EDA for HR comma sep

by Oki Samila Rici

Mind Map


Use Case Summary
Business Understanding
Data Understanding
Data Preparation
Data Cleansing
Exploratory Data Analysis
Conclusion

Use Case Summary

Objective Statement:

- Examine whether the values in the statistical summary make sense.
- Get insights into whether there is an outlier and how the distribution shape on the variables.
- Get an insight into which department has the lowest salary.
- See how the correlation between one variable and another.

Expected Outcome:

- Know whether the values in the statistical summary make sense.
- Know whether there is an outlier and how the distribution shape on the variables.
- Know the name of a low-paying company department.
- Know how the correlation between one variable and a nother.

Business Understanding

- This dataset contains a company that records various employee parameters (example : salary, satisfaction le vel, etc).
- some questions based on the data:
    - whether the values in the statistical summary make sense.?
    - whether there is an outlier and how the distribution shape on the variables?
    - Which department has the lowest salary?
    - How the correlation between one variable and another?

Data Understanding

- Source data by Kaggle
- Data Dictionary:
  - statisfaction_level: satisfaction level at the job of an employee
  - last_evaluation: rating between 0 to 1 received by an employee at this last evaluation
  - number_project: number of projects an employee involved
  - average_monthly_hours: average number of hours in a month, spent by an employee at the company
  - time_spend_company: number of years spent in the company
  - Work_accident : 0 = no accident during employee stay, 1 = accident during employee stay
  - left : 0 = indicates employee stay in the company, 1 = indicates employee left the company
  - promotion_last_5year: number of promotions in his stay
  - Department: department an employee belongs to
  - salary: salary in USD

Data Preparation

- Packages : Pandas, Numpy, Seaborn, Matplotlib
- Use HR comma sep dataset
  - All columns, except sales and salary columns are numeric.
  - Work_accident, left, and promotion_last_5years are binary (0,1).
  - The name of the sales column will be changed to Department.

