Department of Mathematics,
College of Engineering, Design and Physical Sciences
Brunel University

Employee Turnover Modelling
with Survival Analysis

Mays Azeez
Course: Mathematics BSc
Supervisor: Jia Wei Lim
2020/2021

Abstract

Employee turnover is the number of employees who leave a business in a set amount of time. Employee turnover can be a significant obstacle for some companies because employee loss can affect the organisation financially. The loss can also be in terms of productivity and innovation if they are losing their talented hirers. This issue highlights the importance of using analytics to predict the turnover behaviour of employees and be able to prepare for it by identifying the factors that affect it most.

Survival Analysis is one of the methods used to predict turnover. This document describes some basic concepts of survival analysis, such as censoring, survival function, hazard rate and the models used to estimate the survival and risk rates. Cox Model and Kaplan-Meier estimate will be mainly used to predict employees' turnover rate and identify the factors that significantly affect it.

The data is real-world data representing over a thousand employees in various industries and professions in Russia. The data is obtained from Edward Babushkin Blog who is an HR Analyst from Moscow, Russia.

The data includes the length of time an employee has been in a company along with the type of industry in the profession they were in, as well as a description of personal, constructional, and organisational factors like gender, age, wage, innovation, level of anxiety and type of support provided for new employees.

The analysis will look at how the turnover rate various in different industries and identify the factors that affect it most. The analysis will be implanted using the Kaplan-Meier method to determine the overall employee turnover as the Kaplan-Meier model offers the most realistic representation of the data.

The analysis will also use Cox Regression Model to identify the factors with a significant effect on the turnover rate. Then create the predictive Cox model accordingly to calculate the survival and hazard of quitting. The predictions made using the obtained Cox model can allow employers to plan their hiring, implement changes that might retain employees and reduce the financial impact of turnover.

Contents

List of Tables

List of Figures

1 Introduction to Survival Analysis

Survival analysis: is a collection of statistical procedures for data analysis for which the variable of interest is time (or survival time T where $T \geq 0$) until an event occurs. [2]

For example:

- Time till a person receives an organ from an organ waiting list.
- Time till a call centre picks up a phone.
- Time until a machine part fails.

The methods used in survival analysis depend on the survival distribution, which is defined mainly by the survival function $\boldsymbol{S(t)}$ and the hazard function $\boldsymbol{h(t)}$**.** They will be explained in-depth in a later section.

Key Characteristics of survival analysis data are:

- The responses variable is a non-negative random variable representing time from a well-defined origin to a well-defined event. The time could be in years, days, or hours
- Censoring arises when the starting or ending events are not precisely observed [1]

The goals of survival analysis are:

- Evaluate the survival distribution
- Compare two or more survival distribution
- Assess the effects of factors on the occurrence of an event and the survival time

2 Basics of Survival Analysis

2.1 Censoring

Censoring: The event was not observed at its occurrence. The most common is the **right censoring,** where the event does not occur during the observed time (right side of the observed time). For example, consider a clinical trial with four patients observed over five years, as illustrated in **Figure 1**.

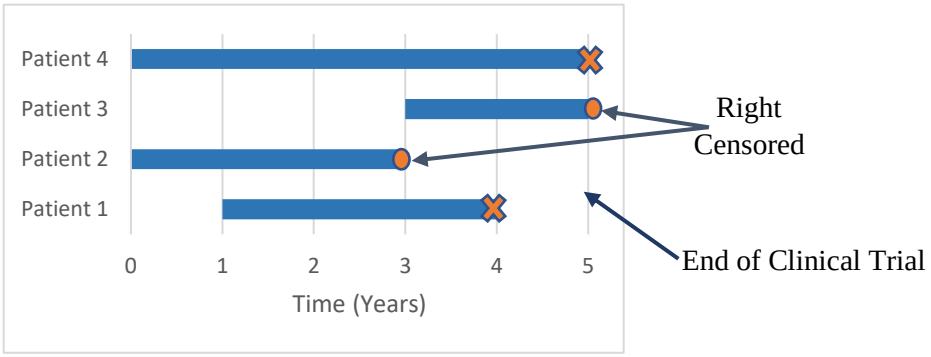

Figure 1: Clinical Trial Follow-up Time Illustration

Patient 1 enters the trial at the one-year mark, and the event occurs at four years. Patient 4 was there from the start of the trial, and the event occurs at the end of the study. Patient 2 enters at the beginning of the trial and leaves it at the three-year mark with the event not occurring. Similarly, for patient 3, they join at the three-year mark until the end of the study with an event not occurring. These two patients (2 and 3) are right-censored as we do not know if the event has happened after they left the study.

