Data Analysis For Employee Churn Prediction

Statistical Analysis Report

Business Problem

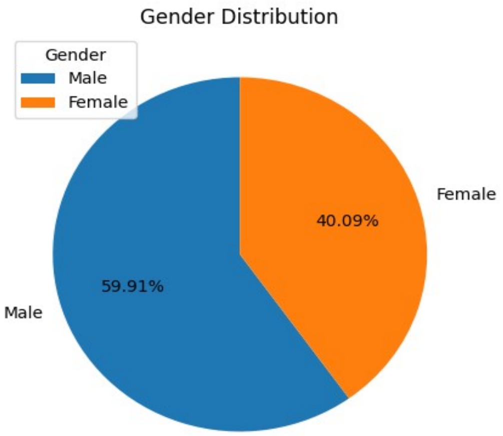
The goal is to analyze how many employee's are leaving the company, and what reasons why this might be the case and, determining the causes of employees leaving. Afterwards to create a machine learning model to predict which of the employees who have not left are likely to leave.

Table of Contents

1. Demography

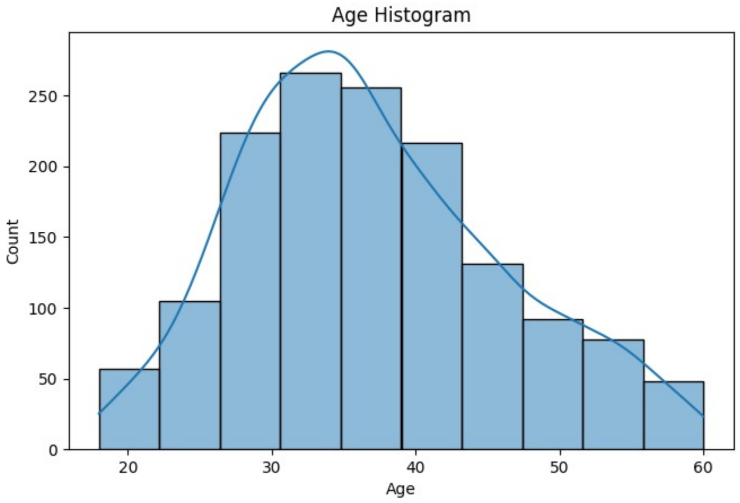
1.1 Gender

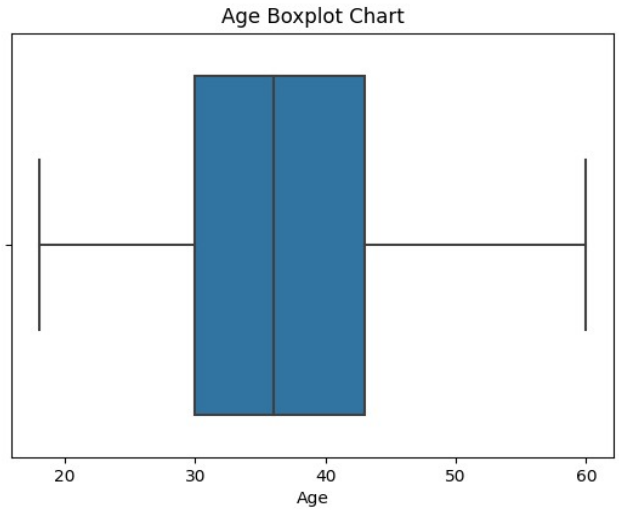
Gender	Count	Ratio%
Male	883.00	59.91
Female	591.00	40.09
Total	**1474.00**	**100.00**



1.2 Age

	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Age	1474.00	**36.95**	9.15	18.00	30.00	**36.00**	43.00	60.00





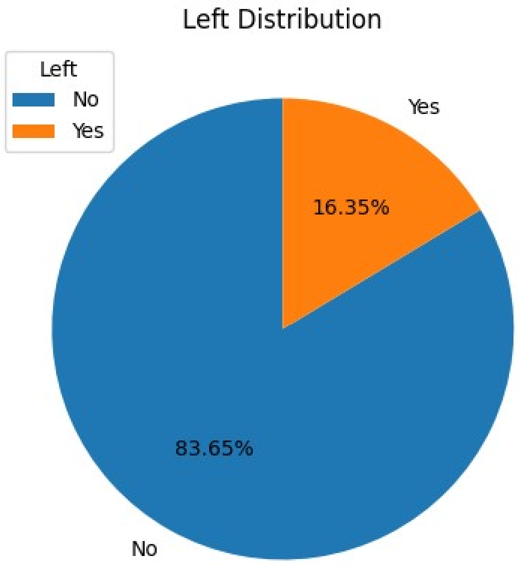
1.3 Gender vs Age

		Age				
		Count	**Mean**	**Median**	**Min**	**Max**
Gender	**Female**	591.00	**37.35**	36.00	18.00	60.00
	Male	883.00	**36.67**	35.00	18.00	60.00

Mann-Whitney U test (Test Statistic = 271923.00, p-value = 0.16951) p > 0.05
With 95% reliability, there is no significant difference between gender averages.

1.4 Left

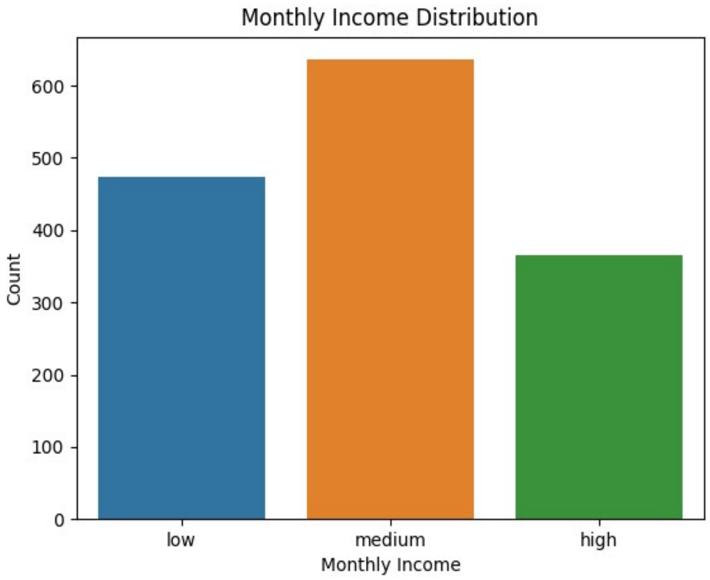
Left	Count	Ratio%
No	1233.00	83.65
Yes	241.00	16.35
Total	**1474.00**	**100.00**



The total number of Left is 241 and the rate is 16.35%

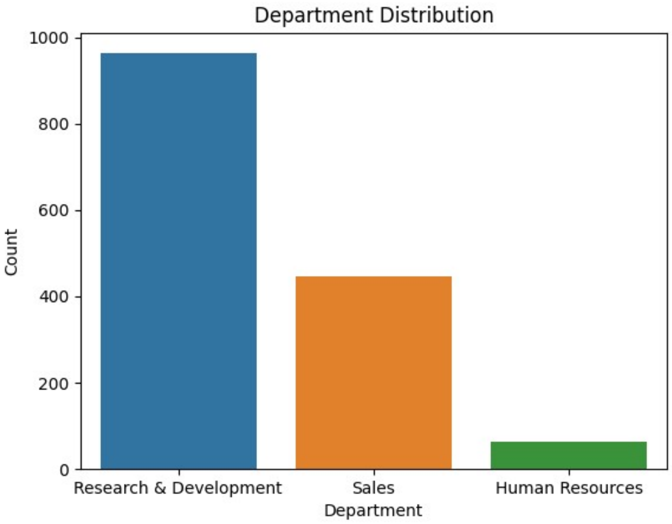
1.5 Monthly Income

Monthly Income	Count	Ratio%
medium	636.00	43.15
low	473.00	32.09
high	365.00	24.76
Total	**1474.00**	**100.00**



