Assessment 1 Template: Dataset preparation report .

Michael Teixeira S4133975

Introduction

This report outlines the process of preparing and exploring a dataset for a fictional organization. It represents a crucial step in the data science, focusing on data cleaning, preprocessing, and initial exploratory analysis. The case study involves analysing employee data to understand factors affecting retention. This report aims to provide insights that could help the organization improve employee retention strategies.

Data preparation

Overview

The dataset contains information about 1482 employees, including various attributes such as age, job satisfaction, income, and work-life balance. This data is crucial for understanding the factors that may influence employee retention within the organization.

Process

1. *Imported necessary libraries (pandas, numpy, seaborn, matplotlib) and loaded the dataset.*
2. *Conducted initial data inspection using df.info() and df.head() to understand the structure and content of the dataset.*
3. *Checked for unique values in each column to identify potential inconsistencies or errors.*
4. *Cleaned and standardized categorical variables (e.g., 'Resigned', 'BusinessTravel', 'Gender', 'MaritalStatus').*
5. *Handled missing values by replacing them with median for numerical columns and mode for categorical columns.*
6. *Corrected data types for appropriate columns (e.g., changing some columns to 'category' type).*
7. *Performed sanity checks on logical relationships between variables (e.g., years at company vs. total working years).*
8. *Detected and handled outliers, in the 'AverageWeeklyHoursWorked' column.*

Issues discovered

#	Issue name	Location	Code to identify	Rationale and solution
1	*Data entry error*	*Age;*	*df['variable'].unique();* *df['variable'].value_counts()*	*36a is not a valid age number. We assume that the correct number is 36 and use 'df['variable'].replace()' to replace.*
2	*Data entry errors and inconsistencies*	*Resigned;*	*df['variable'].unique();* *df['variable'].value_counts()*	*We changed ['Y', 'NO', 'N', 'no'] to ['Yes', 'No']. We used 'df['variable'].replace()' to replace values for correct ones*

				‘df[‘variable’].title()’ to keep values concise.
3	Data entry errors and inconsistencies	BusinessTravel;	df[‘variable’].unique(); df[‘variable’].value_counts()	We change ['Travels_Rarely' 'TRAVEL_RARELY' 'rarely_travel'] to 'Travel_Rarely'. We use ‘df[‘variable’].replace() to keep values consistent.

4	Data entry errors and inconsistencies	BusinessUnit;
		Gender;
			df[‘variable’].unique();
			df[‘variable’].value_counts()
				Identified that ‘BusinessUnit’ and ‘Gender’ columns had a value swapped at the same row.
				We used a mask to identify value and then ‘.replace()’ to values ‘Sales’ and ‘Female’. We changed ['MMale', 'M'] to ‘Male’. We also used .str.strip() to strip white spaces and ‘.str.title()’ to keep names concise.

5	Data entry errors and inconsistencies	MaritalStatus;	df[‘variable’].unique(); df[‘variable’].value_counts()	We assume and change ['D'] to ‘Divorced’. We use ‘df[‘variable’].replace() to keep values consistent. We used .str.strip() to strip white spaces.

6	Missing values	Resigned;
		EducationLevel;
		JobSatisfaction;
		MonthlyIncome;
		OverTime;
		WorkLifeBalance;
			.isnull().sum()	Missing values were filled with mean/mode to retain data integrity.

7	Define the data types	Age;
		Resigned;
		BusinessTravel;
		BusinessUnit; Gender;
		MaritalStatus; OverTime;
			df['Age'] = pd.to_numeric(df['Age'])	Age to numeric and the others to category.

8	Sanity checks: Years at company should not exceed total working years	YearsAtCompany;
		TotalWorkingYears;
			df[(df['YearsAtCompany'] > df['TotalWorkingYears'])]	Replace the invalid total working years with the correct value. df.loc[1472, 'TotalWorkingYears'] = 1

9	*Detect outliers using Boxplot*	*AverageWeeklyHoursWorked;*	*sns.boxplot()*	*Assume that value 400 was a data entry error and used .replace() to change to 40*

Data exploration

Overview

In this exploration, we focused on several key features from the employee dataset: OverTime, Years at Company, Business Travel, Job Satisfaction, and their relationships with employee resignations. These features were chosen to understand the factors that might influence an employee's decision to resign.

