SURVIVAL ANALYSIS FOR EMPLOYEE TURNOVER

Author: Minh Nguyen

Date: 21 October 2024

Contents

Project objectives

The objective of this analysis is to predict employee turnover using Survival Analysis. This approach will help identify key factors, such as age and wage category (greywage), that influence how long employees stay with the company. By modeling time-to-event data, the analysis provides actionable insights for improving employee retention strategies, helping stakeholders make informed decisions to reduce churn and enhance workforce management.

Overview about the dataset

The dataset used for this analysis contains information on employee turnover, with a focus on employee tenure and whether they have left the company (churned) or remained. The dataset can be found here on Kaggle. Key attributes include:

- stag: The number of days an employee has worked (tenure).
- event: A binary indicator where 1 represents an employee who has churned, and 0 represents an employee who has remained.
- age: The age of the employee.
- greywage: A classification of employees’ wage category, either “grey” or “white.”
- Additional variables related to demographics and personality traits, such as gender, coach status, and psychological factors like extraversion and anxiety.

The goal of the analysis is to use these attributes to build a model that accurately predicts employee churn, helping us identify the most important factors that influence retention and provide insights for reducing turnover.

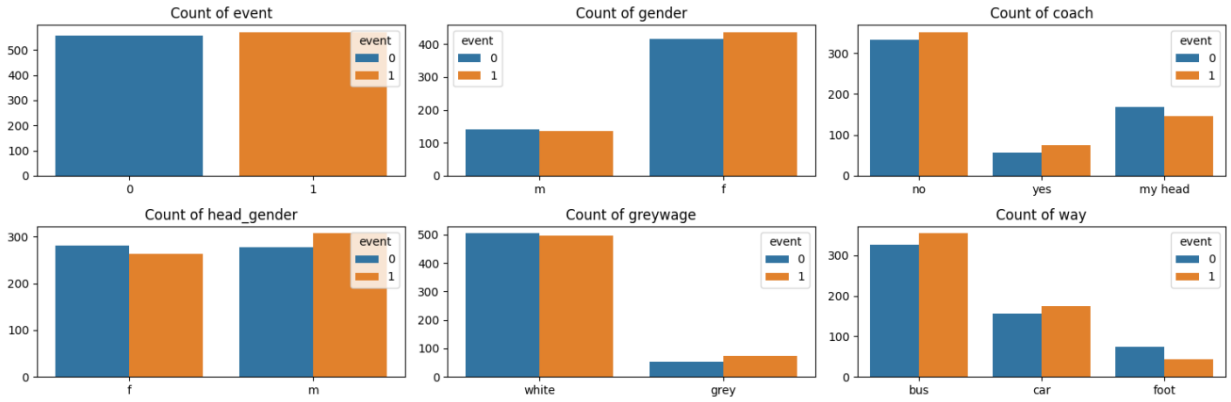
Summary of data exploration

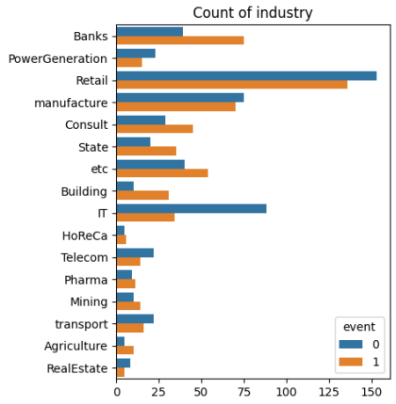
The dataset consists of 1,129 observations with features capturing employee demographics, tenure, and behavioral traits, along with the event variable indicating whether an employee has churned (1) or remained (0). Below is a summary of the key findings and steps taken during data exploration.

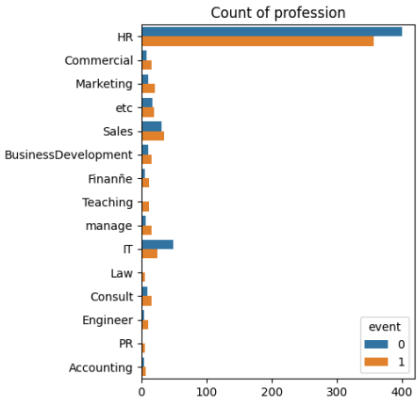
	stag	event	age	extraversion	independ	selfcontrol	anxiety	novator
count	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000	1129.000000
mean	36.627526	0.505757	31.066965	5.592383	5.478034	5.597254	5.665633	5.879628
std	34.096597	0.500188	6.996147	1.851637	1.703312	1.980101	1.709176	1.904016
min	0.394251	0.000000	18.000000	1.000000	1.000000	1.000000	1.700000	1.000000
25%	11.728953	0.000000	26.000000	4.600000	4.100000	4.100000	4.800000	4.400000
50%	24.344969	1.000000	30.000000	5.400000	5.500000	5.700000	5.600000	6.000000
75%	51.318275	1.000000	36.000000	7.000000	6.900000	7.200000	7.100000	7.500000
max	179.449692	1.000000	58.000000	10.000000	10.000000	10.000000	10.000000	10.000000