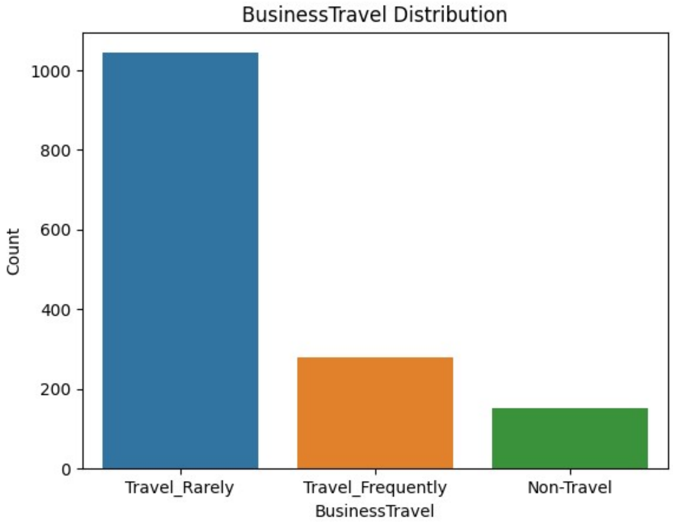
1.6 Department

Department	Count	Ratio%
Research & Development	963.00	65.33
Sales	447.00	30.33
Human Resources	64.00	4.34
Total	**1474.00**	**100.00**



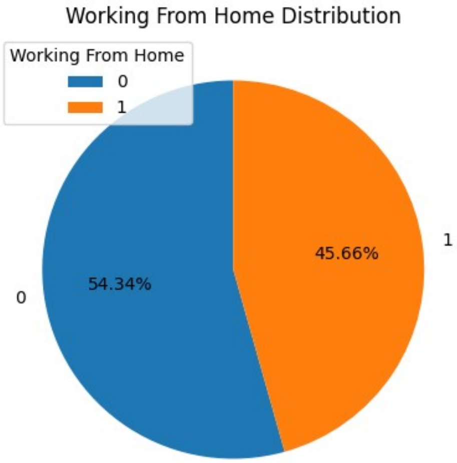
1.7 Business Travel

Business Travel	Count	Ratio%
Travel_Rarely	1044.00	70.83
Travel_Frequently	278.00	18.86
Non-Travel	152.00	10.31
Total	**1474.00**	**100.00**



1.8 Working From Home

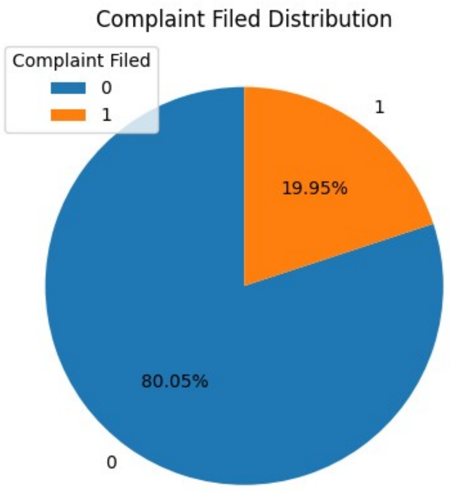
Working From Home	Count	Ratio%
0 (No)	801.00	54.34
1 (Yes)	673.00	45.66
Total	**1474.00**	**100.00**



1.9 Complaint Filed

Complaint Filed	Count	Ratio%
0 (No)	1180.00	80.05
1 (Yes)	**294.00**	19.95
Total	**1474.00**	**100.00**

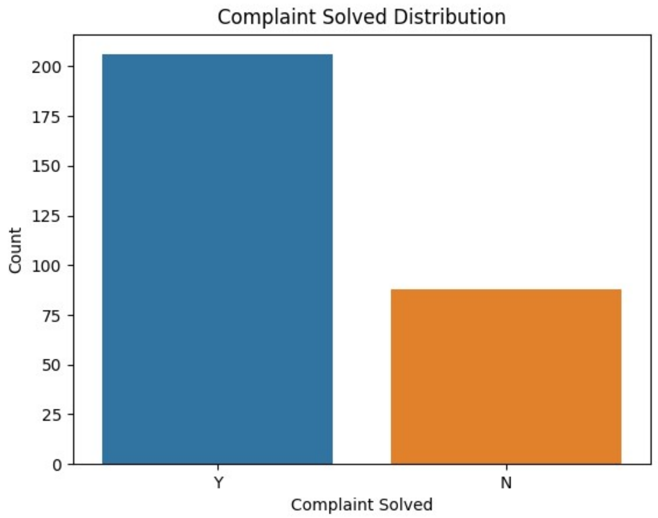
Number of complaint filed is 294.



1.10 Complaint Resolved

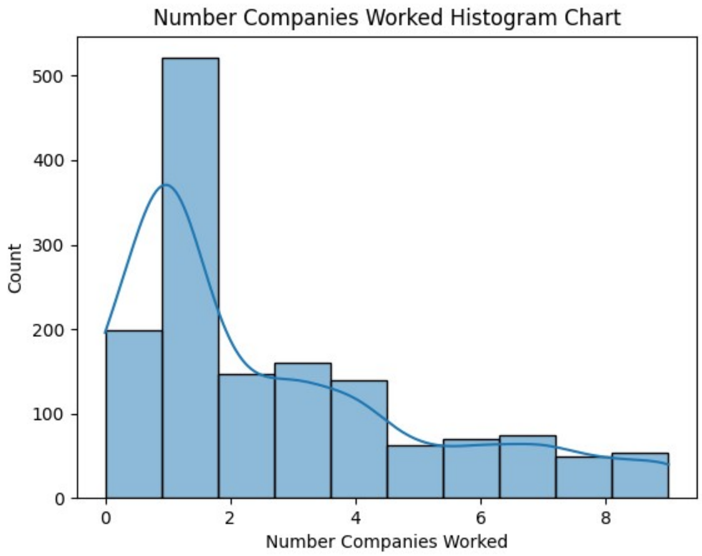
Complaint Resolved	Count	Ratio%
Yes (Y)	206.00	13.98
No (N)	88.00	5.97
Missing Value	1180.00	80.05
Total	**294.00**	**19.95**

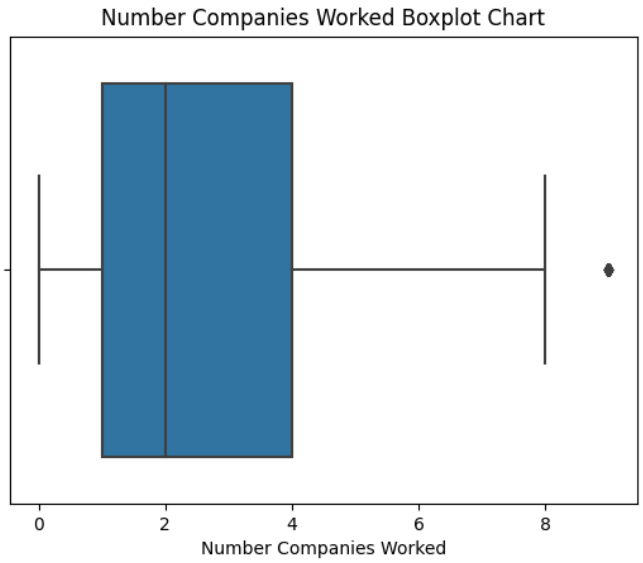
206 out of 294 complaints have been resolved. 88 complaints were not resolved.



