Determinant Factors Contributing to Staff Attrition in a Company

People are the backbone of an organization. Any company needs employees who are fit to do their work. But there's a reality that every organization, big or small, has to face: movement within the workforce is inevitable. Employees come into a job, learn various aspects and gain experience from it, and decide to leave if better opportunities arise. This isn't always a bad thing like when low performing employees are let go. However, if a company experiences a large number of top-performing employees leaving constantly, it may be an indication of inherent problems within the company. That's where attrition in HR may emerge as a noteworthy concern.

There are several disadvantages to a company when attrition occurs, which include:

1. Staff attrition can lead to a considerable reduction in the size of the overall workforce of an organization.
2. A reduction in the number of people can mean there will be an increase in the workload for people who stay back.
3. Staff attrition can increase costs associated with paying existing employees, hiring new employees, as well as training them.
4. Since providing sufficient training to new hires can become an issue with attrition, the organization might witness a decline in its overall performance.
5. High staff attrition rates may paint a negative image of the organization.
6. High staff attrition rates can negatively affect the organization's culture.
7. It puts excess pressure on HR to conduct onboarding and training activities.

Given the various drawbacks of attrition on a company, it is important to understand factors that contribute to employees leaving a company so that the company can put measures in place to mitigate this.

Link to gitHub repo: [Mamba GitHub Repository]

Link to Notebook: [Mamba Python Notebook]

Link to Slides: [Google Slides Presentation]

Link to Tableau: [Tableau Presentation]

The Team members are: Lynn Njoroge, Elsie Felab, Edwin Kutsushi and Alex Twenji.

1.BUSINESS UNDERSTANDING

1.1. PROBLEM STATEMENT

This analysis seeks to find out reasons that prompt employees with key positions in the company to leave and find alternative employment so as to allow an organization to design and implement an effective staff retention strategy.

1.2. DATA MINING GOAL

The goal is to use the data provided and see if we can identify various factors that lead to staff attrition.

1.3. ANALYSIS

1.3.1. Data

The main dataset contains employee attrition data of an organization.

1.3.2. Risks

Financial Risks are negligible since the project will be carried out on open source software services such as Google Colab, GitHub, Tableau Public, Jupyter Notebook and Atom.

Other risks and their contingency plans are described in the table below:

Risk	Contingency
Unavailability of a Team Member	If a team member is unavailable for any scheduled meeting, he or she will communicate to the rest of the team and the meeting will be rescheduled.
Unavailability of one or several team members.	If a team member is unavailable for a task he or she is scheduled to perform, he or she should communicate to the team leader so that task may be reassigned to another member.
Possible Crashing of Colab Notebook	If this happens, we will use Jupyter Notebook on our local machines.

1.3.3. Limitations

- Lack of an overview on the data.

1.3.4. Assumptions

- All the data provided is accurate to the best of our knowledge.
- This is the latest data from the company on employment.
- All the employees were under permanent or on contract of over a year.

1.3.5. Implementation Plan

To implement this analysis, we will need to follow a laid out plan in order to finish our analysis on time.

RESPONSIBILITIES

1. DATA PREPARATION - Lynn Njoroge (Tuesday)
2. DATA CLEANING - Elsie Felab (Tuesday)
3. EXPLORATORY DATA ANALYSIS
    - ❖ UNIVARIATE ANALYSIS - Edwin Kutsushi (Tuesday)
    - ❖ BIVARIATE ANALYSIS - Alex Twenji (Tuesday)
    - ❖ MULTIVARIATE ANALYSIS - Alex Twenji (Tuesday)
4. SAMPLING PLAN - Alex Twenji (Tuesday)
5. HYPOTHESIS TESTING - Edwin Kutsushi (Wednesday)
6. DATA REPORT - Lynn Njoroge (Wednesday)
7. TABLEAU VISUALIZATION - Elsie Felab (Wednesday)
8. PRESENTATION SLIDES - Alex Twenji. (Wednesday)

2. DATA UNDERSTANDING

2.1. DATA COLLECTION

This data was extracted from Data.World which is a data catalog that is free and open to the public.

Dataset Files:

Employee Attrition Data: [Link]

2.2. DATA DESCRIPTION

Column	**Description**
Age	Age of Employee.
BunisessTravel	Does the Employee travel frequently or not?
DailyRate	Daily Rate of Pay.
Department	Section of the Company Employee works in.
DistanceFromHome	How far an employee lives from the workplace in miles.
Education	Labelled 1-5
EducationField	Course taken in University.
EmployeeCount	1 for each Employee.
EmployeeNumber	Unique Employee ID
EnvironmentSatisfaction	Rating 1-4 about Employee's satisfaction with the work environment.
Gender	Male or Female.
HourlyRate	Rate of Pay per Hour.
JobInvolvement	Rating 1-4 about Employee's involvement with the company.
JobLevel	Rating 1-5 depending on position held in the company.

JobRole	Position held in company.
JobSatisfaction	Rating 1-4 about Employee’s satisfaction with the job.
MaritalStatus	Single, Married or Divorced.
MonthlyIncome	Amount paid per Month.
MonthlyRate	Rate of Pay per Month.
NumCompaniesWorked	Number of different Companies Employee has worked for.
Over18	If Employee is of Age i.e., an Adult.
OverTime	If Employee receives Overtime pay.
PercentSalaryHike	Annual percentage salary increases.
PerformanceRating	Rating from 1-5.
RelationshipSatisfaction	Rating 1-4 about Employee’s satisfaction with their personal relationships.
StandardHours	Typically, 80 hours.
StockOptionLevel	0-3 rating of stock compensation.
TotalWorkingYears	Number of years worked.
TrainingTimesLastYear	Number of times trained in the last year.
WorkLifeBalance	Rating of 1-5 about Employee’s work schedule mixed with general life activities.
YearsAtCompany	Number of years worked in the company.
YearsInCurrentRole	Number of years worked in the company, in the current role.