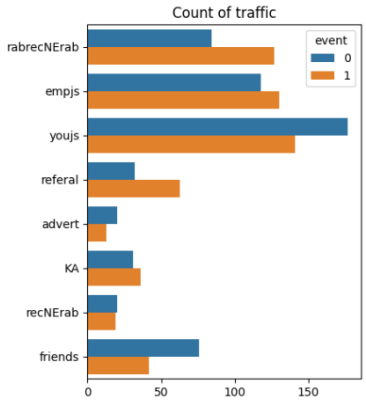
Basic statistics:

- The average tenure (stag) of employees is around 36 days, with a standard deviation of 34 days, indicating a wide variation in how long employees stay with the company.
- Employees have an average age of around 31 years, and personality traits such as extraversion, independence, self-control, and anxiety are recorded on a 1-10 scale.









Event: The churn rate is balanced, with about 50% of employees having churned (event = 1) and 50% remaining (event = 0).

Gender: The dataset is predominantly female, with around 75% of employees identifying as female. There is almost an equal proportion of churned (event = 1) and retained (event = 0) male employees. A higher number of female employees churn compared to male employees.

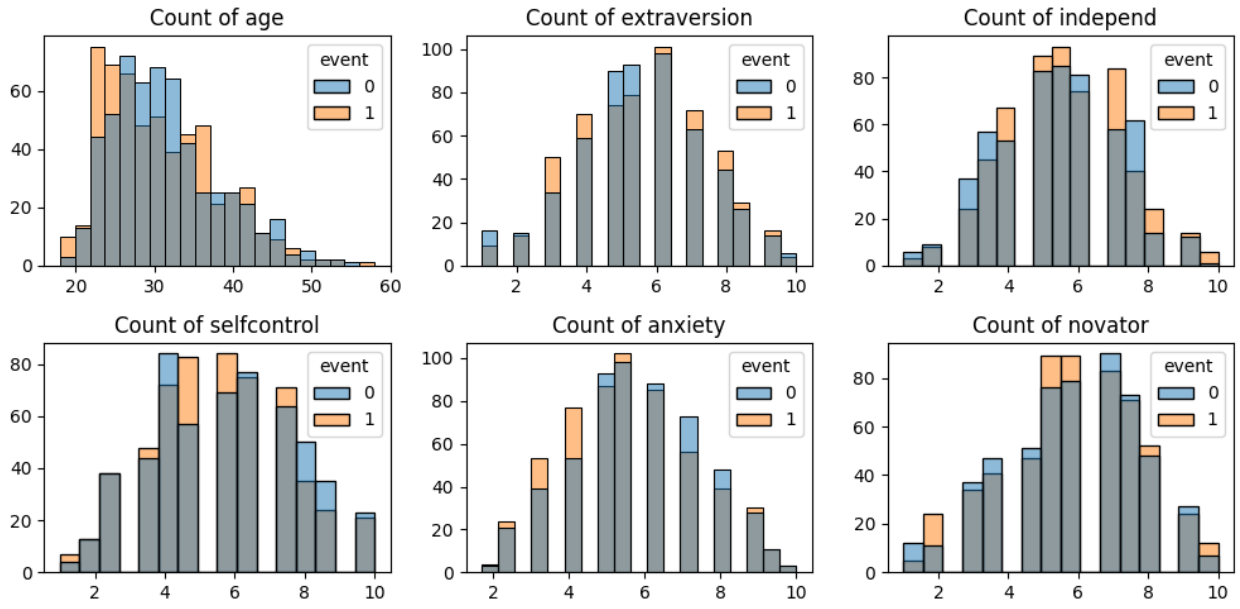
Coach: Most employees do not have a coach, but some are coached by "my head" or another coach. Employees with no coach have a slightly higher churn rate. Employees with a coach or who are coached by "my head" show more retention, though the difference is modest.

Head Gender: There is a slight skew in churn rates based on the gender of the manager (head_gender), but it is not particularly pronounced. Employees under both male and female supervisors have similar churn rates.

Greywage: The majority of employees fall into the "white" wage category, with only a small portion in the "grey" category. Employees classified as grey have a much higher churn rate compared to those classified as white.This **suggests that greywage is a strong predictor of churn**, as those in the "grey" wage category are leaving the company at a higher rate.

Way: Employees using buses tend to churn more frequently, while those who use cars or go on foot seem to show less churn.

Industry & Profession: The data covers employees from various industries, with retail being the largest sector, and HR being the dominant profession. **Retail industry has the highest churn rate**, followed by banks. These industries show a strong correlation with higher employee churn. Other industries, such as IT, show a much lower churn rate.