1.11 Number of Companies Worked

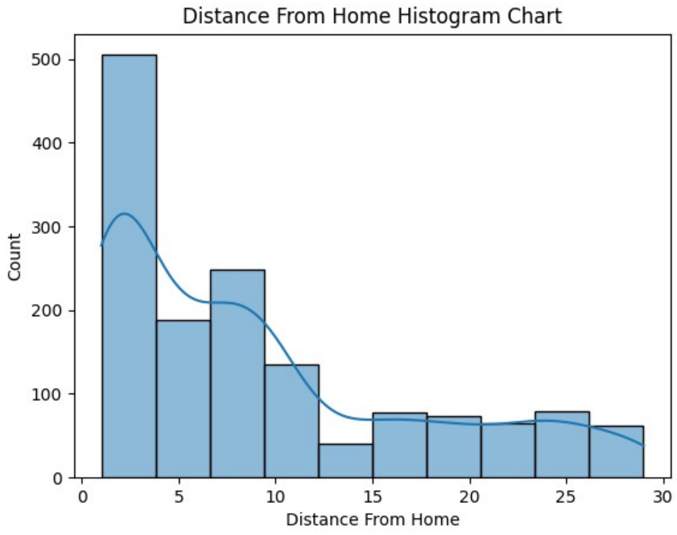
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Number of Companies Worked	1474.00	**2.70**	2.50	0.00	1.00	2.00	4.00	9.00

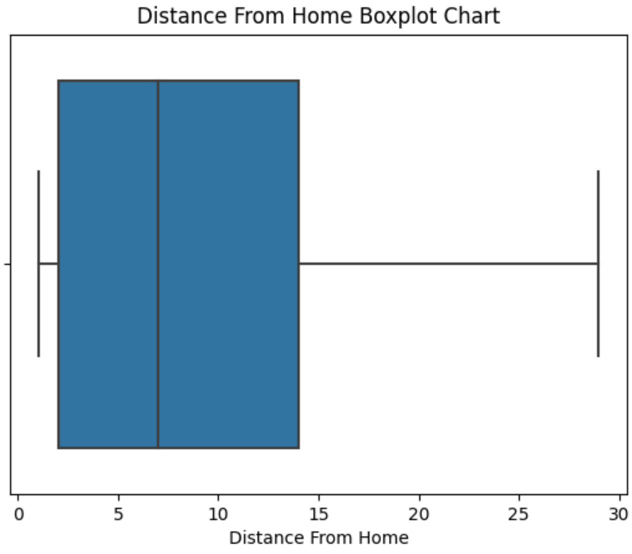




1.12 Distance From Home

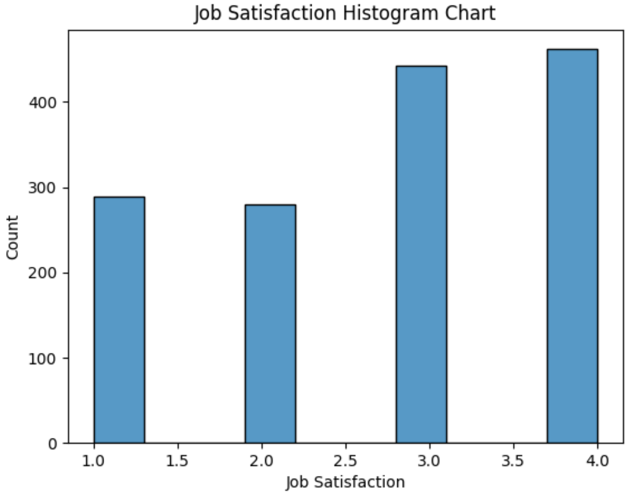
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Distance From Home	1474.00	**9.20**	8.12	1.00	2.00	7.00	14.00	29.00

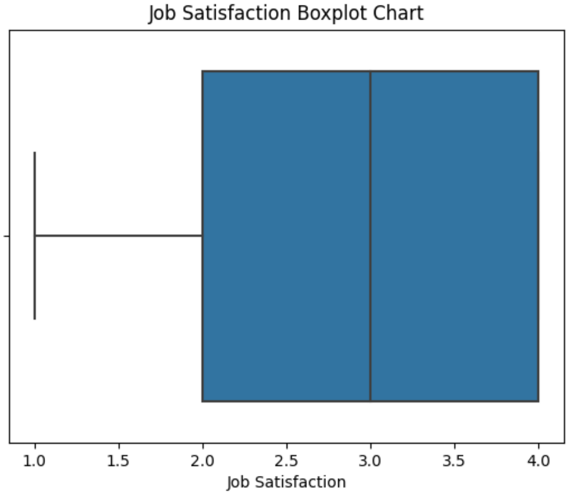




1.13 Job Satisfaction

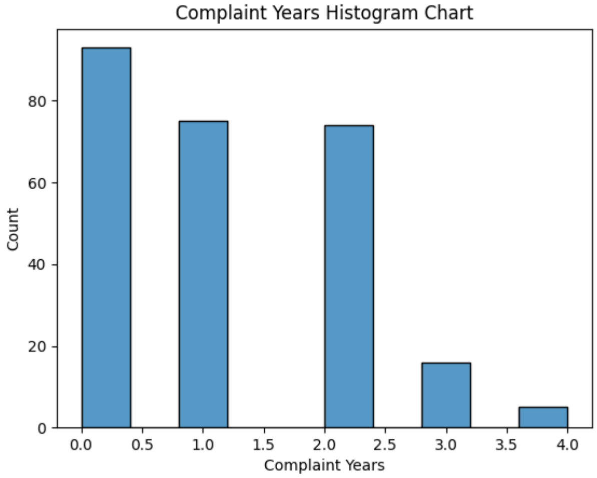
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Job Satisfaction	1474.00	**2.73**	1.10	1.00	2.00	3.00	4.00	4.00

