Employee Attrition & Performance

Group 8

Kathryn Falvo, Lorenzo Pepito & Rebecca Bennett

BIA-5301-0LA

Dr Salam Ismaeel

October 17th, 2022

Introduction and Review of Related Literature

How companies manage and use their human capital is crucial to employee turnover and satisfaction (1). In fact, among the most significant factors affecting employee performance in the workplace are management support and job environment (2). An earlier study found that working relationships are crucial to employee engagement and significantly affect performance (3). The evidence highlights just how important it is for managers to understand their employees' backgrounds and roles at the company. Not only does this people-centric approach impact overall job experience and satisfaction, but it also contributes to employee turnover rates.

In the workplace, personal and professional development is a significant driver of employee performance (4). Hameed and Wheed's (2011) research integrated development opportunities made available to employees by employers. Development opportunities listed include business travel, training to expand skills inventories, and promotions for career progression. This study highlights how important it is to consider internal and external factors when investigating employee performance, satisfaction and attrition.

This report seeks to reinforce these findings by providing statistical summaries and visualizations of publicly sourced employee data and analyzing select variables similar to those previously mentioned with greater depth.

Methodology

Our group has selected a dataset that holds employee data from Kaggle, measuring employee attrition, performance, satisfaction, compensation, period of time in current role, work-life balance, salary increase possibilities, and so on. Use cases for this data set may include providing internal demographic insights to human resources professionals to enable them to fine-tune their future talent acquisition efforts. It may also aid in accurately identifying cases that will contribute to attrition for appropriate intervention by line managers and other stakeholders. To begin analyzing this dataset, we made sure to import numpy and pandas libraries for the ability to use data manipulation, analysis, and matrix tools. Once we had installed these, we imported our data and named our dataset “hr_df”.

We then wanted to identify the variables in the dataset, and their types. This is needed in order to know in which ways we can manipulate and sort these variables. We first went through each column manually and began identifying whether or not it would be considered qualitative or quantitative, and then further deciding if it was categorical, ordinal, nominal, or ratio. We further confirmed each variable by using the dtypes function to ensure that the variables we thought would be quantitative matched up as integer type, and the qualitative as object. These results are shown in Figure 1 below along with our identifications of each variable.

Figure 1 - Data Types by Variable

```
#Age = numerical (discrete - expressed as an integer with no decimals) - ordinal
#Attrition = qualitative - categorical data
#BusinessTravel = qualitative - ordinal
#DailyRate = numerical (discrete) - ratio
#Department = qualitative - nominal
#DistancefromHome = numerical (discrete) - ratio
#Education = qualitative - ordinal
#EducationField = qualitative - nominal
#EmployeeNumber - numerical (discrete) - nominal
#EnvironmentSatisfaction = numerical (discrete) - ordinal
#Gender - qualitative - categorical data
#HourlyRate - numerical (discrete) - ratio
#JobInvolvement - numerical (discrete) - ordinal
#JobLevel - numerical (discrete) - ordinal
#JobRole - qualitative - nominal
#JobSatisfaction = numerical (discrete) - ordinal
#MaritalStatus = qualitative - nominal
#MonthlyIncome = numerical (discrete) - ratio
#MonthlyRate = numerical (discrete) - ratio
#NumCompaniesWorked = numerical (discrete) - nominal
#Over18 = qualitative - categorical
#OverTime = qualitative - categorical
#PercentSalaryHike = numerical (discrete) - ratio
#PerformanceRating = numerical (discrete) - ordinal
#RelationshipSatisfaction = numerical (discrete) - ordinal
#StockOptionLevel = numerical (discrete) - ordinal
#TotalWorkingYears = numerical (discrete) - ratio
#TrainingTimesLastYear = numerical (discrete) - ordinal
#WorkLifeBalance = numerical (discrete) - ordinal
#YearsAtCompany = numerical (discrete) - ratio
#YearsInCurrentRole = numerical (discrete) - ratio
#YearsSinceLastPromotion = numerical (discrete) - ratio
#YearsWithCurrManager = numerical (discrete) - ratio
```

```
Age                          int64
Attrition                   object
BusinessTravel              object
DailyRate                    int64
Department                  object
DistanceFromHome             int64
Education                    int64
EducationField              object
EmployeeCount                int64
EmployeeNumber               int64
EnvironmentSatisfaction      int64
Gender                      object
HourlyRate                   int64
JobInvolvement               int64
JobLevel                     int64
JobRole                     object
JobSatisfaction              int64
MaritalStatus               object
MonthlyIncome                int64
MonthlyRate                  int64
NumCompaniesWorked           int64
Over18                      object
OverTime                    object
PercentSalaryHike            int64
PerformanceRating            int64
RelationshipSatisfaction     int64
StandardHours                int64
StockOptionLevel             int64
TotalWorkingYears            int64
TrainingTimesLastYear        int64
WorkLifeBalance              int64
YearsAtCompany               int64
YearsInCurrentRole           int64
YearsSinceLastPromotion      int64
YearsWithCurrManager         int64
dtype: object
```