Age: Employees **aged 20 to 35 show the highest churn** (orange bars), especially in the age group 25 to 30, where churn is most frequent. Employees over 35 are less likely to churn (lower orange bars), particularly those in the 40-50 range, where fewer employees leave. Younger employees (particularly under 35) are more prone to churn, while older employees (over 35) have higher retention rates.

Extraversion: Employees with **mid-level extraversion scores (5 to 7) show the highest churn** rates. Employees with very low or very high extraversion scores tend to remain, as seen by the lower churn (orange bars) in these groups. Employees with moderate extraversion tend to churn more, while those on the extremes (either very introverted or very extroverted) show better retention.

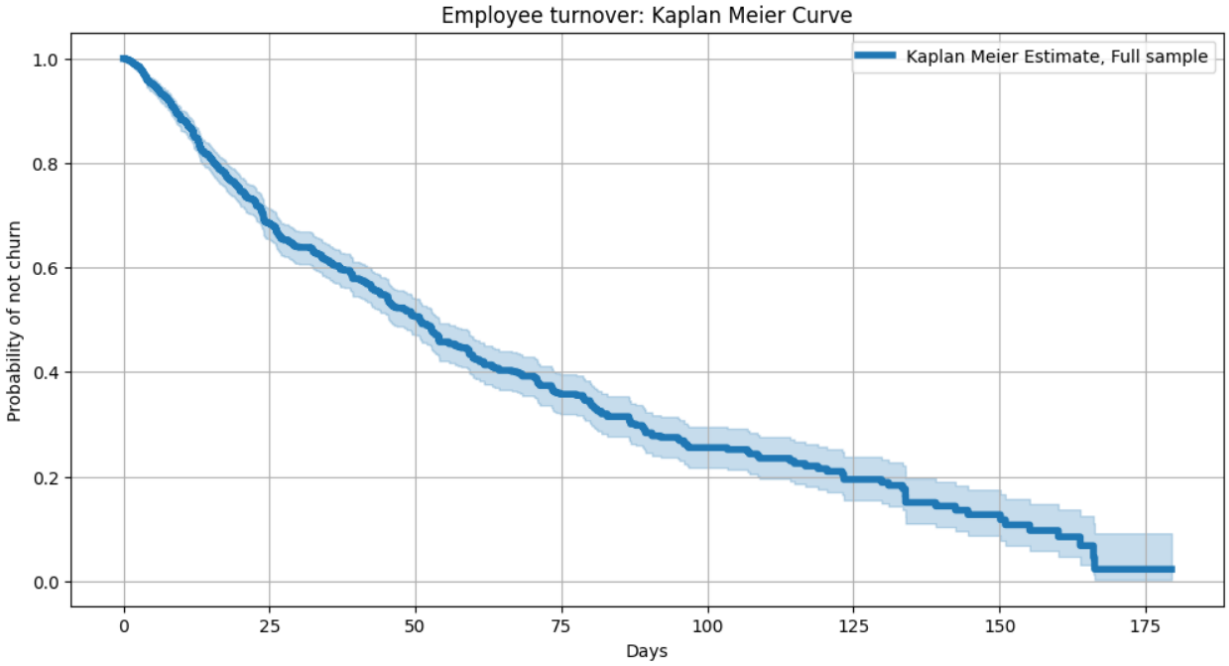
Independence: Churn is **concentrated among employees with independence scores between 5 and 7**. Employees with low independence scores (below 4) and high scores (above 8) churn less often. Employees with moderate levels of independence are more likely to churn, while those who are either very dependent or very independent have higher retention rates.

Self-control: Employees with self-control scores between **5 and 7 show the most churn**, with fewer churn cases at both low and high extremes. Moderate self-control scores correspond with a higher likelihood of churn, while those with either very low or very high self-control are more likely to stay.