Table 1 shows survival time for each patient and their censoring status as 1 if the event has occurred or 0 if censored. The censoring indicator is denoted by $\boldsymbol{\delta}$, which is defined as a random variable that only takes the values 0 and 1.

$$\delta = \begin{cases} 0 & censored \\ 1 & event\ occurred \end{cases}$$

Patient no.	Survival Time	Indicator δ
1	3	1
2	3	0
3	2	0
4	5	1

Table 1: Example Survival Data

This example also shows the most common reasons for censoring, which are:

- The study ends, and the event has not occurred (similar to patient 3)
- The subject left the study before it ends or there was a failure to follow up (similar to patient 2), so it is unknown if the event occurred or not

The other type is **left censoring** is where events are known to have occurred before a specific time (left side of the observed time). For example, if we follow patients with HIV, we may start follow-up when a subject first tests positive for HIV, but we may not precisely know the time of first exposure to the virus. Thus, the survival time is censored on the left side because there is an unknown follow-up time from first exposure until the first positive HIV test. [2]

Also, censoring can be **non-informative**. The event being censored does not relate to the probability of an event occurrence or **Informative censoring,** where it shows the probability or the likelihood of an event occurring. We will always assume non-informative when analysing data unless stated otherwise.

2.2 The Survival and Hazard Functions

The **survival function *S(t)*** represents the probability of survival until at least time ***t***.

$$S(t) = P(T > t) = 1 - F(t) \quad (1)$$

where $0 < t < \infty, \quad T \geq 0, \ 0 \leq S(t) \leq 1$

T is the survival time, and *t* is a value taking by *T* at some point of the analysis. *F(t)* is the cumulative distribution function (CDF) of the survival time *T*, also known as the cumulative risk function, and it is right continuous.

$$F(t) = P(T \leq t), \ 0 < t < \infty \quad (2)$$

The **hazard function *h(t)*** represents the instantaneous risk of death at time *t*, conditional on surviving up to *t*. For example, a patient survived until the end of a drug trial, what is the probability per unit time they die straight after.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}, \qquad 0 \leq h(t) \leq \infty \quad (3)$$

These two functions are related, and by knowing one, the other can be calculated, as can be seen in the following calculus formulas. [1]

$$S(t) = exp\left[-\int_0^t h(u)\, du\right] \quad (4)$$

$$h(t) = -\left[\frac{d\, S(t)/dt}{S(t)}\right] \quad (5)$$

Another way these two functions are related is by the probability density function *f(t)*, which is defined as the rate of change in CDF or minus the rate of change o the survival function. [1]

$$f(t) = -\frac{d}{dt} S(t) \ = \ \frac{d}{dt} F(t) \quad (6)$$

$$h(t) = \frac{f(t)}{S(t)} \quad (7)$$

Figure 2 illustrates the connection between the two functions. Where two cases are considered, in the first case, the hazard is initially very high. Such a hazard might be appropriate to describe the lifetimes of animals with high mortality early in life. **Figure 2** illustrates such a hazard function (a) and the corresponding survival function (b). In the second case, the hazard is initially low and increases later in life. Such a hazard would describe organisms with a low initial hazard of death. This is illustrated in **Figure 2** (c) and the corresponding survival in (d). [1]

