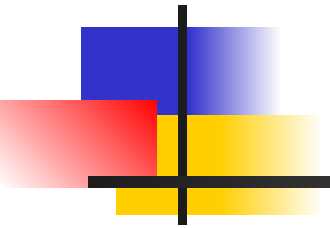
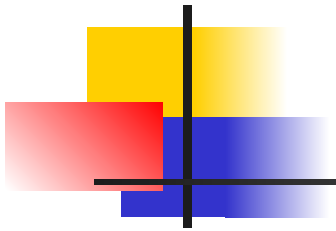


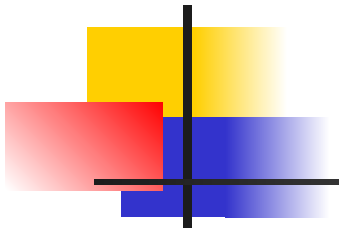
Data Wrangling



- Identify and handle missing values
- - Identify missing values
- - Deal with missing values
- - Correct data format
- Data standardization
- Data Normalization (centring/scaling)
- Binning
- Indicator variable

Data Wrangling

- Data Wrangling is the process of converting data from the initial format to a format that may be better for analysis.
- The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.



Steps for working with missing data:

- identify missing data
- deal with missing data
- correct data format

Steps for working with missing data:

- **1. Identify and handle missing values**

Convert "?" to NaN

Use the function: .replace(A, B, inplace = True)

- **Evaluating for Missing Data**

There are two methods to detect missing data:

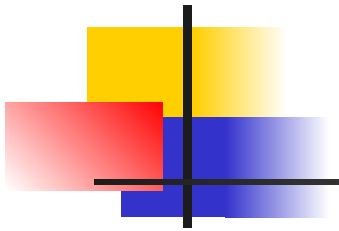
- **.isnull()**
- **.notnull()**
- The output is a boolean value indicating whether the passed in argument value are in fact missing data.

Deal with missing data

- How to deal with missing data?

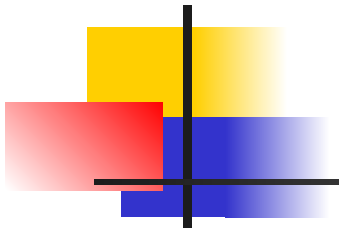
- 1. drop data
- a. drop the whole row
- b. drop the whole column
- 2. replace data
- a. replace it by mean
- b. replace it by frequency
- c. replace it based on other functions

Correct data format



- Making sure that all data is in the correct format (int, float, text or other).
- In Pandas, we use
- **.dtype()** to check the data type
- **.astype()** to change the data type

Data Standardization

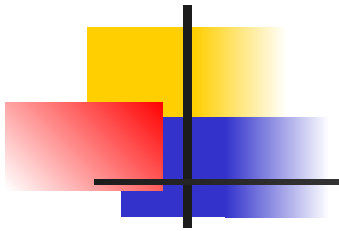


- Data is usually collected from different agencies with different formats. (Data Standardization is also a term for a particular type of data normalization, where we subtract the mean and divide by the standard deviation)
- **What is Standardization?**
- Standardization is the process of transforming data into a common format which allows the researcher to make the meaningful comparison.

Data Normalization

- **Why normalization?**
- Normalization is the process of transforming values of several variables into a similar range. Typical normalizations include scaling the variable so the variable average is 0, scaling the variable so the variable variance is 1, or scaling variable so the variable values range from 0 to 1

Standardisation (Z-score Normalization)	Max-Min Normalization
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

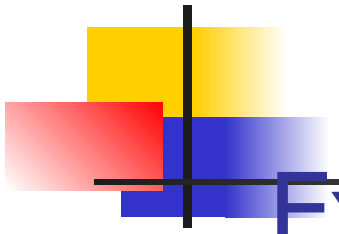


Binning

- **Why binning?**
- Binning is a process of transforming continuous numerical variables into discrete categorical 'bins', for grouped analysis.
- Normally, a histogram is used to visualize the distribution of bins created

Indicator variable (or dummy variable)

- **What is an indicator variable?**
- An indicator variable (or dummy variable) is a numerical variable used to label categories. They are called 'dummies' because the numbers themselves don't have inherent meaning.
- **Why we use indicator variables?**
- So we can use categorical variables for regression analysis

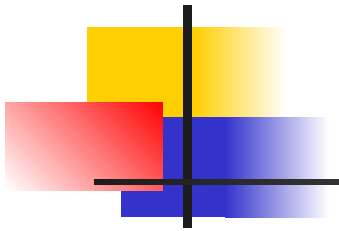


Exploratory Data Analysis (EDA)

- For data analysis, Exploratory Data Analysis (EDA) must be your first step. Exploratory Data Analysis helps us to –
- To give insight into a data set.
- Understand the underlying structure.
- Extract important parameters and relationships that hold between them.
- Test underlying assumptions.

Classification of EDA

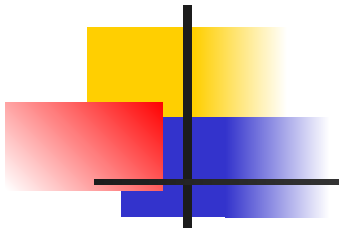
- Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).
- Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way.
- Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables. It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.



Data Frames attributes

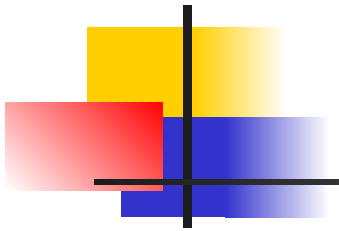
Python objects have *attributes* and *methods*.

df.attribute	description
dtypes	list the types of the columns
columns	list the column names
axes	list the row labels and column names
ndim	number of dimensions
size	number of elements
shape	return a tuple representing the dimensionality
values	numpy representation of the data



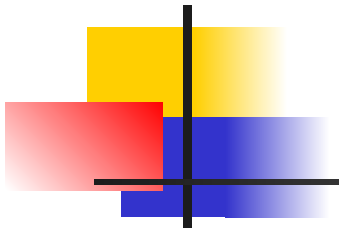
Data Frames methods

df.method()	description
head([n]), tail([n])	first/last n rows
describe()	generate descriptive statistics (for numeric columns only)
max(), min()	return max/min values for all numeric columns
mean(), median()	return mean/median values for all numeric columns
std()	standard deviation
sample([n])	returns a random sample of the data frame
dropna()	drop all the records with missing values



Basic Descriptive Statistics

df.method()	description
describe	Basic statistics (count, mean, std, min, quantiles, max)
min, max	Minimum and maximum values
mean, median, mode	Arithmetic average, median and mode
var, std	Variance and standard deviation
sem	Standard error of mean
skew	Sample skewness
kurt	kurtosis



Grouping

- The "groupby" method groups data by different categories. The data is grouped based on one or several variables and analysis is performed on the individual groups.

Analysis

- Univariate Analysis:If we analyze data over a single variable/column from a dataset, it is known as Univariate Analysis. Categorical Unordered Univariate Analysis and Categorical Unordered Univariate Analysis
- Bivariate Analysis:If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.
- Multivariate Analysis:If we analyze data by taking more than two variables/columns into consideration from a dataset, it is known as Multivariate Analysis.