

OESON Project 2

Data analysis of HR dataset using Python (seaborn, plotly, scipy and pandas)

Shuvam Mondal

Dataset, Intent and Tools Used

Dataset:

An employee dataset from a company with the following employee details collected during a survey:

'MarriedID', 'MaritalStatusID', 'GenderID',
'EmpStatusID', 'DeptID', 'PerfScoreID',
'FromDiversityJobFairID', 'Salary', 'TermID',
'PositionID', 'Position', 'State', 'Zip', 'DOB', 'Sex',
'MaritalDesc', 'CitizenDesc', 'HispanicLatino',
'DateofHire', 'DateofHire.1', 'TermReason',
'EmploymentStatus', 'Department', 'ManagerName',
'ManagerID', 'RecruitmentSource',
'PerformanceScore', 'EngagementSurvey',
'EmpSatisfaction', 'SpecialProjectsCount',
'LastPerformanceReview_Date', 'DaysLateLast30',
'Absences'

Intent:

To create visualizations based on employee information to assess:

- 1) Salaries across departments and genders
- 2) Employee satisfaction and productivity levels
- 3) Productivity levels for different teams
- 4) Employee absence and leave statistics
- 5) Engagement in special projects

Tools:

Violin Plots: `seaborn.violinplot`

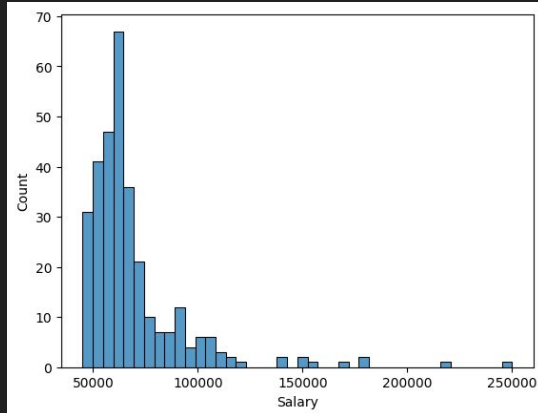
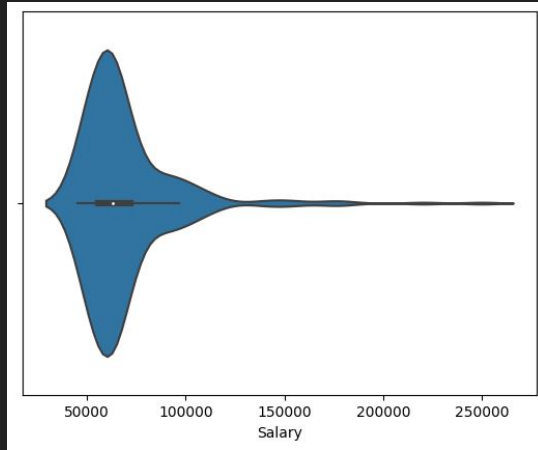
Box Plots: `seaborn.boxplot`

Histograms: `seaborn.histplot`

Scatter Plots: `matplotlib.pyplot`

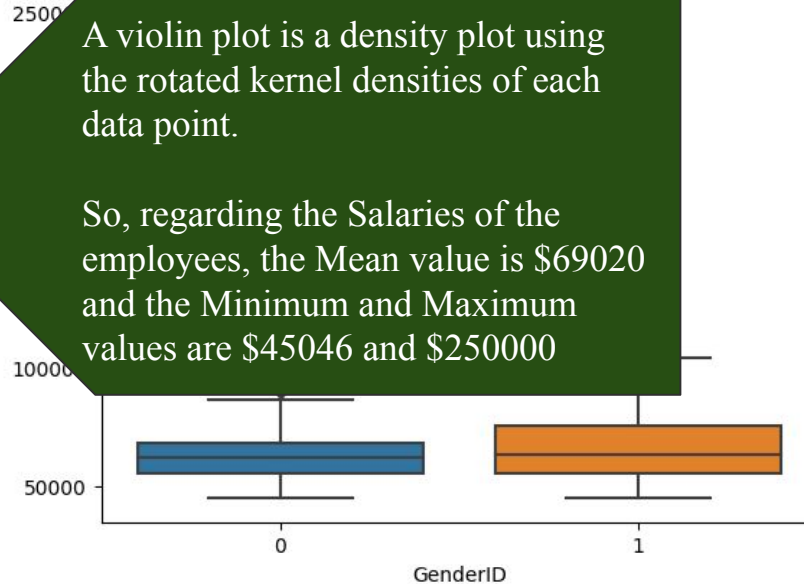
Descriptive Statistics: `scipy.stats.describe`

Salary analysis



A violin plot is a density plot using the rotated kernel densities of each data point.

So, regarding the Salaries of the employees, the Mean value is \$69020 and the Minimum and Maximum values are \$45046 and \$250000



Descriptive Statistics (Scipy.stats):

(Min,Max)=(45046, 250000)

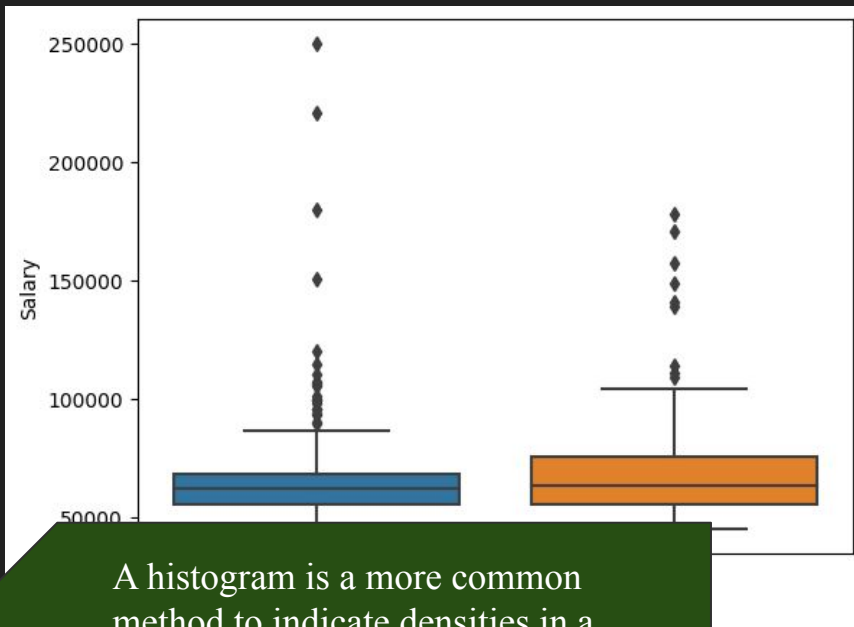
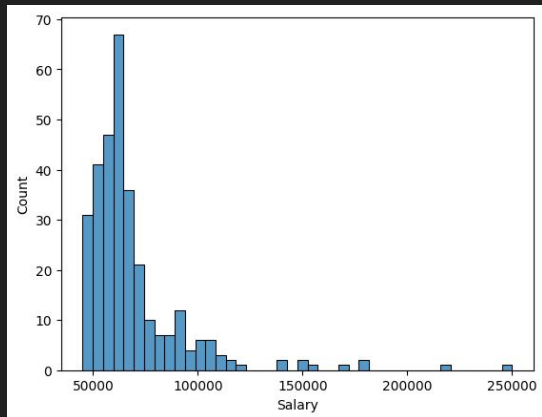
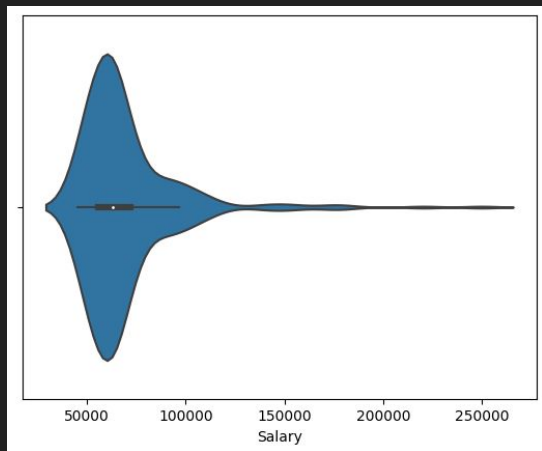
Mean=69020.6848874598

Variance=632856381.6100613

Skewness=3.290213187177172

Kurtosis=15.1856070456835

Salary analysis



A histogram is a more common method to indicate densities in a dataset and conveys the same information using predefined bins of the dataset.