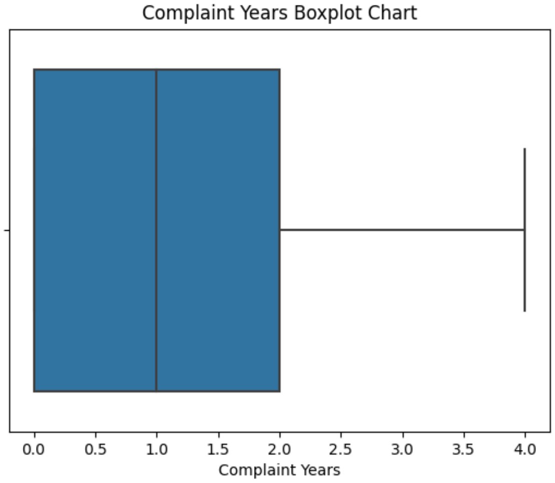




1.14 Complaint Years

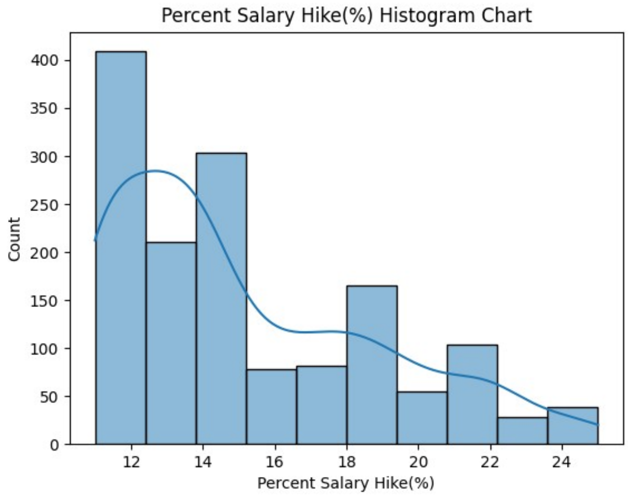
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Complaint Years	263.00	**1.11**	1.02	0.00	0.00	1.00	2.00	4.00

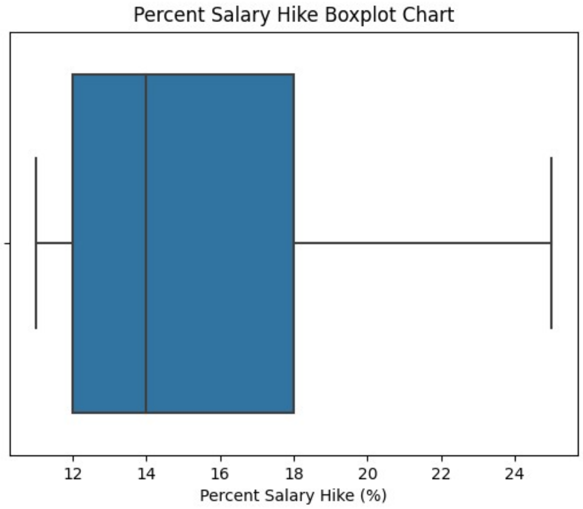




1.15 Percent Salary Hike

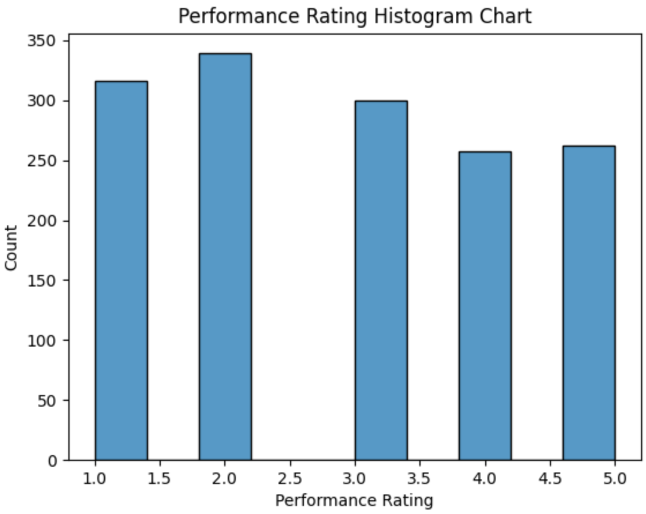
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Percent Salary Hike	1474.00	**15.20**	3.66	11.00	12.00	14.00	18.00	25.00

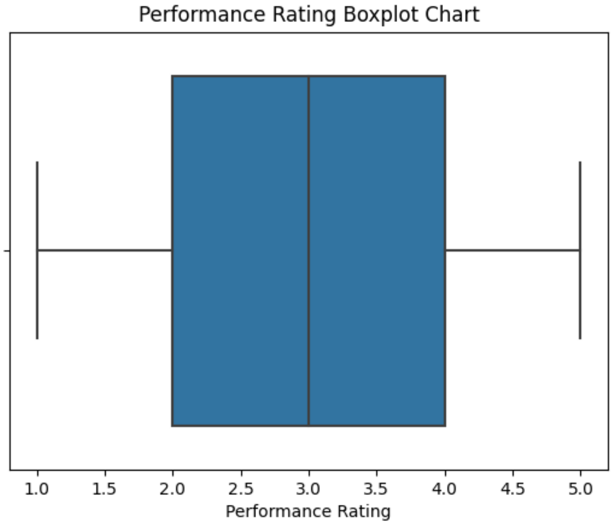




1.16 Performance Rating

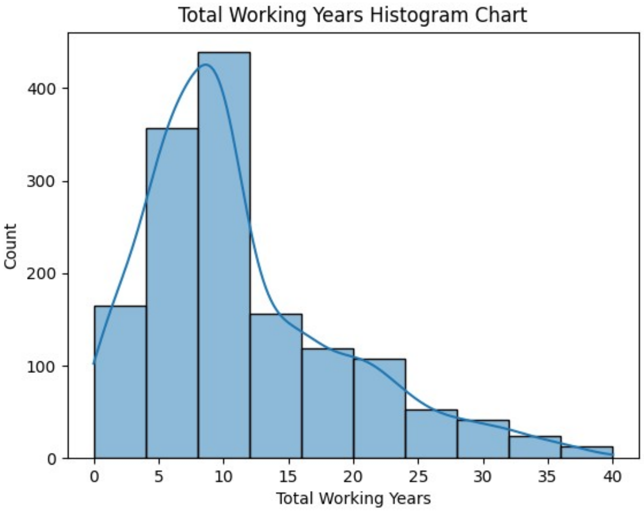
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Performance Rating	1474.00	**2.87**	1.40	1.00	2.00	3.00	4.00	5.00

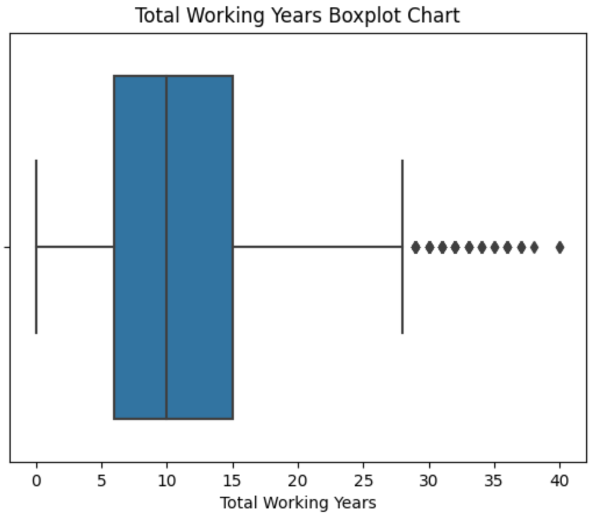




1.17 Total Working Years

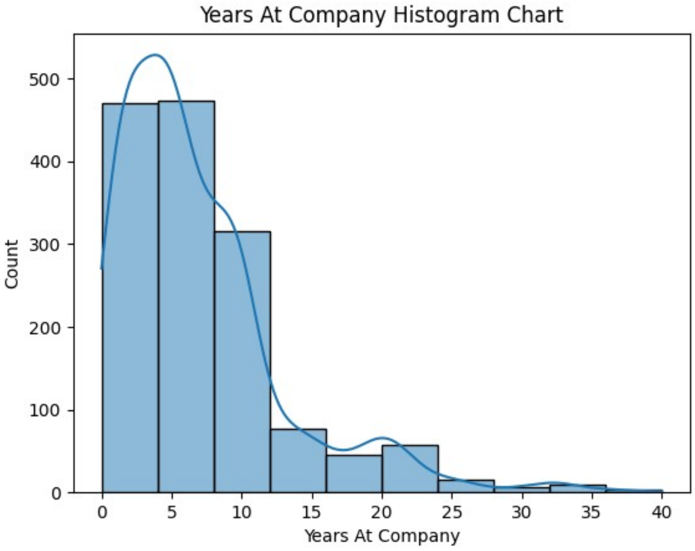
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Total Working Years	1474.00	**11.29**	7.79	0.00	6.00	10.00	15.00	40.00

