## Front End Engineering-II /Artificial

## Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

**Case Study** :Employee Salary

[Url:-](file:///C:\Users\spars\Desktop\AIML%20Project\-) https://drive.google.com/file/d/10vyVq4CQiCAoy5YCrpw1mZDj1al34cdm/view?usp=drive\_link

A red and white sign

Description automatically generated with low confidence­

**Supervised By: Submitted By:**

**Faculty Name: Mr Rajeev Bhardwaj** **Name:Taranpreet Kaur**

**Roll no:2210990908**

**Group:G-13**

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

**Description about Case Study: -**

* Display Top 10 rows of dataset salaries
* Display the Last 10 rows
* Find shape of our dataset(number of rows and columns)
* Getting information about our dataset.
* Check null values in the Dataset
* Drop id, notes, agency and status columns
* Find occurrence of the employee name(top 5)
* Find number of unique job titles.
* Total number of job titles that contains captain.
* Display all the employee names from fire department.
* Find minimum, maximum and average basepay.
* Replace ‘NOT PROVIDED’ in employeename column to NaN.
* Drop the rows having more than 5 missing values.
* Find the job title of ALBERT PARDINI.
* How much albert pardini make (include benefits)?

**Library: -**

**PANDAS:-** In this file, we used the pandas library. Pandas is a powerful and popular open-source Python library for data analysis and manipulation. It provides data structures such as Series (one-dimensional) and Data Frame (two-dimensional) to handle various types of data. Pandas simplify data wrangling, reading/writing data from various file formats, filtering, sorting, merging, handling missing data, reshaping, aggregating, and time series analysis. It is widely used in data science, machine learning, and data engineering projects, often in conjunction with other libraries like NumPy, Matplotlib, and Scikit-learn. Its simplicity and efficiency make it indispensable for working with data in Python, enabling data-driven decisions and insights.

**Methods: -**

1. **read\_csv():**

Description: Reads a CSV file and converts it into a data frame.

1. **tail():**

Description: Displays the last few rows of the data frame.

1. **head():**

Description: Displays the first few rows of the data frame.

1. **shape():**

Description: we can calculate the shape(i.e rows and columns) by using len() function. For rows we use data.index and for columns we use data.columns.Shape()returns the shape of the data frame.

1. **info():**

Description: Provides basic information about the data frame, such as column types and missing values.

1. **isnull():**

Description: Returns True/False for each value in the data frame, indicating whether the value is missing (NaN) or not.

1. **sum():**

Description: Calculates the sum of values in each column of the data frame.

1. **drop():**

Description: Removes specific rows or columns from the data frame.

1. **value\_counts():**

Description: Counts the unique values in a specific column of the data frame.

1. **nunique():**

Description: Returns the count of unique values in a specific column of the data frame.

1. **str.contains():** Checks if a specified substring or value is present in a column of the data frame.
2. **max():** Returns the maximum value in a column of the data frame.
3. **min():** Returns the minimum value in a column of the data frame.
4. **mean():** Calculates the mean (average) value of a column in the data frame.
5. **len():** Returns the number of rows in the data frame
6. **value\_counts():** Counts the unique values in a specific column of the data frame.
7. **apply():** Applies a function to transform the values in the data frame.
8. **replace():** method in Pandas is used to replace values in a Series with other values. It is a versatile method that allows you to perform replacements based on various criteria.
9. **dropna():** method in Pandas is used to remove missing (NaN) values from a Series. It returns a new Series with the missing values removed.