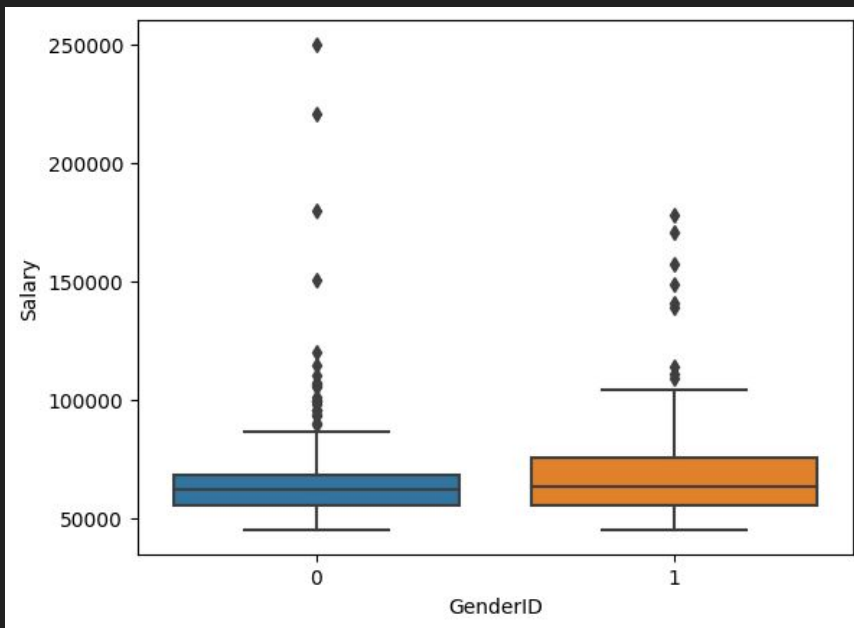
Skewness=3.290213187177172
Kurtosis=15.1856070456835

Salary analysis

Box plots definitely convey the most amount of information since they show:

- 1) The median (The central line of each box)
- 2) The Interquartile range (Values from 25% to 75% of the data are enclosed in the box)
- 3) The Maximum and Minimum values demonstrated by horizontal lines.
- 4) Outliers plotted as dots.

And the box plot clearly indicates that the median of the salaries across genders is quite similar ~\$69000. The interquartile range however, might be different because of over-representation by a gender.



Descriptive Statistics (Scipy.stats):

(Mix,Max)=(45046, 250000)

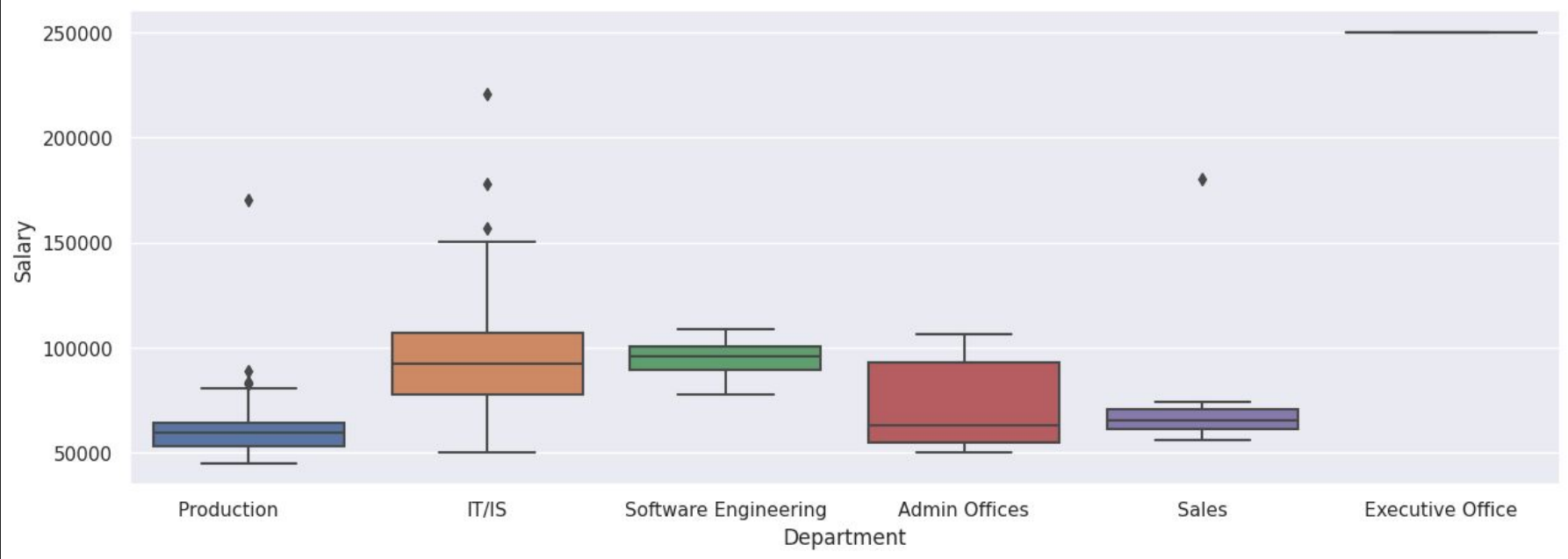
Mean=69020.6848874598

Variance=632856381.6100613

Skewness=3.290213187177172

Kurtosis=15.1856070456835

Salary analysis across departments



Box plots can also be used to plot the salaries of employees across departments.

As is evident by the plot, Software engineering and IT sectors are paid the most followed by Administration and sales. The width of the box also gives a clear indicator to the amount of people working in a sector in the company.

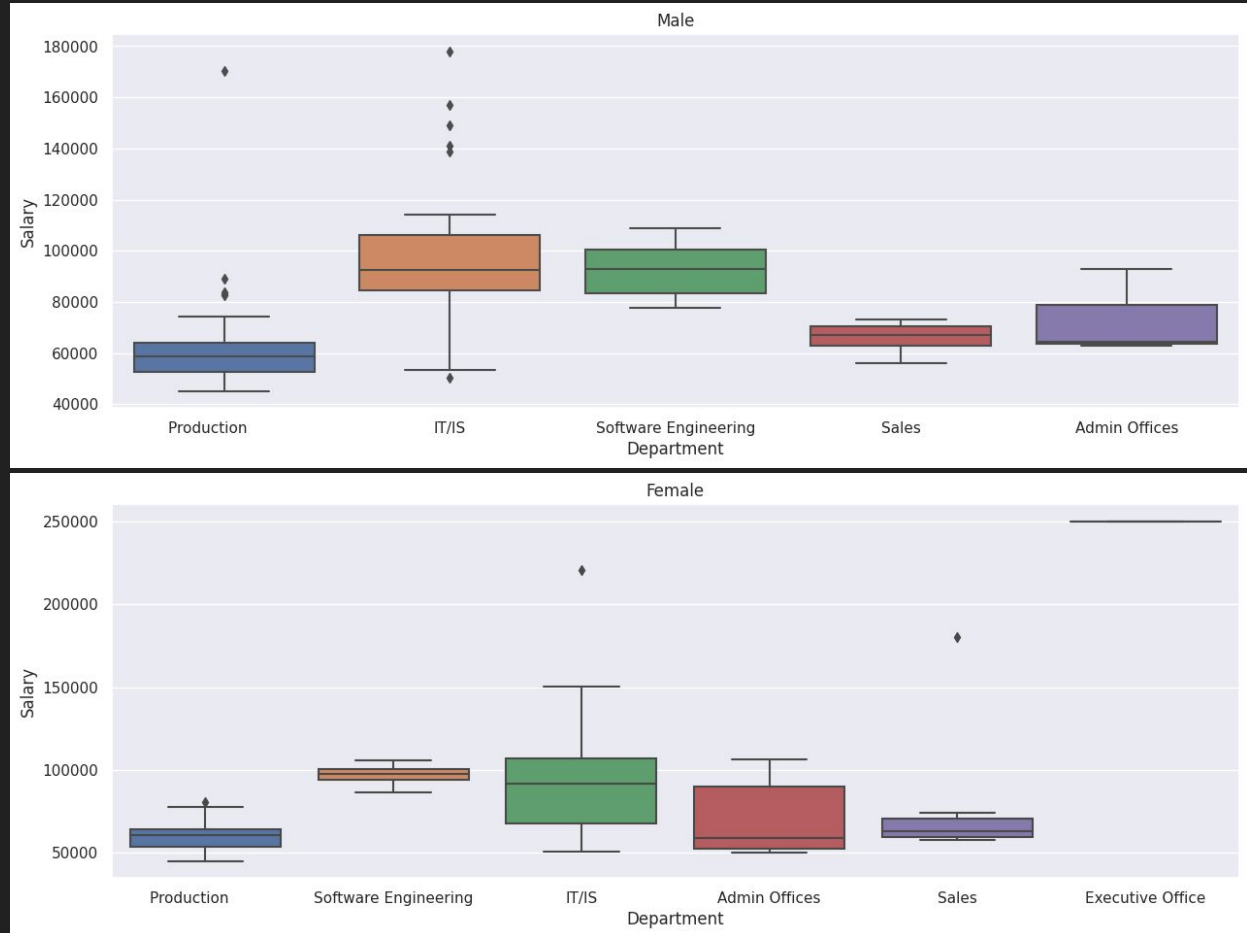
Salary across departments for different genders

Similarly,
Box plots for employees of
different genders across
departments can be plotted.

