# Project 01

Taylor Saunders

2023-11-17

## Correlation between Employee Traits and Attrition

Employee attrition is attributed to the departure of a previous employee from a workplace for any reason. Simply put, it is the reduction of staff, and often results in the phenomenon of employers reducing positions available, meaning there are less employees at any given time due to less roles needing to be filled.

Although at first consideration attrition may be seen as a potential positive for the company due to pay cuts and other reasons, it is generally negative as attrition overall results in the cost of hiring new employees. Additionally. the challenge of finding individuals suited for the role that a possibly very experienced employee abandoned is an extreme loss for any company.

Of course, some situations present a positive side to attrition as the departure of one or more employees allows the opportunity for the company to potentially find better suited people for the roll, attributing but not limited to better suited talent, fresh perspectives, and culture.

However, employee retention has overall been described to be positive for corporations. Employee retention creates a good reputation for new hires when employees inevitably age out of the field, and employees who may leave due to unforeseen circumstances not following the trend of attrition (meaning they are an outlier) will have a positive view of the company and may recommend the company to people in their network. Other reasons include reduced costs for repeated training for a flow of new employees, having skilled and experienced teams, and creating familiar faces for customers which boosts customer service.

Because of these positive impacts, it is important for companies to try and reduce attrition as much as possible. However, there are many factors as to why individuals may leave a company that may be unidentifiable without data. I am attempting to uncover the correlation between employee characteristics and attrition. This dataset is entirely fictional and was crafted by IBM data scientists, however can be applied to any workforce.

The first step is to load in the dataset:

```
HR_DF <- read.csv("HR_Analytics.csv")
```

There are 1470 individuals being observed in this dataset.

## Considered Factors

The factors considered when analyzing employee lifestyle/demographics in relation to attrition are as follows:

1. Education - Below college, college, bachelor, master, doctor
2. Environment Satisfaction - Low, medium, high, very high
3. Job Involvement - Low, medium, high, very high
4. Job Satisfaction - Low, medium, high, very high
5. Performance Rating - Low, Good, Excellent, Outstanding
6. Relationship Satisfaction - Low, medium, high, very high
7. Work life balance - Bad, good, better, best

Please note that the factors are labeled using numeric values, so for example: Education - 1 (below college), 2 (college), 3 (bachelor), 4 (master), 5 (doctor)

In this project, I will be working heavily with the Education Level of employees, and showcase a final example using Environment Satisfaction.

There are multiple other factors that this project scope does not contain. Following is a full un-abridged list:

```
colnames(HR_DF)
```

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EmployeeCount"            "EmployeeNumber"
## [11] "EnvironmentSatisfaction"  "Gender"
## [13] "HourlyRate"               "JobInvolvement"
## [15] "JobLevel"                 "JobRole"
## [17] "JobSatisfaction"          "MaritalStatus"
## [19] "MonthlyIncome"            "MonthlyRate"
## [21] "NumCompaniesWorked"       "Over18"
## [23] "OverTime"                 "PercentSalaryHike"
## [25] "PerformanceRating"        "RelationshipSatisfaction"
## [27] "StandardHours"            "StockOptionLevel"
## [29] "TotalWorkingYears"        "TrainingTimesLastYear"
## [31] "WorkLifeBalance"          "YearsAtCompany"
## [33] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

The demographic information I will use for this project includes gender and marital status.

#Data Cleaning

Before any data exploration or visualization can be done, I took the necessary steps to prepare the data. I also loaded the necessary packages for this project. NOTE: Even if this dataset does not change after the following steps (meaning the dataset is in good condition/suitable for statistical use), it is still good practice to take precautions.

```
suppressMessages(library(dplyr))
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
HR_DF <- HR_DF %>%
  # removes all duplicate rows
  distinct(.keep_all = TRUE) %>%
  #removes all na/empty values
  na.omit()
```

## Data Exploration

Secondly, visualizing general demographics can give us a thorough understanding of the data we work with. I will begin this process by creating tables and graphs to enable us to visually compare and analyze information before proceeding with further analysis.

```
mean_age <- mean(HR_DF$Age, na.rm = TRUE)
```