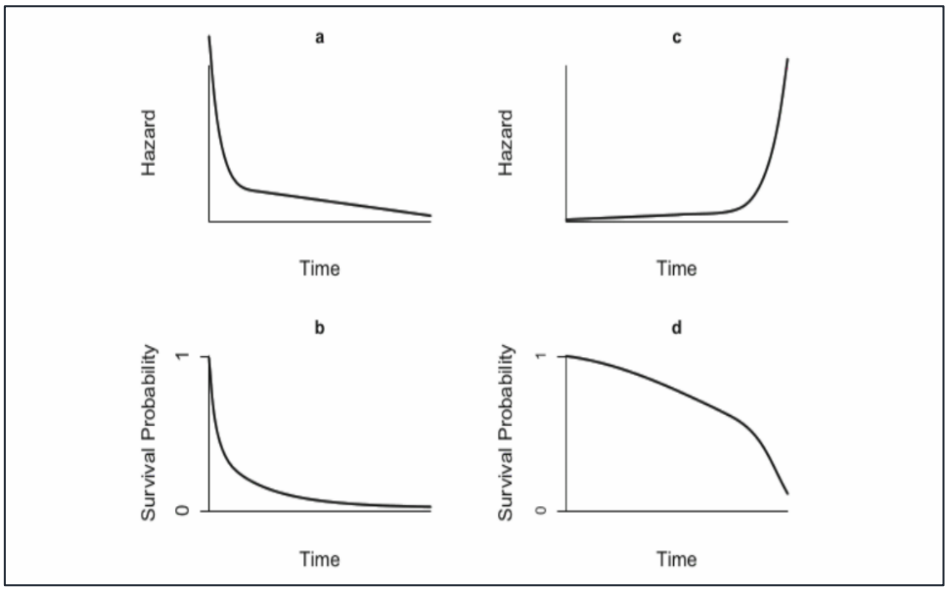

Figure 2: Hazard and Survival Functions with High Initial Hazard (a and b) and Low Initial Hazard (c and d) [1]

The survival function at the beginning of the distribution (where *t=0*) always equals one as the probability of survival at the beginning of the survival analysis model is at its highest. So

$$S(0) = 1$$

2.3 Cumulative Hazard Function

The cumulative hazard function is the integral of the hazard function. The cumulative hazard is commonly used as it is easier to estimate in models of survival analysis. The formula is given in equation (8). [1]

$$H(t) = \int_0^t h(u)\, du \qquad (8)$$

Combining that with equation (4), we get the equation (9)

$$S(t) = exp\big(-\int_0^t h(u)\, du\big) = exp(-H(t)) \qquad (9)$$

Correspondingly,

$$H(t) = -\ln\big(S(t)\big) \qquad (10)$$

2.4 Hazard Ratio:

The hazard ratio (HR) is used to compare the risk (hazard) in two distributions of survival analysis or the hazard corresponding to two different covariates. For example, compare the risk (hazard function) of developing side effects in two groups of patients in a drug trial, where one group is taking the drug tested and the other taking a placebo.

$$HR = \frac{h(t)_{x=1}}{h(t)x=0} \quad (11)$$

The results of HR are interpreted as follows

- If $HR > 1$, then group 1 has an increased hazard than group 0
- If $HR > 1$, then group 1 has a reduced hazard than group 0
- If $HR > 1$, then there is no different effect on hazard between the two groups

3 Models of Survival Analysis

3.1 Non-Parametric

3.1.1 Kaplan-Meier estimator

The most common non-parametric method for modelling the survival function is the **Kaplan-Meier estimator,** also known as the product-limit estimator, which is defined as the product over the failure times of the conditional probability of surviving to the next failure time [1].

It is given by

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (12)$$

Where n_i is the total number of subjects at risk at a time t_i, and d_i is the number of subjects that the event occurred for at a time t_i (failed subjects).[1] Graphically, the survival function when using the Kaplan-Meier estimator is a non-increasing, right continuous step function, making this model very flexible due to the shape of the distribution. Thus, this model is mainly used when modelling human or animal survival as it considers the quirks of the survival of living things.

Consider the clinical trial example with four patients observed over five years mentioned before in section 2.1. Using the Kaplan-Meier estimator, we get the figures shown in **Table 2.**

t_i	n_i	d_i	$\frac{d_i}{n_i}$	$1 - \frac{d_i}{n_i}$	$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$
3	3	1	0.333	0.667	0.667
5	1	1	1	0	0

Table 2: Kaplan-Meier Estimator Calculations for Survival Function S(t)

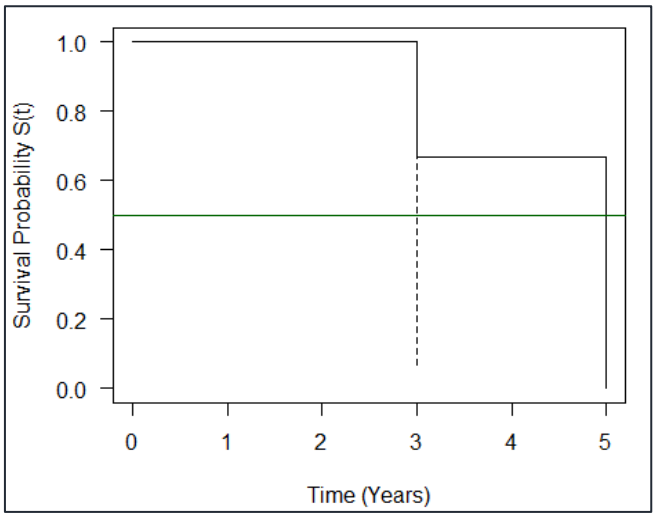

Figure 3: Clinical Trial Example Survival Curve

The green line in **Figure 3** shows the median survival time, which equals 5. Median survival time is the smallest t such that $S(t) \leq 0.5$.