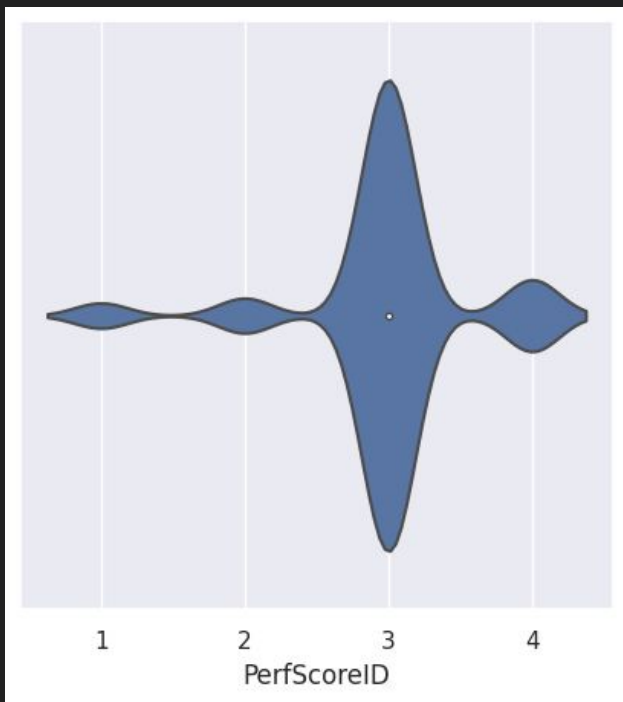
The plots aside suggest an
over-representation by men in
Software engineering, IT
Production.

Whereas, more women were
Employed in Admin and
Executive positions.

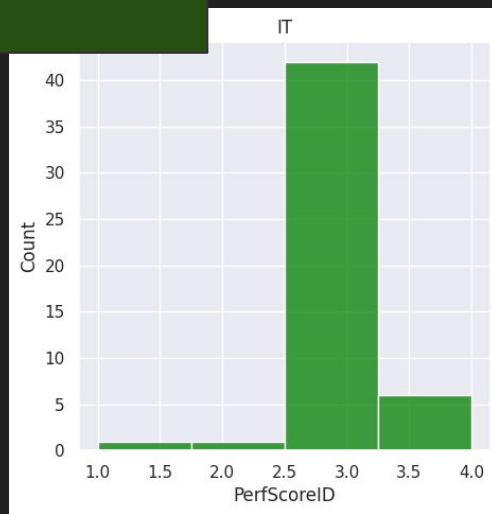
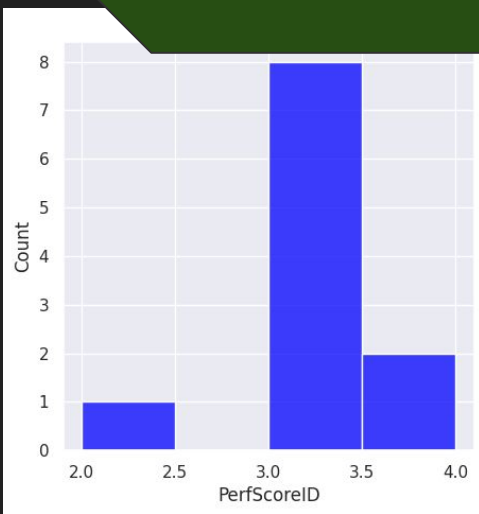
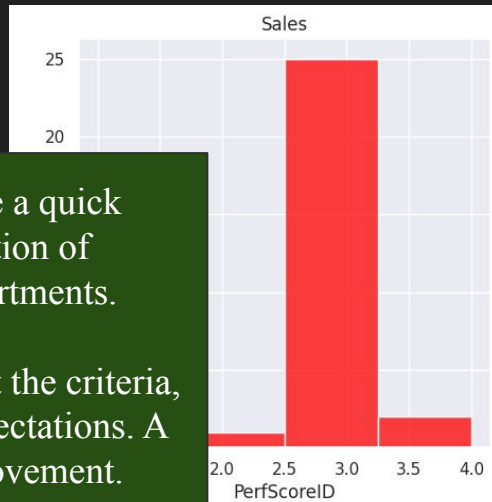
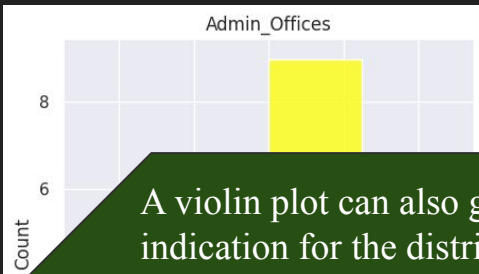
The highest earning employee
was also female.



Performance across Departments



- 1: Severely Underperforming
- 2: Needs Improvement
- 3: Fully meets
- 4: Exceeds



A violin plot can also give a quick indication for the distribution of performances across departments.

Most employees fully met the criteria, while some exceeded expectations. A few of them needed improvement.

Performance across Departments

Histograms for the performances across various departments were plotted individually.

The histograms indicate that most employees fully met the criteria.

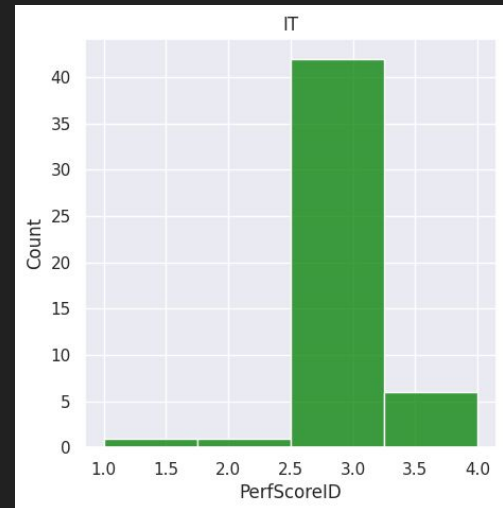
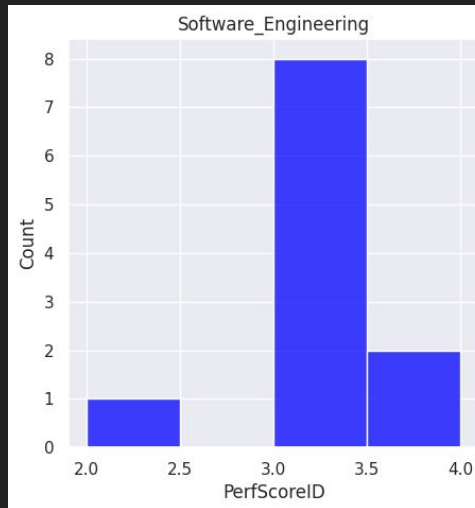
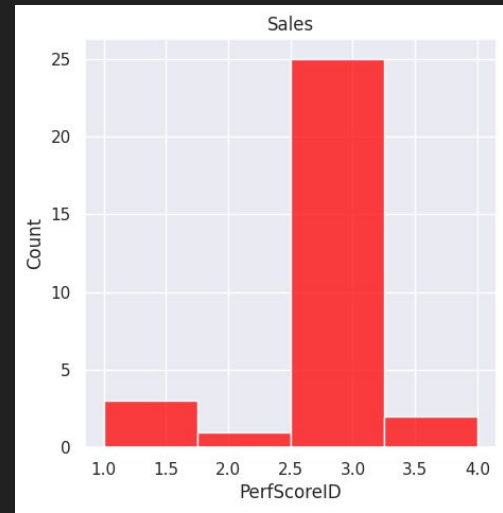
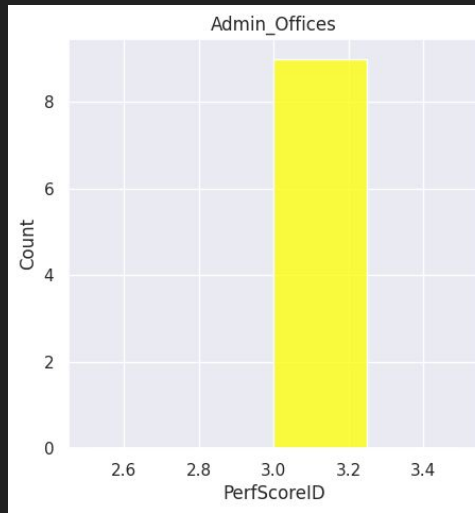
The department that had the most people requiring more attention belonged to sales.

The departments performing best were Software engineering and IT/IS.

