XCCELERATE

DATA COLLECTION AND EXPLORATORY DATA ANALYSIS

LEARNING OBJECTIVES

Collecting Data

Exploratory Data Analysis – Variable Identification, Univariate, Bivariate Analysis

Data Cleaning – Missing Values, Outliers

DATA COLLECTION

Data Collection

Customer Records

Server Logs

Surveys/Interviews

Crowdsourcing

Publicly available datasets

Web APIs

Web Scraping

In this course, we will be using publicly available datasets

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis

Exploratory Data Analysis (EDA) deals with the process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Need for Exploratory Data Analysis

To *listen* to the data:

-to catch mistakes

-to see patterns in the data

-to find violations of statistical assumptions

…and because if you don’t, you will have trouble later

What you are trying to do in EDA

1. What question(s) are you trying to solve (or prove wrong)?

2. What kind of data do you have and how do you treat different types?

3. What's missing from the data and how do you deal with it?

4. Where are the outliers and why should you care about them?

5. How can you add, change or remove features to get more out of your data?

So… What should you do?

- Examine all your variables thoroughly and carefully before you begin analysis

- Use visual displays whenever possible

- Transform each variable as necessary to deal with mistakes, outliers, and distributions
  - If you find a mistake, fix it.
  - If you find an outlier, trim it or delete it (or Keep it if it is important)
  - If your distributions are askew, transform the data.

EDA using Pandas/Seaborn

For EDA with pandas, we will cover the following:

- Variable Identification
- Univariate Analysis
- Bivariate Analysis

Variable Identification

Identify the type of variables – ie its data type. Also Identify whether the data is:

Qualitative - data are measurements that each fall into one of several categories. (hair color, ethnic groups and other attributes of the population)

Quantitative - data are observations that are measured on a numerical scale (distance travelled to college, number of children in a family, etc.)

Quantitative data can be separated into two subgroups:

> **Discrete** (if it is the result of *counting* (the number of students of a given ethnic group in a class, the number of books on a shelf,…)
>
> **Continuous** (if it is the result of *measuring* (distance travelled, weight of luggage…)

Identify the target and predictor variables.

Univariate Analysis

Univariate Analysis deals with exploring variables individually. This also depends on the type of variable:

Continuous Variables :

Find the central tendency and spread of the data for the variable. Visualization techniques such as histograms and Box Plots would work well for this.

Categorical Variables :

Frequency tables can be used to understand distribution of each variable. Visualization techniques such as Bar charts/Pie Charts would work well for this.

Bivariate Analysis

Bivariate Analysis deals with exploring relationships between 2 variables. Different techniques are Used depending on the types of variables being explored:

Both Continuous Variables :

Visualization techniques such as scatterplots would work well for this.

Both Categorical Variables :

Visualization techniques such as Two-Way Tables or Stacked Column Charts.

Categorical and Continuous Variable:

Visualization techniques such as Box Plots for each level of categorical variables.

Missing Values

Missing Values

Identify Missing Values

You can use the Pandas Dataframes following functions to achieve these:

- pandas.Dataframe.isna

Missing Values

To take care of missing values, we can

- Remove rows with missing values
- Remove the entire column with missing values
- Set the values to a meaningful value (zero,mean,median)

You can use the Pandas Dataframes following functions to achieve these:

- pandas.Dataframe.dropna
- pandas.Dataframe.drop
- pandas.Dataframe.fillna
- Imputation of missing values

FEATURE SCALING

Feature Scaling

Standardise
Standard Gaussian distribution with a mean of 0 and a standard deviation of 1