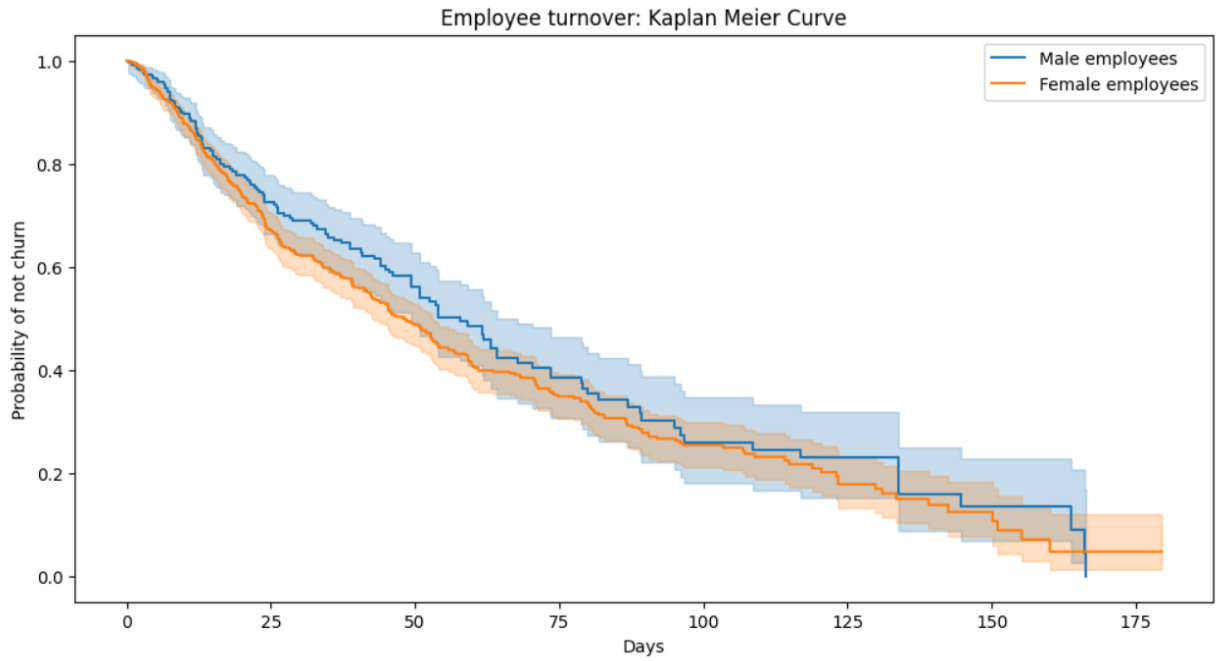
Anxiety: Churn **is more frequent for employees with mid-level anxiety scores (5 to 7)**. Employees with low anxiety levels (1-3) and high anxiety levels (8-10) show fewer churn cases. Employees with moderate anxiety levels are more likely to leave, while those with either very low or very high anxiety levels tend to remain with the company.

Novator: Employees with **mid-level novator scores (5-7) experience the highest churn**, especially those with scores of 6 and 7. Employees with low scores (1-3) show lower churn rates and are more likely to remain. Similarly, those with high scores (8-10) also exhibit lower churn,

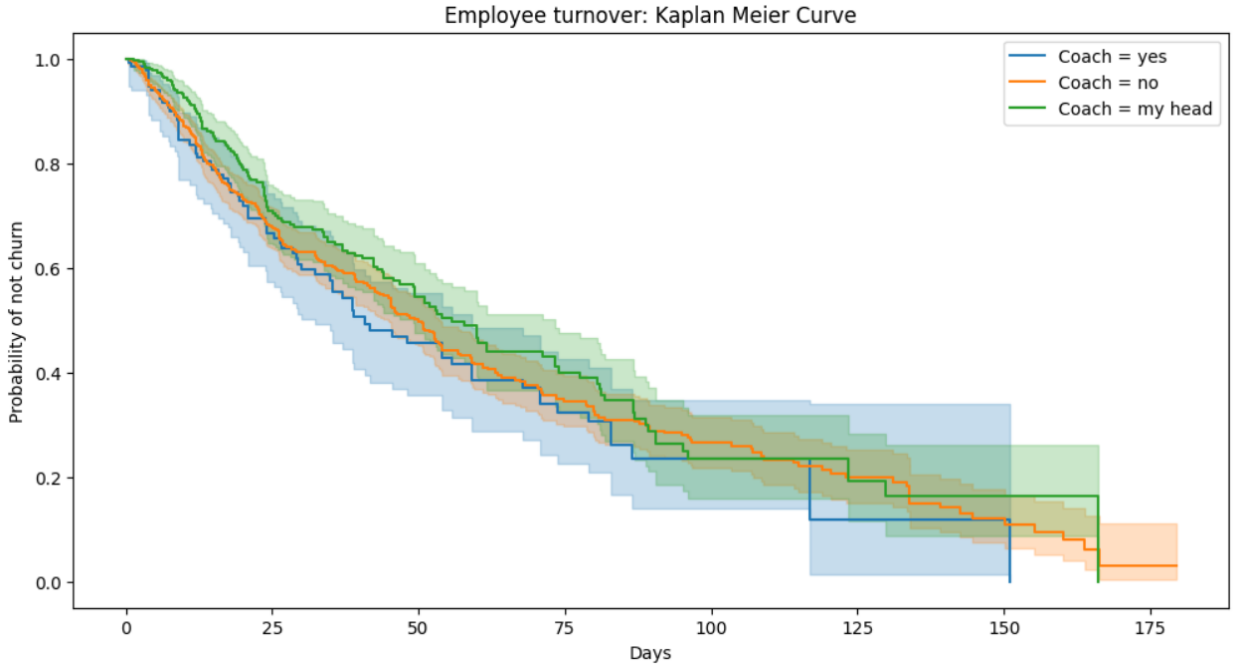
indicating that employees at both the lower and higher ends of innovativeness tend to stay, while those in the middle range are more likely to leave.