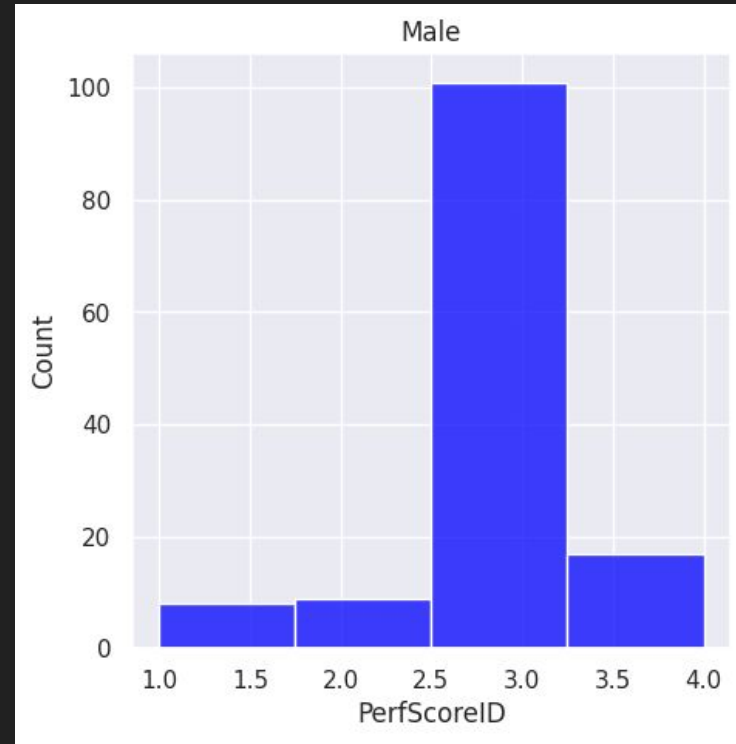
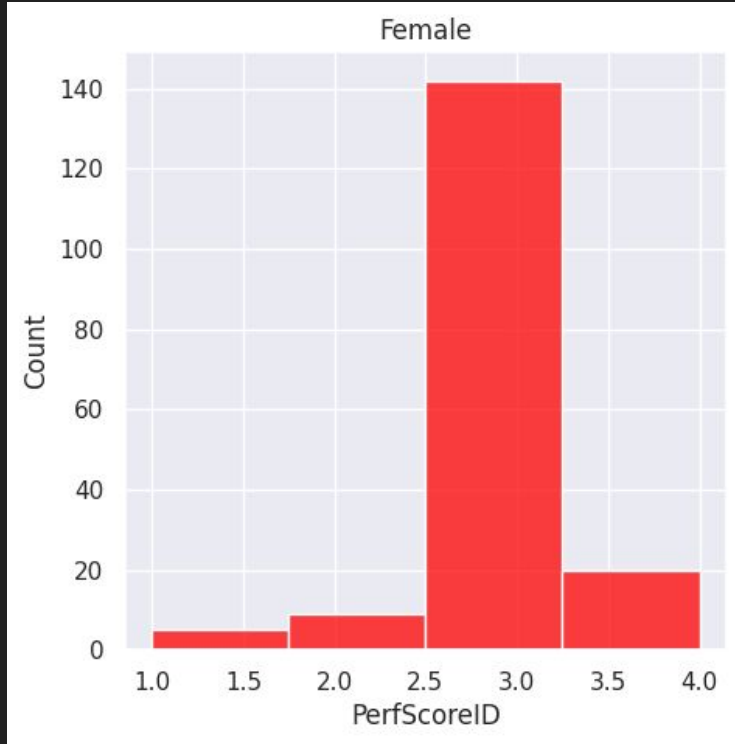
2: Needs improvement

3: Fully meets

4: Exceeds



Performance across Genders



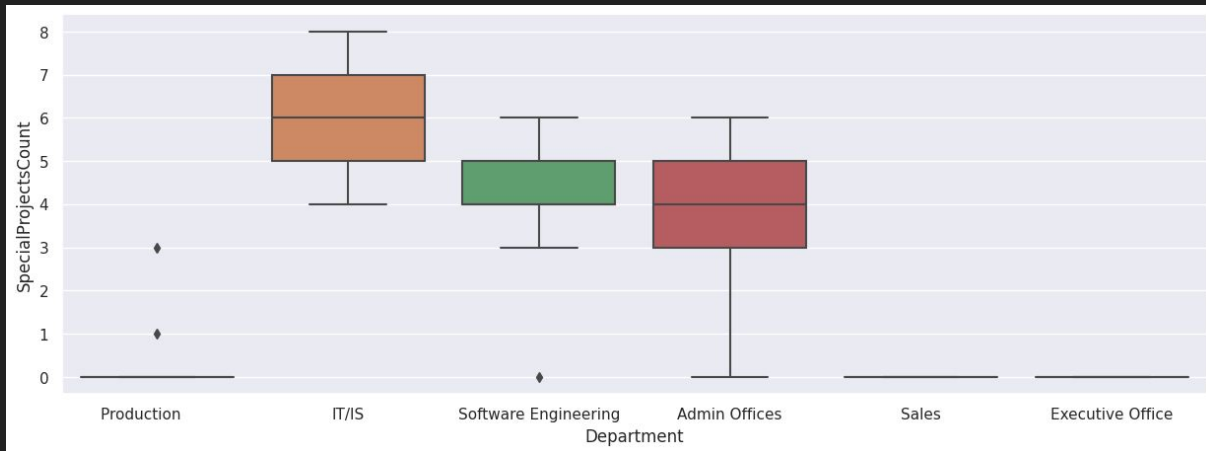
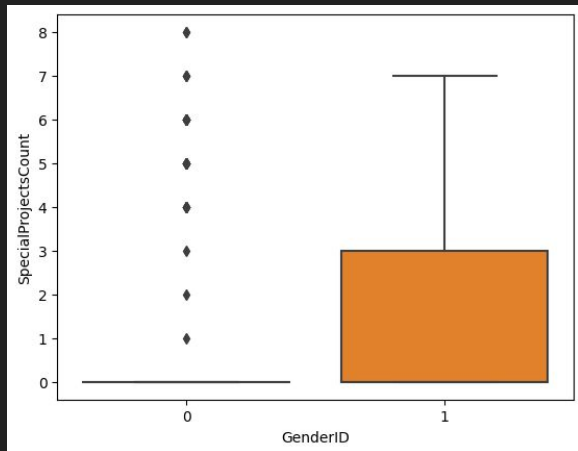
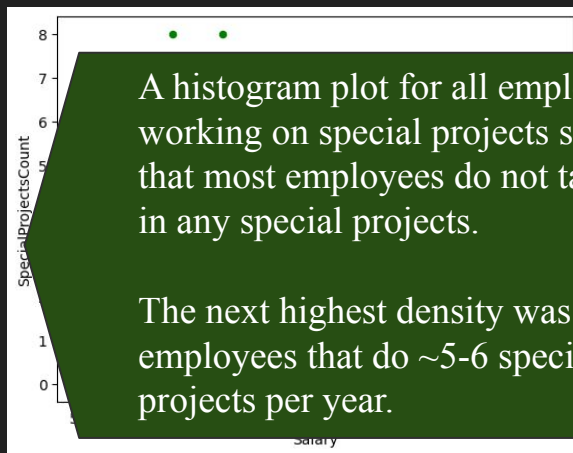
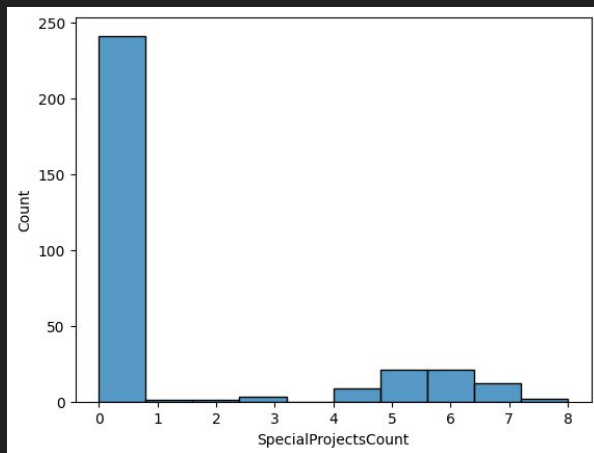
A similar analysis across genders demonstrated that both genders performed relatively equally well with more men requiring attention to meet job criteria.

