Project:IBM HR Analytics Employee Attrition & Performance

Brief description of the data set and a summary of its attributes:

This is a fictional data set created by IBM data scientists, Source from Kaggle Website,

The dataset contains 1470 observation and 35 features, "the task is to Uncover the factors that lead to employee attrition and explore important questions such as ‘show me a breakdown of distance from home by job role and attrition’ or ‘compare average monthly income by education and attrition’ " as stated by dataset contributor on Kaggle.

AGE	Numerical Value
ATTRITION	Employee leaving the company (no, yes)
BUSINESS TRAVEL	(1=No Travel, 2=Travel Frequently, 3=Tavel Rarely)
DAILY RATE	Numerical Value - Salary Level
DEPARTMENT	(1=HR, 2=R&D, 3=Sales)
DISTANCE FROM HOME	Numerical Value - THE DISTANCE FROM WORK TO HOME
EDUCATION	Numerical Value(1 'Below College'/2 'College'/3 'Bachelor'/4 'Master'/5 'Doctor')
EDUCATION FIELD	(1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL)
EMPLOYEE COUNT	Numerical Value
EMPLOYEE NUMBER	Numerical Value - EMPLOYEE ID
ENVIROMENT SATISFACTION	Numerical Value - SATISFACTION WITH THE ENVIROMENT(1 'Low'-2 'Medium'-3 'High'-4 'Very High')
GENDER	(1=FEMALE, 2=MALE)
HOURLY RATE	Numerical Value - HOURLY SALARY
JOB INVOLVEMENT	Numerical Value - JOB INVOLVEMENT(1 'Low'-2 'Medium'-3 'High'-4 'Very High')
JOB LEVEL	Numerical Value - LEVEL OF JOB
JOB ROLE	(1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE)
JOB SATISFACTION	Numerical Value - SATISFACTION WITH THE JOB
MARITAL STATUS	(1=DIVORCED, 2=MARRIED, 3=SINGLE)
MONTHLY INCOME	Numerical Value - MONTHLY SALARY
MONTHY RATE	Numerical Value - MONTHY RATE
NUMCOMPANIES WORKED	Numerical Value - NO. OF COMPANIES WORKED AT
OVER 18	(1=YES, 2=NO)
OVERTIME	(1=NO, 2=YES)
PERCENT SALARY HIKE	Numerical Value - PERCENTAGE INCREASE IN SALARY
PERFORMANCE RATING	Numerical Value - PERFORMANCE RATING
RELATIONS SATISFACTION	Numerical Value - RELATIONS SATISFACTION(1 'Low'-2 'Medium'-3 'High'-4 'Very High')
STANDARD HOURS	Numerical Value - STANDARD HOURS

STOCK OPTIONS LEVEL	Numerical Value - STOCK OPTIONS
TOTAL WORKING YEARS	Numerical Value - TOTAL YEARS WORKED
TRAINING TIMES LAST YEAR	Numerical Value - HOURS SPENT TRAINING
WORK LIFE BALANCE	Numerical Value - TIME SPENT BEWTWEEN WORK AND OUTSIDE(1 'Bad'-2 'Good'-3 'Better'-4 'Best')
YEARS AT COMPANY	Numerical Value - TOTAL NUMBER OF YEARS AT THE COMPNAY
YEARS IN CURRENT ROLE	Numerical Value -YEARS IN CURRENT ROLE
YEARS SINCE LAST PROMOTION	Numerical Value - LAST PROMOTION
YEARS WITH CURRENT MANAGER	Numerical Value - YEARS SPENT WITH CURRENT MANAGER

Initial plan for data exploration:

1- Asking Research questions relating to the data set:

-Compare average monthly income by education and attrition?
-Which Dept has the highest attrition?
-Is there any relationship between who a person works for and their performance score?
-What are the key factors resulting in employee attrition?

2- Import all necessary Libraries for reading and exploring data set.
3- Explore keys factors that lead to employee's attrition starting from univariate variable exploration and building further bivariate and multi-variate exploration and relationships, trying to uncover key factors for employee's attrition, trends and relationships.
4- Feature engineer some of the Keys variables
5- Hypothesis Testing for One variable in the Jupyter Notebook.
6- Conclusion and summary of the key insights.

Actions taken for data cleaning and feature engineering:

1st looking at the data structure, shape, duplicates null values and Statistics: This dataset is clean no duplicates or null values, datatypes are correct, however Attrition variable should be changed to dummy or numerical value.

2nd Identifying main features of interest such as: Attrition(Dependent Variable), Age, Education, Job Level, Monthly Income, Performance Rating, Total Working Years, Years At Company, Years With Current Manager.

3rd Using Polynomial feature engineering on some of the features: features = ['Monthly Income', 'Age']

Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner:

Summary of the findings considering the research questions:
Attrition rate of the company is 16 %.

The Age variable seem to be normally distributed, most of the ages of the employees lies between 30 and 43,The average age of attrition in both males and females is about the same 33 and 34 years old. The higher concentration of attrition is in the younger age and lower income. The lower experienced are more likely to leave in terms of total working years

Male employees are 60 % of the company employment, females make the remaining 40%.There's not that much difference between average Female and male monthly pay except in the highest educational level 5 (doctor)

The distribution of the distance from home is right skewed where most of the data lies nearest to 5 miles.