Process

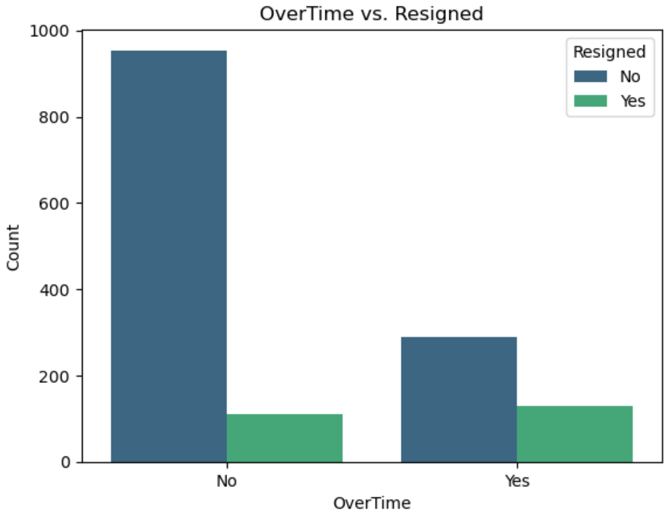
1. *Created visualizations for analysis. Visualizations help in identifying patterns and relationships in the data more easily than raw numbers.*
2. *Examined the relationship between OverTime and Resignations to understand if working overtime affects employee retention.*
3. *Analysed the impact of Years at Company on Resignations to investigate if employee tenure is related to the likelihood of resignation.*
4. *Explored the effect of Business Travel on Resignations to determine if travel requirements influence an employee's decision to leave.*
5. *Investigated the relationship between Job Satisfaction, Years at Company, and Resignations to understand how job satisfaction and tenure interact with resignation rates.*

### Observations

#	Observations	Significance
1	*Employees who work overtime have a higher resignation rate compared to those who don't.*	*This suggests that excessive work hours may lead to burnout and increased likelihood of resignation.*
2	*Resignation rates are higher among employees with fewer years at the company, particularly in the first 5 years.*	*This indicates that retention efforts should focus on newer employees who may be at higher risk of leaving.*
3	*Employees who travel rarely for business have the lowest resignation rate, while those who travel frequently have a slightly higher rate.*	*Travel frequency appears to have some impact on employee retention, with more travel potentially leading to higher turnover.*
4	*There is no clear correlation between job satisfaction and years at the company. However, there are more resignations among employees with lower job satisfaction scores.*	*This suggests that while tenure doesn't necessarily improve job satisfaction, maintaining high job satisfaction could be key to reducing resignations.*

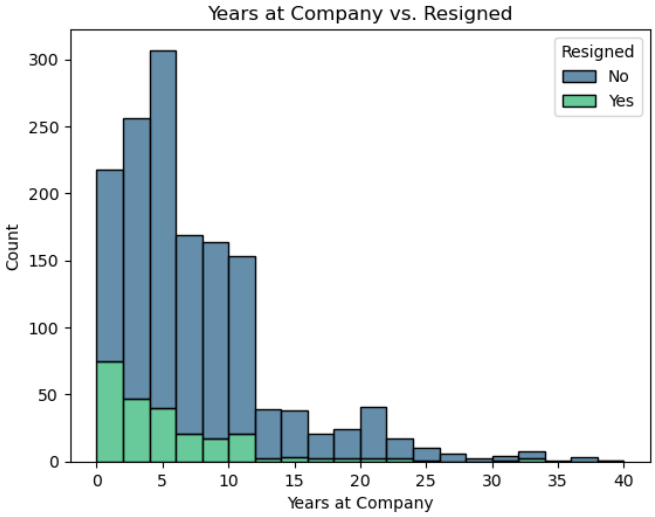
Plots

OverTime vs. Resigned



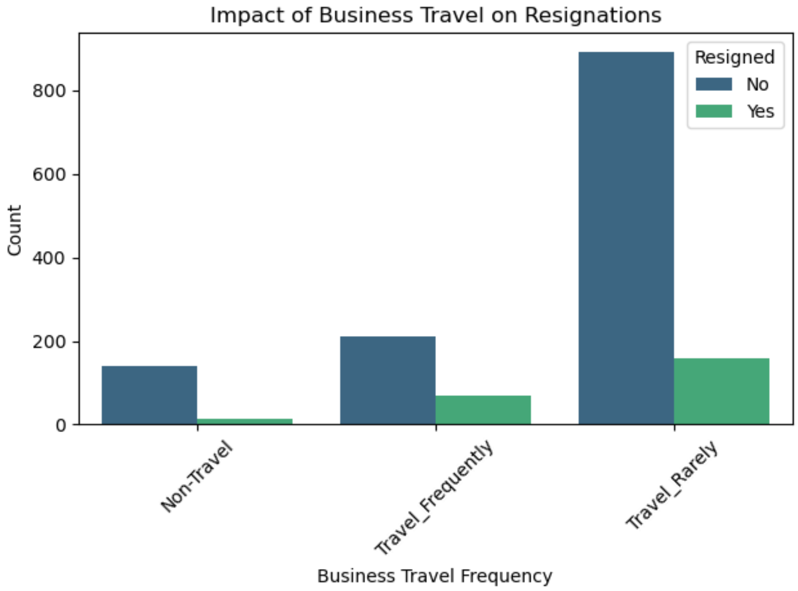
The chart shows that employees who work overtime have a higher proportion of resignations compared to those who don't. This suggests that excessive work hours may be a contributing factor to employee turnover.