Performance across Managers



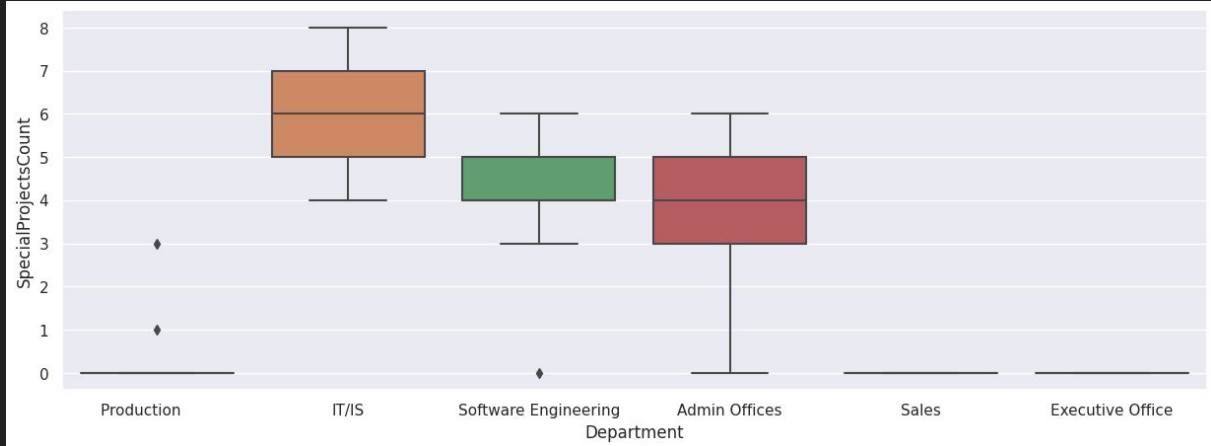
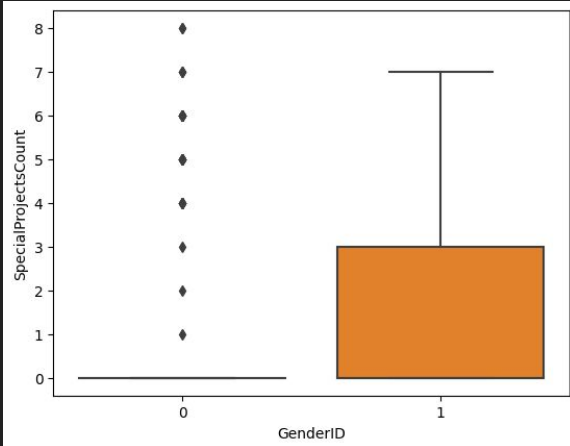
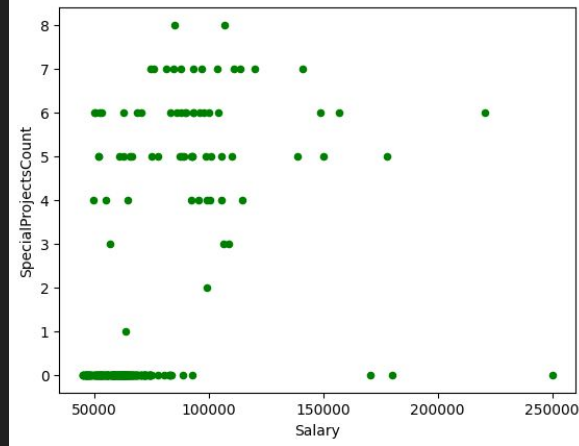
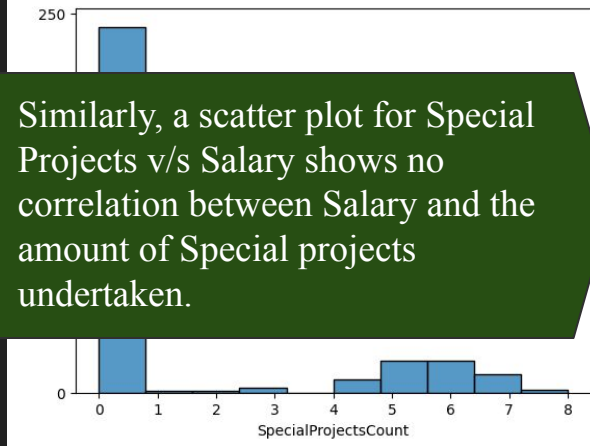
Box plots across all managers showed that Brannon Miller, Jennifer Zamora, Eric Dougall and Debra Houlihan had the largest groups. Among these, Jennifer's team performed the best and Debra's team performed the worst.

Special Projects analysis

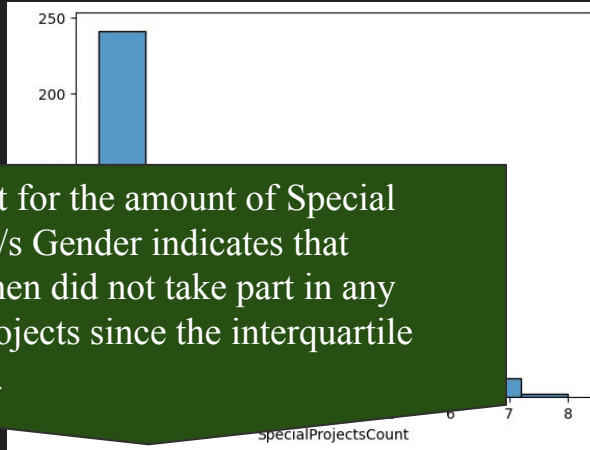


Special Projects analysis

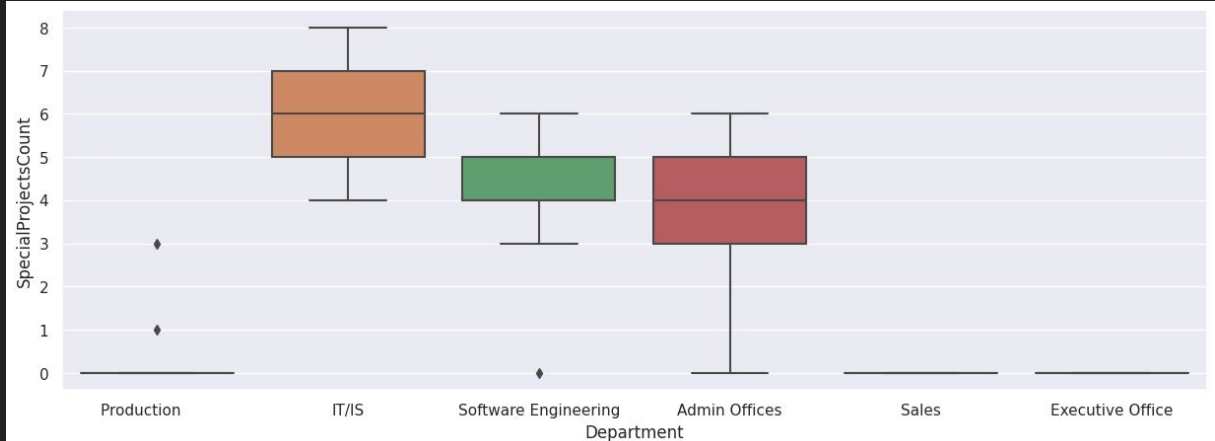
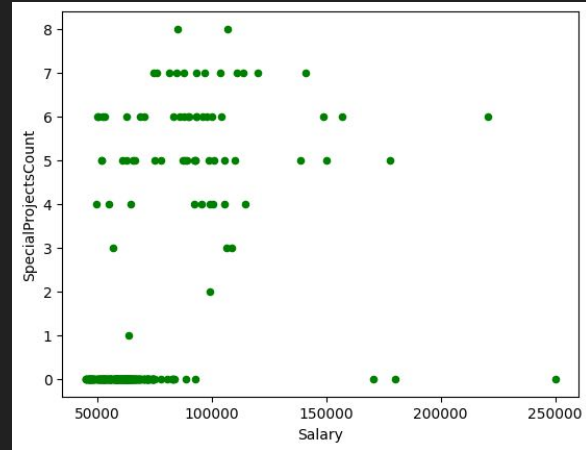
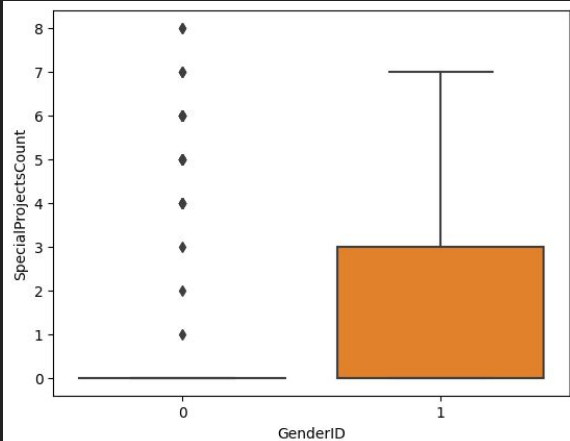
Similarly, a scatter plot for Special Projects v/s Salary shows no correlation between Salary and the amount of Special projects undertaken.