YearsSinceLastPromotion	Number of years worked in the company, in the current role, without promotion.
YearsWithCurrManager	Number of years worked in the company, under the current Manager.

The target variable is the column ‘Attrition’, where it’s either Yes or No.

2.3. DESCRIBING THE QUESTION

2.3.1. Specifying the Question

The research question is to identify various factors that lead to staff attrition.

Additionally, we would like to come up with recommendations to improve staff retention.

2.3.2. Defining the Metric for success

For this analysis to be considered successful, the following areas must be covered:

- Overall Exploratory Data Analysis.
- Univariate Analysis.
- Bivariate Analysis.
- Multivariate Analysis.
- Use Appropriate Visualizations.
- Test the Hypothesis

2.3.3. Understanding the Context

Staff attrition refers to the loss of employees through a natural process, such as retirement, resignation, elimination of a position, personal health, or other similar reasons.

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire training are some of the common expenses of losing employees and replacing them.

Another problem posed by attrition is that it may lead to loss of customers as customers often prefer to interact with familiar people.

Additionally, Errors and issues are more likely if you constantly have new workers.

Given the various drawbacks of attrition on a company, it is important to understand factors that contribute to employees leaving a company so that the company can put measures in place to mitigate this.

2.3.4. Recording the Experimental Design

1. Loading the Datasets (done Above)
2. Data Preparation (done Above)
3. Investigating the Dataset (done Above)
4. Data Cleaning
5. Exploratory Data Analysis (Univariate, Bivariate and Multivariate)
6. Hypothesis Testing
7. Conclusions
8. Recommendations

2.3.5. Data Relevance

This will be discussed after the analysis is complete.

3. DATA PREPARATION

3.1. OVERVIEW OF DATA

We first imported all the libraries we needed for our analysis and hypothesis testing.
Then we loaded both our datasets into our programming environment and created a data frame (df) to hold the data so that we can analyze it as a dataframe.
We determined that the dataset has 1470 rows, and 35 columns.
We checked all the datatypes of the 13 columns and accessed some general information and summary statistics(like mean and standard deviation)of the data.
Additionally, we Checked the entire profile of the dataframe.

3.2. DATA CLEANING

For a sequential approach that strives to cover all aspects of Data cleaning, the following checklist/steps were followed:

3.2.1. Validity of Data

To increase validity of data we dropped columns that we did not need for our analysis.

We dropped the following columns: EmployeeCount, EmployeeNumber, Over18, StandardHours, MonthlyRate. This action left us with 30 columns to work with.

.

3.2.2. Accuracy of Data

This was done to confirm that all the columns were of the appropriate types / dtypes.

3.2.3. Completeness of Data

The data was found to be complete there being no null values in our data

3.2.4. Consistency of Data

The data was found to be consistent there being no duplicated data.

3.2.5. Uniformity of Data

To enhance uniformity of our data, we fixed messy column names by changing all column names to lowercase and stripping whitespaces all around them.

4. DATA ANALYSIS

4.1. EXPLORATORY DATA ANALYSIS

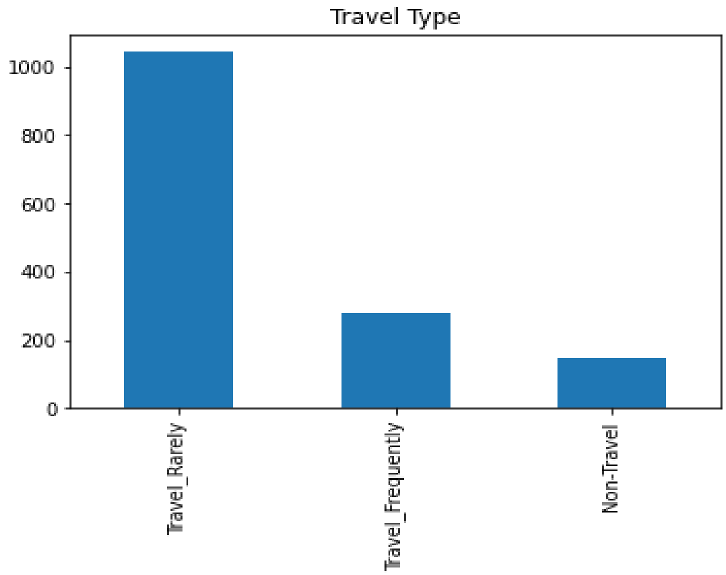
4.1.1. UNIVARIATE DATA ANALYSIS

a) Numerical Data

There are 22 columns of numerical data, i.e. age, dailyrate, distancefromhome, education, environmentsatisfaction, hourlyrate,jobinvolvement,joblevel, jobsatisfaction, monthlyincome, numcompaniesworked, percentsalaryhike, performancerating, relationshipsatisfaction, stockoptionlevel, totalworkingyears, trainingtimeslastyear, worklifebalance, yearsatcompany, yearsincurrentrole, yearssincelastpromotion and yearswithcurrmanager. An investigation was carried out using box plots on all these numerical variables and some were found to contain outliers like; company worked, performance rating, total working years and training time last year. An investigation was carried out to determine if we can remove the outliers from the dataset, when outliers were removed,the dataframe was left with 863 entries out of 1470. We determined that such loss of a huge part of our data would severely affect the validity of our analysis results. Therefore, we decided not to remove the outliers.

