An analysis into what drives employee attrition

Context: employee retention is cruical to an organisation's success, and attrition rate is a metric that can provides insight into how well they are able to retaining their employees. Understanding the key variables which play a factor in employee attrion can help a company develop relevant and effective strategies to reduce attrition and retain the key employees that enable their organisation to succeed.

Goal: a fictitious company based in Europe is interested in understanding which variables, if any, drive employee attrition. They have produced an employee dataset spanning 2005 - 2016 tracking things such as employee gender, location, seniority, age and length of service. The goal of this project is to analyse the data to see if any of these variabl..

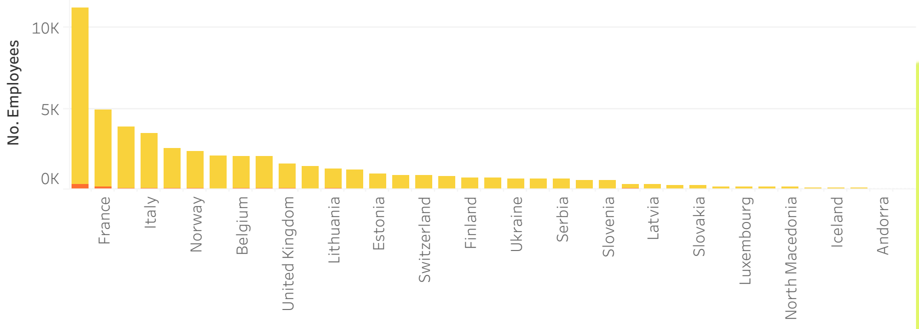
Data source: this source for this project is The source of this dataset is a Kaggle project called Employee Attrition. This is a fictionalised dataset for the purposes of trying to predict employee attrition, and with permis..

Exploratory Data Analysis

Exploratory analysis was used to better understand the dataset and the variables we have c

Employees are spread across **39 European countries**. Click on each country to explore the data more!

Employme.. Acti.. Ter..

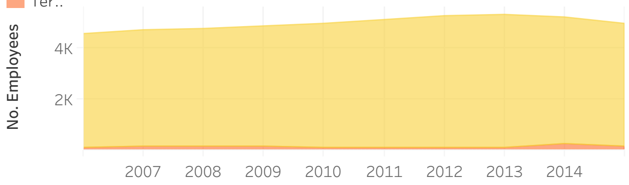


Year Of Rec.. All

Country All

Highlight C.. No items h..

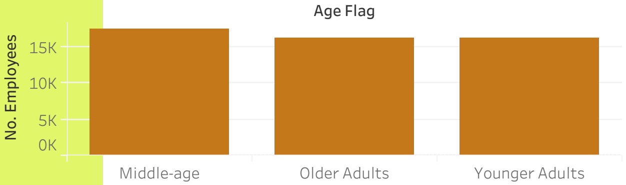
Active vs Terminated Employees



Employment Status
☑ Active
☑ Terminated

Year Of Record
All

We have a ***very imbalanced*** dataset, whereby our potential dependent variable - employment status - is split 97% vs 3%. This is important to note as it may impact some analyses later on. ..



Numbers of employees across each of the three age-groups are broadly the same, although there is a slight increase in middle-age group adults.

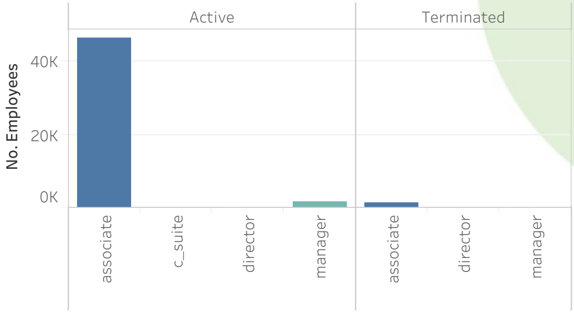
Young adults: 18-34, Middle-age adults: 35-49, Older-adults: 50-65

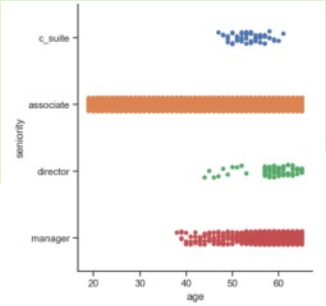
How do seniority and age impact?