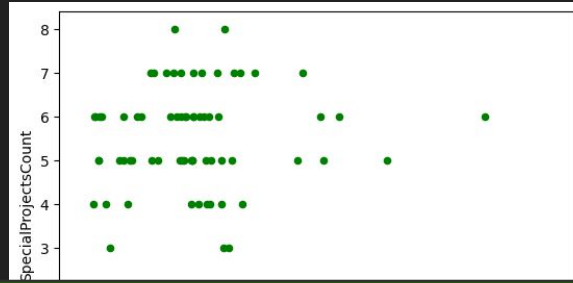
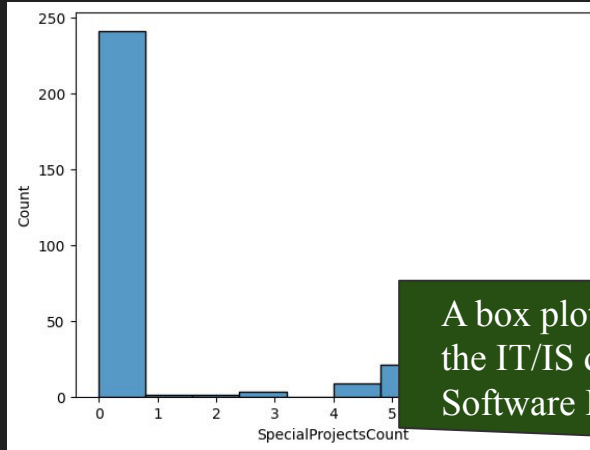
Special Projects analysis



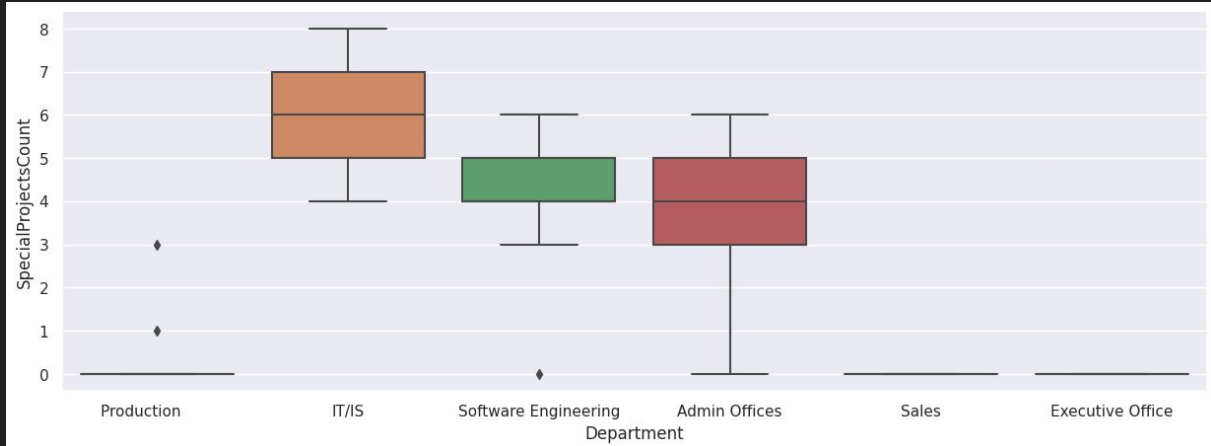
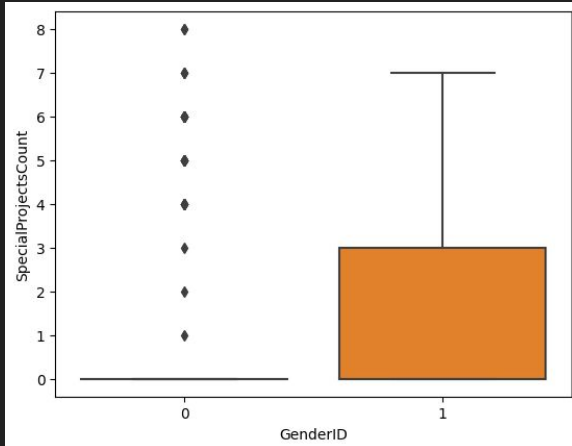
A box plot for the amount of Special Projects v/s Gender indicates that most women did not take part in any special projects since the interquartile range is 0.



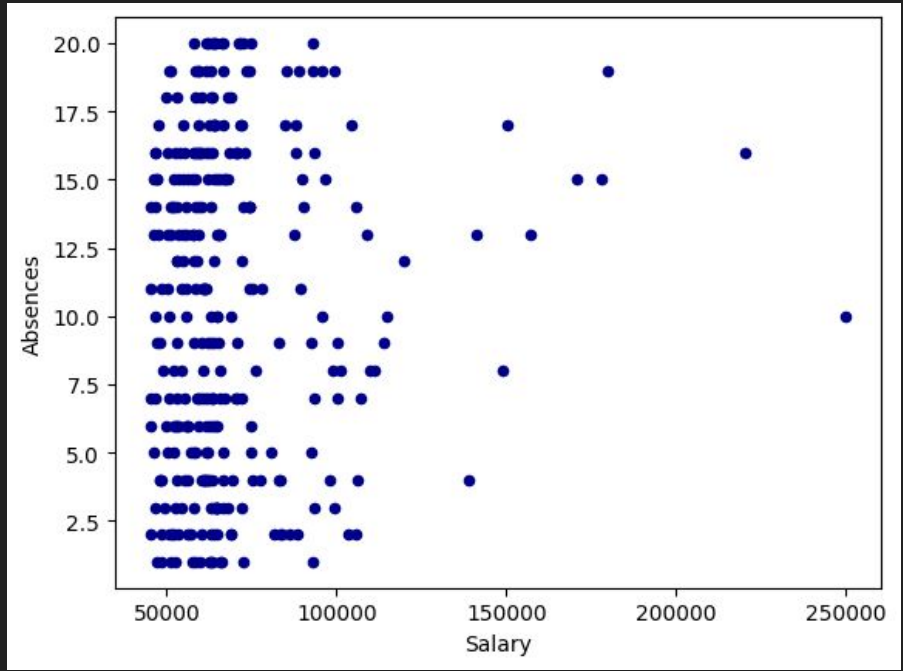
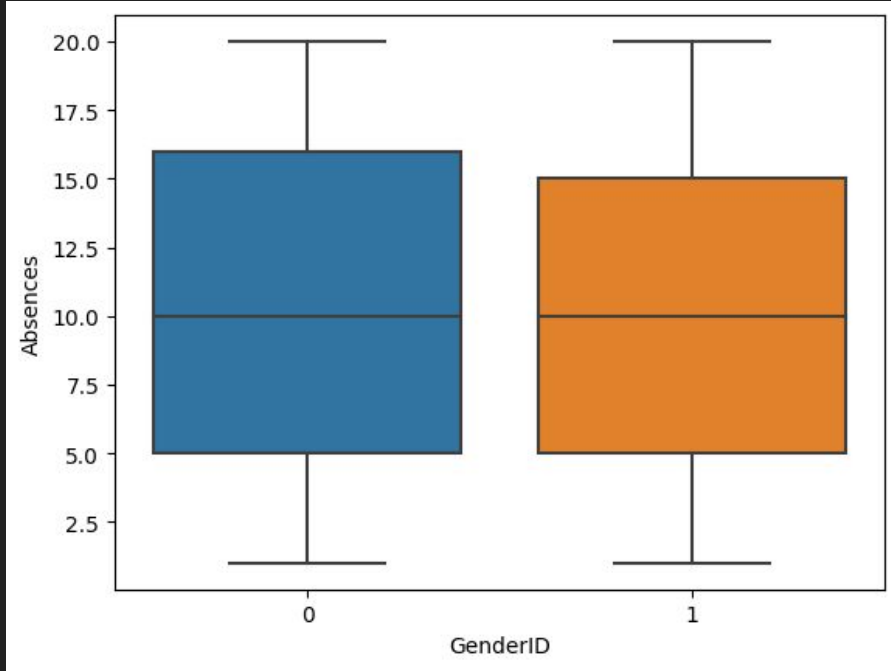
Special Projects analysis



A box plot for the amount of Special Projects v/s Departments showed that the IT/IS department had the most special projects per year followed by Software Engineering and then Admin Offices.