Years at Company vs. Resigned



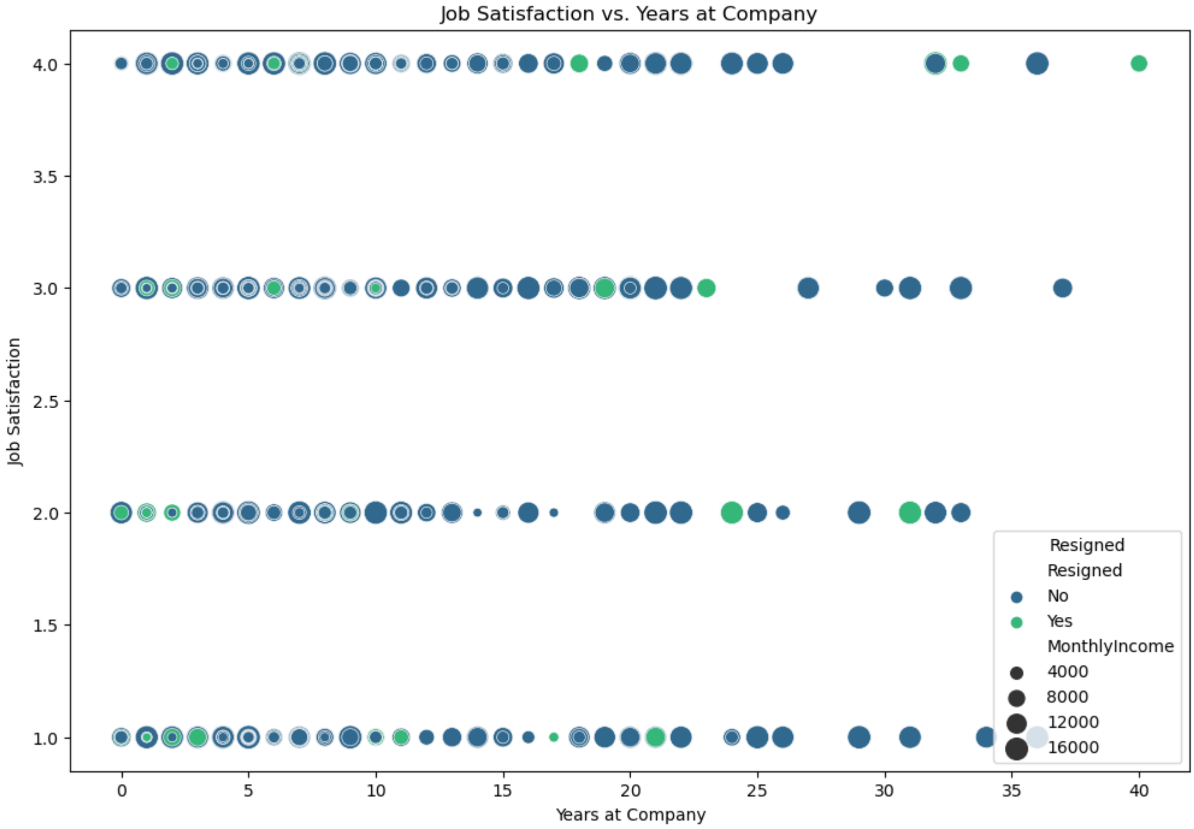
The histogram reveals that resignation rates are higher among employees with fewer years at the company, particularly in the first 5 years. This suggests that retention strategies should focus on newer employees who may be at higher risk of leaving.

Impact of Business Travel on Resignations



The chart shows that employees who travel rarely for business have the lowest resignation rate, while those who travel frequently have a slightly higher rate. This indicates that travel requirements may have some impact on employee retention.

Job Satisfaction vs. Years at Company



The low monthly income could also be a factor specially in cases where job satisfaction is higher.

Conclusion

We can conclude that there are bigger resign rates for people that do overtime, travel frequently, and are new with the company. The low monthly income could also be a factor specially in cases where job satisfaction is higher.

These findings suggest that the company should focus on improving work-life balance, enhancing job satisfaction, and providing extra support for new employees to reduce resignation rates and improve overall employee retention.

References

Iglewicz B, Hoaglin DC (1993) How to detect and handle outliers, Wiley, New York.