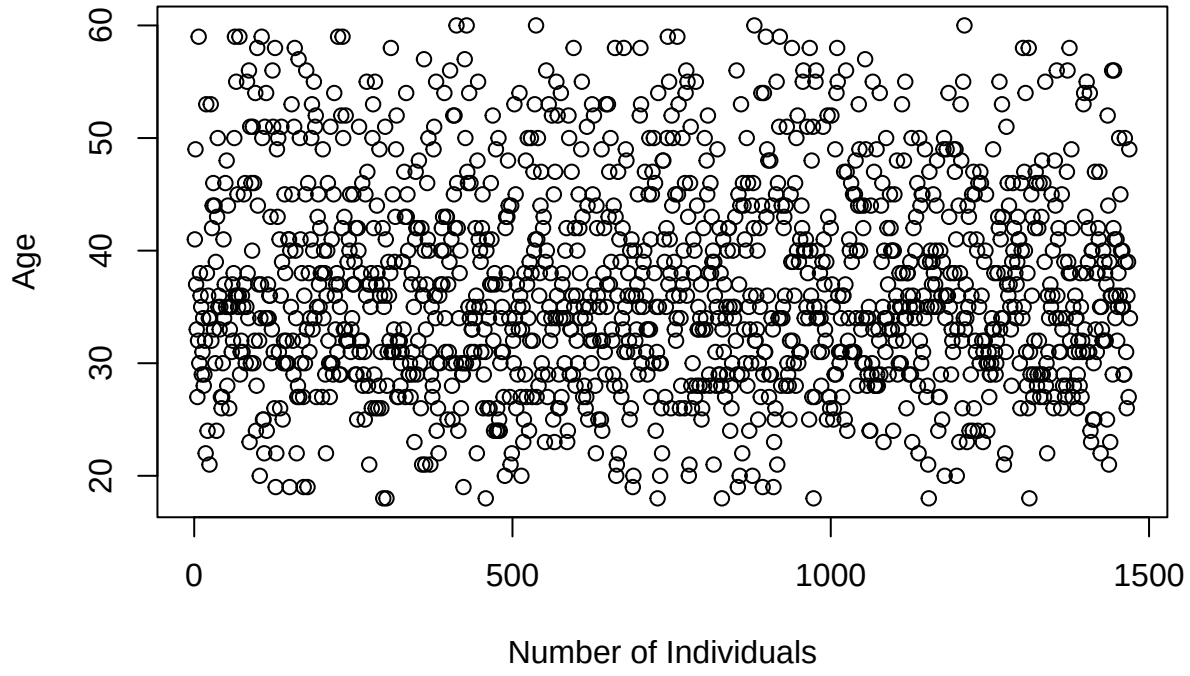
The average age of individuals in our dataset is 36.9 years.

I will be using ggplot to create the graphs for this report.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