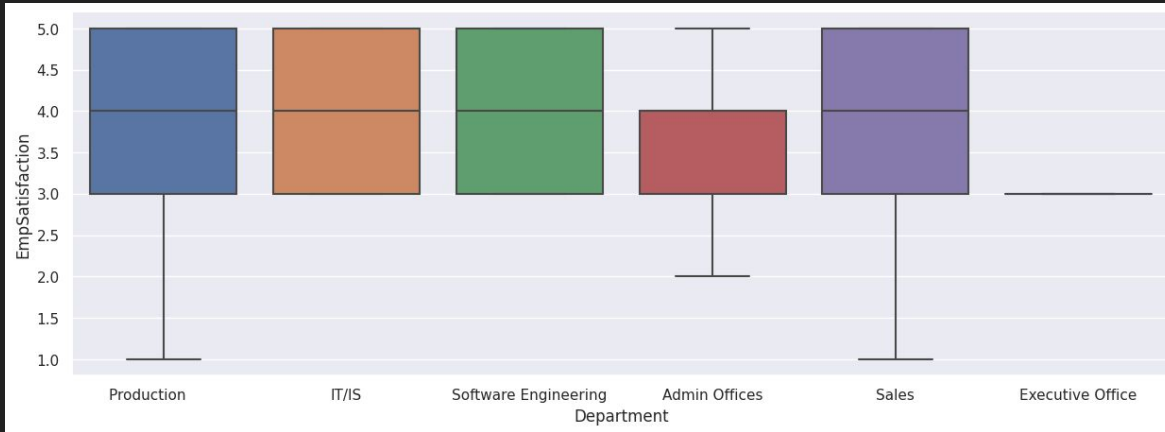


Absence analysis

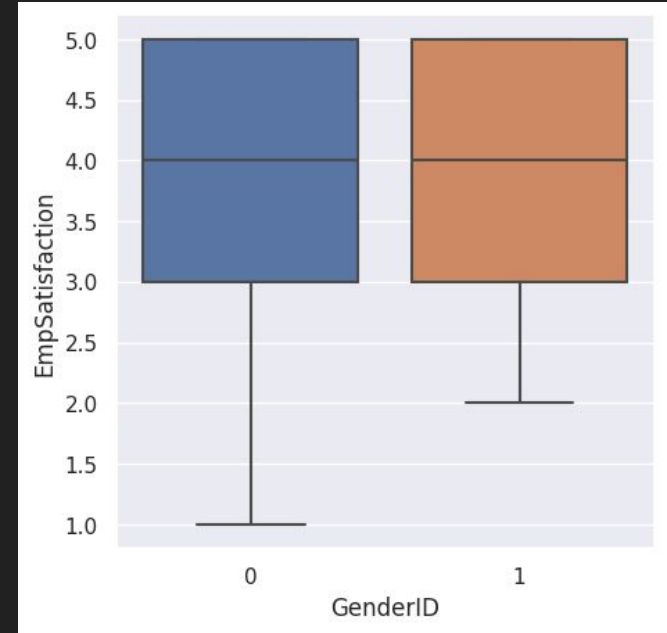


Similarly box plots for Gender and Salary v/s Absence were created and showed no correlation between Gender / Salary and the amount of Absences.

Employee Satisfaction

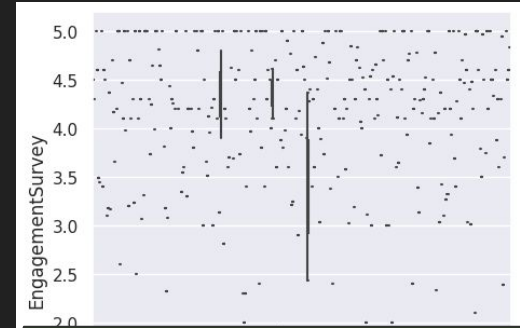
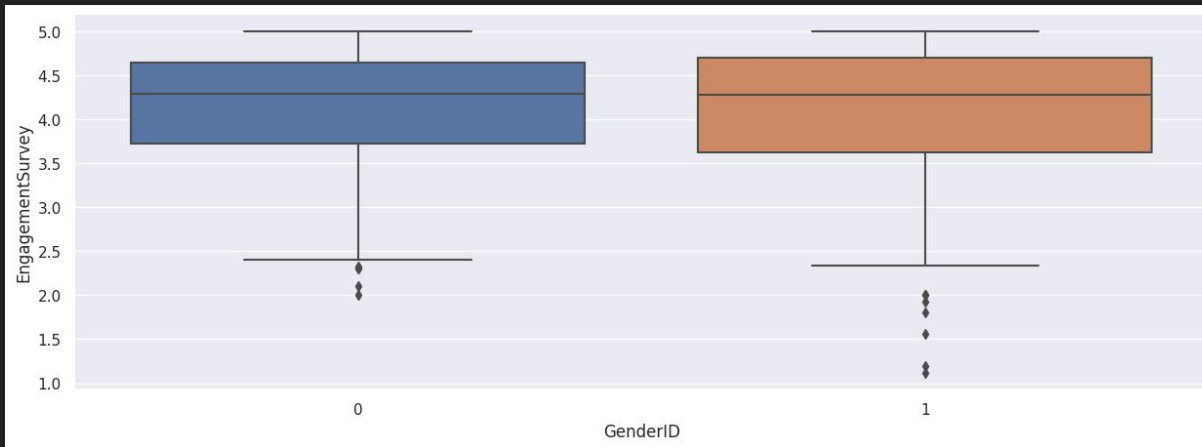
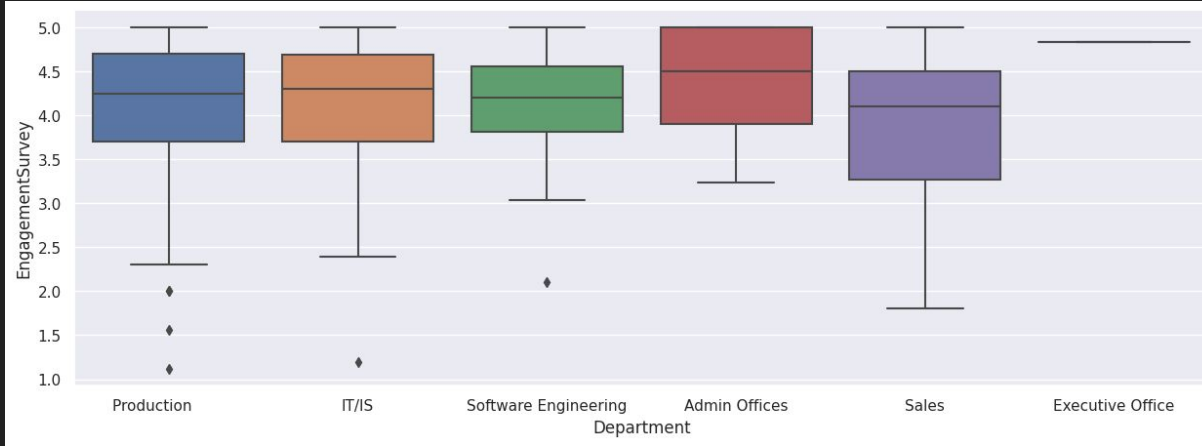


Box plots for Department and Salary v/s Employee Satisfaction also showed that most employees were satisfied with their job / wage irrespective of Department or Salary.

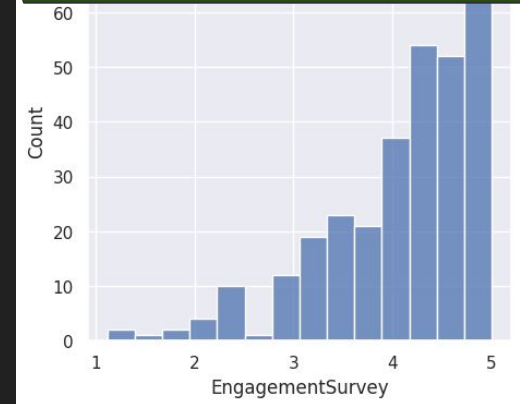


This was also true across Genders.

