Pham Nhat Duc - Troy ID: 1624630

1. How many rows are there ?

There are 15099 rows

2. How many columns are there ? What are they ?

The number of columns: 11

The columns are:

- satisfaction_level
- last_evaluation
- number_project
- average_montly_hours
- time_spend_company
- work_accident
- left
- promotion_last_5years
- is_smoker
- department
- salary

3. How many features? What are they ?

The number of features: 10

The features are:

- satisfaction_level
- last_evaluation
- number_project
- average_montly_hours
- time_spend_company
- work_accident
- promotion_last_5years
- is_smoker
- department
- salary

4. How many duplicates if any ?

The number of duplicate rows: 2840

5. Remove duplicates if there are any.

After removing duplicate, the number of rows is 12259.

6. Print the distributions of each features. What do you see ?

Many outliers typically appear isolated from the main data, make it easy to spot.

7. How many missing values are there ? What features have missing values ?

Total missing values in the dataset: 12542

Features with missing values:

- average_montly_hours: 368 missing values

- time_spend_company: 150 missing values

- is_smoker: 12024 missing values

8. Drop the features with largest number of missing values and fill up the rest of the features with missing values with mean. Can you explain why did we drop the feature with largest portion of missing values in this case ?

We dropped the feature with the largest portion of missing values because it lacks sufficient data to provide meaningful information, and imputing it could introduce noise or bias, negatively impacting the cleanliness of data.

