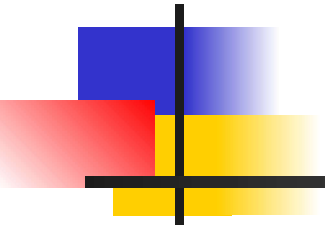
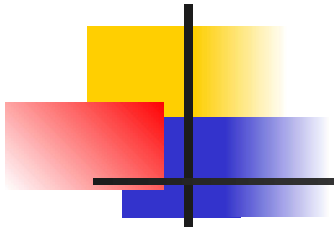
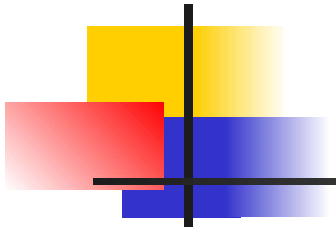


Data Wrangling

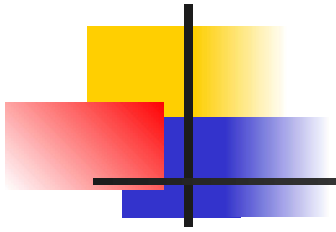


- Identify and handle missing values
- - Identify missing values
- - Deal with missing values
- - Correct data format
- Data standardization
- Data Normalization (centring/scaling)
- Binning
- Indicator variable



Data Wrangling

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



Steps for working with missing da

- 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**

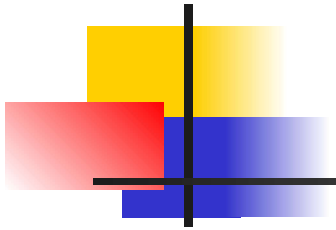
There are two methods to detect missing data:

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

Deal with missing data

- How to deal with missing data?

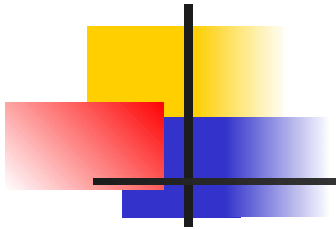
- 1. drop data
- a. drop the whole row
- b. drop the whole column
- 2. replace data
- a. replace it by mean
- b. replace it by frequency
- c. replace it based on other functions



Correct data format

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

Data Standardization

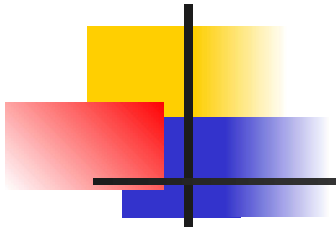


- Data is usually collected from different agencies with
formats. (Data Standardization is also a term for a pa
type of data normalization, where we subtract the m
divide by the standard deviation)
- **What is Standardization?**
- Standardization is the process of transforming data i
common format which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

Data Normalization

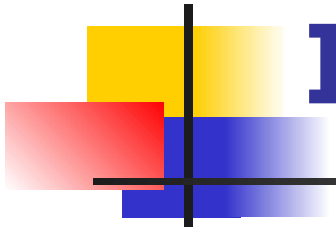
- **Why normalization?**
- Normalization is the process of transforming values o variables into a similar range. Typical normalizations scaling the variable so the variable average is 0, scal variable so the variable variance is 1, or scaling varia variable values range from 0 to 1

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



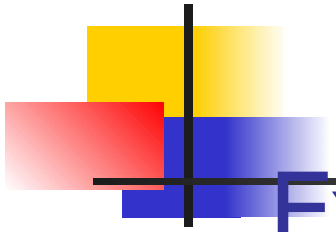
Binning

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



Indicator variable (or dummy va

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

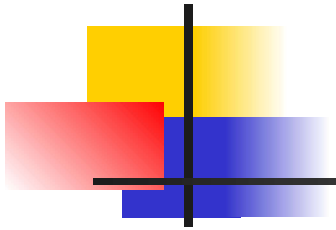


Exploratory Data Analysis (EDA)

- For data analysis, Exploratory Data Analysis (EDA) must 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

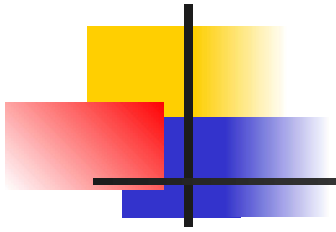
- Exploratory data analysis is generally cross-classified in two each method is either non-graphical or graphical. And secon method is either univariate or multivariate (usually just biva
- Non-graphical methods generally involve calculation of sum statistics, while graphical methods obviously summarize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



Data Frames attributes

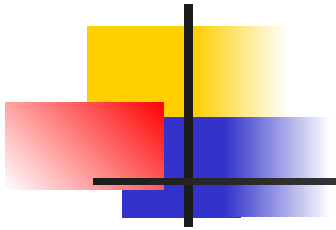
Python objects have *attributes* and *methods*.

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



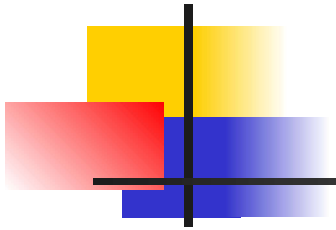
Data Frames methods

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 catego
  The data is grouped based on one or several variables
  analysis is performed on the individual groups.

Analysis

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