Data Cleansing

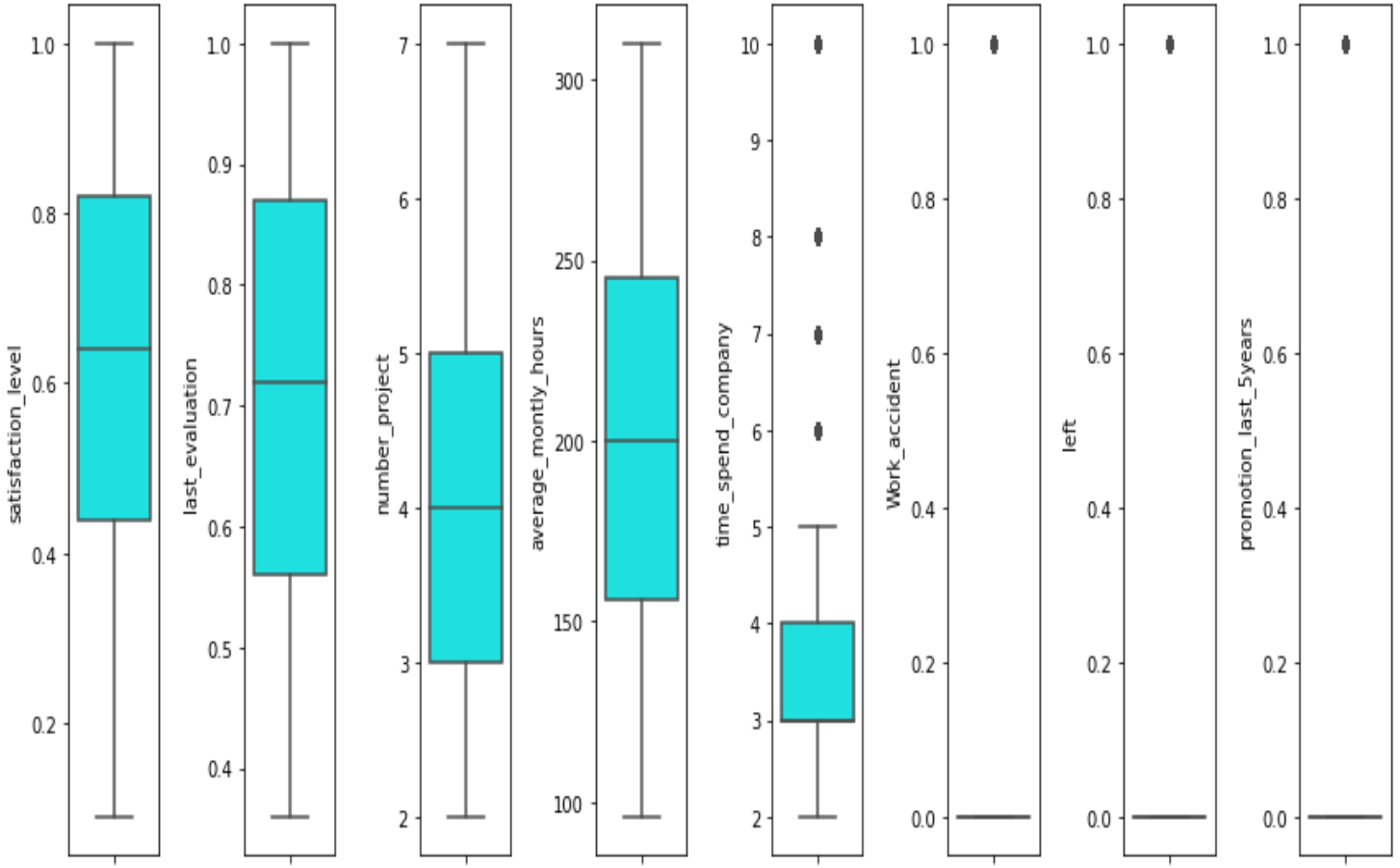
- The dataset has no missing values, and the data type names are all correct.