We then wanted to compute the mean, median, standard deviation, min and max for these variables. Given that we can only compute these values for quantitative variables, we first created a new dataset called "hr2_df", including only the integer type variables. We then used the describe function to show us the mean, median, standard deviation, min and max values for each variable. The results from this function are below in Figure 2. For reference, the value of "50%" is the median.

Figure 2 - Mean, Median, Standard Deviation, Min and Max for Quantitative Variables

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156	2.729932
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.329428	0.711561
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000	1.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000	2.000000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000	3.000000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.750000	3.000000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000	4.000000

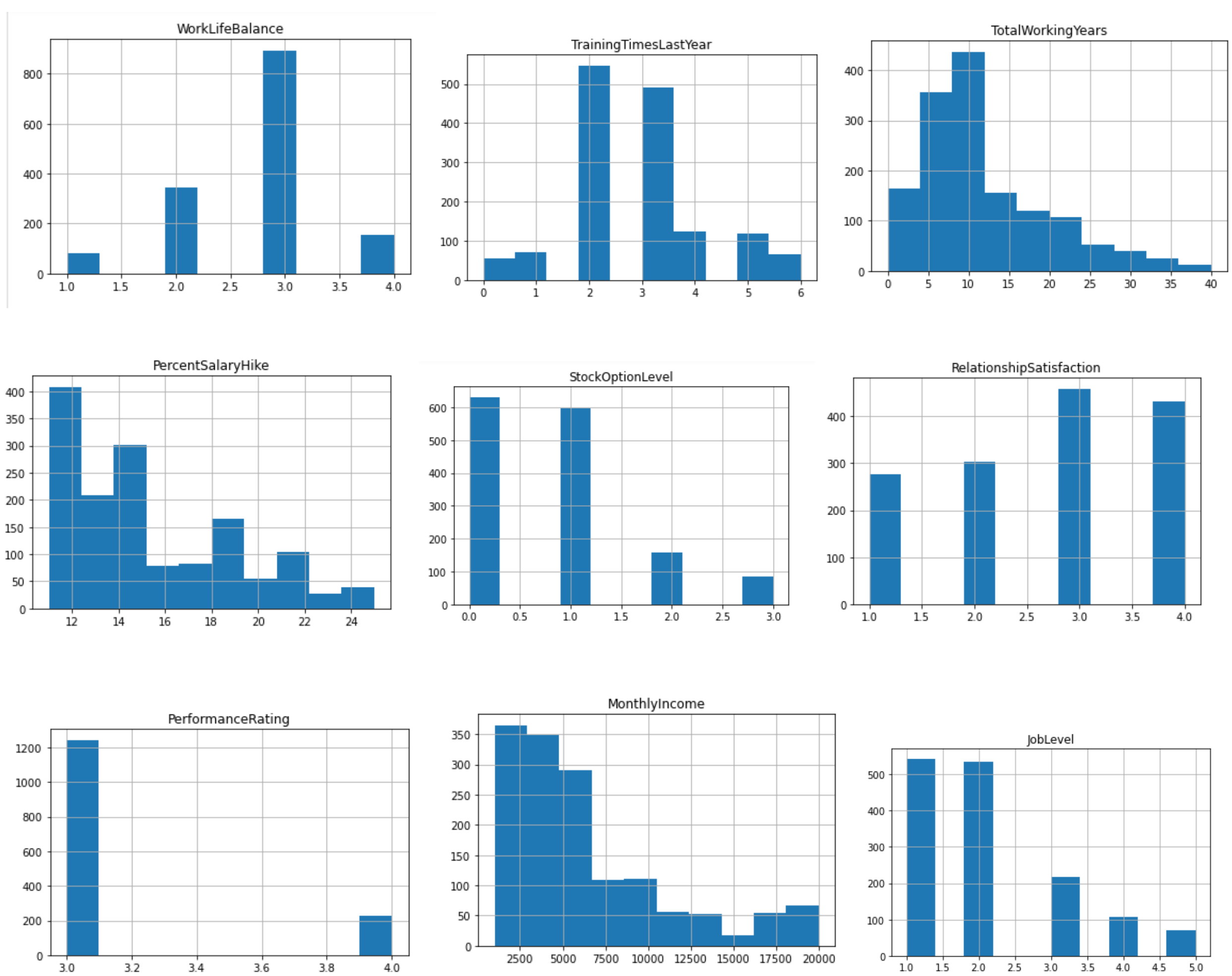
	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate	NumCompaniesWorked	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	2.063946	2.728571	6502.931293	14313.103401	2.693197	15.209524	3.153741	2.712245
std	1.106940	1.102846	4707.956783	7117.786044	2.498009	3.659938	0.360824	1.081209
min	1.000000	1.000000	1009.000000	2094.000000	0.000000	11.000000	3.000000	1.000000
25%	1.000000	2.000000	2911.000000	8047.000000	1.000000	12.000000	3.000000	2.000000
50%	2.000000	3.000000	4919.000000	14235.500000	2.000000	14.000000	3.000000	3.000000
75%	3.000000	4.000000	8379.000000	20461.500000	4.000000	18.000000	3.000000	4.000000
max	5.000000	4.000000	19999.000000	26999.000000	9.000000	25.000000	4.000000	4.000000

