## Artificial Intelligence

## and

## Machine Learning

Semester-IV (Batch-2022)

Case Study: Salaries.csv

<https://drive.google.com/file/d/18TU15RykE4UC5ZnLon_wCjNRkgobH6x5/view?usp=drive_link>

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Faculty Name Name : Simranpreet Kaur

Roll Number (G13 A)

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

Chitkara University, Punjab

**DESCRIPTION**

**Libraries:**

**Pandas:** Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant.

**Numpy:** NumPy is a Python library that provides a multidimensional array object and various routines for processing those arrays. It is the fundamental package for scientific computing with Python, and it is widely used in many fields of science and engineering

**Questions:**

# Display Top 10 Rows of The Dataset

1. Check Last 10 Rows of The Dataset
2. Find Shape of Our Dataset (Number of Rows And Number of Columns)
3. Getting Information About Our Dataset Like Total Number Rows, Total Number of Columns,memory usage
4. Check Null Values In The Dataset
5. Datatypes of Each Column And Memory Requirement
6. Drop ID, Notes, Agency, and Status Columns
7. Get Overall Statistics About The Dataframe
8. Find Occurrence of The Employee Names (Top 5)
9. Find The Number of Unique Job Titles
10. Total Number of Job Titles Contain Captain
11. Display All the Employee Names From Fire Department
12. Find Minimum, Maximum, and Average BasePay
13. Replace 'Not Provided' in EmployeeName' Column to NaN
14. Drop The Rows Having 5 Missing Values
15. Find Job Title of ALBERT PARDINI
16. How Much ALBERT PARDINI Make (Include Benefits)?
17. Display Name of The Person Having The Highest BasePay

**Methods:**

1. pandas.read\_csv(): **p**andas.read\_csv() is a function in the pandas library that reads a comma-separated values (csv) file into a Dataframe. It has many parameters that allow you to customize how the file is read, such as sep, header, index\_col, usecols, dtype, skiprows, nrows, na\_values, and more.
2. Data.head(): data.head(10) is a method that returns the first 10 rows of the DataFrame data. It is useful for getting a quick overview of the data or checking the format and column names. You can change the number of rows by passing a different value to the method, such as data.head(5) or data.head(20)
3. **data.tail(10) :** data.tail(10) is a method that returns the last 10 rows of the DataFrame data. It is useful for getting a quick overview of the data or checking the format and column names. You can change the number of rows by passing a different value to the method, such as data.tail(5) or data.tail(20)
4. **data.isnull().sum():** data.isnull().sum() is a method that returns the number of missing values in each column of the DataFrame data. It is useful for detecting and handling missing data in your analysis. It returns a Series with the column names and the count of null values.
5. **max():** The max() function is a built-in function in Python that returns the maximum value among the given arguments.
6. Info() : the **info()** method is used to display a concise summary of a DataFrame, including the data types of columns, the number of non-null values, and memory usage.
7. Replace(): replace() is a method that can be used to replace a part of a string with another string. It takes two or three arguments: the old substring to be replaced, the new substring to replace it with, and optionally the number of times to perform the replacement.
8. describe() : describe() is a method in the pandas library of Python. It generates descriptive statistics for a DataFrame or a Series of numeric or object values. It summarizes the central tendency, dispersion, and shape of the data distribution, excluding NaN values. It also analyzes the unique, top, and frequency values for object data, such as strings or timestamps
9. Mean() : mean() is a function that can be used to calculate the arithmetic mean of a given set of numbers in Python. The arithmetic mean is the sum of the numbers divided by the number of elements.
10. **Drop():** In the context of the pandas library in Python, the drop() method is used to remove specified labels from rows or columns of a DataFrame. It returns a new DataFrame with the specified labels removed.
11. **nunique() :** In pandas, the nunique() method is used to count the number of unique elements in a Series. It returns the count of distinct elements in a Series, excluding NaN values.
12. **value\_counts() :** In pandas, the value\_counts() method is used to count the occurrences of unique values in a Series. It can be applied to a single column of a DataFrame, which is represented as a Series in pandas.
13. **str.contains :** In pandas, the str.contains() method is used to check whether each element in a Series contains a specified substring or pattern. It's particularly useful when dealing with text data in a DataFrame column.
14. **Len() :** the len() function is a built-in function that is used to get the length (the number of items) of an object. The behavior of len() depends on the type of object it is called on.