Exploratory Data Analysis

Statistical Summary

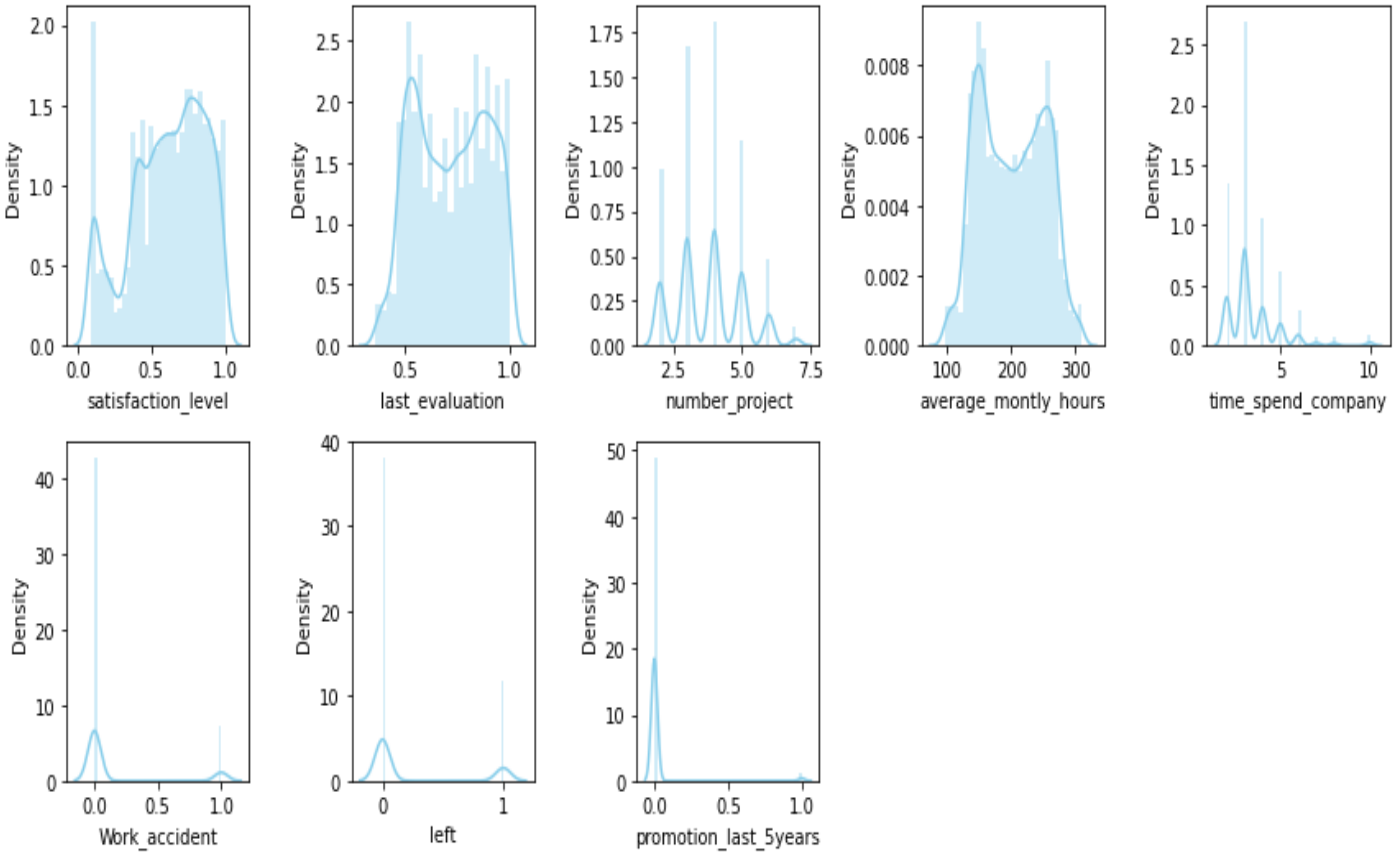
- All statistical summary values have appropriate values.
- Mean < Median in satisfaction_level and number_project, indicating negatively skewed distribution.
- Mean ~ Median in last_evaluation and average_monthly_hours, indicating somewhat a symmetrical distribution.
- Mean > Median in time_spend_company, indicating positively skewed.
- Work_accident, left, and pomotion_last_5years, are binary/boolean columns since the value is 0 or 1.