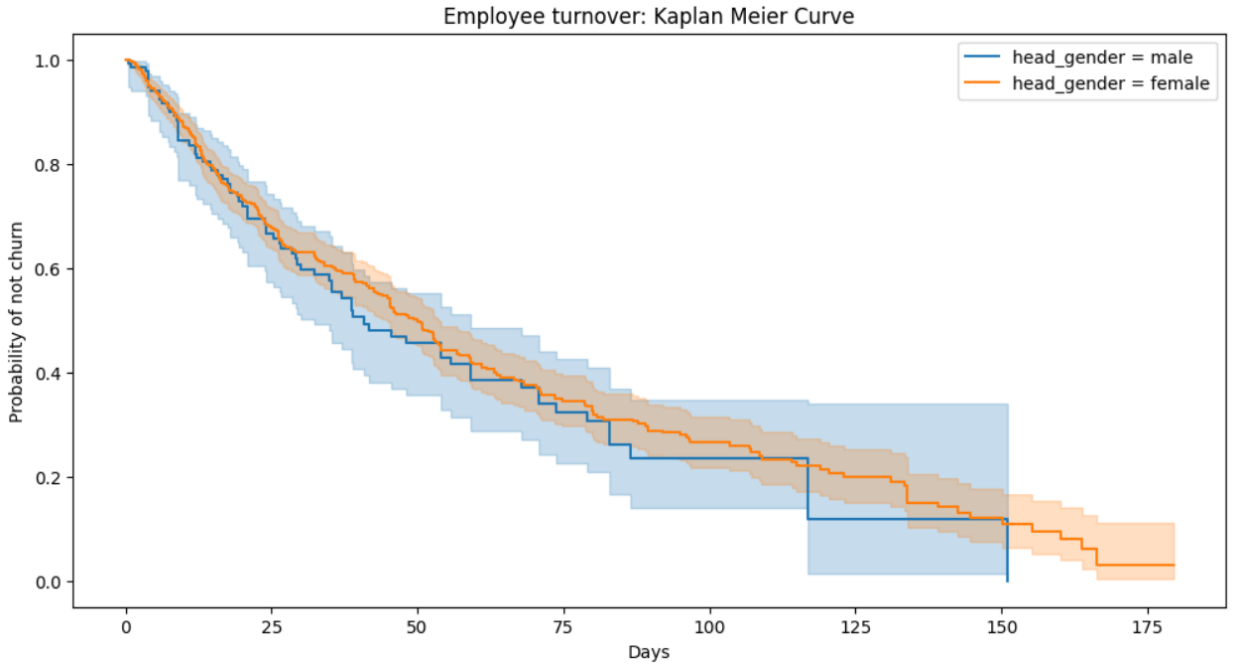
This Kaplan-Meier curve shows the probability of employees staying with the company over time. The retention rate steadily declines, with about 50% of employees churning by 50 days and nearly all employees leaving by 175 days. The shaded area represents the confidence intervals, with more uncertainty as time progresses.



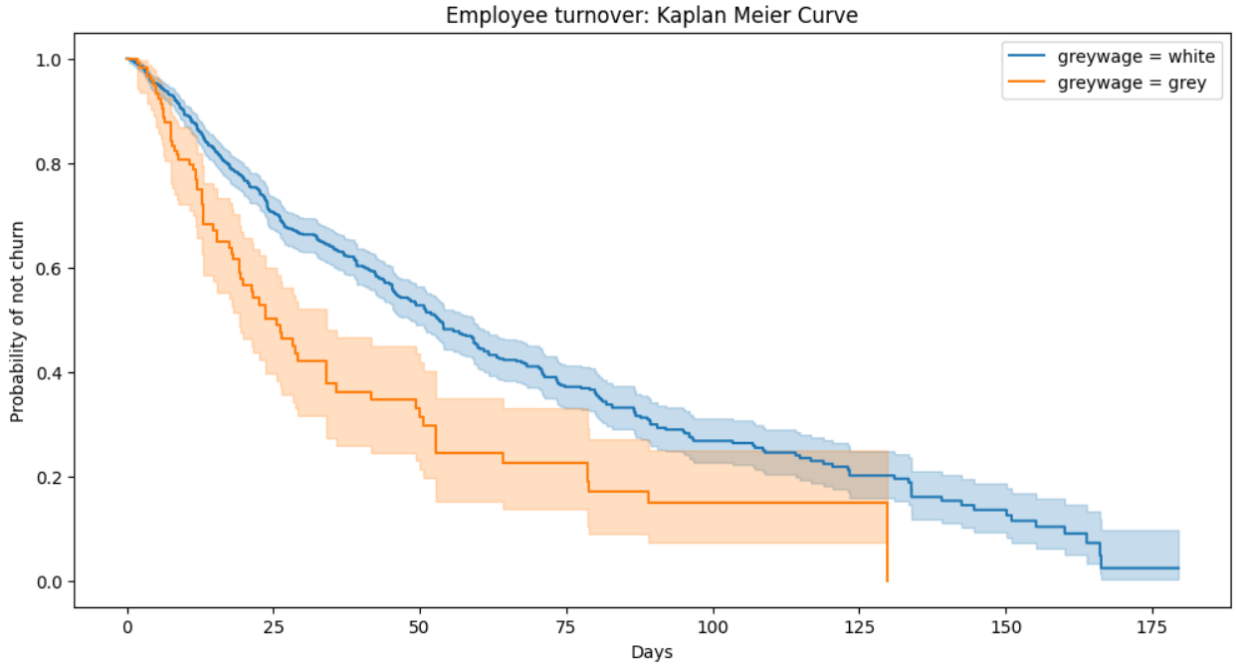
This Kaplan-Meier curve shows that female employees (orange line) churn faster than male employees (blue line). Female retention declines more rapidly, especially within the first 50 days. Male employees consistently have higher retention throughout the time period. Confidence intervals widen toward the end, indicating more uncertainty in the estimates.