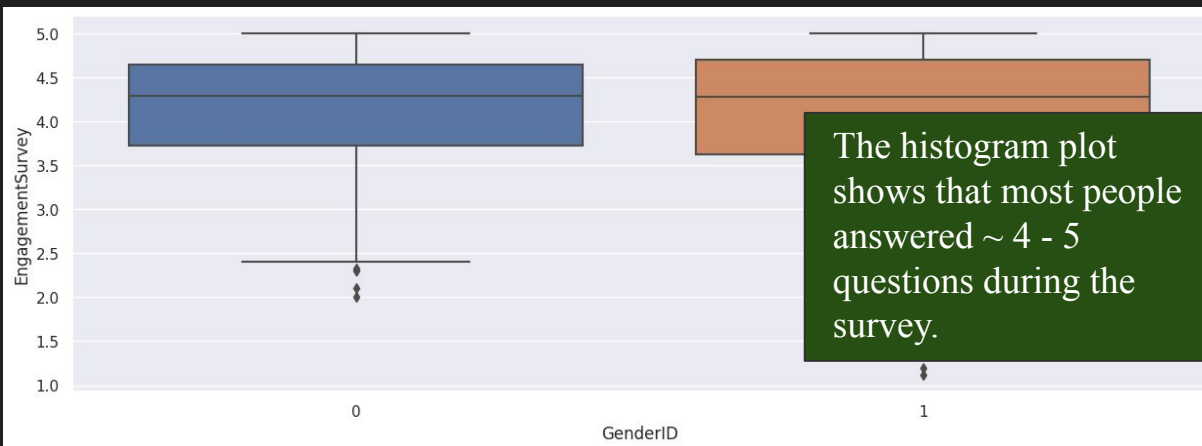
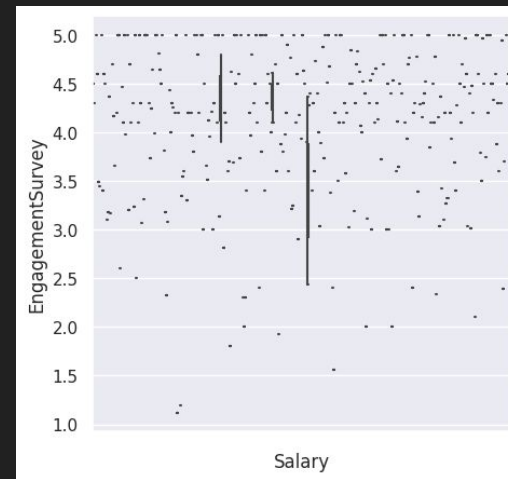
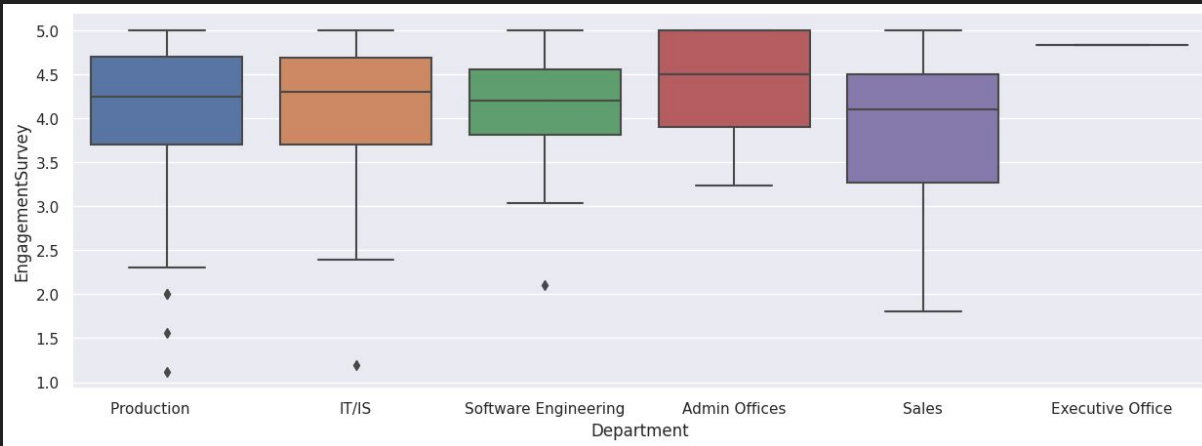
Engagement Survey



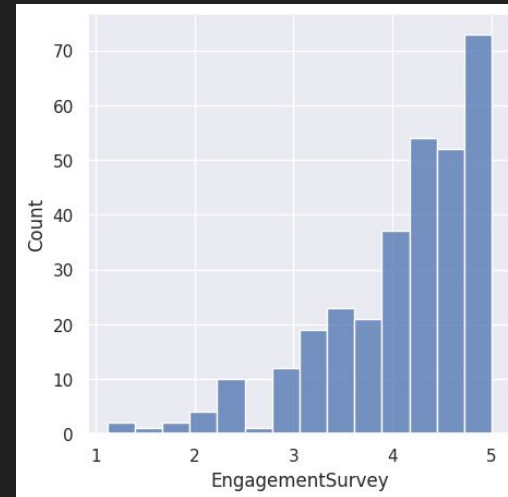
Box plots for Department and Gender v/s Engagement during a survey also showed that most employees agreed to participate in the survey irrespective of Department or Gender.



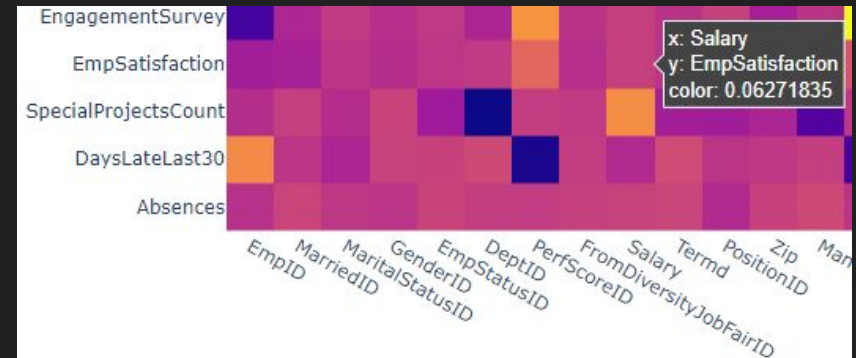
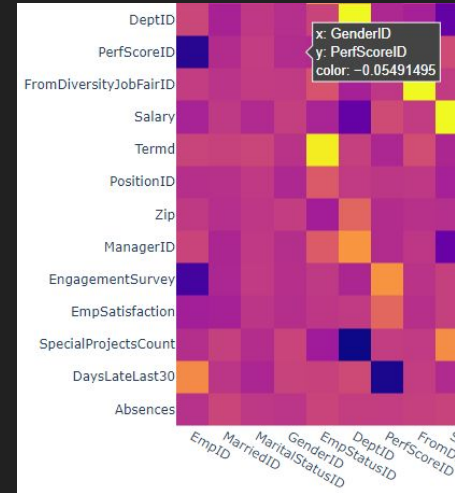
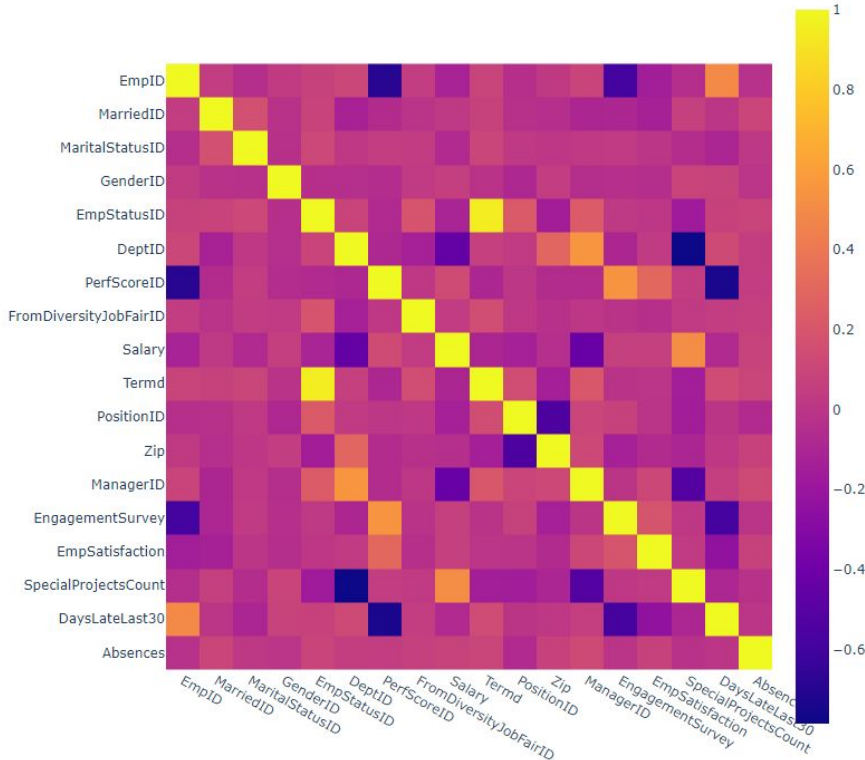
Engagement Survey



The histogram plot shows that most people answered ~ 4 - 5 questions during the survey.



Heat Map



A heat map shows the correlation of multiple values in a database w.r.t each other.

Inferences like

- 1) if the Gender correlates with Performance or
- 2) If the Salary correlates with Employee satisfaction

can be easily visualized using a heat map.

A value closer to zero such as those shown in the figures aside indicate minimal correlation which is ideal if Performance / Employee Satisfaction is to be maximized across different employees.

Conclusion

- The company shows equal inclusion for both genders across departments and pays them a very similar salary.
- Employee satisfaction levels and engagement levels during a survey are quite high.
- There is no trend in terms of the number of absences across the salary range.
- IT/IS and Software teams performed the best while the Sales department needed a little more attention.
- Salaries for Software engineering and IT were the highest, followed by production and administration.