Boxpot to detect outliers



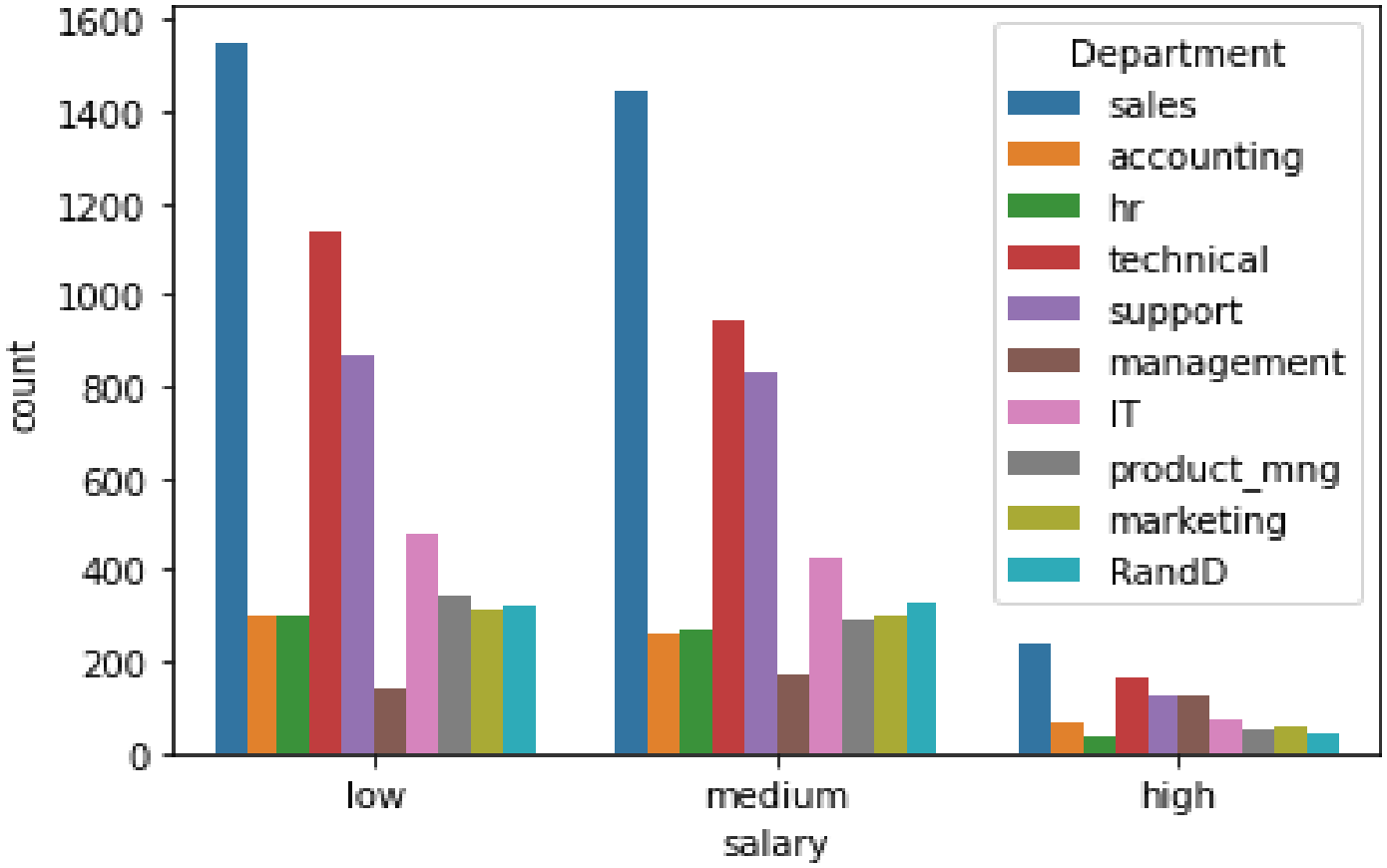
- There are many outliers in time_spend_company, because the average employee worked at the company 2 to 5 year.
- Work_accident have an outlier because more are not work accidents.
- left variable have an outlier because more employees who don't leave the company.
- The outlier promotion_last_5year variable on number 1, because more employees who have not received a promotion in the last five years.