Another example to show the Kaplan-Meier model's graphical representation is generated using lung cancer patients' data. This data is obtained from the R package "Survival" and using the R program ***FYP_examples.R***, where the Kaplan-Meier method estimates the survival probability of patients with advanced lung cancer over approximately 3.5 years, as shown in **Figure 4.** The median survival time for lung cancer patients equals 310.

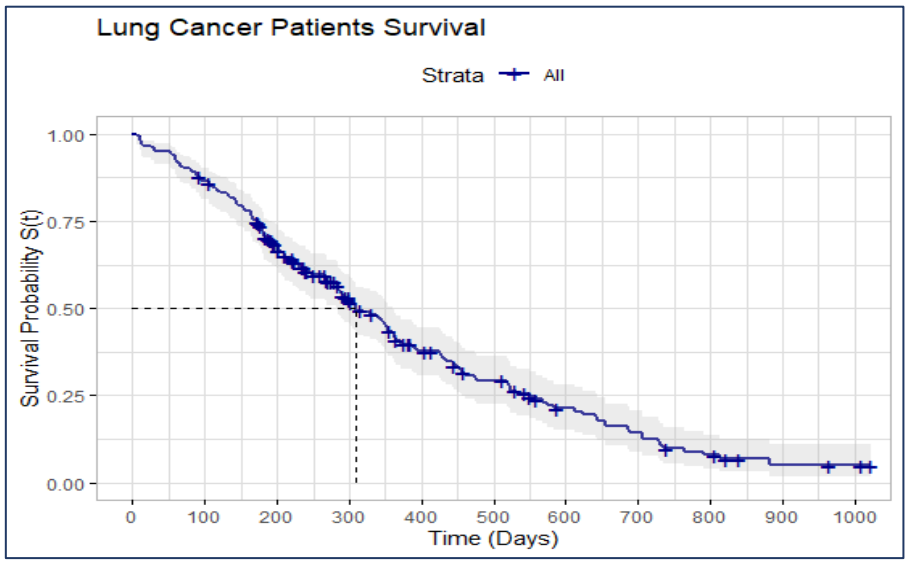

Figure 4: Survival in Lung Cancer Patients

The disadvantages of this distribution are:

- There is no simple mathematical function representing the curve as the survival function is not smooth as it varies at different points
- It is not easy to incorporate covariates as the distribution has a different hazard at various points. Hence, it is difficult to estimate a hazard rate to compare two groups
- It is a univariable analysis

3.1.2 Nelson-Aalen Estimator

An alternative estimator is the **Nelson-Aalen** (also known as the Fleming-Harrington). The estimator relies on the relationship between the hazard and survival functions. The formula (13) is estimating the cumulative hazard function, where it equals the sum of hazards up to the time t_i. [1]

$$\widehat{H}(t) = \sum_{t_i \le t} \frac{d_i}{n_i} \quad (13)$$

So, the survival function estimate is.

$$\hat{S}(t) = exp\left(-\widehat{H}(t)\right) \quad (14)$$

3.1.3 Comparing Non-parametric Distributions

The **log-rank test** is the primary method used when comparing survival curves of two groups with non-parametric distributions. The null hypothesis being tested is that there is no statistical difference between the two survival curves and an alternative hypothesis that states otherwise.

$$H_0: S_0(t) = S_1(t)$$

The log-rank statistics is approximately distributed a chi-square test statistic with one degree of freedom (χ_1^2) thus, a *p*-value for the log-rank test is determined from the table of the chi-square distribution. The formula for the log-rank test is as follows. [2]

$$\sum_{i}^{\#\ number\ of\ groups} \frac{(O_i - E_i)^2}{E_i} \approx X_1^2 \quad (15)$$

