The leadership team wants to understand why employees are leaving as the organization grows. Leadership believes the pace that the company is growing is causing the employees to feel a lack of direct engagement and job satisfaction. The question I will be addressing for this analysis is: Why are employees leaving the organization and what metrics can we use to gauge attrition? In order to address this question, I will need to use HR attrition data as well as other data sets from HR. The parameters that I will use to determine whether the employee has left the company or not is the variable attrition. I will compare this variable with other metrics to determine which variables have a correlation.

There are 15 data sets that will need to be merged into a single exportable .csv file. The data is coming from the HR department of the organization. Many preprocessing methods will be used in order to clean the data and prepare it for analysis. To being the merging process, I used pandas to import all the HR data and simultaneously merge all the data sets into one file. I used the concat() function to concatenate the data.

A screenshot of a computer program

Description automatically generated with low confidence

I used the print() function to display the data frame to ensure all of the material was merged appropriately.

A screenshot of a computer screen

Description automatically generated with low confidence

A picture containing text, screenshot

Description automatically generated

I then saved the merged data into a single .csv file and named it Merged\_HR\_Data.



The merged raw dataset contains 7510 rows and 35 columns. Many of these columns contain variables that are irrelevant to addressing the organizational problem. Some variables use the same value for every row such as *EmployeeCount*, *Over18*, and *StandardHours*. I will use the *drop()* function in python to remove these variables from the set. These are blank values, specifically in the attrition column. I will use the *dropna()* function to remove the rows with blank values. I began cleaning the data by deleting columns that did not pertain to the analysis using the drop() function.

A screenshot of a computer code

Description automatically generated with low confidence

I then removed the rows that contained blank values in the attrition column using the dropna() function.



There are also issues with categorical data. Attrition uses yes and no values, however, I will need numeric values to appropriately conduct the analysis. I will use the *map()* function to make this change so the values will display 1 for yes and 0 for no.



Next, I changed values in the *BusinessTravel* column to numeric values. Non-Travel now is represented by 0, *Travel\_Rarely* is represented by 1, and *Travel\_Frequently* is represented by 3.



I used the head() function to reveal the first 5 rows of the dataset to ensure the changes displayed properly.

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, font, menu

Description automatically generated I will also use the *corr()* function to display the correlation matrix between the variables to assist in finding which variables have a stronger correlation to attrition than others. The closer a value is to 1 or -1, the stronger the correlation.

After viewing the correlation matrix, I dropped the variables that had a lower correlation value.

A picture containing text, font, screenshot

Description automatically generated

This will allow me to reduce the noise by determining which variables are statistically significant for this analysis and remove data that has little to no correlation. Next, I used the corr() function to create the new correlation matrix.

A screenshot of a computer

Description automatically generated with low confidence

Finally, I converted the cleaned dataset to a .csv file entitled Finalized\_HR\_Data.



The organization could update company policies to decrease attrition based on the first analysis. As one issue gets addressed, there may be different factors that lead to attrition. This process will need to be dynamic in order to monitor reports as the data updates. This will assist future analysts to analyze additional data without having to create an entirely different analysis. The goal would be for the analysists to determine what factors lead to attrition as the organization evolves their policies until the attrition rate minimizes to an appropriate rate.

Visualizations allow for analysts to easily identify trends within data quickly and efficiently. Dynamic visualizations are representations of data that are interactive and capable of updating in real-time. This allows analysts to uncover trends, patterns, and insights that may be more challenging to identify in a static visualization. Dynamic visualizations can be created using different visualization tools such as Power BI or Tableau and can be easy to understand by limiting complexity, keeping the visualization intuitive, providing context, and using guided interactivity. I opened the file in Power BI and created a modifiable reporting structure that junior data analysts can manipulate for basic exploratory analysis. The analysts can toggle between attrition between departments or job roles. They can also use any variable within the Finalized HR Dataset. Additionally, the junior analysts can assess key performance indicators between departments and understand what factors contribute to a high level of attrition.

A picture containing text, screenshot, font, circle

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

References

Reddy, E. P. K. (2021, March 25). *Step by step process of feature engineering for machine learning algorithms in Data Science*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/step-by-step-process-of-feature-engineering-for-machine-learning-algorithms-in-data-science/

Sun, E. D., Ma, R., & Zou, J. (2022, December 30). *Dynamic visualization of high-dimensional data*. Nature News. https://www.nature.com/articles/s43588-022-00380-4

Zach. (2021, December 22). *How to create a correlation matrix in python*. Statology. https://www.statology.org/correlation-matrix-python/

Kumar, B., & Bijay KumarPython is one of the most popular languages in the United States of America. I have been working with Python for a long time and I have expertise in working with various libraries on Tkinter. (2023, March 28). *Pandas drop() function in Python*. Python Guides. https://pythonguides.com/pandas-drop/