KDE plot for knowing the distribution form



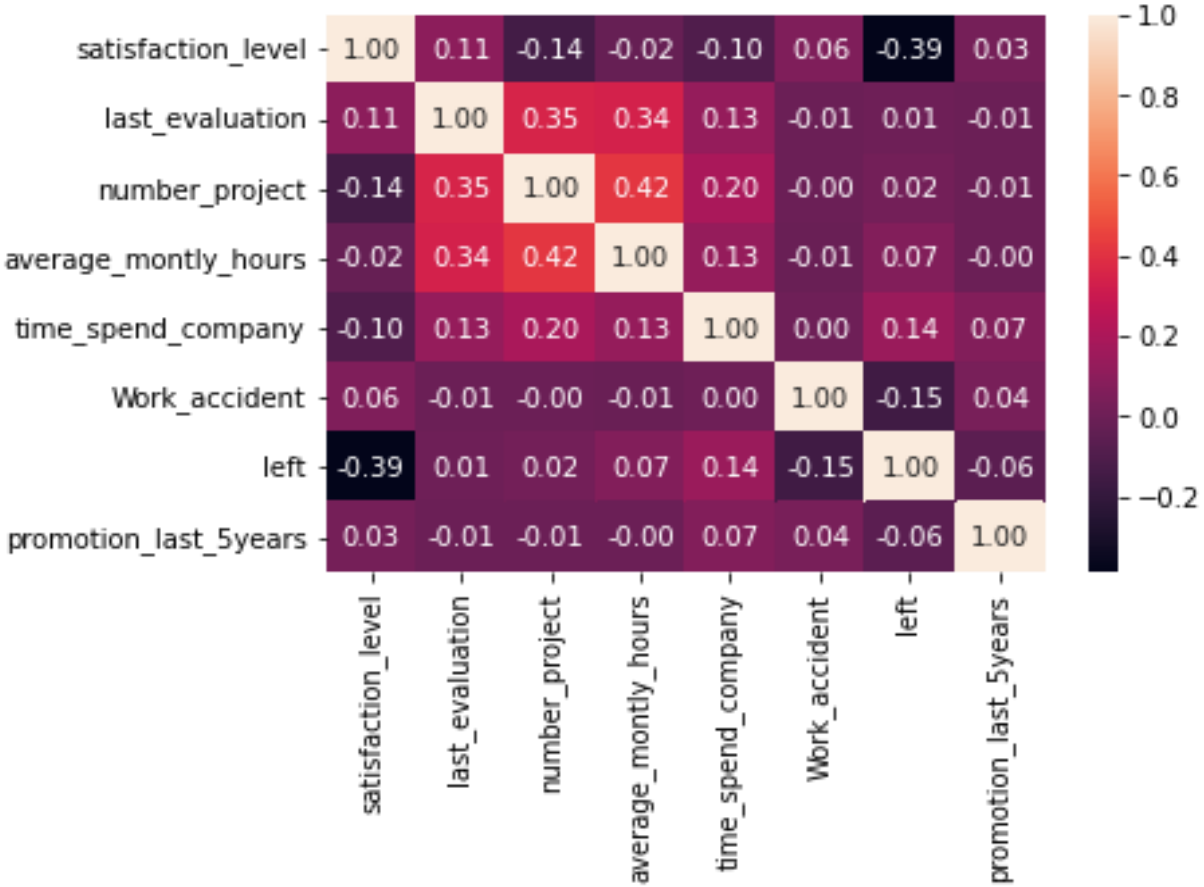
- statisfaction_level is slightly negative skewed.
- last_evaluation and average_monthly_ hours the most symmetric distribution.
- number project and time_spend_ company are slightly positively skewed
- Work_accident, left, and promotion_las t_5years has boolean with value 0 and 1.

Which department has the lowest salary?