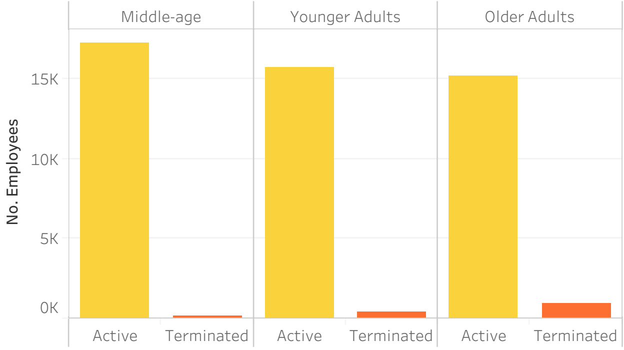
Seniority
☑ associate
☑ c_suite
☑ director
☑ manager

Employment Status
☑ Active
☑ Terminated

When looking at employees by seniority and age, we see that the majority of employees fall within the associate level, and that the age of an associate is evenly distributed across the dataset.
The remaining three seniority groups have far fewer employees, but those employees tend to all be aged 40 or above. There were no terminations at the C-suite level, and very few within Director.



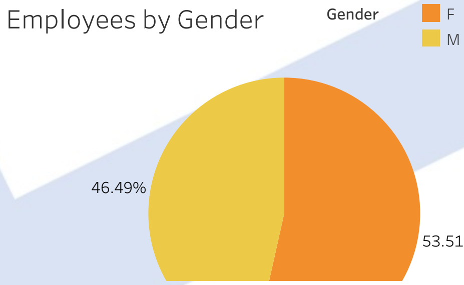


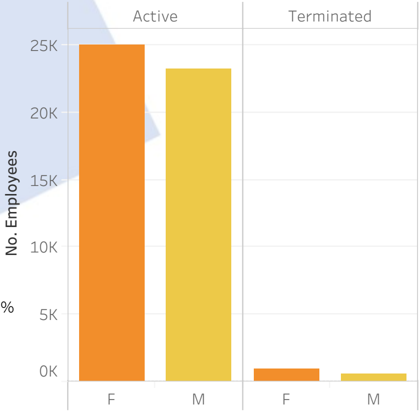


If we look at employment status by age group, we see that the majority of terminations are within the older-adults category. This is perhaps unsurprising when we consider that terminations at the fictional company include retirements as well.
Younger-adults account for the second highest number of terminations, ..

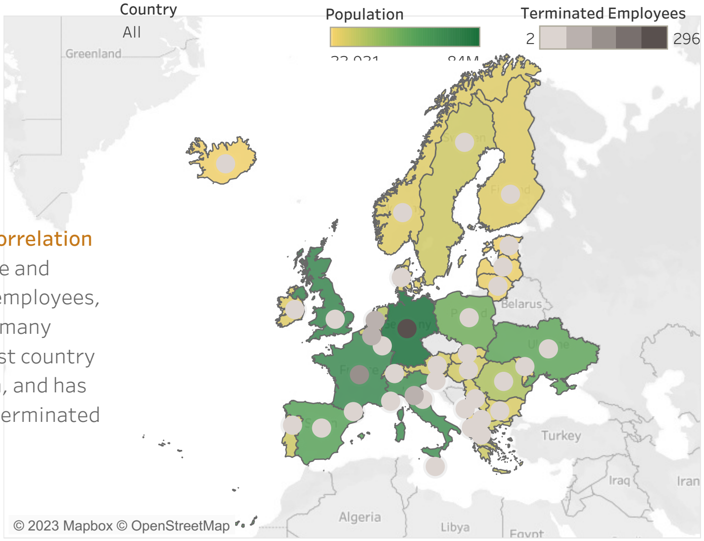
Do gender or population play a part?

