Data analysis for the uncovering the reason among employees while leaving organization

Project Introduction

Human resourcement is a very useful domain and take part to while hiring the people for any organization. In this data analysis, the employess leaving ratio is being assesed using the python languge. The dataset is taken from kaggle which is about the human resource management (Kaggle, 2019). This dataset is regarding the emplooyess that are working in the organization and their relationship is being assesed using the different python functions.The main objective of this data anlytics is to find the reaseaon of people leaving organization. Various types of the python libraires like pandas, matplotlib and seasborn are utilized are to handle this issue (Ari & Ustazhanov, 2014; Summerfield, 2010). Moreover, the python functions which includes the sum, count, describe and plots are utilized here to handle this dataset. This data analysis will be very helpful and have positive impacts.

Data Acquisition and Cleaning

The HR dataset is acquired from Kaggle and used the python language for its analysis (Kaggle, 2019). Data cleaning is a very critical aspect and can be performed on the dataset to make it ready for the further analysis. It also handles the noise and inconsistencies from the dataset and remove the unwanted data from it. Total ‘10’ variables and ‘14999’ rows in the dataset which are going to used here for the analysis is shown in figure 1. There are two categorical and all other numeric variables in the dataset.

Figure 1: Dataset Details

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_montly_hours   14999 non-null  int64
 4   time_spend_company     14999 non-null  int64
 5   Work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
 8   sales                  14999 non-null  object
 9   salary                 14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

No, we performed the data cleaning on the dataset and find the missing values from it (Acock, 2005). There are no missing values in the dataset which can be scene from the below figure 2.

Figure 2: Data cleaning

```
satisfaction_level       0
last_evaluation          0
number_project           0
average_montly_hours     0
time_spend_company       0
Work_accident            0
left                     0
promotion_last_5years    0
sales                    0
salary                   0
dtype: int64
```

Moreover, there is no duplication of the rows in the dataset and also there is unwanted variables in this HR dataset.

Data & Exploratory Analysis

Here, exploratory data analysis is performed to make the analysis more valuable. For this purpose, various questions are configured that are going to answer using this dataset. Various types of the graphs are plotted to compare the attributes for better outcome using the raw dataset.

The HR dataset in analyzed using the describe function in pandas that is used to compute the stats of variables like standard deviation, mean, min or max values from it. Satas for the variables are given below for references in figure 3.

Figure 3: Describe the dataset in statistical shape

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610	0.238083	0.021268
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719	0.425924	0.144281
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

There are three levels of salary and ‘0’ shows left and ‘1’ shows in figure 4 they did not leave the organization. So, the number of people who left the organization is given below for every salary level.

Figure 4

```
salary  left
high    0       1155
        1         82
low     0       5144
        1       2172
medium  0       5129
        1       1317
dtype: int64
```

The highest number of projects for the sales in sales variable that can be scene from the below output of the code in figure 5.

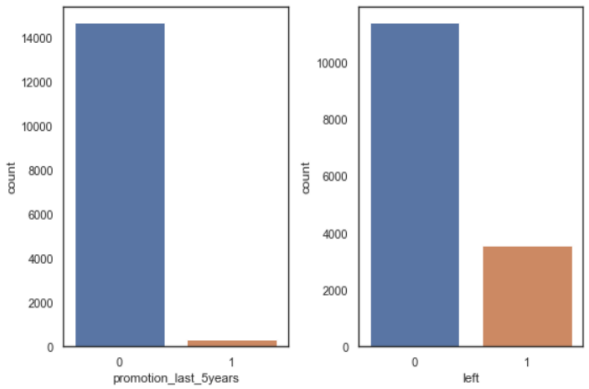
Figure 5

	sales	number_project
0	IT	1227
1	RandD	787
2	accounting	767
3	hr	739
4	management	630
5	marketing	858
6	product_mng	902
7	sales	4140
8	support	2229
9	technical	2720

The rate is displayed here using bar chart. It is very low that can be seen here as a reference

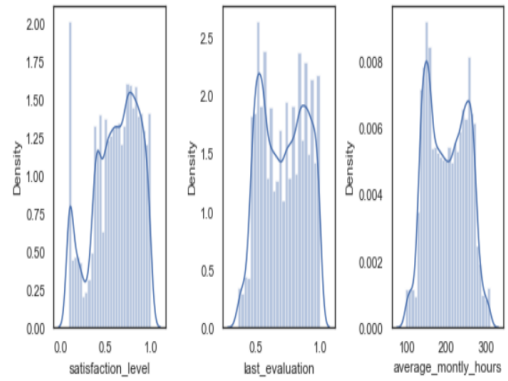
in figure 6. The last 5 years promotion is presented along with the left rate of employees in organization.