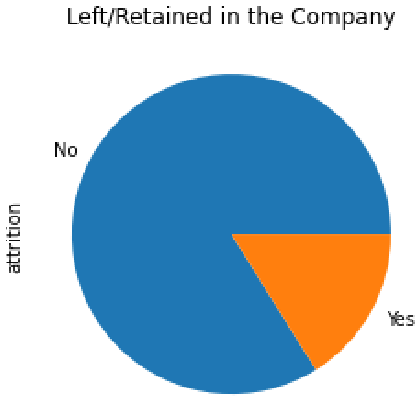
b) Categorical Data

- We plotted a bar graph to visualize travel of employees to work.



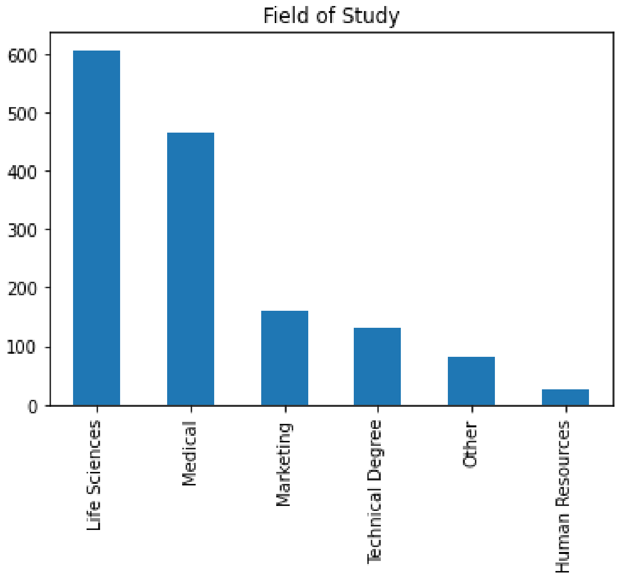
From this we can see that over 1000 employees rarely travel to work, almost 300 travel to work frequently and almost 150 workers never travel to work. Therefore, we can conclude over 80% of the employees work online or stay in or around the company.

- We plotted a pie chart to show the employees' attrition.



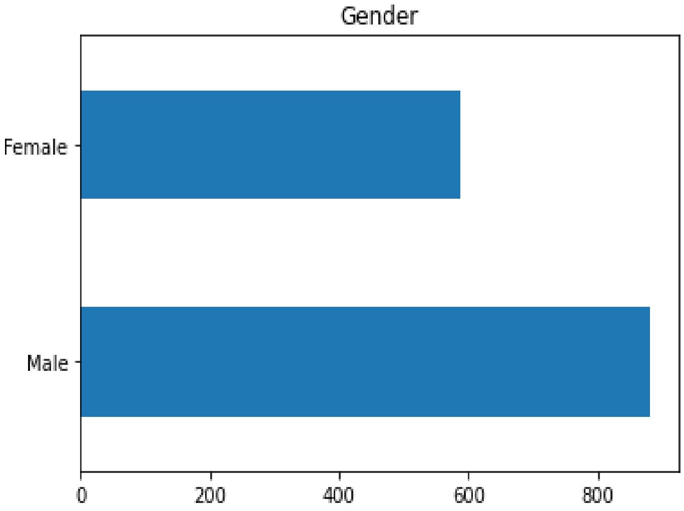
From this we can see that most of the employees stayed at the company, however, some employees also left the company

- A bar chart was constructed on the field of study to determine the field in which most of the employees are.



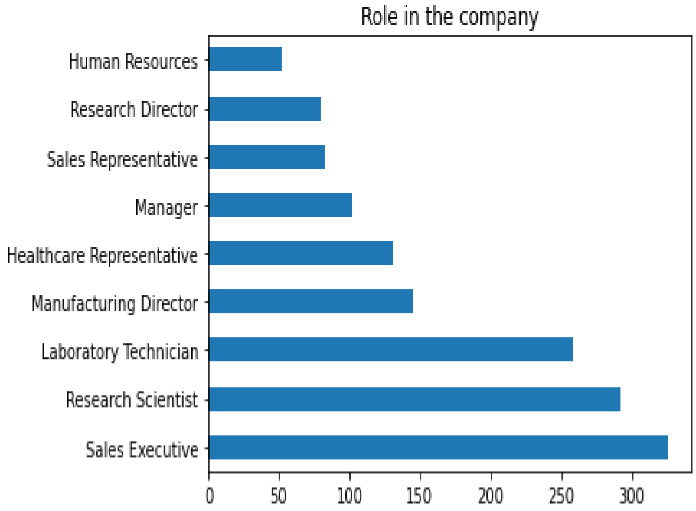
From this, we can see that almost 600 employees are in the field of life sciences, over 450 are in the field of medical. Human resource had a smaller number of almost 50.Other fields were marketing, technical degrees and others not specified.

- We plotted a bar chart showing male and female workers in the company.



The bar chart shows that the majority of the employees are male with over 900 employees and female around 600 employees.

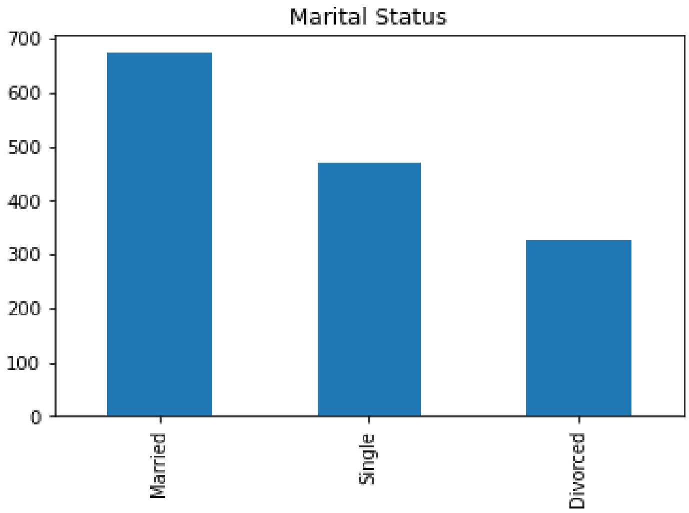
- We plotted a bar chart showing the role of the employees in the company.