Where

O_i is the total number of observed events for group i

E_i is the total number of expected events for group i,

The expected events for group i at different survival times *t*, E_{it} is calculated by

$$E_{it} = \frac{n_{it} O_t}{n_t} \quad (16)$$

Where

n_{it} is the number of subjects at risk for group i at time t

O_t is the number of events observed for all groups at time t

n_t is the total number of subjects at risk for all groups at time t

Consider an example of a clinical trial with six patients assigned to either a controlled or a treatment group. This trial's survival data are shown in **Table 3,** where "T" denotes treatment patient, and "C" denotes controlled patients. [1]

Patient	Survival Time	Censor	Group
1	6	1	C
2	7	0	C
3	10	1	T
4	15	1	C
5	19	0	T
6	25	1	T

Table 3: Survival data of The Clinical Trial Example [1]

The log-rank test is calculated to check if there a statical difference between the survival time for these two groups. **Table 4** shows the calculation needed to find the test statistic χ^2.

Time t	Number at risk in group "C" n_{1t}	Number at risk in group "T" n_{2t}	Total number at risk in n_t	Number of Events in "C" O_{1t}	Number of Events in "T" O_{2t}	Total Number of Events O_t	Expected Number of Events in "C" E_{1t}	Expected Number of Events in "T" E_{2t}
6	3	3	6	1	0	1	0.500	0.500
10	1	3	4	0	1	1	0.250	0.750
15	1	2	3	1	0	1	0.333	0.667
25	0	1	1	0	1	1	0.000	1.000
				2	2		1.083	2.917

Table 4: Calculations to Find the Log-Rank Test Value χ^2

Substituting in equation (15). The calculations are as follows.

$$\chi^2 = \frac{(2-1.083)^2}{1.083} + \frac{(2-2.917)^2}{2.917} = 1.2$$

Compare the value obtained to a chi-square with one degree of freedom and a significant level of 5%, which equals 3.841. We accept the null hypothesis since $\chi^2 < 3.841$ and the p-value = 0.259, so there is no statistical difference between the two groups in terms of the survival time.

If the survival distribution of more than two groups is compared, the log test can still be used with a slight adjustment to the null hypothesis and the test statistic. Suppose there are $p+1$ groups denoted $0,1,\cdots,p$, The null hypothesises would be $H_0: S_0(t) = S_1(t) \ldots S_p(t)$. The exact formula for the test statistic is used to calculate the log-rank test, but the results are compared to a chi-square distribution with p degrees of freedom.

3.2 Parametric Method

This method assumes that survival data approximately a specific parametric form. For example, the data could have an **Exponential distribution** (constant hazard) or **Weibull distribution** (time-varying hazard), which are the most commonly used parametric distributions.

The advantages of this method are:

- The survival function is smooth if a continuous distribution is chosen
- It is easy to incorporate covariates into the model

The distribution chosen must be a good description of the data. So, it either uses formal hypothesis testing or visualisation procedures to test the suitability of the distribution.

3.2.1 Maximum likelihood estimation

Maximum likelihood estimation is used to approximate the unknown parameters of the parametric distributions. The data consists of a series of observations $t_1, t_2, \cdots, t_n$. Each one is linked to a censoring indicator δ_i. For a certain t_i, $\delta_i = 1$ if the event was observed and $\delta_i = 0$ if the event was censored. The likelihood function under general non-informative censoring has the formula (17). [1]

$$L(\lambda) = \prod_{i=1}^{n} h(t_j, \lambda)^{\delta_i} S(t_i, \lambda) \qquad (17)$$

Where, $S(t_i, \lambda)$ is the survival function for observation $t_j$in terms of the distribution parameter λ. Where $h(t_j, \lambda)^{\delta_i}$ is the hazard function for observation $t_j$in terms of the distribution parameter λ to the power of the censoring indicator. This means that when the censoring indicator $\delta_i = 1$, we enter both functions. When $\delta_i = 0$, we use only the survival function.

3.2.2 Exponential Distribution

The exponential distribution assumes a constant hazard, making it the easiest distribution to work with.

$$h(t) = \lambda, \text{ where } \lambda \textit{ is a constant}$$

Thus, the cumulative hazard function becomes.

$$H(t) = \int_0^t h(u)\, du = \int_0^t \lambda\, du = \lambda u|_0^t = \lambda t \qquad (18)$$

$$S(t) = e^{-H(t)} = e^{-\lambda t} \qquad (19)$$

3.2.3 Weibull Distribution

The Weibull distribution offers more flexibility when modelling survival data as the hazard function is proportional with time in this distribution. The hazard function is represented by the formula (20). [1]

$$h(t) = \alpha^{\lambda}(\lambda t)^{\alpha-1} = \alpha\lambda^{\alpha}t^{\alpha-1} \quad (20)$$