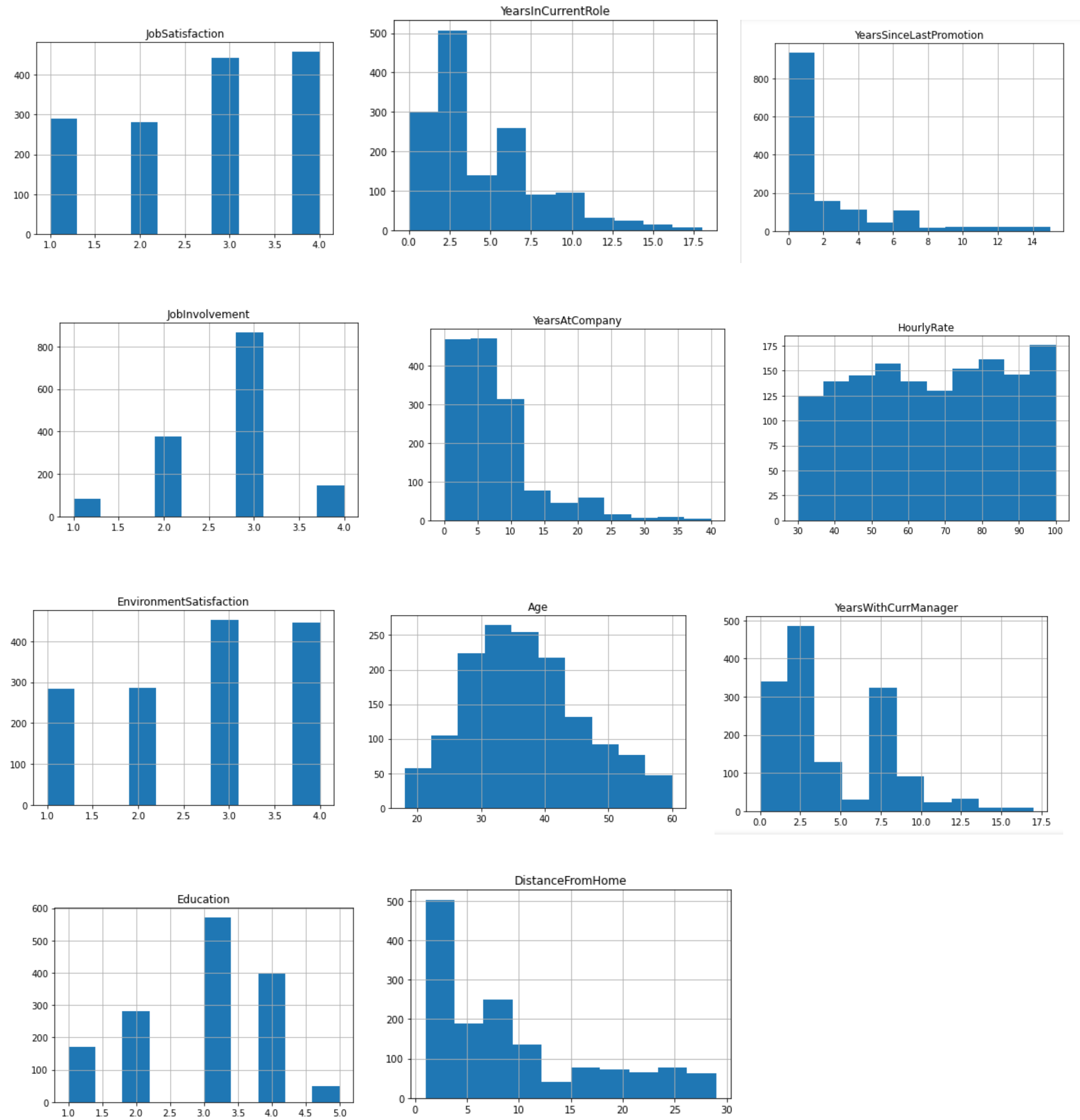
	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole
count	1470.0	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	80.0	0.793878	11.279592	2.799320	2.761224	7.008163	4.229252
std	0.0	0.852077	7.780782	1.289271	0.706476	6.126525	3.623137
min	80.0	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
25%	80.0	0.000000	6.000000	2.000000	2.000000	3.000000	2.000000
50%	80.0	1.000000	10.000000	3.000000	3.000000	5.000000	3.000000
75%	80.0	1.000000	15.000000	3.000000	3.000000	9.000000	7.000000
max	80.0	3.000000	40.000000	6.000000	4.000000	40.000000	18.000000

