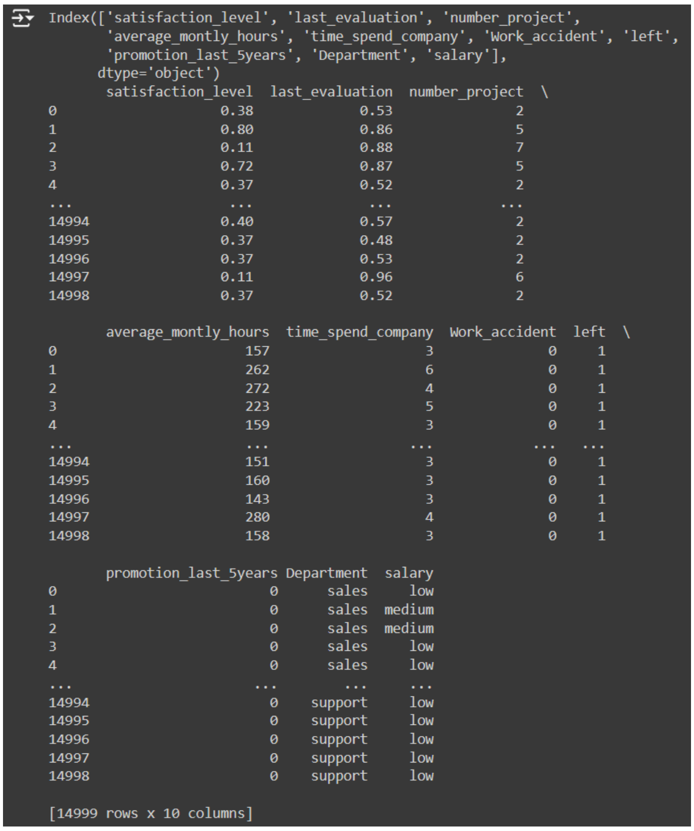
DAT565/DIT407 Assignment 8

Henning Anevik   Joar Forsberg

2024-10-30

Problem 1: Create a Datasheet

With the following code A the datasheet was created. After creation its printed out that gives the following output shown in Figure. 1.



```
Index(['satisfaction_level', 'last_evaluation', 'number_project',
       'average_montly_hours', 'time_spend_company', 'Work_accident', 'left',
       'promotion_last_5years', 'Department', 'salary'],
      dtype='object')
       satisfaction_level  last_evaluation  number_project  \
0                    0.38             0.53               2
1                    0.80             0.86               5
2                    0.11             0.88               7
3                    0.72             0.87               5
4                    0.37             0.52               2
...                   ...              ...             ...
14994                0.40             0.57               2
14995                0.37             0.48               2
14996                0.37             0.53               2
14997                0.11             0.96               6
14998                0.37             0.52               2

       average_montly_hours  time_spend_company  Work_accident  left  \
0                       157                   3              0     1
1                       262                   6              0     1
2                       272                   4              0     1
3                       223                   5              0     1
4                       159                   3              0     1
...                     ...                 ...            ...   ...
14994                   151                   3              0     1
14995                   160                   3              0     1
14996                   143                   3              0     1
14997                   280                   4              0     1
14998                   158                   3              0     1

       promotion_last_5years Department  salary
0                          0      sales     low
1                          0      sales  medium
2                          0      sales  medium
3                          0      sales     low
4                          0      sales     low
...                      ...        ...     ...
14994                      0    support     low
14995                      0    support     low
14996                      0    support     low
14997                      0    support     low
14998                      0    support     low

[14999 rows x 10 columns]
```

Figure 1: Output from datasheet

- **Purpose of dataset creation**: Created to track key employee metrics like performance, satisfaction, and salary, guiding HR processes such as salary evaluation.
- **Dataset creator**: Created by an HR department within a company.

- **Instances represented**: Each instance represents an employee, including ten metrics: employee satisfaction, evaluation score, number of projects, average work hours, employment duration, involvement in workplace accidents, employment status, promotions in the last five years, department, and salary.
- **Total instances**: 14,999 instances.
- **Data in each instance**: Each feature is either numerical (e.g., monthly hours), binary (e.g., employment status), or textual (e.g., department).
- **Label/target for each instance**: Each instance is labeled with a unique employee number.
- **Confidentiality of data**: Potentially confidential depending on local laws; salary data might be considered confidential in some regions.
- **Offensive or sensitive content**: Unlikely to contain offensive or anxiety-inducing content.
- **Identification of subpopulations**: No, the dataset does not identify subpopulations (e.g., age, gender).
- **Possibility of identifying individuals**: Yes, it may be possible to identify specific employees, especially if certain metrics stand out (e.g., involvement in an accident).
- **Sensitive data**: Salary level may be sensitive, as it indicates financial standing, though only general levels (low, medium, high) are provided.
- **Ethical review**: No ethical review was conducted.
- **Data collection source**: Data was directly collected from employees.
- **Notification to individuals**: Employees were notified about some data collection (e.g., performance reviews), but not for all metrics (e.g., salary level).
- **Consent from individuals**: Consent was obtained where employees were notified about data collection.
- **Impact on future uses**: None noted; no issues anticipated from the data's composition or collection.
- **Restricted use**: Intended only for HR purposes; other uses are discouraged.

Problem 2: Ethics

This question hinges on what you believe is the primary object of an HR department. If you are of the belief that the purpose of an HR department is to maximize capital gains (of the company) through extracting as much value as possible through the employees (of said company), it is not so hard to see how the desires of the company may run opposite to that of the employees. In this case the company may use this dataset in such a way to encroach on the rights of the workers in negotiations and other processes in order to serve their own interests.

- In the dataset, entries of employees who left the company are still included. This could be used to punish employees who left by company by giving said employee a bad reputation and spreading it to other companies and as such making it much harder for that employee to get hired somewhere else.
- Collecting data about workplace accidents could be used in salary negotiations in order to keep down the salary of particular employees - even if said employee otherwise has been performing well.
- Tracking workplace satisfaction could also be a problem if a particular entry could be tied to a particular employee. For example, employees that already have a high workplace satisfaction may be passed over for promotions and salary increases because the company believes that they won't quit (since their satisfaction is already high enough).

Problem 3: Data Privacy and the law

a)

No. This would be selling the personal data of students. Unless the students specifically consented to this, this is not allowed. See article 5 and 6.

b)

Yes. First off, the students would most have likely consented to plagiarism controls by attending the course see the Chalmer's guide to Academic integrity and honesty. Secondly, this would also comply with paragraph 5 in article 6, "Processing is necessary to perform a task in the public interest or to carry out some official function." Plagiarism control at a university would most likely fall under condition outlined above.

c)

Yes. This would also most likely fall under the same definition as above, i.e paragraph 5 of article 6. Collecting data is neccessary for the national board of education to do its job properly - that is, it's necessary to perform the a task in the public interest or to carry out some official function.

d)

The university has obligations both to the authorities and the data subjects, as outlined by articles 33 and 34, respectively. Article 33 states that the university must notify the appropriate authorities about the incident.

Article 34 dictates that the university must infrom the data subjects whose information was subject to the leak. The information must contain the nature of the leak and the likely consequences of the leak as well as the measures taken to mitigate the damage caused by the leak.

A Problem 1: Create a Datasheet

```
import pandas as pd

df = pd.read_csv("/content/HR_comma_sep.csv")
print(df.columns)
print(df)
```