Females make up over 53% of employees, versus 46% for males. When comparing active and terminated employees, we see **no significant difference in results when cutting by gender,** the breakdown is broadly in line w..





We see some **positive correlation** between population size and number of terminated employees, as you can see with Germany which is both the largest country in Europe by population, and has the highest number of terminated employees.

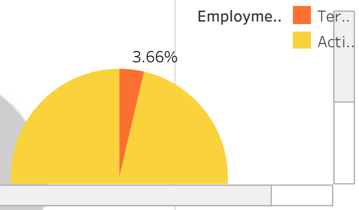


Trying to predict employment status

Based on the results of the exploratory data analysis, age appears to be the variable which may have an impact on employment status, so our initial hypothesis is:

The older an employee is, the more likely they are to be terminated.

The key dependent variable initially looked at is employment status, which tells us whe..

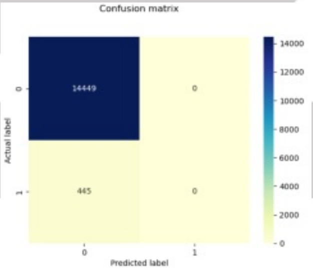


Step one: understanding our variables

Employment status is a categorical variable, which means we need to use a logistic regression model instead of linear. ..

Step two: feature scaling, test and train, and fitting the logistic regression model:

Feature scaling helps to normalise the range of independent variables, whilst splitting the data into test and train will allow us to verify the results of the train set by using the test set. I used a stratify split due to the imbala..



Step three: evaluating our model:

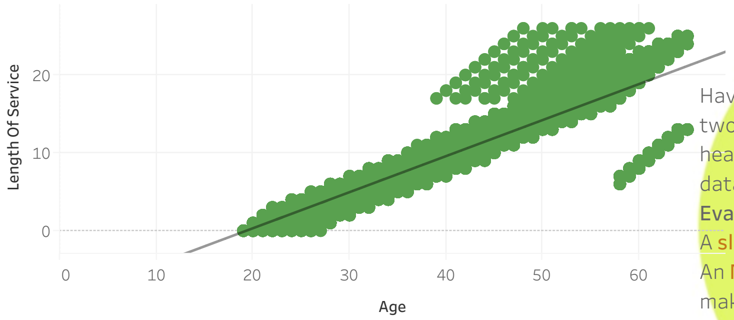
The model achieved an accuracy score of 97%, however, it was only able to predict **active** employees, not **terminated.** The **confusion matrix** states that there at 14,449 true negatives and 445 false negatives as well as 0 false or true predictions. I believe this is due to the imbalanced nature of the dependent variable.

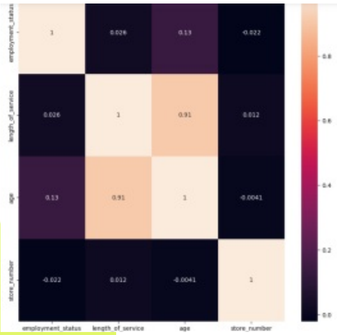
So in conclusion, logistic regression is an unreliab..

Exploring other relationships

Having returned to the initial exploratory analysis, a very strong correlated relationship of 0.91 was observed between age and length of service. A correlation heatmap is demonstrative of a mark of **linear interdependence,** therefore, this result warrants further analysis.

So, could this be an indicator in our dataset:..





Having split the data into test and traing, and fit the linear model to the dataset, a scatterplot between the two variables again shows a strong linear correlation, which supports the findings from the correlation heatmap.It's worth noting that as employees get older, there appears to be more *variance* in the data, with data points less tightly clustered together and further away from the trend line.

Evaluation of the model:

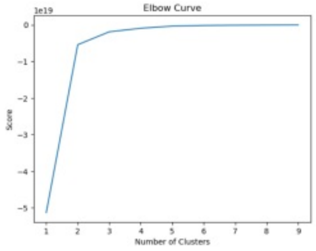
A slope score of 0.46 indicating a *positive* correlation.