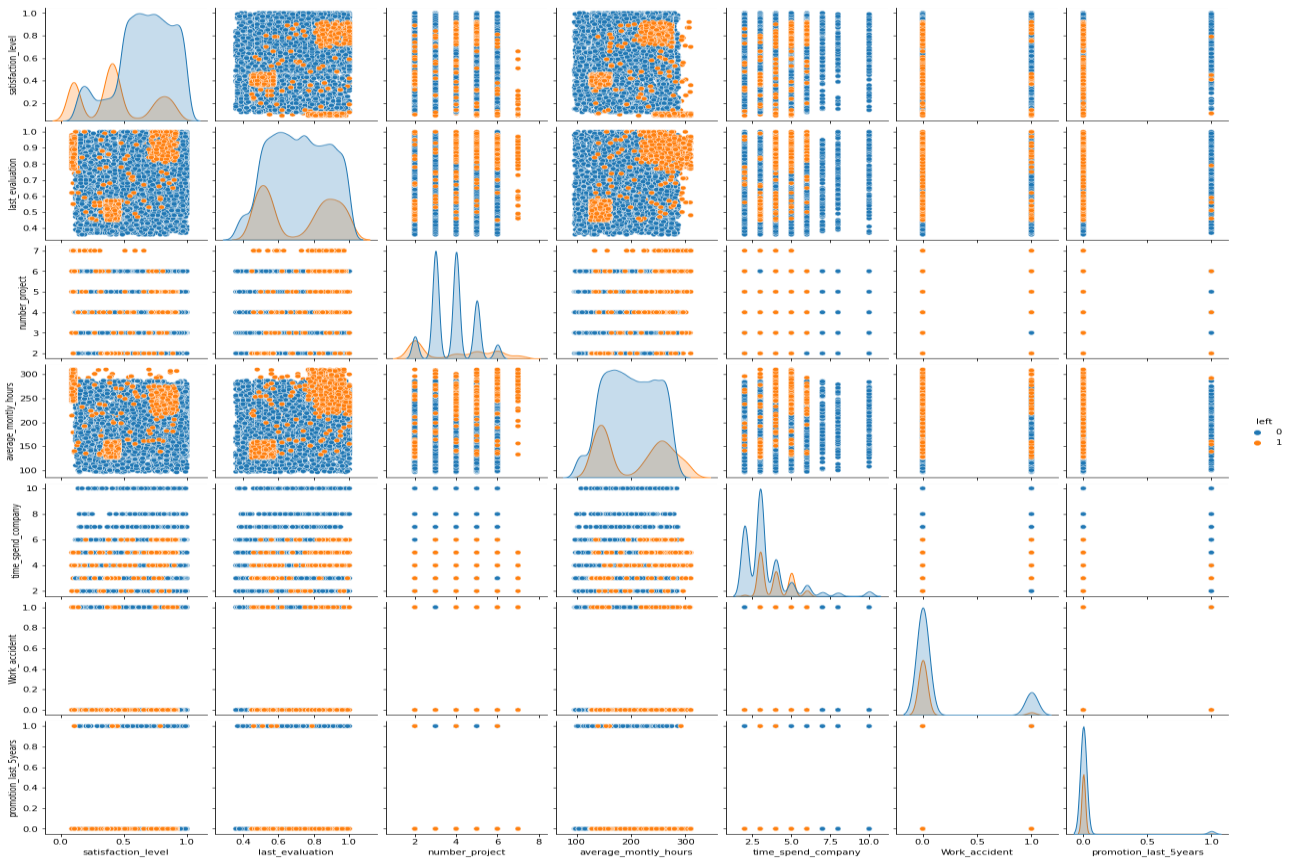
sales, technical and support departments have low and medium salary

Correlation Heatmap



- The correlation between an average_monthly_ hour and number_project is 42%, indicating that the employee's length of time in the company (a matter of months) has a significant impact on the number of projects completed by the employee.
- The correlation between number_project and last_evaluation by 35%, meaning that the number of projects has a high effect on employee evaluation.
- The correlation between average_monthly_hour and last_ evaluation is 34%, indicating that the employee's length of time in the company (month) has a high effect on employee evaluation.
- average_monthly_hour, number_project and last_evaluation are highly correlated each other.

Pairplot of The Data



The blue color represents employees who don't leave the company, while the orange color represents employees who do leave the company.

fewer employees leave the company. Correlations between variables that don't have positive or negative correlations accumulate in the middle.

Conclusion

- The dataset doesn't have a missing value.
- Overall, the minimum and maximum values make sense for each column.
- From the barplot we can see outlier in time_spend_company, work_accident, left, and promotion_last_5years.
- The lowest and medium salaries were found in the sales, technical, and support departments.
- For correlation heatmap, we can see that last_evluation is correlated with average_monthly_hour and number_project.
- For pairplot, correlations between variables that don't have positive or negative correlations and fewer employees leave the company