Where λ and α are unknown, but we use the survival data on hand to estimate these parameters using the maximum likelihood estimator. Thus, the hazard and survival functions are given by (21) and (22), respectively.

$$H(t) = (\lambda t)^{\alpha} \quad (21)$$

$$S(t) = e^{-(\lambda t)^{\alpha}} \quad (22)$$

Figure 5 shows the shape of the hazard for several parameter choices. The exponential distribution is a special case with $\alpha = 1$. It is monotone increasing for $\alpha > 1$and monotone decreasing for $\alpha < 1$. [1]

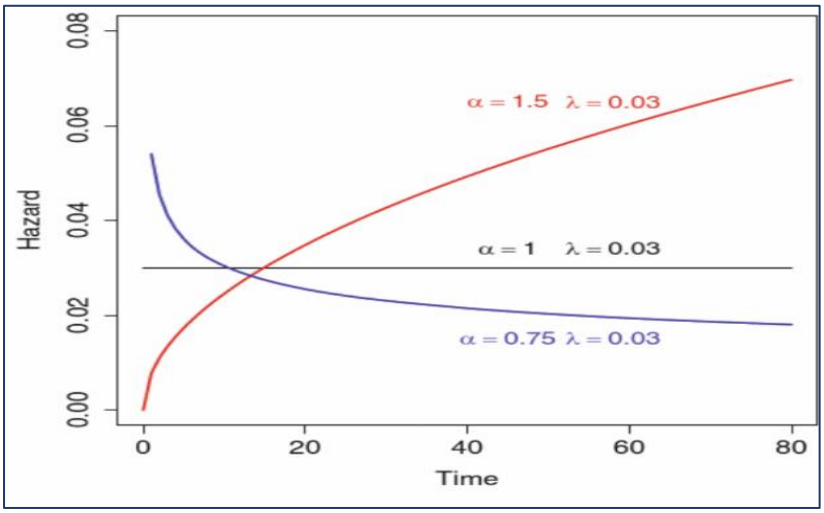

Figure 5: Weibull Hazard Functions [1]

3.3 Cox Proportional Hazard Model

This model consists of both parametric and non-parametric elements, so it is called a semi-parametric method. The **Cox model** is widely used in survival analysis. It models the relationship between a set of one or more covariates and the hazard rate. Covariates may be discrete or continuous. The model decomposes the hazard or instantaneous risk into a non-parametric baseline, shared across all subjects, and a relative risk, which describes how individual covariates affect risk. This model offers the ease of incorporating covariates. The hazard function is given by equation (23). [2]

$$h(t) = \underbrace{h_0(t)}_{\text{baseline}} \underbrace{exp(\sum_{i=1}^{n} \beta_i X_i)}_{\text{relative risk}} \quad (23)$$

Where, $h_0(t)$ is the baseline hazard function, X_i represents explanatory variables used to predict relative risk (It takes values of zero or one for categorical variables and numerical values if the variable is continuous), β_i is a coefficient for each X_i. This allows for time-varying baseline risk, like in the Kaplan Meier model, while allowing subjects to have different survival functions within the same model, making the Cox model a robust model that can be used to approximate a parametric model. [2]

There are a few assumptions made when using the Cox model, which are:

- Constant hazard ratio over time (Proportional hazard)
- The variables X_i are assumed to be time-independent unless stated otherwise
- $h_0(t)$ is taken to be non-parametric (Unspecified function)
- Censoring is assumed to be non-informative unless stated otherwise
- Survival times *t* are independent

Sometimes the model is expressed differently, relating the relative hazard to the baseline hazard, which is the hazard ratio at time *t*. As seen in the following equation. [6]

$$\frac{h(t)}{h_0(t)} = exp\left(\sum_{i=1}^{n} \beta_i X_i\right) \quad (24)$$

Notice that this equation is very similar to the equation of the hazard ratio (11) in section 2.4. So, $\boldsymbol{e^{(\sum_{i=1}^{n} \beta_i X_i)}}$ gives an estimate for the hazard ratio (HR). We can rewrite the hazard ratio for subject i as

$$\psi_i = e^{\beta_i X_i} \quad (25)$$

Notice that there are two unknown parameters in equations for the Cox model, the coefficients β_i and the baseline hazard $h_0(t)$. These parameters are estimated using the methods explained in the upcoming sections.