Most of the employees are in the field of sales executive, research scientists and laboratory technicians.

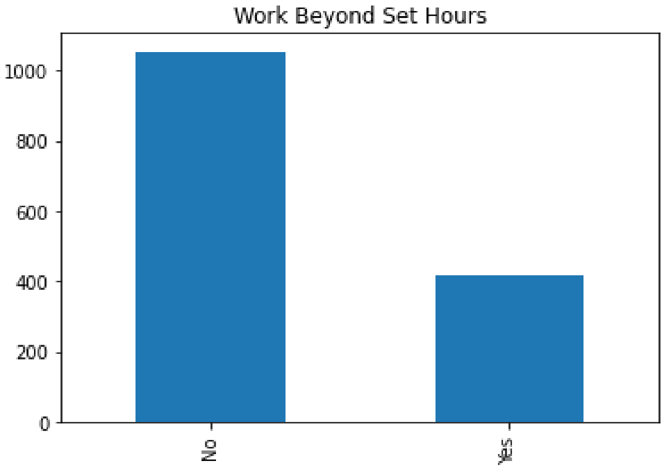
The least occupied roles are in the human resource and others are as research directors, sales representatives, managers, healthcare representatives and manufacturing directors.

- We plotted a bar chart to visualize the marital status of employees in the company.



Most of the employees are married, i.e. over 650.
Employees who are single are around 450.
Divorced employees are around 300 employees.

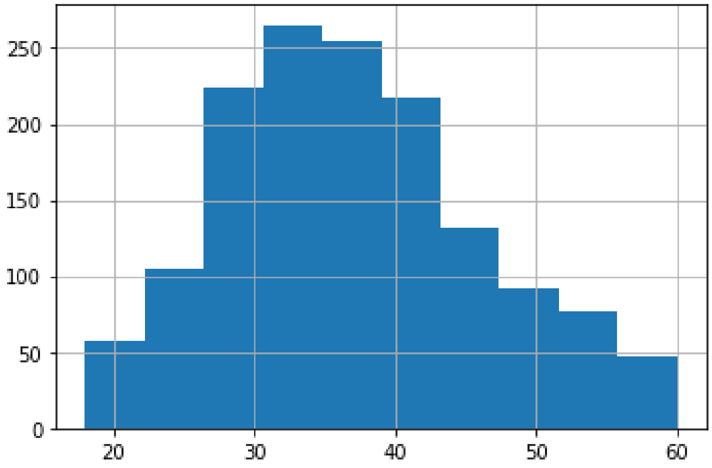
- We plotted a bar chart showing employees working overtime.



Most of the employees do not work beyond set hours (i.e. over 1000 employees do not work overtime)

Only about 400 employees work over time

- We plotted a histogram to show the age distribution of all employees.



Most of the employeesare aged between 30 and 40 years.

c) Summary Statistics

- The standard deviation was calculated for numerical columns of the dataset and most of the columns had small standard deviation meaning, most of the dataset was concentrated around the mean.
  Some variables like monthly rate and daily rate had large standard deviation which means these columns had values deviating away from the mean.
- Variance of the columns was performed and similar to the standard deviation above, the smaller the variance, the more the data is concentrated around the mean and the greater the variance, the greater the spread in the data about the mean.

- When skewness of the data set was conducted, we got negative skewness and positive skewness. Positive values of skewness indicates that the tail of the data is right-skewed. Negative values of skewness indicates that the tail of the data is left-skewed
- When kurtosis was performed most of the data had negative kurtosis meaning the data had less extreme outliers while for the positive kurtosis the data had extreme outliers, but the data comes from the normal distribution.

4.1.2. BIVARIATE DATA ANALYSIS

For our bivariate analysis, we calculated the Pearson coefficient correlation. From this calculation and additional information from our profile report, the columns found to have some correlation were: Attrition, BusinessTravel, Department, Education, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, OverTime, PerformanceRating, RelationshipSatisfaction, StockOptionLevel and WorkLifeBalance.

4.1.3. MULTIVARIATE DATA ANALYSIS

We conducted a Principal Component Analysis. PCA is a statistical technique used to convert high dimensional data to low dimensional data by selecting the most important features that capture maximum information about the dataset. The features are selected on the basis of variance that they cause in the output.

From the Multivariate Analysis, using principal component analysis with a cut-off point of 80%, as is the standard recommendation, the columns found to have some correlation were:

1. ‘age'
2. 'Businesstravel'
3. 'Dailyrate'
4. 'Department'

5. 'Distancefromhome'
6. 'Education'
7. 'Educationfield'
8. 'Environmentsatisfaction'
9. 'Gender'
10. 'Hourlyrate'
11. 'Jobinvolvement'
12. 'Joblevel'
13. 'Jobrole’
14. 'Jobsatisfaction'
15. 'Maritalstatus'
16. 'Monthlyincome',
17. 'Numcompaniesworked'

4.1.4. DATA ANALYSIS OF FACTORS LEADING TO STAFF ATTRITION

Considering the features found with correlation in Bivariate and those found important in Multivariate Analysis with the Target Variable being Attrition, the features to analyze are:

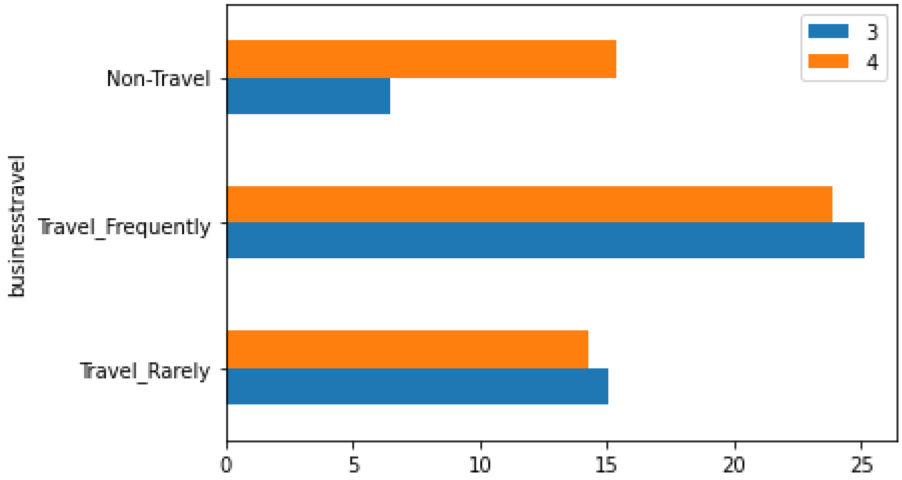
1. BusinessTravel
2. Department
3. Education
4. EducationField
5. EnvironmentSatisfaction
6. Gender
7. JobInvolvement
8. JobLevel
9. JobRole
10. JobSatisfaction
11. MaritalStatus
12. MonthlyIncome

We will however drop Education, EducationField, JobInvolvement and JobRole since they are ordinal with no explanation what their values stand for. We will group them in terms of PerformanceRating where possible to give more insight into the data.

1. How Does Business Travel Affect Attrition considering Employee Performance.

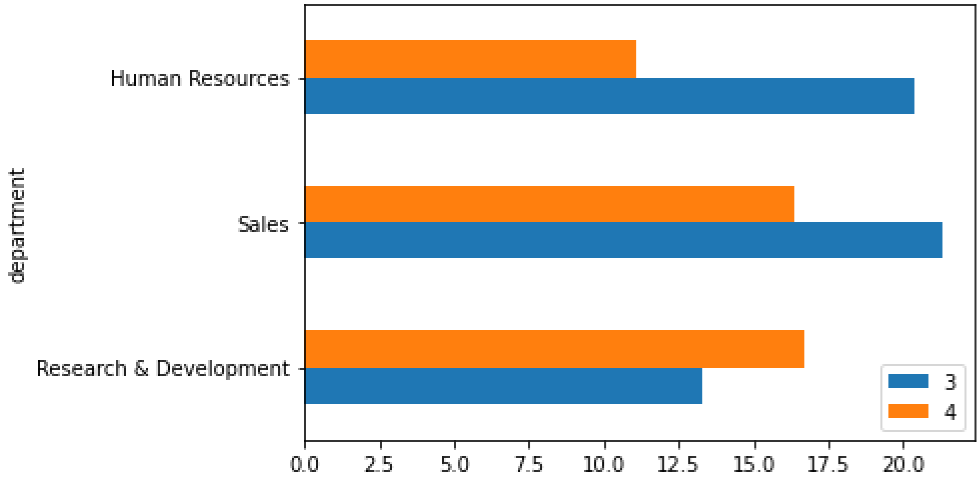
Employees who travelled frequently experienced higher attrition rates in general (about 25%). This could be caused by experiencing different cultures that widen their view of the world and

the workplace, thereby challenging them to change their job to fit the new mold they are molding their habits into. Employees who did not travel and had a lower performance ranking (3) had the least attrition (about 6%) and this could be due to contentment in the workplace due to the lack of an outside view of work from different areas and cultures.
This is visually represented by the graph below



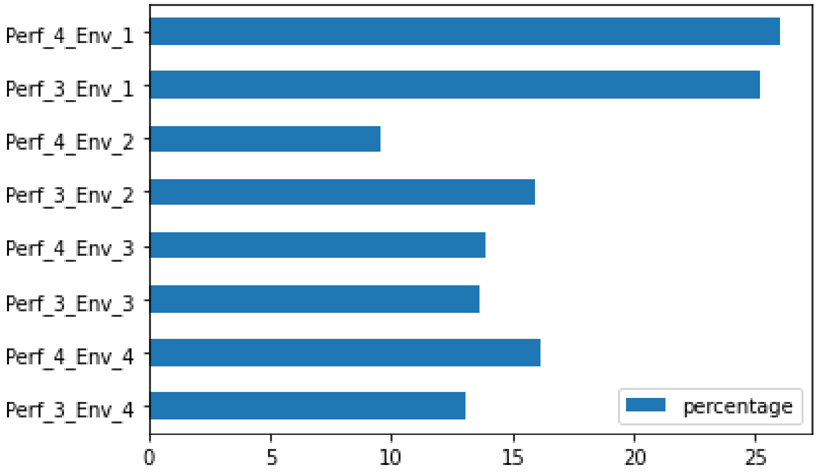
2. How Does Department affect Attrition considering Employee Performance

Employees with performance rank 3 showed high attrition rates in the Human Resources (20%) and Sales (23%) departments. These are less technical departments. A reason for this could be that a department like sales is one where target meeting performance is crucial which puts a lot of pressure on an employee. This constant stress to meet targets could be a reason for the high attrition rate.
Those with rank 4 had the highest attrition in Sales and Research & Development departments (17%). Research and Development is more technical.
This is visually represented by the graph below