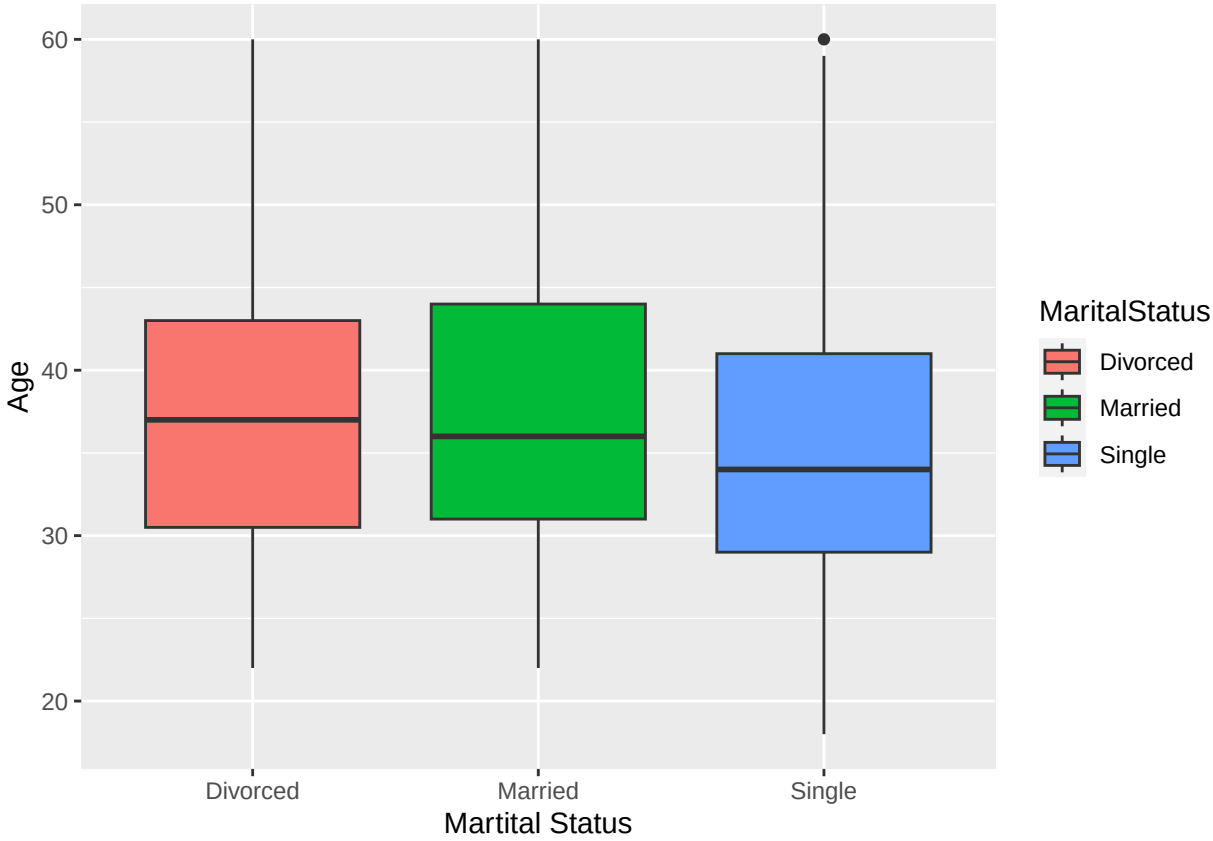
```
min_age <- min(HR_DF$Age, na.rm = TRUE)
max_age <- max(HR_DF$Age, na.rm = TRUE)
plot(HR_DF$Age,
     xlab = "Number of Individuals",
     ylab = "Age")
```




The range of ages in our data frame is 18 to 60 years old.

To analyze the distribution of the Matrital Status of employees by their age, we can proceed by makin a boxplot:

```
ggplot(HR_DF, aes(MaritalStatus, Age, fill = MaritalStatus)) +
  geom_boxplot() +
  xlab("Martital Status")
```