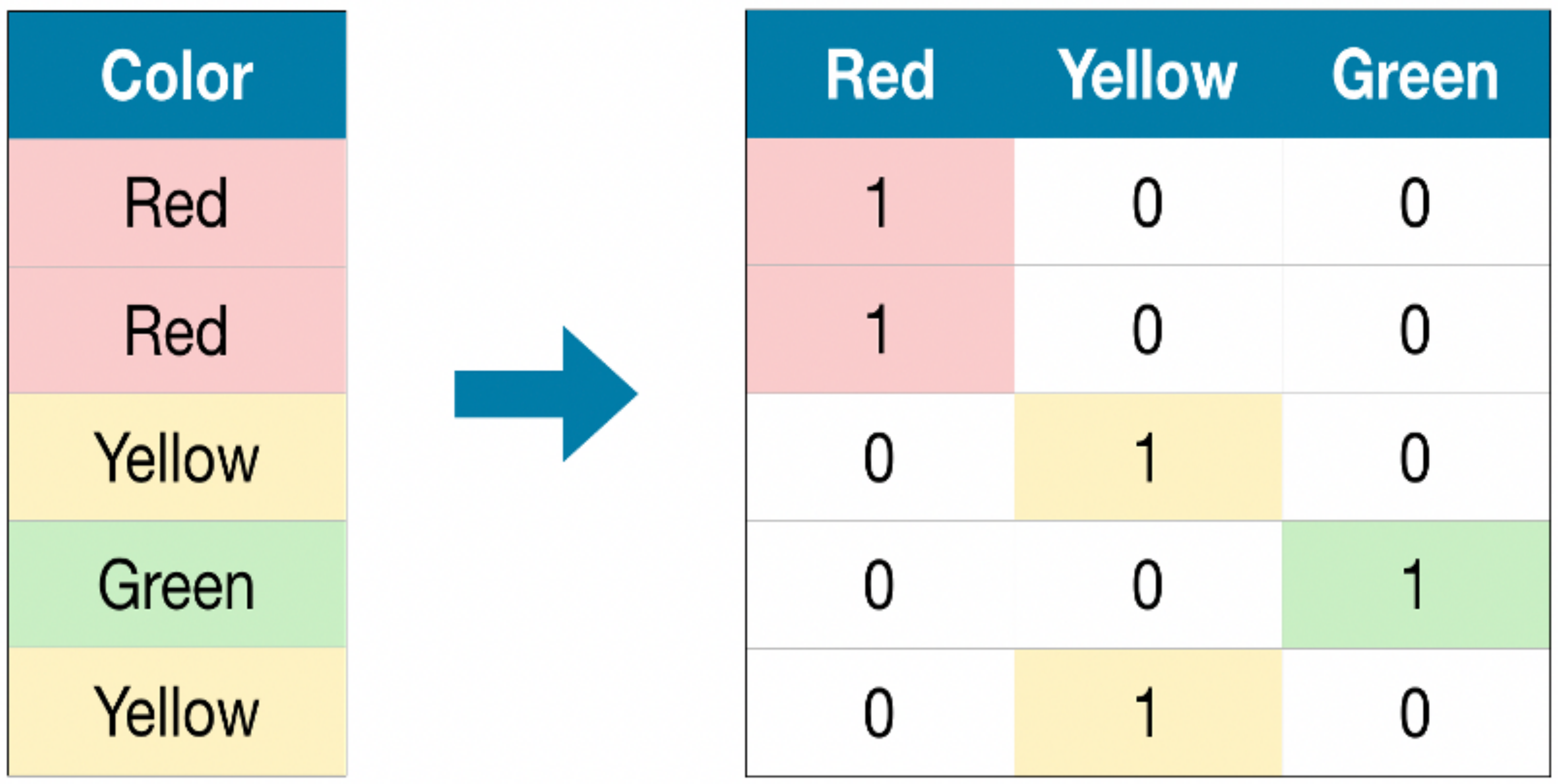
Modules: StandardScalar (from scikit learn)

Pre-processing libraries or API
Fit – prepare the params to transform once Transform – prepare for modelling

HANDLING TEXT and CATEGORICAL VALUES

Handling Text and Categorical Data

Pandas get_dummies



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

SUMMARY