This Kaplan-Meier curve shows that employees with no coach (orange line) churn faster than those with a coach (blue) or coached by "my head" (green). The "my head" group has the highest retention, while the no coach group has the lowest. Confidence intervals widen over time, indicating increased uncertainty.



This Kaplan-Meier curve compares employee retention based on the gender of their supervisor (head). Employees with female supervisors (orange line) show slightly better retention than those with male supervisors (blue line), particularly in the earlier days. However, the difference is small, and by 175 days, both groups experience similar churn rates. The shaded areas indicate confidence intervals, which widen over time, reflecting greater uncertainty in the estimates.



This Kaplan-Meier curve compares employee retention based on wage category: greywage = white (blue) and greywage = grey (orange). Employees in the grey wage category churn significantly faster than those in the white wage category, with a sharp decline in retention within the first 50 days. In contrast, employees in the white wage category show better retention, with a more gradual decline over time. By 175 days, the majority of employees from both groups have churned, but grey wage employees churn at a much higher rate throughout the period.

Conclusion: The data exploration reveals several key insights regarding employee churn. Age, wage category (greywage), industry, and certain personality traits, such as extraversion, self-control, and anxiety, appear to significantly influence retention rates. Younger employees, those in the "grey" wage category, and those working in retail or with mid-level personality scores are more likely to leave the company earlier. Kaplan-Meier curves further highlight differences in churn based on gender, coaching status, and supervisor gender, with female employees, those without coaches, and those in the "grey" wage category showing higher churn rates. These findings provide a foundation for building a Cox proportional hazards model to further investigate and quantify the effects of these factors on employee turnover.

Summary of the training process

First model with “greywage” variable

From the above exploration, we can see that **greywage could be a strong predictor of churn** The first variation of the Cox model includes only the greywage variable as a predictor of employee churn.

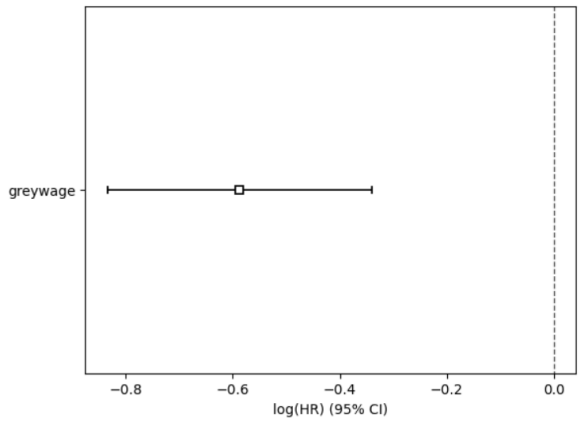
- **Coefficient (coef)**: -0.59. This negative coefficient indicates that being in the "greywage = white" category is associated with a reduced risk of churn compared to "greywage = grey."
- **Hazard Ratio (exp(coef))**: 0.56. The hazard ratio of 0.56 means that employees in the white wage category have a 44% lower risk of churn compared to those in the grey wage category.
- **Confidence Interval (95%)**: The confidence interval for the hazard ratio is between 0.43 and 0.71, confirming the significance of the relationship. The entire interval is below 1, which supports that being in the white wage category significantly reduces churn risk.
- **p-value**: < 0.005. The p-value is highly significant, meaning the impact of greywage on churn is statistically significant and unlikely to be due to chance.
- **Concordance**: 0.53. The concordance score is 0.53, indicating moderate predictive power for this model.
- **Visual Summary (Forest Plot)**: The forest plot shows the log hazard ratio (HR) for greywage, with the 95% confidence interval not crossing zero. This reinforces that greywage is a significant factor in predicting employee churn, with "white" wage category employees having lower churn risk.