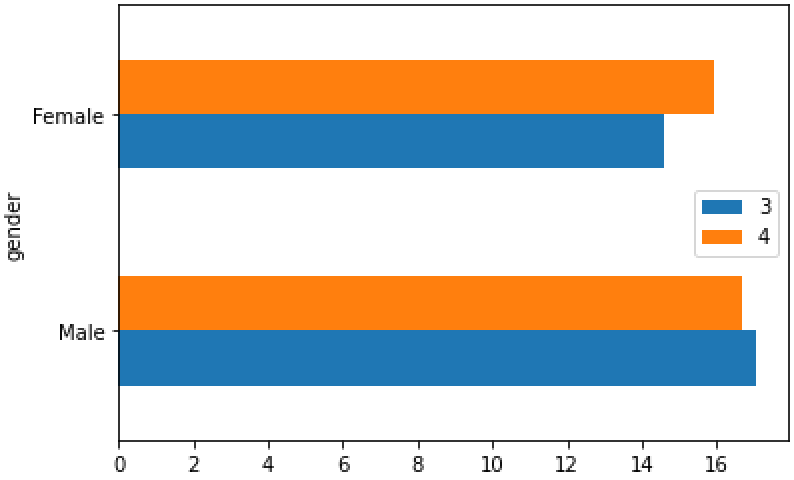
3. How Does Environment Satisfaction affect Attrition considering Employee Performance

Both performance rating groups i.e. 3 and 4, had high attrition rates (25%) if their environment satisfaction was low i.e. 1. This could be dissatisfaction with their work environment including the work atmosphere of the workplace, compared to their performance.
This is visually represented by the graph below



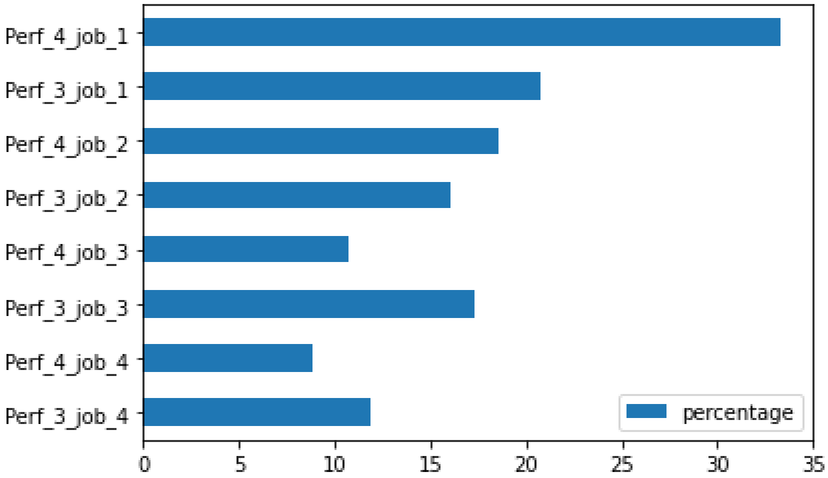
4. How Does Gender affect Attrition considering Employee Performance

Males generally had more attrition rates, although only 1-2% more than the females.
This is visually represented by the graph below



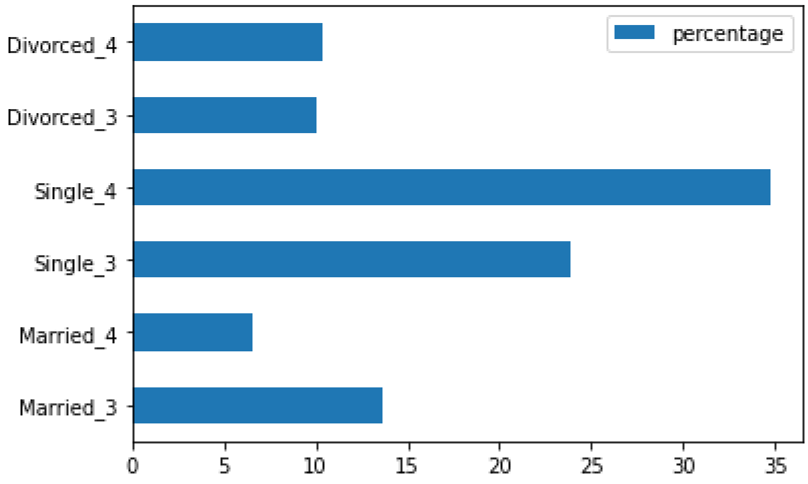
5. How Does Job Satisfaction affect Attrition considering Employee Performance

Performance rating group 4 had high attrition rates (25%) if the job satisfaction was low i.e. 1. This could be dissatisfaction with their day to day activities including compared to their performance. Whereas those in group 4 with high job satisfaction i.e. 4 having the least attrition at 7% indicating relative contentment with their day to day activities including compared to their performance.
This is visually represented by the graph below



6. How Does Marital Status affect Attrition considering Employee Performance

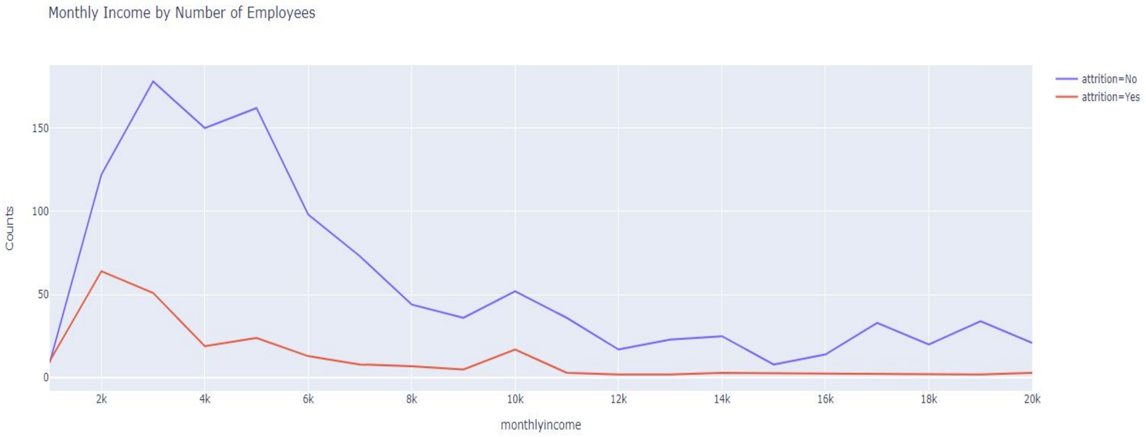
Single people in general had the highest attrition rate with those with a performance rating of 4 having an attrition rate as high as 35%. This could be due to their confidence in their ability to find other jobs as well as the lack of dependents, therefore more risk averse. Married people with a performance rating of 4 had the least attrition rate at 6% and could be due to being less risk averse due to them having dependents such as a family.
This is visually represented by the graph below