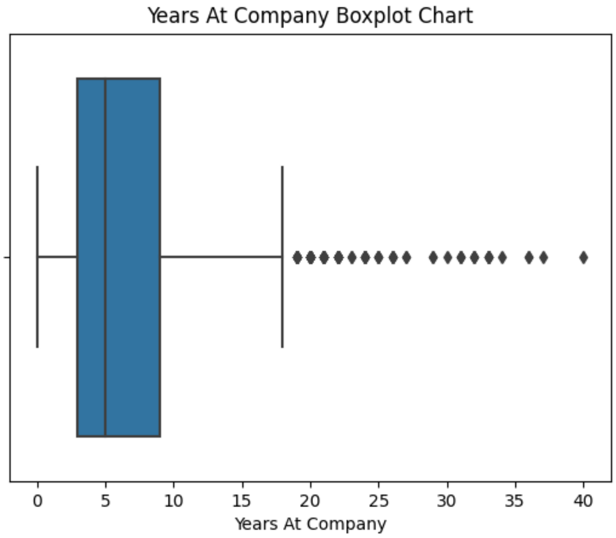




1.18 Years At Company

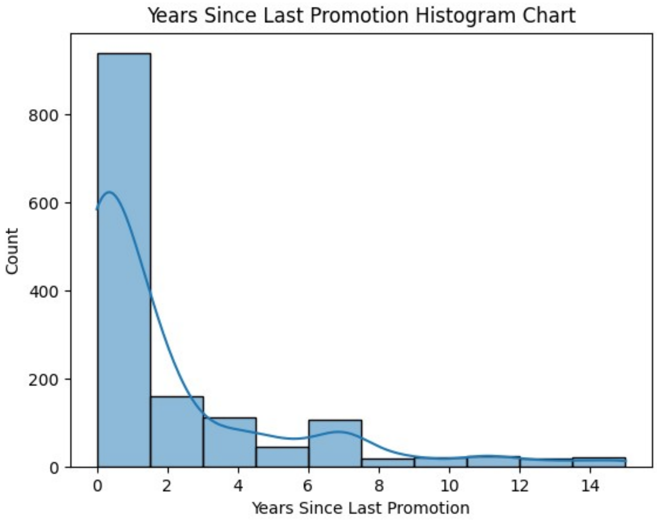
	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Years At Company	1474.00	**7.01**	6.12	0.00	3.00	5.00	9.00	40.00

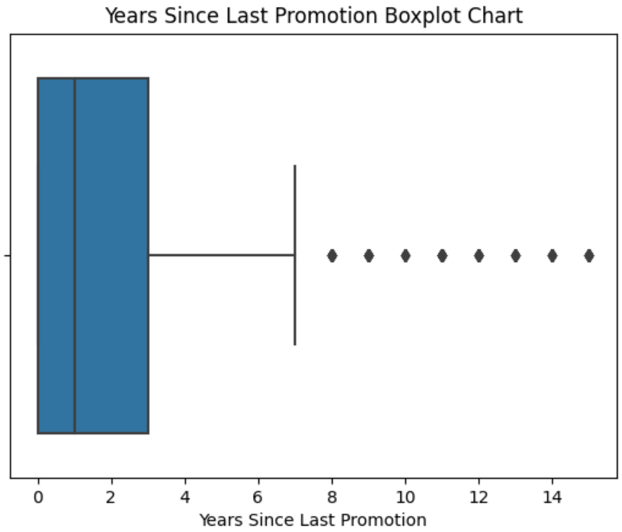




1.19 Years Since Last Promotion

	Count	Mean	Std.Dev.	Min	%25	%50	%75	Max
Years Since Last Promotion	1474.00	**2.19**	3.22	0.00	0.00	1.00	3.00	15.00





Analysis of Target Variable (Left)

2.1. Left & Age

		Age(Mean)
Left	**No**	37.56
	Yes	**33.79**

Mann-Whitney U test (Test Statistic = 187058.5000, p-value = 1.8753305644285697e-10) p < 0.05
With 95% reliability, there is a significant difference between the age mean of those who Left and those who did not Left.

Gender	Left	Age(Mean)
Female	**No**	38.16
	Yes	**32.88**
Male	**No**	37.15
	Yes	**34.34**

We observe that the average age of leavers is lower than non-leavers. In addition, avarage age of company employees : **36.95 ,** average age of female : **37.35 ,** average age of male : **36.67 .**

2.2. Left & Distance From Home

	Left	Distance From Home(Mean)
Left	**No**	8.92
	Yes	**10.66**

Mann-Whitney U test (Test Statistic = 130349.5000, p-value = 0.00247) p < 0.05
With 95% reliability, there is a significant difference between the mean Distance From Home of those who Left and those who did not Left.
We observe that those who left their jobs have a high avarage distance from their workplace.

2.3. Left & Total Working Years

		Total Working Years(Mean)
Left	**No**	11.86
	Yes	8.36

Mann-Whitney U test (Test Statistic = 187058.5000, p-value = 1.061100903516452e-13) p < 0.05
With 95% reliability, there is a significant difference between the mean Total Working Years of those who Left and those who did not Left.
We observe that those who left their jobs have a low avarage Total Working Years ..

2.4. Left & Years At Company

		Years At Company(Mean)
Left	No	7.37
	Yes	5.17

Mann-Whitney U test (Test Statistic = 187058.5000, p-value = 6.803538201944542e-13) p < 0.05
With 95% reliability, there is a significant difference between the mean Years At Company of those who Left and those who did not Left.
We observe that those who left their jobs have a low avarage Years At Company.

2.5. Left & Years Since Last Promotion

		Years Since Last Promotion(Mean)
Left	No	2.23
	Yes	1.97

Mann-Whitney U test (Test Statistic = 159316.0000, p-value = 0.06435) p > 0.05
With 95% reliability, there is no significant difference between the mean Years Since Last Promotion of those who Left and those who did not Left.
We observe that those who left their jobs have a low avarage Years Since Last Promotion ,but the difference between the averages is not large.

2.6. Left & Percent Salary Hike

		Percent Salary Hike(Mean)
Left	No	15.23
	Yes	15.07

Mann-Whitney U test (Test Statistic = 154428.5000, p-value = 0.33002) p > 0.05
With 95% reliability, there is no significant difference between the mean Percent Salary Hike of those who Left and those who did not Left.

We observe that the average Percent Salary Hike for leavers is the same as for non-leavers.

2.7. Left & Performance Rating

		Performance Rating(Mean)
Left	No	2.88
	Yes	2.83

Mann-Whitney U test (Test Statistic = 151975.0000, p-value = 0.56580) p > 0.05
With 95% reliability, there is no significant difference between the mean Performance Rating of those who Left and those who did not Left.
We observe that the average Performance Rating for leavers is the same as for non-leavers.