This first variation of the Cox model shows that greywage is a strong and statistically significant predictor of employee churn. Employees in the "white" wage category have a considerably lower risk of leaving the company compared to those in the "grey" wage category. The model, while simple, provides meaningful insights into the impact of wage classification on churn risk.

model	lifelines.CoxPHFitter
duration col	'stag'
event col	'event'
baseline estimation	breslow
number of observations	1129
number of events observed	571
partial log-likelihood	-3460.89
time fit was run	2024-10-21 06:28:39 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
greywage	-0.59	0.56	0.13	-0.83	-0.34	0.43	0.71	0.00	-4.66	<0.005	18.27

Concordance	0.53
Partial AIC	6923.78
log-likelihood ratio test	18.82 on 1 df
-log2(p) of ll-ratio test	16.09



Second model with all variables

In this second variation, the model includes all available variables, such as age, greywage, personality traits, and coaching status. Here's a breakdown of the key results:

Age:

- Coefficient (coef): 0.02
- Hazard Ratio (exp(coef)): 1.02 – For each additional year of age, the risk of churn increases by 2%.
- p-value: < 0.005 – Statistically significant, showing that age is a significant predictor of churn.

Greywage_white:

- Coefficient (coef): -0.59
- Hazard Ratio (exp(coef)): 0.55 – Employees in the white wage category have a 45% lower risk of churn compared to the grey wage category.
- p-value: < 0.005 – Highly significant, confirming the strong impact of greywage on churn.

Coach_yes:

- Coefficient (coef): 0.25
- Hazard Ratio (exp(coef)): 1.28 – Employees with a coach are 28% more likely to churn, though this result is borderline significant with a p-value of 0.09.

Other Variables: Extraversion, independence, self-control, anxiety, gender_m (male), and head_gender_m (male) have p-values above 0.05, meaning these factors are not statistically significant in predicting churn in this model. These variables show no strong effect on employee turnover based on the confidence intervals (crossing zero in the forest plot).

Concordance: 0.60 – The model's predictive power is better than the first variation, with a moderate ability to correctly classify churn risk.

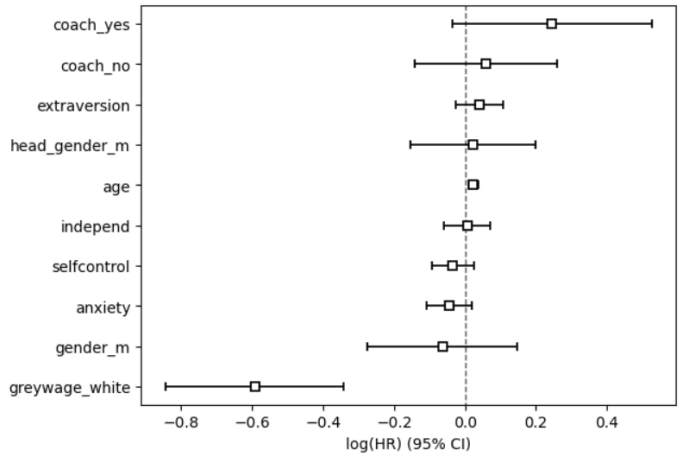
Visual Summary (Forest Plot): The forest plot shows the log hazard ratios for all variables, with greywage_white and age being significant predictors. The confidence intervals for other variables, such as extraversion, coach status, and gender, cross zero, indicating their effects are less certain.

In this model, age and greywage_white remain significant predictors of employee churn. Employees in the white wage category and older employees are less likely to churn. Coaching appears to increase churn risk, although it is borderline significant. Other variables, including personality traits and gender, do not significantly impact churn. The model shows improved fit and predictive power compared to the first variation, though not all variables are contributing meaningfully.

model	lifelines.CoxPHFitter
duration col	'stag'
event col	'event'
baseline estimation	breslow
number of observations	1129
number of events observed	571
partial log-likelihood	-3445.98
time fit was run	2024-10-21 06:58:12 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
age	0.02	1.02	0.01	0.01	0.04	1.01	1.04	0.00	3.47	<0.005	10.92
extraversion	0.04	1.04	0.03	-0.02	0.11	0.98	1.11	0.00	1.21	0.23	2.15
independ	0.01	1.01	0.03	-0.06	0.07	0.94	1.07	0.00	0.17	0.87	0.21
selfcontrol	-0.03	0.97	0.03	-0.09	0.03	0.91	1.03	0.00	-1.13	0.26	1.96
anxiety	-0.04	0.96	0.03	-0.11	0.02	0.90	1.02	0.00	-1.36	0.17	2.52
gender_m	-0.06	0.94	0.11	-0.27	0.15	0.76	1.16	0.00	-0.60	0.55	0.87
coach_no	0.06	1.06	0.10	-0.14	0.26	0.87	1.30	0.00	0.57	0.57	0.82
coach_yes	0.25	1.28	0.14	-0.04	0.53	0.97	1.69	0.00	1.71	0.09	3.53
head_gender_m	0.02	1.02	0.09	-0.15	0.20	0.86	1.22	0.00	0.26	0.80	0.33
greywage_white	-0.59	0.55	0.13	-0.84	-0.34	0.43	0.71	0.00	-4.62	<0.005	18.02

