Assessment 1: Dataset preparation report

Introduction

This Dataset which we analyzed is from the HR department of an organization. They need a detailed analysis about job satisfaction of their Employees and want to know the problems which employees face according to the previous data.

Basically this report is the result of gathering raw data from different sources, cleaning it afterwards and applying different rules & algorithms to understand more about the company. This dataset gives an insight about the employee salaries,salary hike, their perfomance rating and working duration in the company.

This report is meant for senior stakeholders mainly used at the end of fiscal year to understand where the company is going in terms of growth and any gender discrepancy with salary offering.

Data preparation

Overview

This dataset is all about employees' education level which correlates with the employee performance rating and salary hike ,their working experience and the working duration in this company. This dataset also gives an insight of employees age, gender and marital status which might show relationship with jobSatisfaction and work life balance.To understand the problems of working environmnet and prdicting the solutions to improve it, we obtained information from working years with the current manager, fortnight working hours, overtime and business travel for different employees.

1. Load Dataset

I have loaded the dataset as a csv file with my student ID as per requirement.

2. Column Names

I want an insight to read data, so explore column names using code functions().

3.DataType

Data types are the main part to figure out the type of column, which helps in working on data with different algorithmic functions

4.Typos

There are some typo mistakes which need to be fixed for making a graph like Male is typed as M and SSale, it has been fixed by using mask function.

5.Whitespaces

I can see some whitespaces for the string values which I removed using the strip function. The numeric values were fixed using the fillNA function.

6. Sanity Checks

Observations in each column might have some unusual values which I have checked with sanity check algorithms and then fix them.

7.Nan Replaced

There are some Nan in a few of the observations which I have fixed by masking them with the right value.

8.Converting to UpperCase

I have converted all of the string values into uppercase to make it more clear and understandable.

9.Replace NAN with Mean/Median

In integer columns, there are some Nan which can be fixed by figurig out their mean and median values, whatever suits according to the observations values.

10.Plotting Columns

To understand more in detail about the dataset, I have plotted different graphs, as graphs are easy to present data that are too numerous or complicated to be described adequately in the text and in less space.

11. Relationship Between Columns

Graphs are a common method to visually illustrate relationships in the data. This can also help in making predictions and interpretations to stakeholders for the future.

Issues Discovered

#	Issue name	Location	Code to identify	Rationale and solution
1	Typos like Divorced in typed as D/uppercase, lowercase letters, typo like N instead of No	MaritalStatus/ Resigned	.mask() function to fix this typo .replace()	When we saw this column in the dataset, every observation had 3 values, Married, Single and Divorced except this which is typed as D.It wasn't easily understandable A Replacement function is used to fix these N, no, NO observations which does not look good and we can't make proper plots to explore data and it doesn't count in value for that column.
2	Whitespaces/ missing values	Regined MonthlyIncome	.fillNa() function .mean() .median()	FillNa is a function that can help in fixing integer missing values by using mean and median functions.Mean or median gives an integer number to fill related to the observations, which can help in making good graphs

3	Uppercase letters for string values	BusinessUnit,Marital Status,Gender,Business Travel and Resigned	.uppercase() function is used to convert lower case string to uppercase string	Upper case values give us easy understanding of data.

Data exploration

Overview

This dataset refers to employee development programs, their mental health, work life balance and the working environment of this organization based on the gender, Age,MonthlyIncomeand working hours. According to my analysis, there is no gender discremination in this organization as data shows both genders have similar salary ranges.

Data gathered from different surveys tells us employee work life balance. I have explored the work life balance column to check whether the male or female gender have spent more balanced life while doing jobs in this organization. Employees have very Balanced life decpite of any gender equality.

I have explored data for column working with current manager time duration and how it affects employees job satisfaction as it is very important in every employee's working life. Based on the data we can imply that if employees are more comfortable with their manager, they will have a more satisfied life and give a healthy output to the organization.

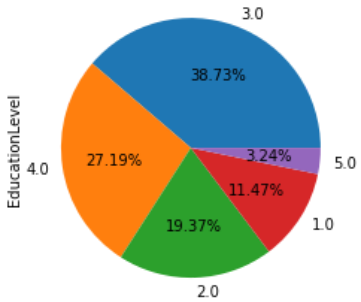
Process

1. Converting string to integer

Age attribute is a string data type. When I tried to make a histogram for the age column it started giving an error. Upon close inspection it turned out that one of the rows had a “36a” in the data that needed data fix. After doing data fix, I converted the column to number and the column was ready for plotting.