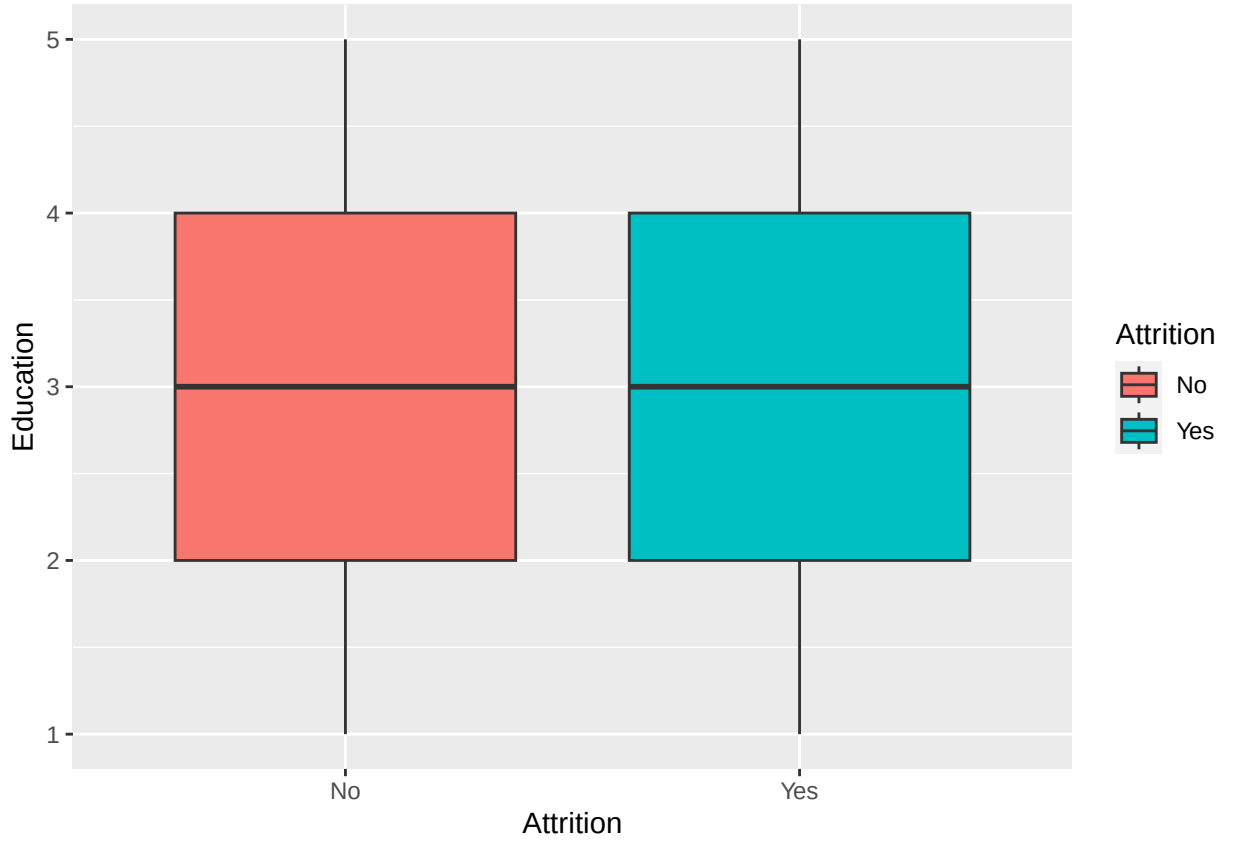



As we can see, divorced or married employees are above 25, while single employees cover the full range of 18-60 years of age. This is not something we will be evaluating to determine if it leads to employee attrition, however in a higher-scope project this could be a valuable insight.

## Correlation Testing

Now, we will determine the correlation between Education and Attrition

```
ggplot(HR_DF, aes(Attrition, Education, fill = Attrition)) +
  geom_boxplot() +
  xlab("Attrition")
```




Using the above plot, it appears that the mean education of employees is a bachelors degree. We will verify this using the mean function: Based on this plot, the mean

```
mean(HR_DF$Education)
```

```
## [1] 2.912925
```

Additionally, there appears to be no correlation between education and attrition when using a box plot as both have equality.

However, to be entirely sure we will perform a correlation test on Education and Attrition to determine if there is a factual correlation.

## Pearson Correlation Formula Attempt

To do so, we first attempt will follow the t-score of a correlation coefficient (r) using the Pearson correlation formula:

$t = r * \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$

This equation always results in a value between -1 and 1, where: -1 = negative linear correlation 0 = no linear correlation 1 = positive linear correlation

For some clarification, a positive correlation is identified as when one variable increase, the other also increases. A negative correlation is when while one variable increases, the other decreases. A 0 correlation means there is no relationship between the two variables.

To determine the correlation between Education and attrition, we proceed as follows:

```
library("ggpubr")
```

```
## Warning: package 'ggpubr' was built under R version 4.3.2
```

```
library("rstatix")
```

```
## Warning: package 'rstatix' was built under R version 4.3.2
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
result <- cor.test(HR_DF$Education, as.numeric(HR_DF$Attrition == "Yes"), method = "pearson")
result
```