	YearsSinceLastPromotion	YearsWithCurrManager
count	1470.000000	1470.000000
mean	2.187755	4.123129
std	3.222430	3.568136
min	0.000000	0.000000
25%	0.000000	2.000000
50%	1.000000	3.000000
75%	3.000000	7.000000
max	15.000000	17.000000

Moving onto visualization, we first decided to plot a histogram for these values. This is able to show us the frequency distributions of each variable, what are the most common and least common values, which variables may be skewed, where there may be inconsistencies starting to show, and so on. Each of these histograms is shown below in Figure 3.

Figure 3 - Histograms for Quantitative Values

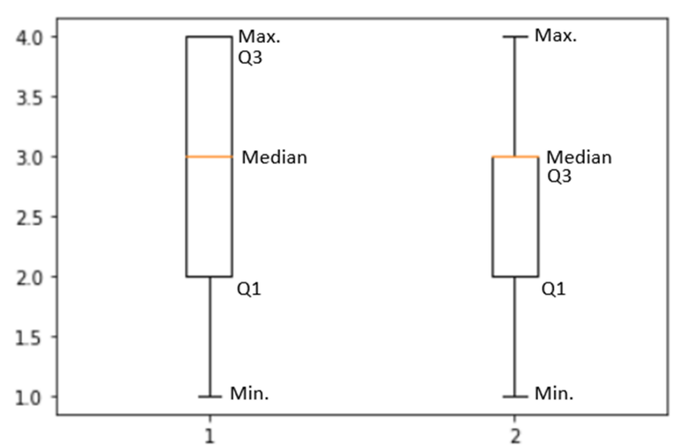




From this we had noticed that the Hourly Rate has a large variability from $30 to $100, as well as Total Working Years and Years at Company ranging from 0 to 40 years. These variabilities are understandable given the range of different roles in this dataset. We also noticed that some of these histograms show skewing in certain variables. For example, we noticed that Distance from Home, Years in Current Role, Years Since Last Promotion and Years with Current Manager were left skewed, meaning there were higher frequencies at lower values. Using Distance from Home as an example, this would mean that most employees tend to live close to the office. We also saw that Work Life Balance is slightly skewed to the right, showing that most employees have a relatively good work-life balance, along with job satisfaction, job involvement, and environment satisfaction all being slightly skewed right. Given that these variables would work hand-in-hand with each other, we can say this makes logical sense.