7. How Does Monthly Income affect Attrition

The rate of attrition is highest at the lowest income levels less than 4,000 monthly. The attrition rate decreases with increase in income from around 5,000 monthly, with a minor spike at 10,000 monthly. These individuals represent the midpoint of the payscale and could be looking for better opportunities to earn better from other jobs. The attrition rate flat lines from about 11,000 monthly to 20,000 monthly.

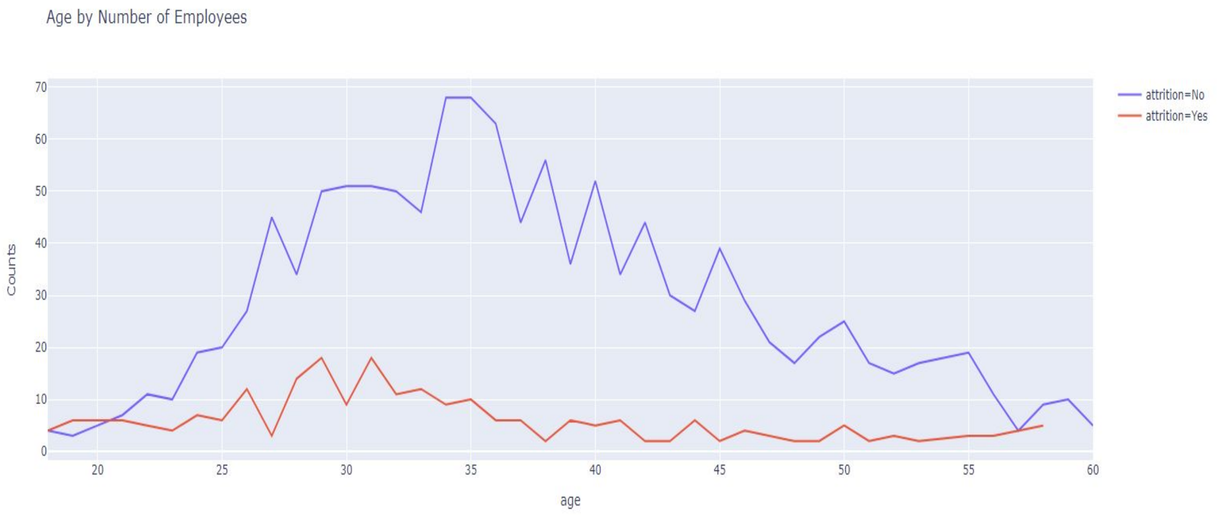
This is visually represented by the line graph below



8. How Does Age affect Attrition

The attrition rate is highest between 28-32 year olds. It falls with increasing age from there on, suggesting employees looking for job stability possibly due to having families. From 53 years old, the attrition rate starts increasing again and could be due to people approaching their retirement age.

This is visually represented by the line graph below



5. IMPLEMENTING SOLUTION

We shall implement our solution by performing hypothesis testing.

Hypothesis testing is a statistical method that is used in making statistical decisions using data.It is an assumption that we make about the population parameter from a sample.

In our case, we want to see if there is a significant difference in the means of satisfaction level between employees who had left the company and employees who had retained in the company?

For our Hypothesis testing, we followed the following steps:

Step 1*: Formulate the null hypothesis and the alternative hypothesis.*

The hypotheses we have formulated are:

H0: Null Hypothesis: There is no difference in satisfaction level between employees who had retained and those who left.

Ha: Alternate Hypothesis: There is a difference in satisfaction level between employees who retained and those who left.

In statistical terms:

$H0$: $\mu1 = \mu2$

Ha : $\mu1 \neq \mu2$

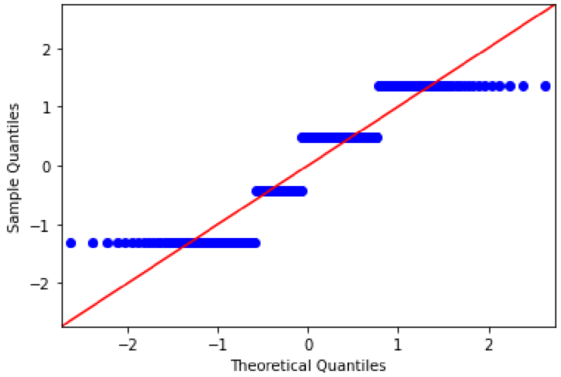
Because the alternate hypothesis had a not equal to sign, we concluded that our test would be two tailed.