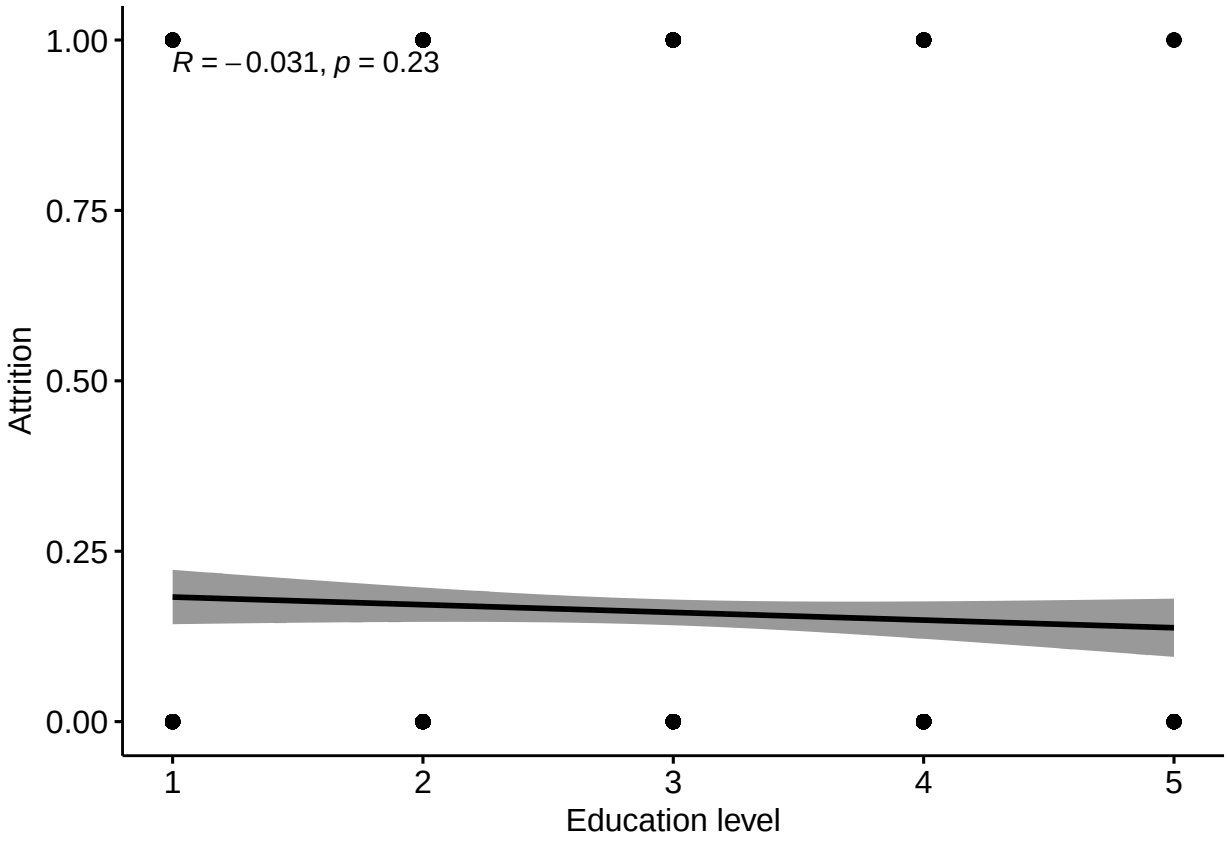
```
##
##  Pearson's product-moment correlation
##
## data:  HR_DF$Education and as.numeric(HR_DF$Attrition == "Yes")
## t = -1.2026, df = 1468, p-value = 0.2293
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08236816  0.01978637
## sample estimates:
##         cor
## -0.03137282
```

The correlation coefficient is -0.0314. This suggests a weak negative correlation, meaning as education increases, attrition decreases by a small amount. However, since our P-value is 0.223, which is greater than 0.05 (chosen significance level), we do not have evidence to suggest a strong correlation between the level of education an employee has and their attrition.

Here is a visual representation of the test that just took place:

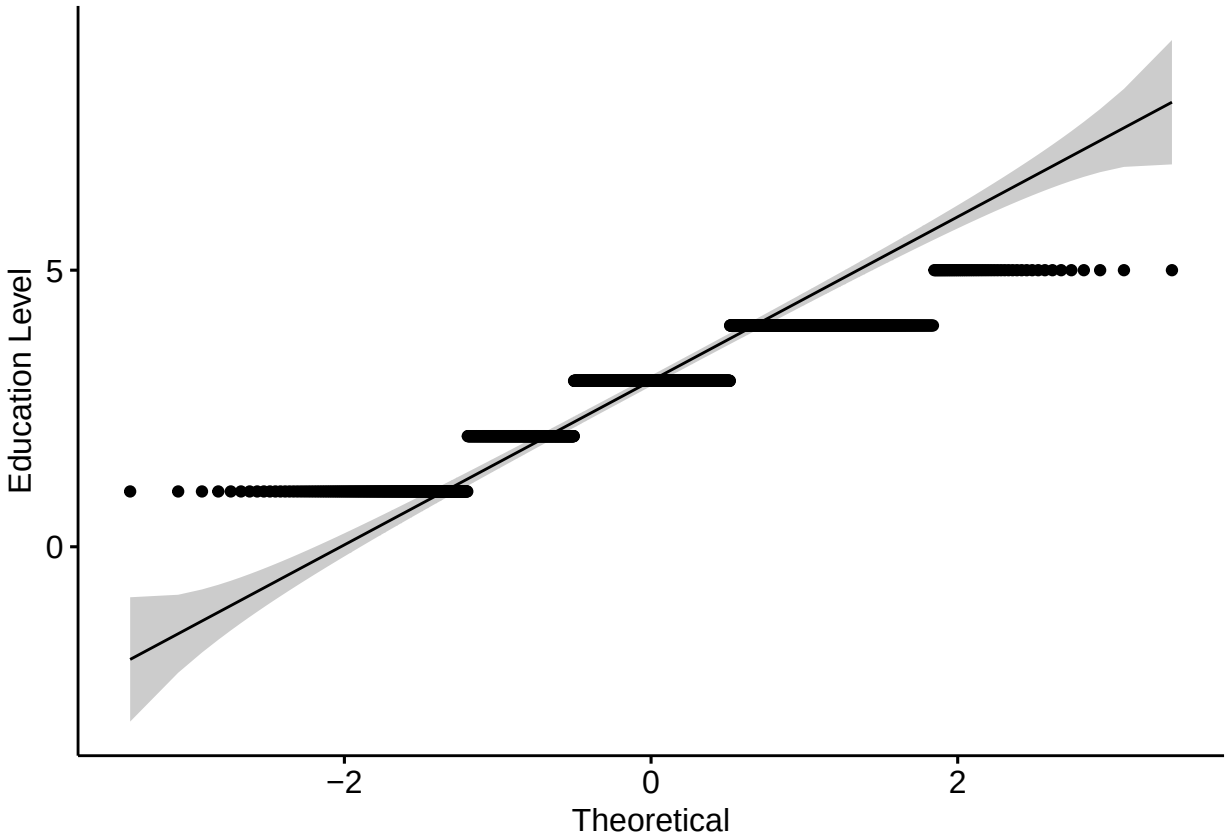
```
HR_DF_2 <- HR_DF %>%
  mutate(Attrition_Numeric = ifelse(Attrition == "Yes", 1, ifelse(Attrition == "No", 0, NA)))

ggscatter(HR_DF_2, x = "Education", y = "Attrition_Numeric",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Education level", ylab = "Attrition")
```




We seem to draw a solid conclusion using the Pearson Formula. However, our data does not have a normal distribution:

```
ggqqplot(HR_DF$Education, ylab = "Education Level")
```




Our data showcases a rank based distribution. Because of this, we will get more accurate results using the Kendall rank correlation test, however not very different. This is because the Kendall rank correlation coefficient is used to estimate a rank-based measure of correlation, which is what our variables Attrition (0-1) and Education (1-5) represent.

## Kendall Rank Correlation Test

The Kendall Correlation Formula is as follows:

tau = $\frac{c-d}{\frac{1}{2}(n-1)}$

Where: c = total number of concordant pairs d = total number of discordant pairs n = size of x and y

A positive tau value indicates a positive monotonic relationship, which indicates as one variable increases, so does the other. A negative tau value indicates a negative monotonic relationship, which indicates that as one variable increases, the other decreases. The rate of increase/decrease depends on the magnitude of the resulting tau value.

```
result2 <- cor.test(HR_DF_2$Attrition_Numeric, HR_DF_2$Education,  method="kendall")
result2
```

