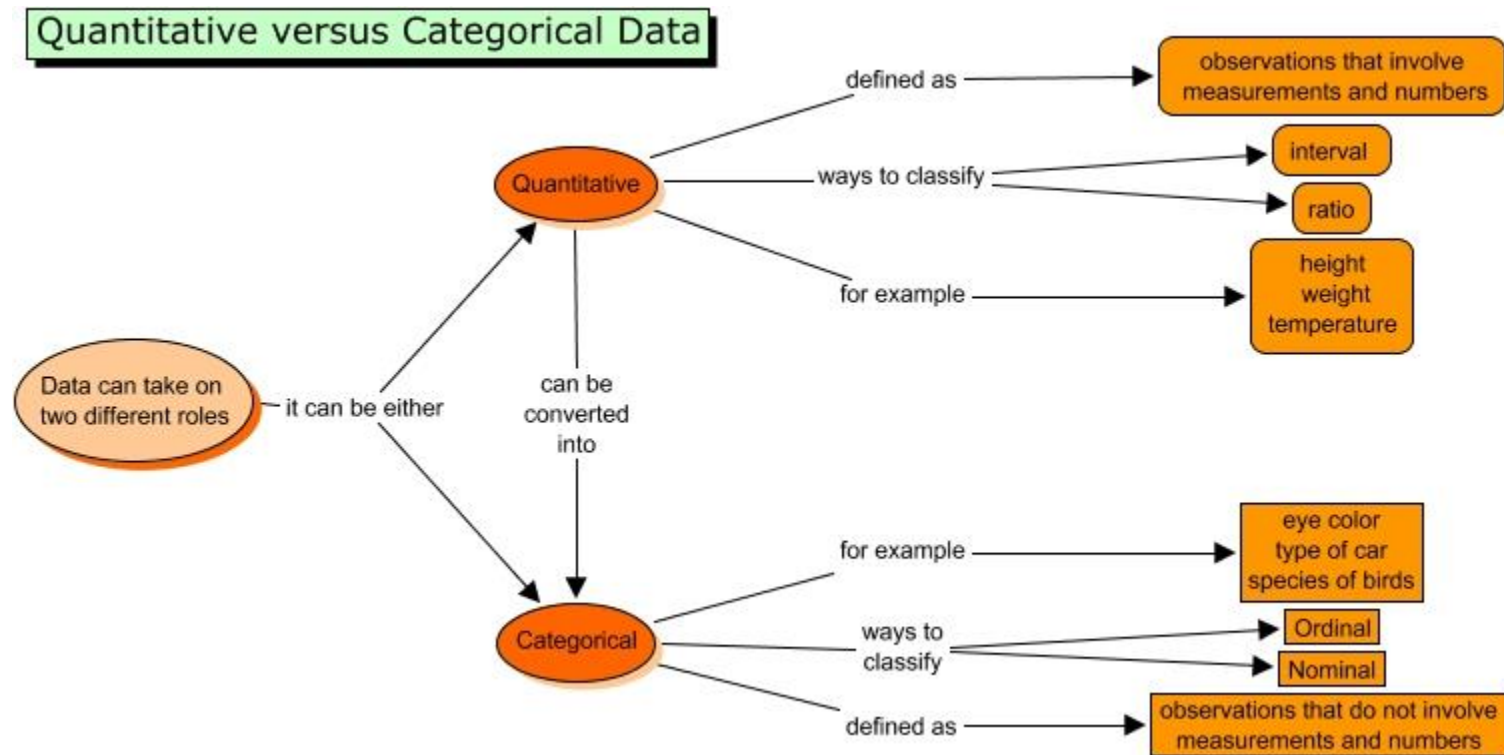
In statistics, a categorical variable is a variable that contains limited, and usually a fixed number of possible values. They take values which are normally names or labels. Examples are:

- The color of a wall, like red, blue, pink, grey, etc.,
- Gender of people, like male, female and transgender
- Blood group of a person: A, B, O, AB, etc.,

These variables are used to assign each individual or another unit of observation to a particular group or nominal category based on some qualitative property. Generally, each of the potential values of a categorical variable is said to be as a level. The probability distribution linked with a random categorical variable is known as categorical distribution.

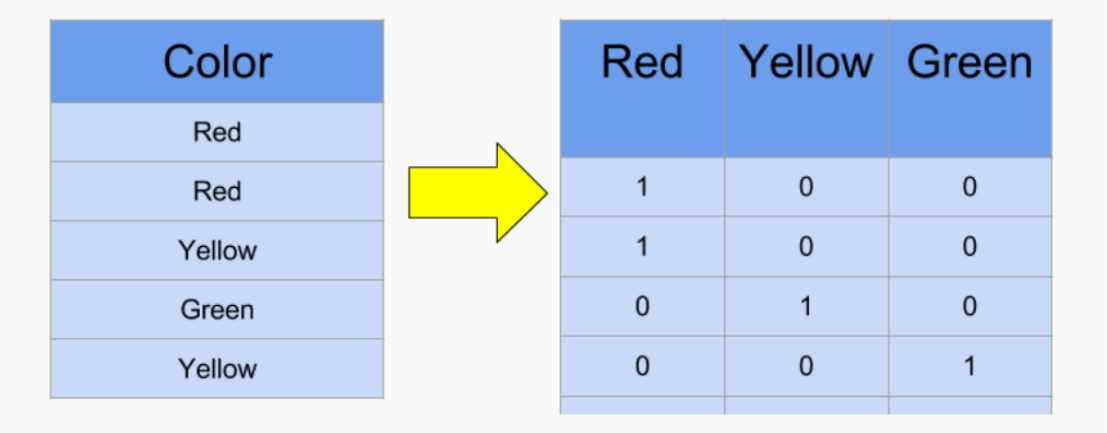
Why we need encoding?