The highest Job Involvement rate is 3 which is a high Job Involvement level. Work Life Balance also has 3 as the highest value.

Research and Dev. has the most employees then comes sales and Human Resources
HR Dept has the lowest job satisfaction level among attritions.

People who have resigned have less avg monthly income than those who didn't. The highest attrition by Job role is 'Laboratory Technician', 'Sales Executive' then 'Research Scientist'. Life sciences are in more demand that's why attrition here is higher then comes medical educational field Attrition is highest when working with the same manager for over 5 years the highest peak in attrition was when years with current manager reached 10 years

Formulating at least 3 hypotheses about this data:

Hypothesis 1:

* Hypothesis 0: Avg Monthly Salary of attrition = avg Monthly Salary of non_attrition μ1=μ2

* Hypothesis 1: Avg Monthly Salary of attrition ≠ avg Monthly Salary of non_attrition μ1≠μ2

Hypothesis 2:

* Hypothesis 0: Job Level of Attrition = Job Level of non_attrition

* Hypothesis 1: Job Level of attrition <= Job Level of non_attrition

Hypothesis 3:

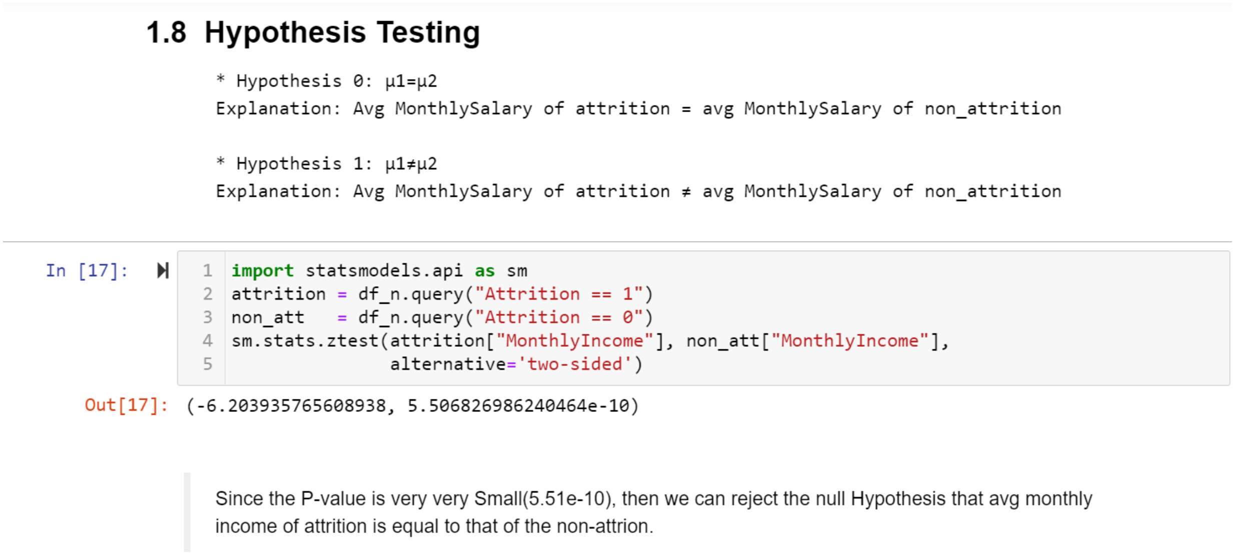
* Hypothesis 0: Years with Current Manager of attrition = Years with Current Manager of non_attrition

* Hypothesis 1: Years with Current Manager of attrition ≠ Years with Current Manager of non_attrition

Conducting a formal significance test for one of the hypotheses and discuss the results

Suggestions for next steps in analyzing this data:

I did a formal test on the 1st hypothesis as follows:



1.8 Hypothesis Testing

```
* Hypothesis 0: μ1=μ2
Explanation: Avg MonthlySalary of attrition = avg MonthlySalary of non_attrition

* Hypothesis 1: μ1≠μ2
Explanation: Avg MonthlySalary of attrition ≠ avg MonthlySalary of non_attrition
```

In [17]:

```python
import statsmodels.api as sm
attrition = df_n.query("Attrition == 1")
non_att   = df_n.query("Attrition == 0")
sm.stats.ztest(attrition["MonthlyIncome"], non_att["MonthlyIncome"],
               alternative='two-sided')
```

Out[17]:

```
(-6.203935765608938, 5.506826986240464e-10)
```

> Since the P-value is very very Small(5.51e-10), then we can reject the null Hypothesis that avg monthly income of attrition is equal to that of the non-attrion.

A paragraph that summarizes the quality of this data set and a request for additional data if needed:

Quality of data:
This Data set is Clean, no duplicates, no null values and the datatypes are mostly right.

Limitations:
Some of the dataset variables are ambiguous, like monthly rate, also extra data should be provided like the year of employment and the year of resignation, this should give extra information on the peaks of hiring, and attrition. reasons of attrition for each employee as provided before leaving the company.