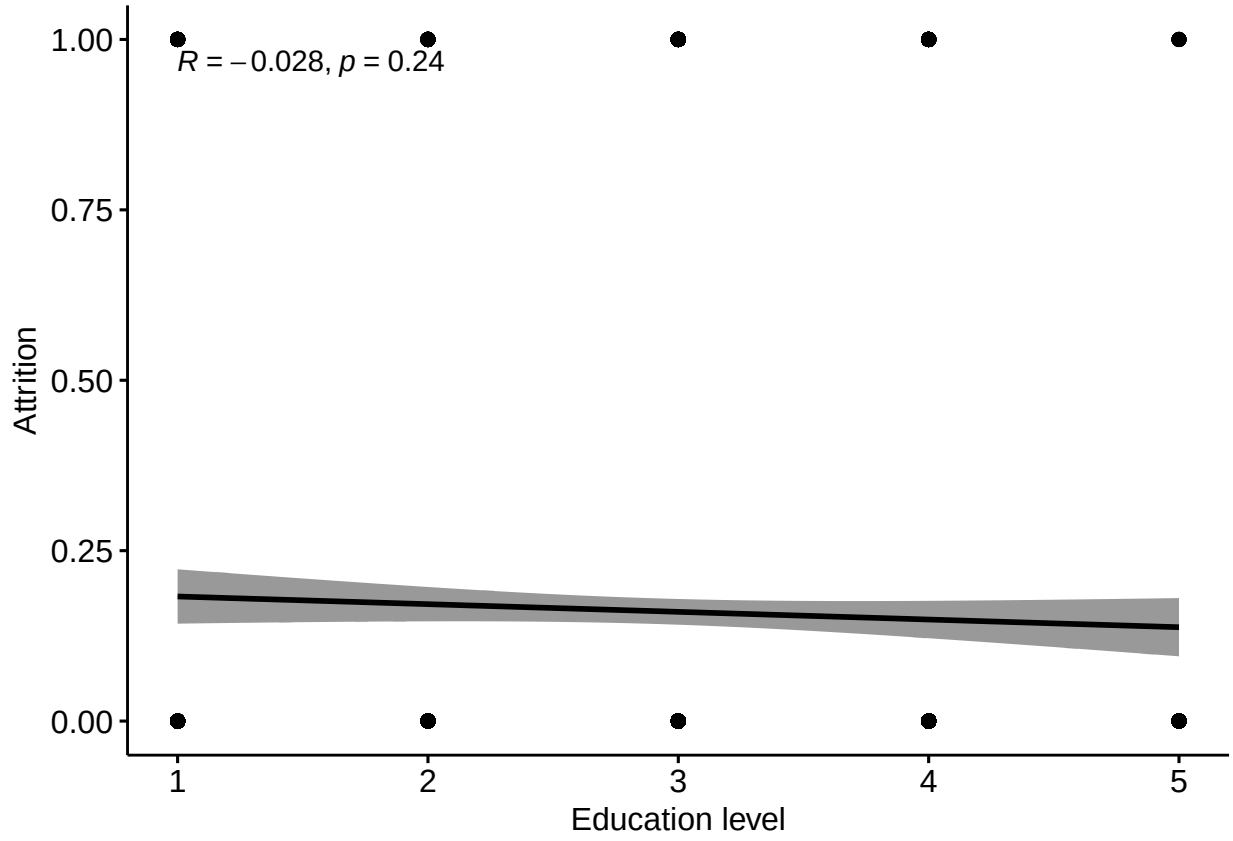
```
##
##  Kendall's rank correlation tau
##
## data:  HR_DF_2$Attrition_Numeric and HR_DF_2$Education
## z = -1.1631, p-value = 0.2448
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
```

```
## -0.02781708
```

Here, we achieve very similar results to when we used the Pearson formula. Our p-valye is still 0.2, which means we cannot derive the conclusion that there is a correlation between Education level and employee attrition.

Visualization of the above test:

```
ggscatter(HR_DF_2, x = "Education", y = "Attrition_Numeric",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "kendall",
          xlab = "Education level", ylab = "Attrition")
```




In conclusion, the two tests I performed (Kendall and Pearson) provided very similar results. The conclusion derived about the correlation between Education level and employee attrition is that there is a weak negative correlation, meaning as education increases, attrition decreases by a small amount. However, since our P-value is 0.24 (using Kendall), which is greater than 0.05 (chosen significance level), we do not have evidence to suggest a strong correlation between the level of education an employee has and their attrition.

## Final Example

Finally, I will provide a final example using Environment Satisfaction and Attrition,

```
result3 <- cor.test(HR_DF_2$Attrition_Numeric, HR_DF_2$EnvironmentSatisfaction,  method="kendall")
result3
```