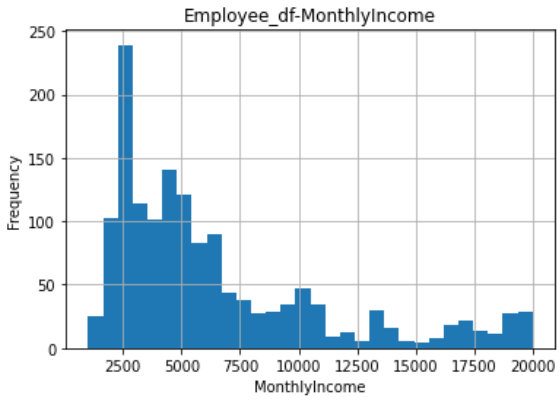
2. Education level

Employees have different education levels. In this organization, there are very few highly qualified PHD level employees and less qulified employees whose education level is below the job requirement level, instead they have graduated level employees more in number.



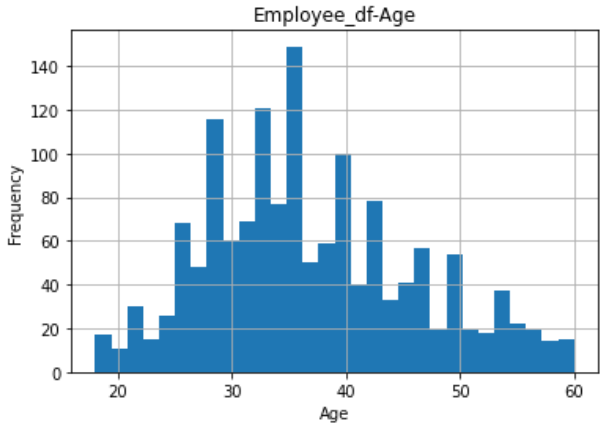
3. MonthlyIncome

MonthlyIncome has a different salary range , this organization has a high number of employees having a salary range of around 2500$.



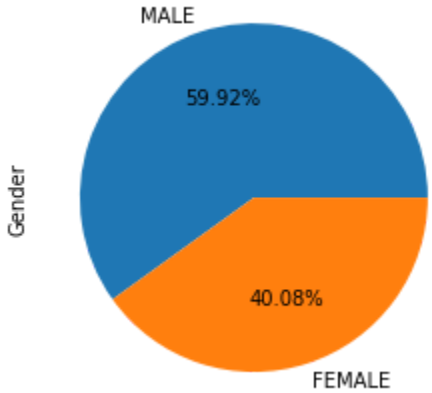
4. Age

I have explored the data of Age attribute by making histogram which shows the most of the employees in this company is between 27 to 40 age group



5. Gender

Males are more in number then females in this organization. By plotting the pie chart,I have obtained the information that there are 40.08% of females and 59.92% males. We can imply that we should hire more females to show gender equality.