An MSE of 6.81 which is a small number, meaning the regression line passes fairly close to the observations, making it a good fit.

An R2 score is 0.82 indicating a strong fit...

Machine learning: clustering

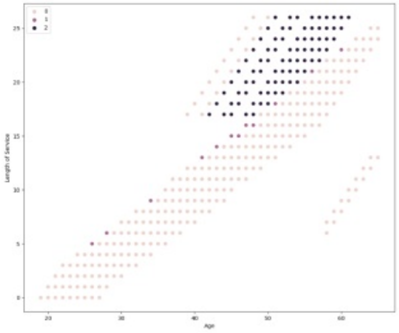
To further explore the relationship between Age and Length of Service, clustering was applied using the k-means algorithm. Clustering helps identify any trends and patterns characterised by similarities which aren't always obvious in other visualisations or analysis.



An elbow curve plot indicates that 2 or 3 clusters would be an appropriate number of groupings before we plot the results on a scatterplot.

The scatterplot reveals a very clear cluster of employees - shaded in dark purple - who are between ~42 and 60, and have been with the company for between 17 and 25 years. Supporting the suggestion that older employees are more likely to stay with the organisation for long periods of time.

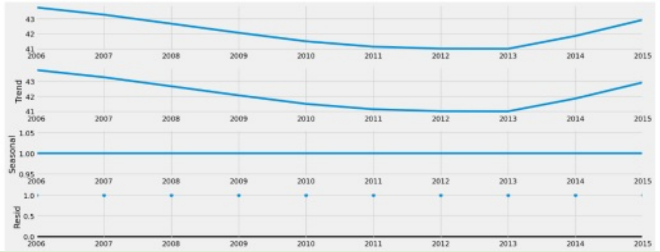
When looking at the rest of the results, the distinctions are less clear. The majority of data points fall within the light purple cluster, and then ..

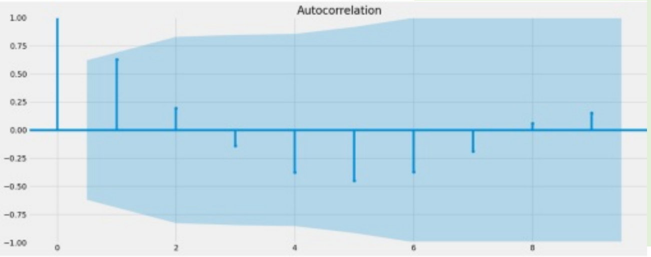


Time-Series Analysis

Finally, **decomposition** of the time series was used to allow for the assessment of individual components, such as seasonality, to identify possible trends.

The analysis demonstrated that there was *very limited* , if any, seasonality and autocorrelation within the dataset.





Conclusions and Recommendations

Conclusions and Recommendations

In conclusion, age appears to impact an employee's length of service, meaning that generally speaking, the older an employee is the more likely they are to have been with the organisation for a significant period of time.

The caveat to this of course, is that terminations are still higher in the older-age group cateogry compared to middle and younger age groups, likely because a number of those older employees are retiring from the company.

Interestingly, other variables such as gender and country population appear to have no statistically significant impact on either the length ..

Limtations of the dataset, and what's next?

The biggest challenge in this project was the imbalance of the dataset. The initial dependent variable was employment status, and being able to accurately predict that based on other variables would have been a successful outcome for this project. The split of 97% vs 3% however, made it challenging for the logistic regression model to accurately predict terminated employees, as seen by the results of the confusion matrix.

In terms of next steps, it would be great to:

1. Collect more data, in particular, having more terminated employees could help us reveal greater insights. We could even focus solely on those employees to understand which variables are most impactful.
2. Collect more data points, perhaps these variables don't have any impact on employee attrition, so collecting difference information e.g., performance reviews, family makeup, salary, or educational background, could help reveal trends and patterns which we haven't explored here.

..