3. Measures of Association

Pearson’s Chi-Square is a statistical hypothesis test for independence between categorical variables.

3.1 Left & Gender

```
crosstab =pd.crosstab(df["Left"],df["Gender"])
```

Chi-Square Test, Test Statistic = 0.78, p-value = 0.37844

Independent (H0 holds true) , p > 0.05

With 95% reliability, that is our variables (Left & Gender) do not have a significant relation.

			Left		
			No (0)	**Yes(1)**	**Total**
Gender	**Female**	**Count**	501.00	90.00	**591**
		Ratio%	84.77	15.23	**100**
	Male	**Count**	732.00	151.00	**883**
		Ratio%	82.90	17.10	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

3.2 Left & Complaint Filed

```
crosstab =pd.crosstab(df["Left"],df["complaintfiled"])
```

Chi-Square Test, Test Statistic = 1.34, p-value = 0.24692

Independent (H0 holds true) , p > 0.05

With 95% reliability, that is our variables (Left & Complaint Filed) do not have a significant relation.

			Left		
			No (0)	**Yes(1)**	**Total**
Complaint Filed	**0 (No)**	**Count**	980.00	200.00	**1180**
		Ratio%	83.05	16.95	**100**
	1 (Yes)	**Count**	253.00	41.00	**294**
		Ratio%	86.05	13.95	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

3.3 Left & MonthlyIncome

crosstab =pd.crosstab(df["Left"],df["MonthlyIncome"])

Chi-Square Test, Test Statistic = 58.58, p-value = 1.902504899634253e-13

Dependent (reject H0) , p < 0.05 - Cramer's V = 0.20

With 95% reliability, that is our variables (Left & Monthly Income) have a significant relation.

			Left		
			No (0)	**Yes(1)**	**Total**
Monthly Income	**low**	**Count**	345.00	128.00	**473**
		Ratio%	72.94	**27.06**	**100**
	medium	**Count**	562.00	74.00	**636**
		Ratio%	88.36	11.64	**100**
	high	**Count**	326.00	39.00	**365**
		Ratio%	89.32	10.68	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

We observe that low-income earners are the most likely to Left at 27,06 %.

3.4 Left & Working From Home

crosstab =pd.crosstab(df["Left"],df["workingfromhome"])

Chi-Square Test, Test Statistic = 1.43, p-value = 0.23140

Independent (H0 holds true) , p < 0.05

With 95% reliability, that is our variables (Left & Working From Home) do not have a significant relation.

			Left		
			No (0)	**Yes(1)**	**Total**
Working From Home	**0 (No)**	**Count**	679.00	122.00	**801**
		Ratio%	84.77	**15.23**	**100**
	1 (Yes)	**Count**	554.00	119.00	**673**
		Ratio%	82.32	**17.68**	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

3.5 Left & Department

crosstab =pd.crosstab(df["Left"],df["Department"])

Chi-Square Test, Test Statistic = 11.05, p-value = 0.00399

Dependent (reject H0) , p < 0.05 - Cramer's V =   0.09

With 95% reliability, that is our variables (Left & Department) have a significant relation.

			Left		
			No (0)	**Yes(1)**	**Total**
Department	**Research & Development**	**Count**	828.00	135.00	**963**
		Ratio%	85.98	14.02	**100**
	Sales	**Count**	354.00	93.00	**447**
		Ratio%	79.19	**20.81**	**100**
	Human Resources	**Count**	51.00	13.00	**64**
		Ratio%	79.69	**20.31**	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

3.6 Left & Business Travel

crosstab =pd.crosstab(df["Left"],df["BusinessTravel"])

Chi-Square Test, Test Statistic = 22.83, p-value = 0.00001

Dependent (reject H0) , p < 0.05 - Cramer's V =   0.12

With 95% reliability, that is our variables (Left & Business Travel) have a significant relation.

			Left		
			No (0)	**Yes(1)**	**Total**
Business Travel	**Travel_Rarely**	**Count**	887.00	157.00	**1044**
		Ratio%	84.96	15.04	**100**
	Travel_Frequently	**Count**	208.00	70.00	**278**
		Ratio%	74.82	**25.18**	**100**
	Non-Travel	**Count**	138.00	14.00	**152**
		Ratio%	90.79	9.21	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

We observe that frequent travelers have the highest left rate 25.18%

3.7 Left & Complaint Resolved

crosstab =pd.crosstab(df["Left"],df["complaintresolved"])

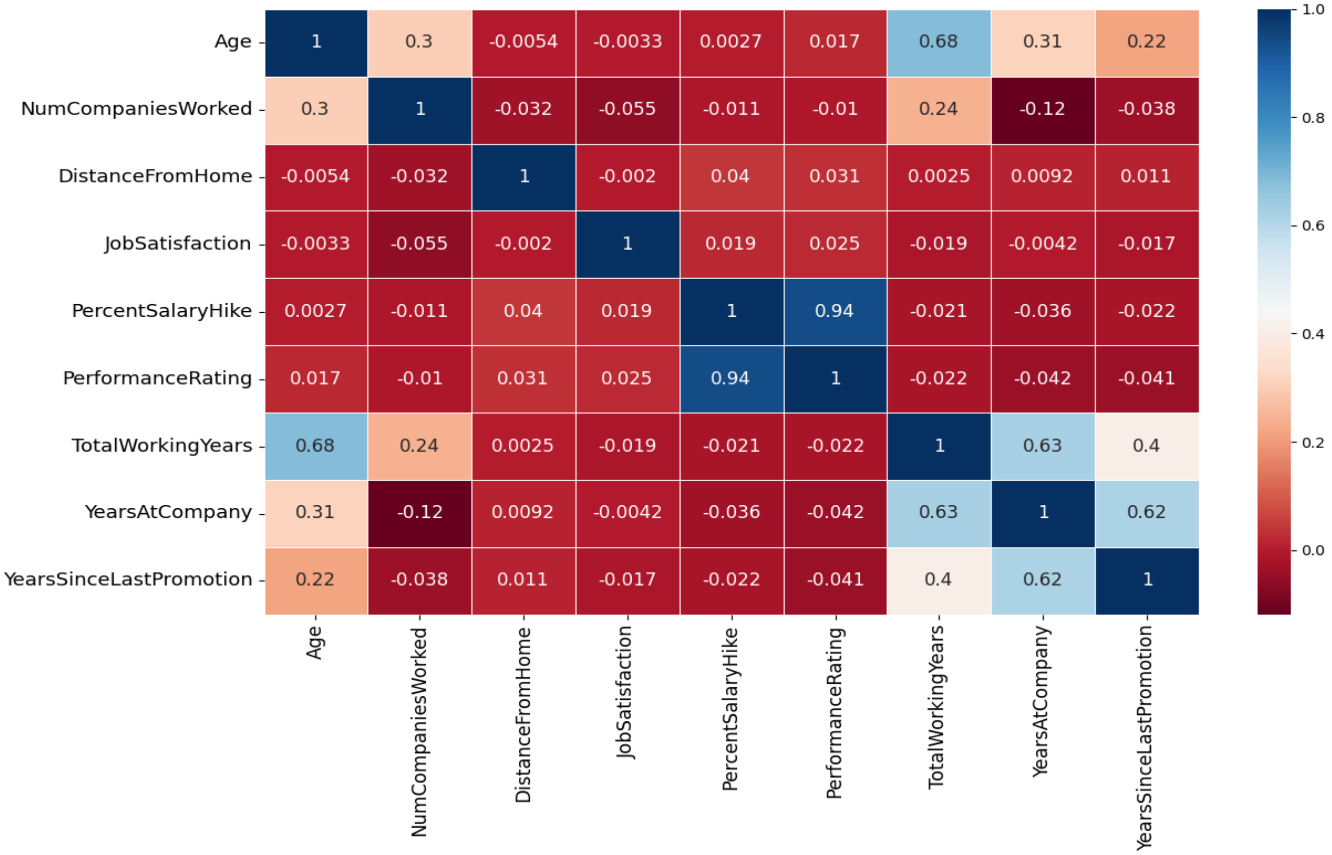
Chi-Square Test, Test Statistic = 4.20 , p-value = 0.12227 ,

Independent (H0 holds true) , p < 0.05

With 95% reliability, that is our variables (Left & Complaint Resolved) do not have a significant relation.