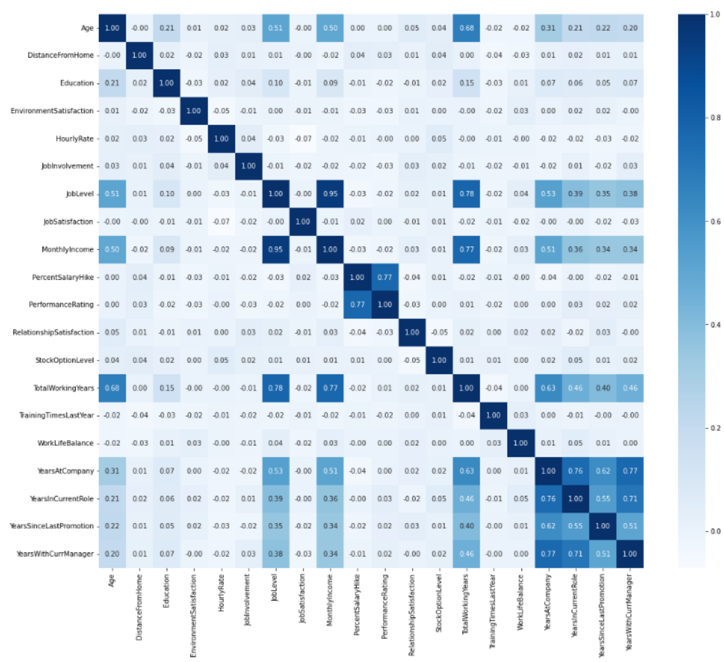
We then created the "satis_bal_df" data frame to plot a side-by-side boxplot comparing the two variables Job Satisfaction and Work Life Balance, shown below in Figure 4. In each scenario, the minimum was 1.0 which was to be expected because these variables were ordinal with employees rating job satisfaction and work-life-balance from 1 to 4 ('1' being low/bad and '4' being very high/best respectively) (5). The first quartile shows that 25% of observations are below 2.0 in each scenario with the median sitting at 3.0. The plots show that there is a negative skewing for job satisfaction which suggests that there are more respondents who enjoy their jobs despite not having the same feelings regarding work-life-balance. We will further investigate this by generating a correlation matrix.

Figure 4 - Boxplots created for JobSatisfaction (1) and WorkLifeBalance (2).



To continue interpreting the correlation between these two variables, we imported the Seaborn library to generate a heatmap to determine the correlation between each of the variables. The initial heatmap that we generated in Figure 5 below, was based on the original data which was not normalized. This said, it resulted in reduced pairwise correlation.

Figure 5 - Heatmap Generated from Raw Data.



To account for this, we imported preprocessing from sklearn, shown below in Figure 6, and then created the "norm_df" and normalized the dataframe. From the first heatmap we would conclude that there was no correlation between JobSatisfaction and WorkLifeBalance, however, after normalizing the data we observed a strong correlation at 0.72. This highlights how important it is to scale values. In Figure 5 PerformanceRating and PercentSalaryHike has a 0.77 correlation, whereas after the data has been normalized it increased to 0.96. WorkLifeBalance and PerformanceRating has a high correlation in Figure 7**,** but the original heatmap had zero correlation. Some of the other variables that were strongly correlated (>0.80) after normalization include: PerformanceRating and

HourlyRate, WorkLifeBalance and PerformanceRating, JobInvolvement and PerformanceRating. Logically this makes sense because we could assume that the more involved an employee is in the job, the better their performance rating. Employees also appear to have an hourly rate that correlates to their performancing rating and work life balance. We further explored the correlation between JobSatisfaction and WorkLifeBalance using PCA.

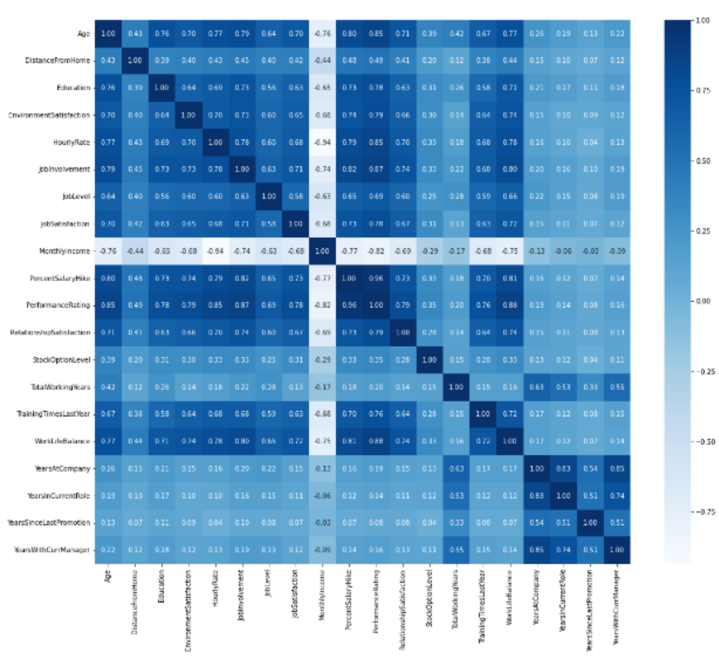
Figure 6 - Normalizing the Dataframe

In [13]:
```
from sklearn import preprocessing
norm_hr = preprocessing.normalize(hr3_df)
hr4_df = pd.DataFrame(norm_hr, columns=hr3_df.columns)
hr4_df.head()
```