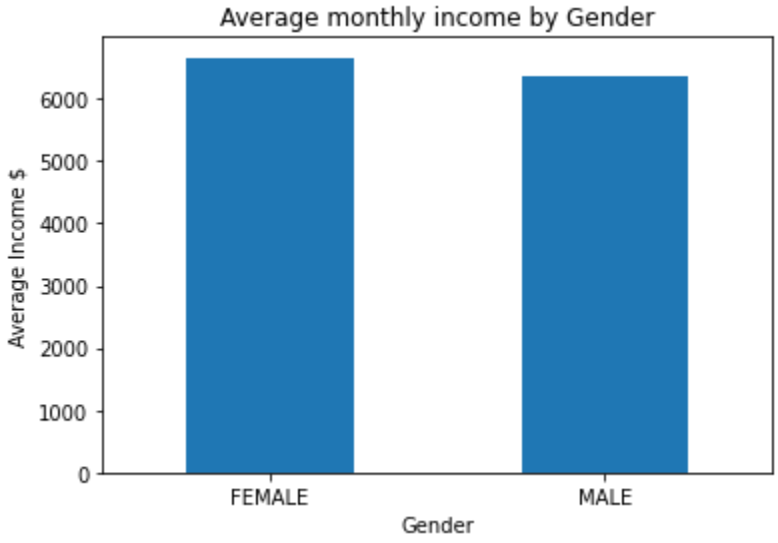


Observations

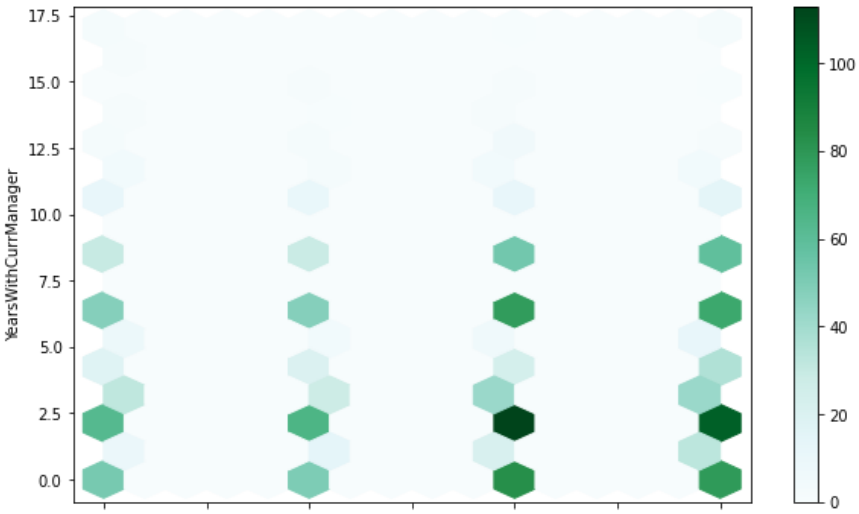
#	Observations	Significance
1	Monthly income, salary hike	As per my analysis monthlyIncome and salary Income have very significant effects on job satisfaction and work life balance of employees.
2	Number of working hours, performance rating, Overtime	Perfomance of employees would be better if we implement regular working hours and avoid overtime to reduce unsatisfied job outcomes and unhealthy family life.

Relationship between Columns

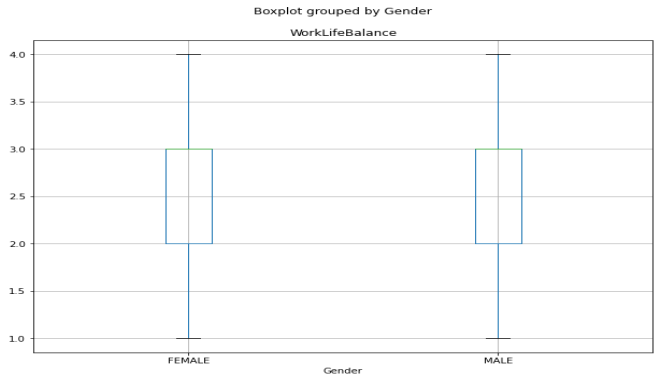
This Bar chart shows the monthly income for both the genders are almost same which indicates there is no gender discrimination in this organisation on monthly salary.



This hexbin chart between Job satisfaction and years of working with the current manager shows that most of the employees are most satisfied after 2-3 years of working with the current manager. This imply that ideally after 3 years of working under the same manager, HR should think about changing employee department or manager for better productivity and less chances of people leaving the organization.



This boxplot between gender and work life balance gives us an insight that both genders have the same kind of work experience, we can imply that male or female both are equal in their work life ba



Conclusion

This report would be useful for HR to plan for next fiscal year in terms of hiring & performance reviews. Also main observations conclude that the salary gap in terms of gender is negligible which needs to be maintained for a healthy organization.

References

1- DapperDuck,Feb2,2021,answered,[Id]:https://stackoverflow.com/questions/65985187/use-of-count-in-jupyter-notebook-i-do-not-understand-why-i-either-get-no-outp

2- answered Aug 15, 2013 at 10:05 waitingkuo[id]:https://stackoverflow.com/questions/18250298/how-to-check-if-a-value-is-in-the-list-in-selection-from-pandas-data-frame