9. Show that there are no missing values in the data at this point.

Result from notebook:
satisfaction_level: 0

last_evaluation: 0

number_project: 0

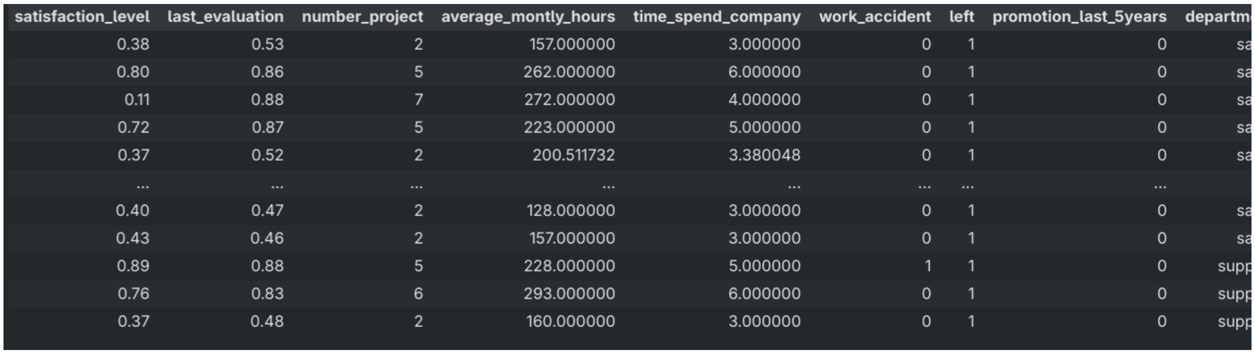
average_montly_hours: 0

time_spend_company: 0

work_accident: 0

promotion_last_5years: 0

10. Feature "left" has values 'yes' and 'no'. Convert those values into integer values, 1 and 0.



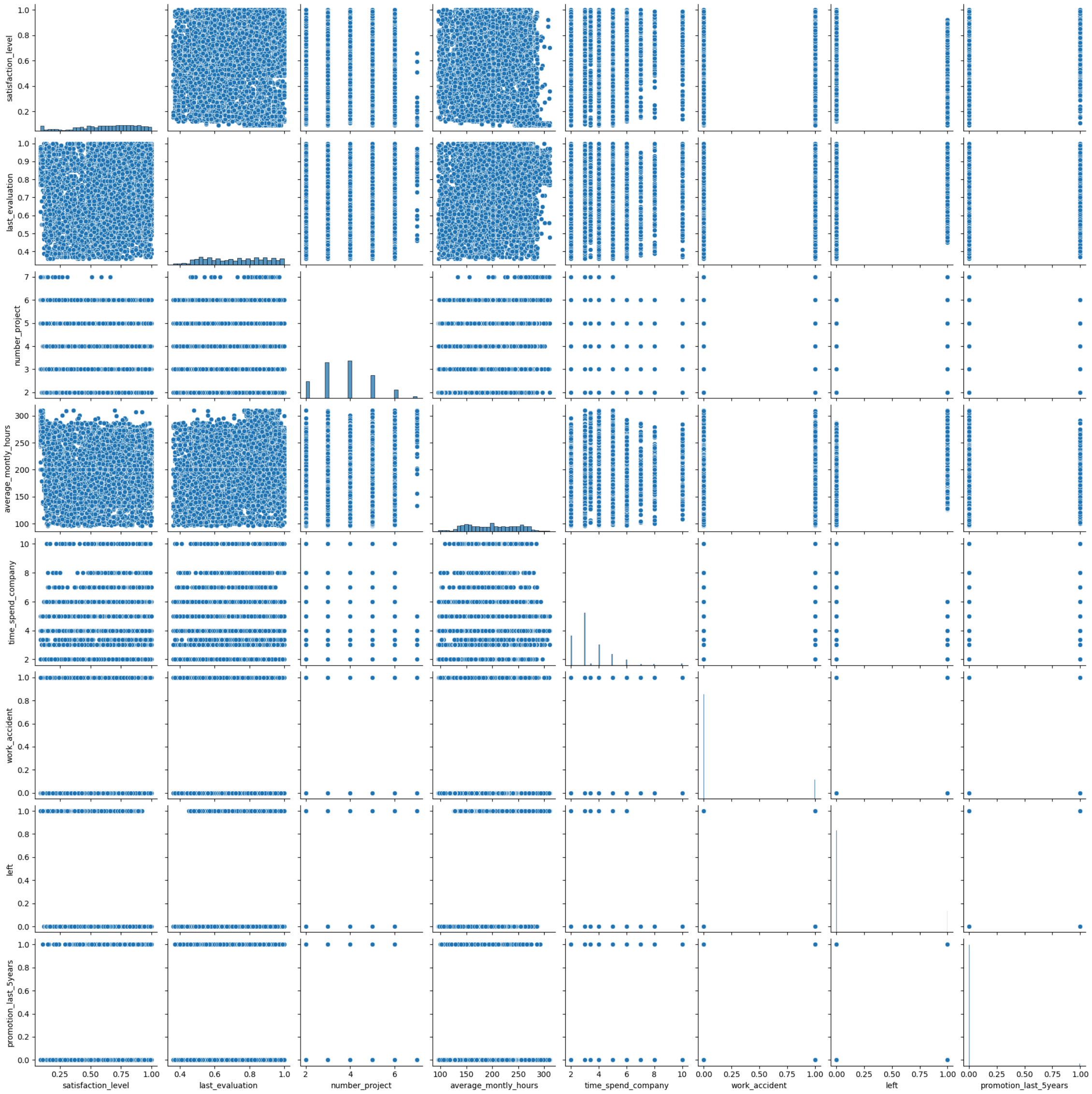
satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	left	promotion_last_5years	departm
0.38	0.53	2	157.000000	3.000000	0	1	0	sa
0.80	0.86	5	262.000000	6.000000	0	1	0	sa
0.11	0.88	7	272.000000	4.000000	0	1	0	sa
0.72	0.87	5	223.000000	5.000000	0	1	0	sa
0.37	0.52	2	200.511732	3.380048	0	1	0	sa
...	...	...	...	...	...	...	...	
0.40	0.47	2	128.000000	3.000000	0	1	0	sa
0.43	0.46	2	157.000000	3.000000	0	1	0	sa
0.89	0.88	5	228.000000	5.000000	1	1	0	supp
0.76	0.83	6	293.000000	6.000000	0	1	0	supp
0.37	0.48	2	160.000000	3.000000	0	1	0	supp

11. Save the resulting data into a file.

We can use this code:

```
clean_data.to_csv('cleaned_data.csv', index=False)
```

12. What else do you observe in in this data set.



13. Divide the data into training set(80%) and test set(20%) and show the number of data points in each by following

1. Uniform sampling
    - In uniform sampling, the dataset is randomly split into an 80% training set and a 20% test set without considering the distribution of any specific feature. This can be done using a function like train_test_split from Python's scikit-learn library

2. Stratified sampling based on the ratio of "yes" and "no" values of feature "left"
    - In stratified sampling, the dataset is split into training and test sets while preserving the proportion of "yes" and "no" values in the "left" feature. Using train_test_split with the stratify parameter ensures that the training and test sets have the same ratio of "yes" to "no" as the original dataset.
3. How would you do the test set and training set sampling and why ?
    - Use uniform sampling for simplicity when the "left" feature is balanced or when class distribution is not a concern.
    - Use stratified sampling based on the "left" feature to ensure proportional representation of "yes" and "no" in both sets, which is critical for imbalanced datasets or when accurate evaluation of both classes is important.