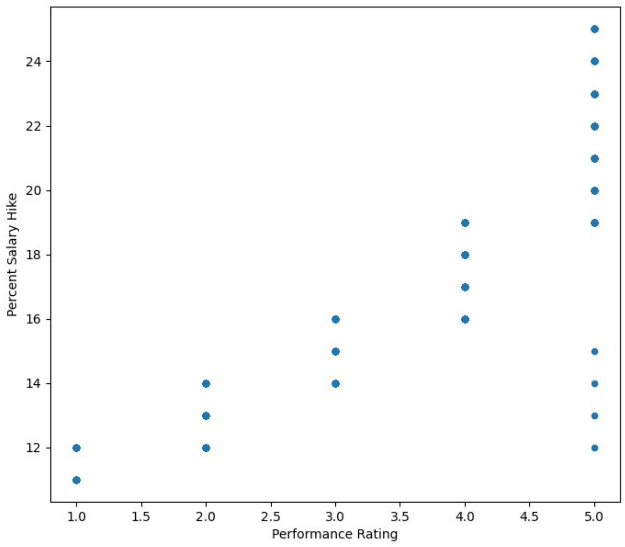
			Left		
			No (0)	**Yes(1)**	**Total**
Complaint Resolved	**Y (Yes)**	**Count**	182.00	24.00	**206**
		Ratio%	88.35	**11.65**	**100**
	N (No)	**Count**	71.00	17.00	**88**
		Ratio%	80.68	**19.32**	**100**
	missing	**Count**	980.00	200.00	**1180**
		Ratio%	83.05	16.95	**100**
	Total	**Count**	1233.00	241.00	**1474**
		Ratio%	83.65	16.35	**100**

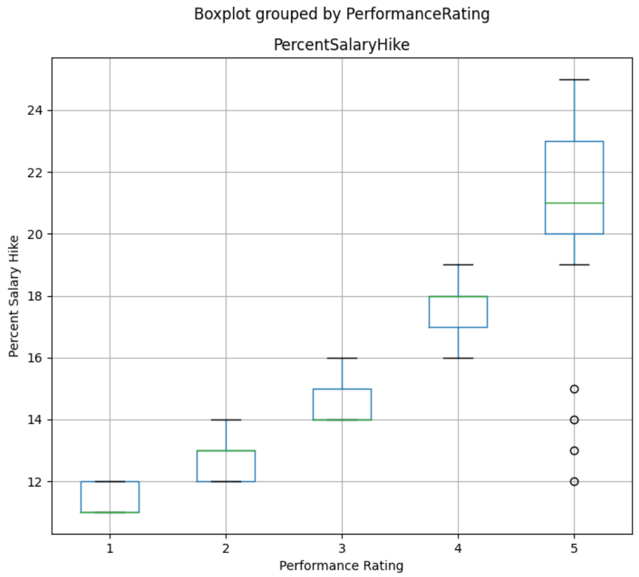
4. Analysis of Correlation



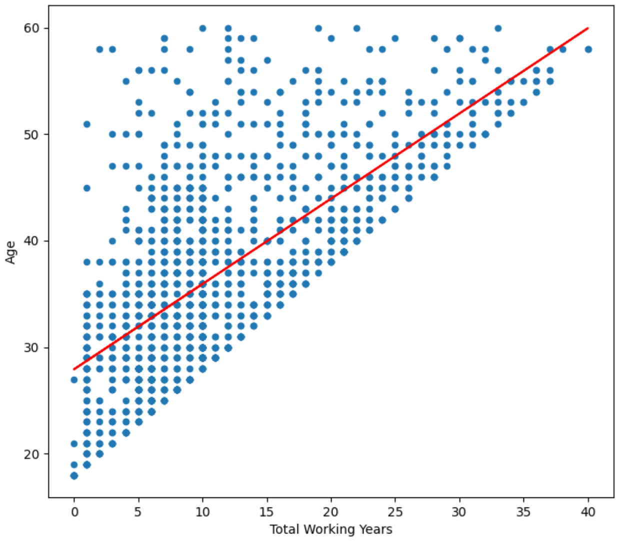
	Age	NumCompaniesWorked	DistanceFromHome	JobSatisfaction	PercentSalaryHike	PerformanceRating	TotalWorkingYears	YearsAtCompany	YearsSinceLastPromotion
Age	1	0.3	-0.0054	-0.0033	0.0027	0.017	0.68	0.31	0.22
NumCompaniesWorked	0.3	1	-0.032	-0.055	-0.011	-0.01	0.24	-0.12	-0.038
DistanceFromHome	-0.0054	-0.032	1	-0.002	0.04	0.031	0.0025	0.0092	0.011
JobSatisfaction	-0.0033	-0.055	-0.002	1	0.019	0.025	-0.019	-0.0042	-0.017
PercentSalaryHike	0.0027	-0.011	0.04	0.019	1	0.94	-0.021	-0.036	-0.022
PerformanceRating	0.017	-0.01	0.031	0.025	0.94	1	-0.022	-0.042	-0.041
TotalWorkingYears	0.68	0.24	0.0025	-0.019	-0.021	-0.022	1	0.63	0.4
YearsAtCompany	0.31	-0.12	0.0092	-0.0042	-0.036	-0.042	0.63	1	0.62
YearsSinceLastPromotion	0.22	-0.038	0.011	-0.017	-0.022	-0.041	0.4	0.62	1

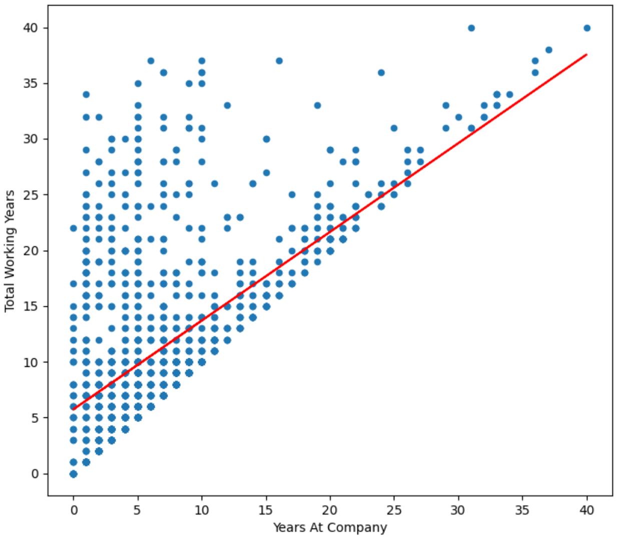
4.1 There is a 94% (0.94) very strong positive correlation between **Performance Rating** and **Percent Salary Hike**