Out[13]:

	Age	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	PercentSalaryHike
0	0.006840	0.000167	0.000334	0.000334	0.015683	0.000501	0.000334	0.000667	0.999849	0.001835
1	0.009550	0.001559	0.000195	0.000585	0.011889	0.000390	0.000390	0.000390	0.999865	0.004483
2	0.017683	0.000956	0.000956	0.001912	0.043968	0.000956	0.000478	0.001434	0.998837	0.007169
3	0.011341	0.001031	0.001375	0.001375	0.019245	0.001031	0.000344	0.001031	0.999727	0.003780
4	0.007785	0.000577	0.000288	0.000288	0.011533	0.000865	0.000288	0.000577	0.999892	0.003460

Figure 7 - Heatmap with Normalized Data



Finally, we conducted a principal components analysis (PCA) to determine if there is enough overlap of information between the JobSatisfaction and WorkLifeBalance to remove one of the variables from the data frame (Figure 8). The weights of Z1 are (0.765803, 0.643076) and for Z2 they are given as (0.643076, 0.765803). Z1 accounts for 86% of the variance whereas Z2 accounts for the remaining 14%.

Figure 8 - Principal Components Analysis

In [23]:
```
# f. Apply of PCA for any two variables

import numpy as np
from sklearn.decomposition import PCA

pcs = PCA(n_components=2)
pcs.fit(hr4_df[['JobSatisfaction','WorkLifeBalance']])
pcs_summary = pd.DataFrame({'Standard Deviation' : np.sqrt(pcs.explained_variance_),
                            'Ratio' : pcs.explained_variance_ratio_,
                            'Cumulative Proportion' : np.cumsum(pcs.explained_variance_ratio_)})

pcs_summary = pcs_summary.transpose()
pcs_summary.columns = ['PC1', 'PC2']
pcs_summary.round(2)
```

Out[23]:

	PC1	PC2
Standard Deviation	0.00	0.00
Ratio	0.86	0.14
Cumulative Proportion	0.86	1.00

In [25]:
```
pcs_comp_df = pd.DataFrame(pcs.components_.transpose(), columns=['PC1','PC2'], index=['JobSatisfaction','WorkLifeBalance'])
pcs_comp_df
```

Out[25]:

	PC1	PC2
JobSatisfaction	0.765803	-0.643076
WorkLifeBalance	0.643076	0.765803

Conclusion

After completing the statistical analysis and visualizations from the "*IBM HR Analytics Employee Attrition & Performance"* dataset we are able to draw some conclusions about the employees. First, it appears that many of the employees are relatively new to the company, which has resulted in employees occupying more entry level positions, having lower monthly incomes, salary hikes, and overall less job experience (TotalWorkYears). We noted that employee PerformanceRating was highly correlated with HourlyRate, WorkLifeBalance and JobInvolvement so IBM should look into improving these areas in order to reduce attrition**.** Additionally, length of time spent at the company remains an issue.This said, we recommend that IBM try to put more focus on making clear to employees the growth opportunities available. With this, they can maintain their employees and lengthen time spent at the company, improving employee-employer relationship and reducing attrition.

References

1. Hoffman, M., & Tadelis, S. (2021). People management skills, employee attrition, and manager rewards: An empirical analysis. Journal of Political Economy, 129(1), 243-285.
2. Diamantidis, A.D. and Chatzoglou, P. (2019). Factors affecting employee performance: an empirical approach. International Journal of Productivity and Performance Management, Vol. 68 No. 1, pp. 171-193. https://doi.org/10.1108/IJPPM-01-2018-0012
3. Anitha, J. (2014). Determinants of employee engagement and their impact on employee performance. International Journal of Productivity and Performance Management, Vol. 63 No. 3, pp. 308-323. https://doi.org/10.1108/IJPPM-01-2013-0008
4. Hameed, A., & Waheed, A. (2011). Employee development and its affect on employee performance, a conceptual framework. International journal of business and social science, 2(13).
5. Pavansubhash. (2016). *IBM HR analytics employee attrition & performance*. Kaggle. Retrieved October 17, 2022, from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
6. ProjectPro. (2022, July 20). *How to drop out highly correlated features in Python?* Retrieved October 17, 2022, from https://www.projectpro.io/recipes/drop-out-highly-correlated-features-in-python