Figure 6: Employee left ratio during last 5 years



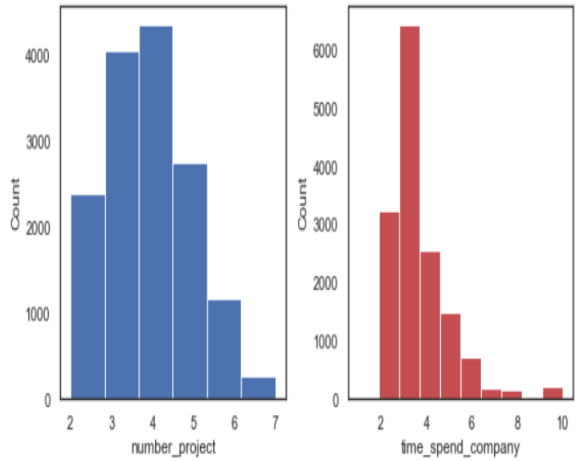
The relationship among various attributes like level of satisfaction, last evaluation and monthly average hours are shown using the density plot is presented in figure 7. Outcome shows that these variables have very strong relationship which means that are positive to each other.

Figure 7: Relationship among different attributes



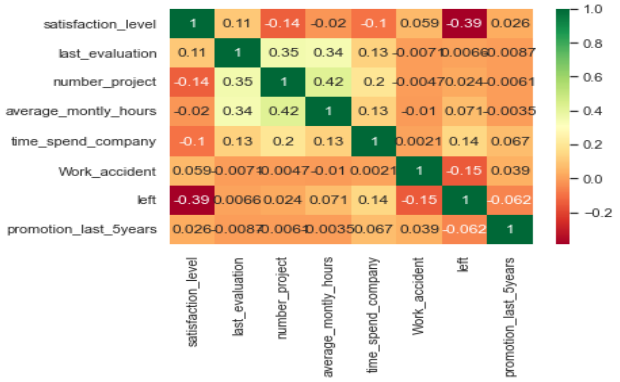
The time spend on each no of the project is given in figure 8. The findings shows that the most of the time is spends for 3-4 number of the projects during the whole journey. It means that the ratio of these projects is very high as compared to the others.

Figure 8: Relationship between time and no of projects



The correlation score is also determined for each variable and it can be scene here easily from the below figure 9. The value of the correlation in positive and near to 1` shows that they are strongly related and the value 0 means that they have no correlation. The values for each attribute is computed.

Figure 9: The correlation score of the dataset



Executive Summary

Finally, the data analysis is performed on HR dataset to find that why employees are leaving the organization. There are various types of the factors that are leaving the organization on the basis of the various factors. The findings of this data analysis shows that sales have the high number as compared to the others. People are leaving organization because people have low salary, and they are not feeling comfortable in the organization. The people who have promotion in last five years are very less as compared to those who left them. The ratio of the people who have the low salary are also leaving the organization. Finally, this data analysis highlights the main issues for the organization about those employees who are leaving.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and family, 67*(4), 1012-1028.

Ari, N., & Ustazhanov, M. (2014). *Matplotlib in python.* Paper presented at the 2014 11th International Conference on Electronics, Computer and Computation (ICECCO).

Kaggle. (2019). HR analytics for employees in organization, https://www.kaggle.com/jacksonchou/hr-analytics/data. Retrieved from https://www.kaggle.com/jacksonchou/hr-analytics/data

Summerfield, M. (2010). *Programming in Python 3: a complete introduction to the Python language*: Addison-Wesley Professional.

https://github.com/zShanzy/data-science-Assignment-1