4.2 There is a 68% (0.68) positive correlation between Total Working Years and Age - There is a 63% (0.63) positive correlation between Total Working Years and Years at Company

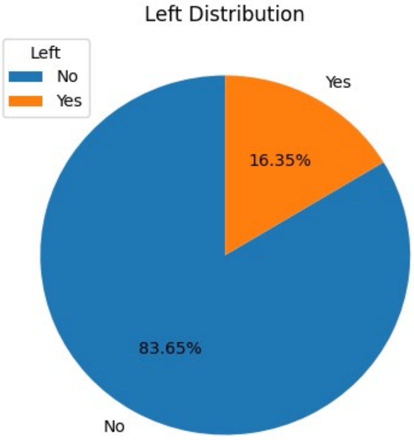




5. Conclusion

1. The total number of Left is 241 and the rate is 16.35%

Left	Count	Ratio%
No	1233.00	83.65
Yes	241.00	16.35
Total	**1474.00**	**100.00**



2. The 5 most important factors among the reasons for leaving a job are **Age, Total Working Years, Years At Company , Monthly Income, Distance From Home.**

- The average **Age** of leavers is **33.79**, while the average age of non-leavers is **37.56**. The average age of leavers is lower than that of current employees.

- The average **Total Working Years** and the average **Years At Company** are lower than those who did not leave their jobs. The average **Total Working Years** for those who didn't leavers is **11.86** , the average **Total Working Years** for leavers is **8.36.**

The average **Years At Company** for those who didn't leavers is **7.37** , the average **Years At Company** for leavers is **5.17**.

- In terms of **Monthly Income**, the highest left rate is in the low **Monthly Income** class with 27.06%. In the medium class, the left rate is **11.64%** and in the high class is **10.68%.**

- The average **Distance From Home** for those who didn't leavers is **8.92** , the average **Distance From Home** for leavers is **10.66.**

In addition, we observe that **department** and **business travel** have a statistically significant relationship with left.

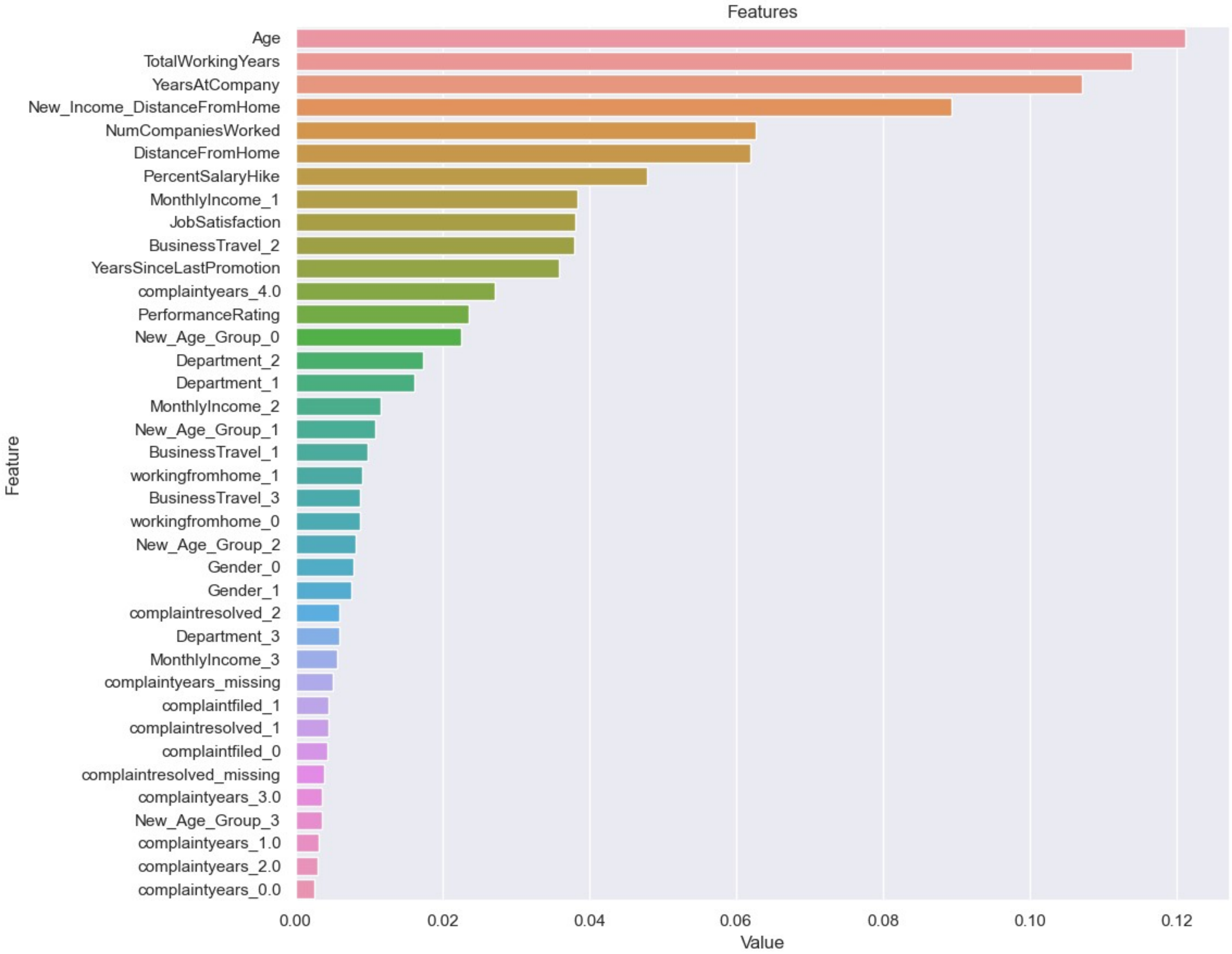
- **Travel_Frequently** have the highest left rate with **25.18%**. The left rate for **Travel_Rarely** is **15.04%**, **Non-Travel** is **9.21%**.

- The left rate for **Sales department** is **20.81%** , **Human Resources** is 20.31% , **Research & Development** is **14.02%**.

6. Machine Learning

After model building and hyperparameter optimization, the **Random Forest** algorithm gave the best measure of success.

accuracy = 0.85 **-** **f1** = 0.17 **-** **roc_auc** = 0.73 .


Features
Age
TotalWorkingYears
YearsAtCompany
New_Income_DistanceFromHome
NumCompaniesWorked
DistanceFromHome
PercentSalaryHike
MonthlyIncome_1
JobSatisfaction
BusinessTravel_2
YearsSinceLastPromotion
complaintyears_4.0
PerformanceRating
New_Age_Group_0
Department_2
Department_1
MonthlyIncome_2
New_Age_Group_1
BusinessTravel_1
workingfromhome_1
BusinessTravel_3
workingfromhome_0
New_Age_Group_2
Gender_0
Gender_1
complaintresolved_2
Department_3
MonthlyIncome_3
complaintyears_missing
complaintfiled_1
complaintresolved_1
complaintfiled_0
complaintresolved_missing
complaintyears_3.0
New_Age_Group_3
complaintyears_1.0
complaintyears_2.0
complaintyears_0.0
Feature
Value