3.3.1 Partial Likelihood

The estimate of the Cox model parameter β_i are derived by maximising a likelihood function, denoted as $\boldsymbol{L}(\boldsymbol{\beta})$. It is called the "partial" likelihood function because the likelihood formula will only consider the probability of subjects who fail (event occurs) and does not explicitly consider censored subjects. The partial likelihood function can be written as the product of several likelihoods, one for each failure time. [2]

All the subjects in the dataset are considered at risk before the first failure time. R_i "the risk set" is defined as the set containing all the individuals at risk for failure. So, at i-th failure time, L_i denotes the likelihood of failing at a time t_i for the i-th subject, given survival up to this time.

$$L_i(\beta) = P \text{ (individual } i \text{ fails| one failure from } R_i)$$

$$= \frac{P \text{ (individual } i \text{ fails| at risk at } t_i)}{\sum_{k \in R_i} P \text{ (individual } k \text{ fails| at risk at } R_i)}$$

$$= \frac{h_i(t_i)}{\sum_{k \in R_i} h_k(t_i)}$$

from substituting equation (23), we get

$$L_i(\beta) = \frac{h_0(t_i)\psi_i}{\sum_{k \in R_i} h_0(t_i)\psi_k}$$

The baseline hazard term $h_0(t_i)$ cancels out in the numerator and the denominator to get the final form of the partial likelihood function. [1]

$$L(\beta) = \prod L_i(\beta) = \prod_{i=1}^{D} \frac{\psi_i}{\sum_{k \in R_i} \psi_k} \quad (26)$$

Where D is the total number of failures. The log partial likelihood function for equation (26) is given as

$$l(\beta) = \sum_{i=1}^{D} \left[log(\psi_i) - log(\sum_{k \in R_i} \psi_k) \right]$$

$$= \sum_{i=1}^{D} X_i \beta - \sum_{i=1}^{D} log(\sum_{k \in R_i} e^{\beta X_k}) \quad (27)$$

The first derivative of the log partial likelihood function is called the score function $U(\beta)$. It is given as

$$U(\beta) = \frac{d}{d\beta} l(\beta) \; = \; \sum_{1}^{D} \left(X_i \; - \; \frac{\sum_{k \in R_i} X_k \, e^{\beta X_i}}{\sum_{k \in R_i} e^{\beta X_i}} \right) \qquad (28)$$

The maximum partial likelihood estimate $\hat{\beta}$ can be found by solving $U(\beta) = \; 0$.

3.3.2 Partial Likelihood Hypothesis Tests

There are three statistical tests used to test the hypothesis $H_0 : \beta = 0$ in survival analysis for the partial likelihood. The tests are:

- Wald test
- Score test
- Likelihood ratio test

The test statistics are constructed using two functions. The score function and the information matrix $I(\beta; X)$. Which is minus the second derivative of the log partial likelihood. [1]

$$I(\beta; X) = -\frac{d^2 l(\beta)}{d\beta \, d\beta'} \; = \; -\frac{dU(\beta)}{d\beta} \qquad (29)$$

The observed information matrix can be found by substituting $\hat{\beta}$ into equation (29), which is also known as the Hessian matrix.

The test statistics for the Wald, Score, and likelihood ratio test statistics are as follows, respectively. [1]

$$\chi_w^2 = \hat{\beta}' I(\hat{\beta}; X) \hat{\beta} \qquad (30)$$

$$\chi_s^2 = U'(\beta = 0, X) \cdot I^{-1}(\beta = 0, X) \cdot U(\beta \, ; X) \qquad (31)$$

$$\chi_1^2 = 2\{l(\beta = \hat{\beta}) - l(\beta = 0)\} \qquad (32)$$

All three are under the null hypothesis $H_0 : \beta = 0$ and have a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of covariates being investigated.

3.3.3 Estimating The Baseline Hazard and Survival Functions

The estimation of the baseline hazard $h_0(t)$ is given by

$$\hat{h}_0(t_i) = \frac{d_i}{\sum_{j \in R_i} exp(x_j \hat{\beta})} \qquad (33)$$

Where d_i is the number of subjects that the event occurred for a time t_i (failed subjects). [1]

Notice that when $\beta = 0$, equation (33) is reduced to the Nelson-Aalen estimator mentioned in equation (13), section 3.1.2. The baseline survival function is

$$\hat{S}_0(t) = exp[-H_0(t)] \quad (34)$$

The estimate of the survival baseline function can obtain by estimating $H_0(t)$ as a cumulative sum of the estimated hazard functions $\widehat{h}_0(t_i)$ for $t_i \leq t$. [1]

3.3.4 Checking The Proportional Hazard Assumption

The proportional hazards assumption (constant HR) is key to constructing the partial likelihood, as this property allows the cancelling out of the baseline hazard function from the partial likelihood factors. **Schoenfeld residuals** provide a helpful way to assess this assumption. [1]
Schoenfeld residuals provide statistical tests and graphical diagnostics to assess the proportional hazard assumption.