- Most machine learning algorithms cannot handle categorical variables unless we convert them to numerical values
- Many algorithm's performances even vary based upon how the categorical variables are encoded

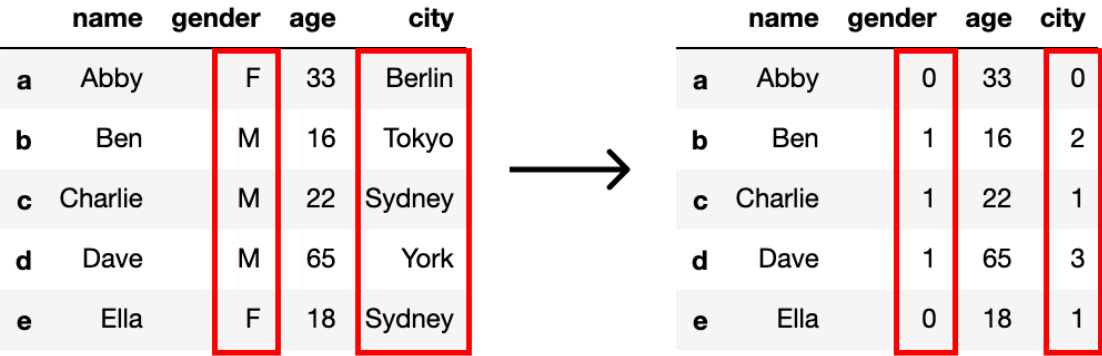


Techniques to deal with categorical data

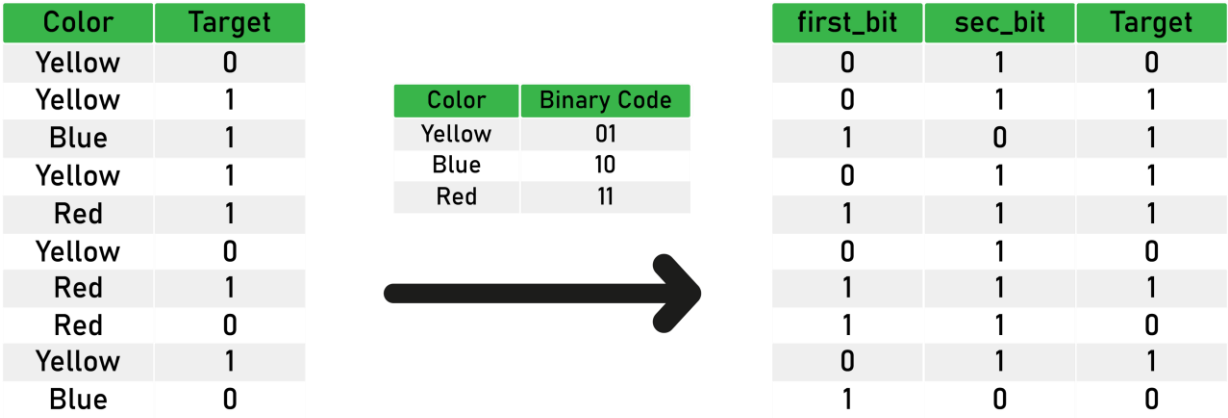
1. **One hot encoding:** In this method, each category is mapped to a vector that contains 1 and 0 denoting the presence or absence of the feature. The number of vectors depends on the number of categories for features.
2. **Label encoding:** In label encoding, each category is assigned a value from 1 through N where N is the number of categories for the feature. There is no relation or order between these assignments.
3. **Binary encoding:** Binary encoding converts a category into binary digits. Each binary digit creates one feature column.



One hot encoding




Label encoding



Binary encoding

4. **Ordinal encoding:** Ordinal encoding's encoded variables retain the ordinal (ordered) nature of the variable. It looks similar to label encoding, the only difference being that label coding doesn't consider whether a variable is ordinal or not; it will then assign a sequence of integers. Example: Ordinal encoding will assign values as Very Good(1) < Good(2) < Bad(3) < Worse(4)



Color	Color
Green	1
Red	2
Blue	3

5. **Dropping categorical columns:** The easiest approach to dealing with categorical variables is to simply remove them from the dataset. This approach will only work well if the columns did not contain useful information.

Resources and References

- [Machine Learning with Categorical Data | Pluralsight](#)
- [How to Deal with Categorical Data for Machine Learning – Kdnuggets](#)
- [Categorical Data & Qualitative Data \(Definition and Types\) \(byjus.com\)](#)
- [How to Handle Categorical Values? | by Aryan Chaudhary | Big Data Center of Excellence | Medium](#)
- [https://miro.medium.com/max/1552/1*kPKvp4c462GB1NS8UotXYQ.png](#)
- [https://th.bing.com/th/id/OIP.gLdye1KcEpEdjs7pFXBd0gHaDR?w=310&h=154&c=7&r=0&o=5&dpr=1.5&pid=1.7](#)
- [https://th.bing.com/th/id/OIP.iOToARpeXMMvVN6pgyYhrAHaCm?w=303&h=122&c=7&r=0&o=5&dpr=1.5&pid=1.7](#)
- [https://th.bing.com/th/id/OIP.6R4l16PtM7WXuAiK4hF6sQHaC5?w=327&h=136&c=7&r=0&o=5&dpr=1.5&pid=1.7](#)
- [https://th.bing.com/th/id/OIP.BfGjPxZexlO3hKlqgti-pQHaDt?w=288&h=175&c=7&r=0&o=5&dpr=1.5&pid=1.7](#)