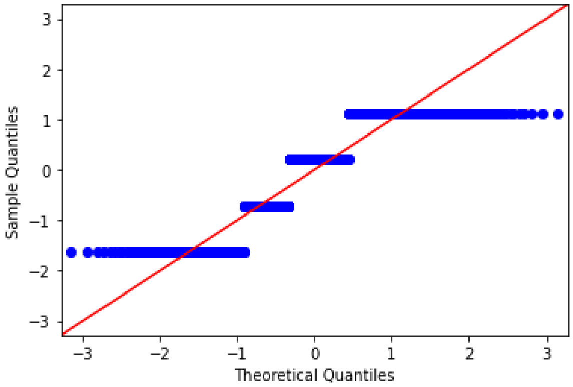
This hypothesis was formulated because, from our analysis, job satisfaction was responsible for the highest rate of attrition among employees.

Step 2*: Identify a test statistic and significance level.*

Before conducting a valid two-sample t-test, the following requirements must be met:

1. Data values must be independent. Measurements for one observation do not affect measurements for any other observation. The data was obtained from two independent populations, one where employees had left the company and the other where employees had stayed at the company. These two populations are independent of one another. (this condition has been satisfied)
2. Data in each group must be obtained via a random sample from the population. We used simple random sampling to select a sample from each population from our filtered attrition data frames. (this condition has been satisfied.).
3. Data in each group are normally distributed. To test for normality of our independent groups, we plotted a q-q plot. The plots verified that samples are normally distributed.





Now that all our conditions are satisfied, we can choose a significance level. The significance level (α), is a measure of the strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically significant. We chose a significance level of 0.05 since it is the most commonly used in statistical tests. This means that there is a 5% risk of concluding that a difference exists when there is no actual difference.

***Step 3**: Computing the test-statistic and P-value*

After conducting the test in our programming environment, the results were:

t statistic is: 2.0189741204714458

p value is: 0.04897887708736376

***Step 4**: Analyze the results and either accept or reject the null hypothesis.*

We used the p-value to analyze the results. Since our p-value was less than the stated significance level, we rejected the null hypothesis.

***Step 5**: Interpreting the Results.*

Rejecting the null hypothesis means that we have enough statistical evidence to state that, There is a statistically significant difference in satisfaction level between employees who retained and those who left.

Discussion of Test Sensitivity

After completion of an analysis, the researcher has to evaluate the chance of making errors in his hypothesis test results. The analyst establishes the maximum chance of making type I and type II errors. The probability of committing a type I error (rejecting the null hypothesis when it is actually true) is called α (alpha) which is the level of statistical significance. We had already stated our α as 0.05 before starting our tests. This means that 5% is the maximum chance of incorrectly rejecting the null hypothesis (and erroneously inferring that there is a statistical difference between the mean number of blue cars taken from two postal codes). This further means that we are 95% confident in the results of our tests. The probability of making a type II error (failing to reject the null hypothesis when it is actually false) is called β (beta). The quantity (1 - β) is called power. Here we set our β as 0.10 meaning that we are willing to accept a 10% chance of missing a difference in the mean number of blue cars taken from two postal codes.

Ideally *alpha* and *beta* errors would be set at zero, eliminating any possibility of Type I or Type II error. In reality though, this is not possible, and the goal is to make them as small possible. However, in order to reduce *alpha* and *beta,* one would need to increase the sample size. Therefore the effect of increasing our sample size (to say a higher number-greater than 30) would reduce the possibility of committing either a Type I or Type II error.

6. CONCLUSIONS

We set out to identify determinant factors that lead to staff attrition in a company. From our analysis we have found that:

- There is a high number of staff attrition among employees with low job satisfaction levels.
- The attrition rate is highest between 28-32 year olds. Young people tend to move from company to company. It seems that people switch jobs at the earlier years of their careers.
- The rate of attrition is highest among employees with the lowest income levels i.e. less than $4,000 monthly.
- Single people in general had a higher attrition rate compared to married or divorced employees.
- Employees in the Human Resources and Sales departments had a higher attrition rate compared to other departments.
- Employees who travelled frequently experienced higher attrition rates in general.
- Males generally had more attrition rates, although the difference to females was not that large.

From this analysis, we can conclude that low job satisfaction level, low monthly income, marital status, age and long distance to and from work are among the main causes of staff attrition in the company.

7. RECOMMENDATION

Now that we have identified the factors affecting attrition in the organization, let us make workable recommendations so that the company can increase their staff retention rate.

Since the rate of attrition is higher among low income earners in our analysis, we would recommend that the management review the performance of these low income employees and consider raising their pay so that they can opt to stay with the company instead of leaving. If the company is unable to increase their pay at this time, they could always result in using non-financial motivators. These would include things like praise from the boss, recognition for doing outstanding work and opportunities to lead projects for top performers in this group. These kinds of non-financial incentives motivate employees to do better and increase their job satisfaction level even though they may not earn much.

Attrition rate is higher in the sales department. A reason for this could be that this department is target driven and employees are monitored highly to ensure they meet their targets. This means that employees are micromanaged and face repercussions when they fail to deliver satisfactory sales. Most employees do not like being micromanaged, and may leave a position if they feel too controlled. Top performing sales reps particularly resent this type of control as they feel more empowered when they are allowed to manage their own time and allowed to make decisions that are in the customer’s best interest. Lack of rewards and recognition is also a major cause of sales reps leaving a company. Regardless of whether they make a sale or not, sales reps like to be recognized for their contribution to the company. Giving them this recognition would increase their job satisfaction level, leading them to stay longer with the company.

Any company serious about staff retention must know that good interpersonal relationships at the workplace increase an employees job satisfaction level and environment satisfaction as well. A reason for this is because great work relationships among employees reduces stress and increases productivity. Additionally, if employees feel that they have made meaningful friendships at work, they are more likely to stay. A way to build these bonds and keep them strong is to invest in and organize effective team building activities. These activities could include things like weekly or monthly team lunches, informal meetings and outings where employees get to socialize and know each other better.

With these strategies put in place, we believe that an organization may be able to increase their staff retention rate.