Concordance	0.60
Partial AIC	6911.96
log-likelihood ratio test	48.64 on 10 df
-log2(p) of ll-ratio test	21.01



Third model with “greywage” and “age” variables

In this third variation, the Cox model includes both “greywage” and “age” as predictors of employee churn. Here's a breakdown of the results:

Age:

- Coefficient (coef): 0.02

- Hazard Ratio (exp(coef)): 1.02 – For each additional year of age, the risk of churn increases by 2%.
- p-value: < 0.005 – Age is statistically significant, meaning older employees are less likely to churn, but the effect is relatively small.

Greywage_white:

- Coefficient (coef): -0.57
- Hazard Ratio (exp(coef)): 0.57 – Employees in the white wage category have a 43% lower risk of churn compared to those in the grey wage category.
- p-value: < 0.005 – Greywage is highly significant, confirming the strong influence of wage category on churn risk.

Concordance: 0.57 – The model's predictive power is better than the first variation (0.53), though lower than the full model (0.60), suggesting that these two variables alone provide a reasonable, but not the strongest, predictive capability.

This model confirms that both age and greywage are significant predictors of employee churn. While the effect of age is small but consistent, employees in the white wage category have a much lower risk of leaving the company. This variation balances simplicity with predictive accuracy, improving on the single-variable model while maintaining interpretability. However, it doesn't perform as well as the full model, which includes additional factors.

model	lifelines.CoxPHFitter
duration col	'stag'
event col	'event'
baseline estimation	breslow
number of observations	1129
number of events observed	571
partial log-likelihood	-3455.24
time fit was run	2024-10-21 06:56:33 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
age	0.02	1.02	0.01	0.01	0.03	1.01	1.03	0.00	3.41	<0.005	10.58
greywage_white	-0.57	0.57	0.13	-0.81	-0.32	0.44	0.73	0.00	-4.48	<0.005	17.01

Concordance	0.57
Partial AIC	6914.49
log-likelihood ratio test	30.11 on 2 df
-log2(p) of ll-ratio test	21.72

Model selection

Model	Strengths	Weaknesses	Concordance	AIC	Use Case	Recommendation
First Model (Greywage Only)	Simple and highly interpretable. Significant predictor of churn.	Limited predictive power with only one variable.	0.53	6923.78	Best for quick analysis focused solely on wage category but lacks deeper insights.	Not recommended: Lacks predictive depth.
Second Model (All Variables)	Highest concordance and lowest AIC, fitting the data well and explaining more variance.	Includes many non-significant variables, leading to unnecessary complexity and potential overfitting.	0.6	6911.96	Best for maximizing accuracy, but not ideal when interpretability and simplicity are needed.	Not recommended: Too complex.
Third Model (Greywage + Age)	Balanced approach with significant predictors (greywage and age). Improves accuracy over Model 1.	Slightly lower accuracy compared to the full model but avoids overfitting and complexity.	0.57	6914.49	Ideal for actionable insights and interpretability with sufficient predictive power.	Recommended: Best balance of accuracy and simplicity.

The third model is recommended due to its balance between accuracy, simplicity, and interpretability, making it the best fit for achieving the project's goals.

Summary key findings and insights

The analysis identified several key factors influencing employee turnover, with greywage (wage category) and age emerging as the most significant predictors of churn. Employees in the grey wage category are at a higher risk of leaving the company compared to those in the white wage category, with a hazard ratio of approximately 0.57, indicating a 43% lower churn risk for white wage employees. Age also plays a role, with each additional year reducing the risk of churn by 2%, suggesting that younger employees are more prone to leave.

While some other factors, such as coach status, showed a potential influence on churn, the effect was not statistically significant. Variables like extraversion, self-control, and gender were not significant contributors to churn in the models. The Kaplan-Meier survival curves highlighted stark differences in retention based on wage category, gender, and coaching status. Notably, employees without coaches and those in the grey wage category showed higher churn rates.

Suggestions for next steps

Some suggested next steps could be:

- Include additional data such as work performance or job satisfaction to improve model accuracy and insights.
- Add time-dependent factors to capture changes in employee characteristics over time for better churn predictions.
- Test alternative survival models, such as time-varying or deep learning-based models, for improved accuracy.
- Focus on retention efforts for younger employees and those in the grey wage category, and reassess their impact using updated models.

These actions will enhance the model's predictive power and guide more effective retention strategies.