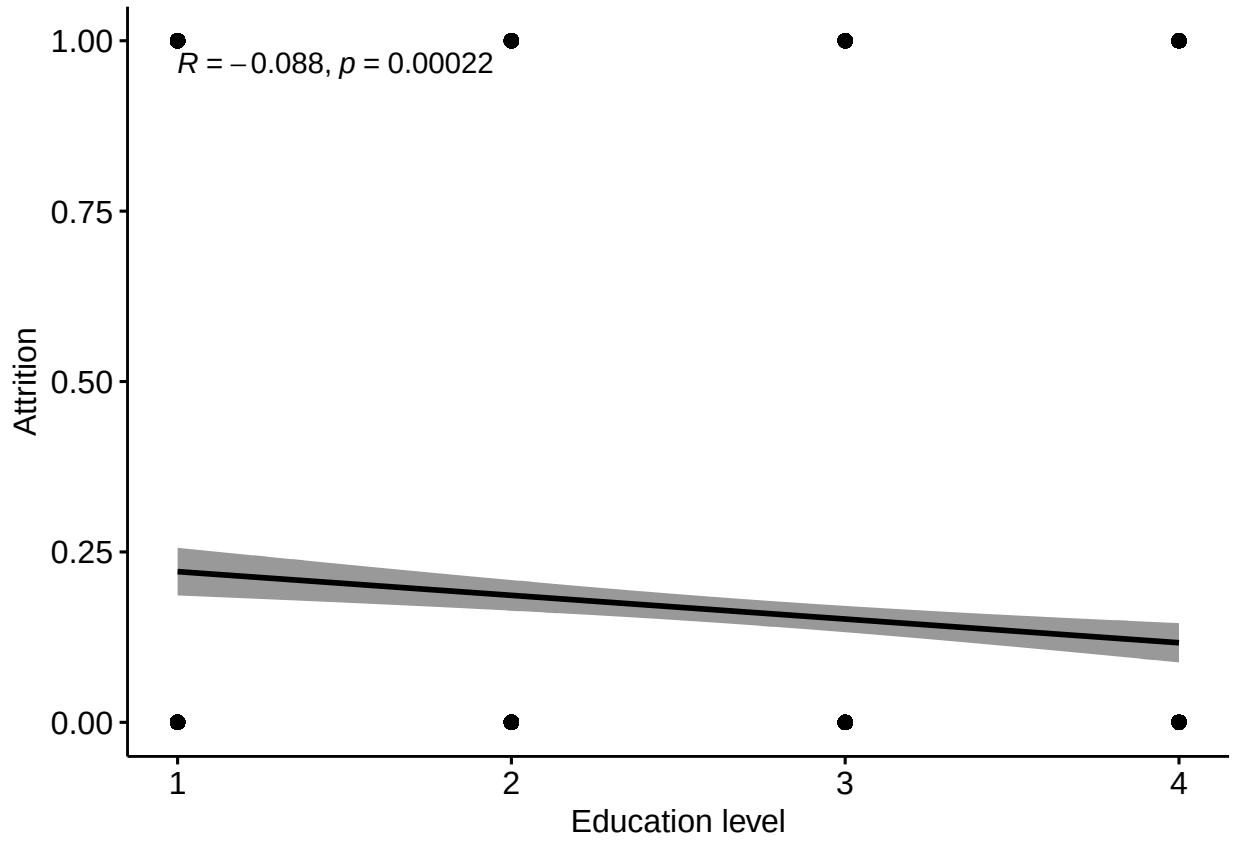
```
##
##  Kendall's rank correlation tau
##
## data:  HR_DF_2$Attrition_Numeric and HR_DF_2$EnvironmentSatisfaction
```

```
## z = -3.6981, p-value = 0.0002172
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.08837668
```

Here, we see that our tau is -0.088 and our p-value is 0.0002. Our p-value is extremely small, meaning we have strong evidence that there is a correlation between Environment Satisfaction and Attrition. Put simply, as an employee's satisfaction with their environment increases, their attrition decreases. This makes sense and can be logically derived, as a happy employee is more likely to stay with their employer compared to an employee who is unhappy.

Visualization:

```
ggscatter(HR_DF_2, x = "EnvironmentSatisfaction", y = "Attrition_Numeric",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "kendall",
          xlab = "Education level", ylab = "Attrition")
```




## Conclusion

Based on the material covered in this project, we can confidently uncover correlation between various variables employees hold and that same employee's attrition. Our obtained values found using different formulas have different assumptions and measure different aspects of the data, so it is important to explore your data to determine which formulas fit your needs best. Future steps will include creating predictive models to predict which variables account for an employees attrition, which can be useful to employers in determining how to ensure employee retention.

## Data Sources

IBM HR Analytics Employee Attrition & Performance. (2017, March 31). Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

## Background Research Source

The advantages of employee retention. (2021, November 20). Bizfluent. https://bizfluent.com/info-7980915-advantages-employee-retention.html