Schoenfeld residuals are defined as the individual terms of the score function, and each term is the observed value of the covariate for patient i minus the expected value of the covariate ($E(x_i) = \bar{x}(t_i)$), which is a weighted sum. [1]

The weights are given by

$$P(\beta, x_k) = \frac{e^{\beta x_k}}{\sum_{j \in R_k} e^{\beta x_j}} \quad (35)$$

Each weight can be defined as the probability of selecting a particular person from the risk set at a time t_i. The residual for the i-th failure time and using the estimate $\hat{\beta}$ is

$$\hat{r}_i = x_i - \sum_{k \in R_i} x_k \cdot P\left(\hat{\beta}, x_k\right) = x_i - \bar{x}(t_i) \quad (36)$$

The plot of these residuals against the covariates x_i will produce a pattern of points that are centred at zero if the assumption of proportional hazard is correct. This is because the residuals are asymptotically uncorrelated and have an expectation of zero under the Cox model. Each covariate has its residual plot defined only for the failure time (event occurring) and not censored times.
Consider the clinical trial example with six patients assigned either to a controlled or a treatment group mentioned in section 3.1.3. The data is shown in **Table 3**. The partial likelihood coefficient estimate is found to be $\hat{\beta} = -1.326$.[1]

Patient	Survival Time	Censor	Group
1	6	1	C
2	7	0	C
3	10	1	T
4	15	1	C
5	19	0	T
6	25	1	T

Table 3: Survival data of The Clinical Trial Example

First, the weights are computed for the expected covariate values before calculating the Schoenfeld residuals, as shown in **Table 5**.

Time t_i	Number at risk in group "C" n_{0i}	Number at risk in group "T" n_{1i}	Weight for group "C" $P(\hat{\beta}, x_k = 0)$	Weight for group "T" $P(\hat{\beta}, x_k = 1)$
6	3	3	$\frac{1}{3+3e^{-1.326}}$	$\frac{e^{-1.326}}{3+3e^{-1.326}}$
10	1	3	$\frac{1}{1+3e^{-1.326}}$	$\frac{e^{-1.326}}{1+3e^{-1.326}}$
15	1	2	$\frac{1}{1+2e^{-1.326}}$	$\frac{e^{-1.326}}{1+2e^{-1.326}}$
25	0	1	$\frac{1}{2e^{-1.326}}$	$\frac{e^{-1.326}}{2e^{-1.326}} = 1$

Table 5: Weights Calculations for Schoenfeld Residuals [1]

Notice that the groups within the covariate take the values zero and one. The controlled group has the value 0, and the treatment group takes the value 1. The final calculations for the Schoenfeld residuals are shown below in **Table 6.**

Time t_i	Expected covariate Value $\bar{x}(t_i) = \sum_{k \in R_i} x_k \cdot P(\hat{\beta}, x_k)$	Observed covariate value x_i	Residuals $\hat{r}_i = x_i - \bar{x}(t_i)$
6	$3 \times 0 \times \frac{1}{3+3e^{-1.326}} + 3 \times 1 \times \frac{e^{-1.326}}{3+3e^{-1.326}} = 0.2098$	0	-0.2098
10	$1 \times 0 \times \frac{1}{3+3e^{-1.326}} + 3 \times 1 \times \frac{e^{-1.326}}{3+3e^{-1.326}} = 0.4434$	1	0.5566
15	$1 \times 0 \times \frac{1}{3+2e^{-1.326}} + 2 \times 1 \times \frac{e^{-1.326}}{3+2e^{-1.326}} = 0.3468$	0	-0.3468
25	1	1	0

Table 6: Schoenfeld Residuals Calculations [1]

Plotting the residuals against the covariate gives the Schoenfeld Residuals plot as shown in **Figure 6**. The y- axis represents the covariate trt (type of treatment group). This plot was generated in R using program ***FYP_examples.R***.

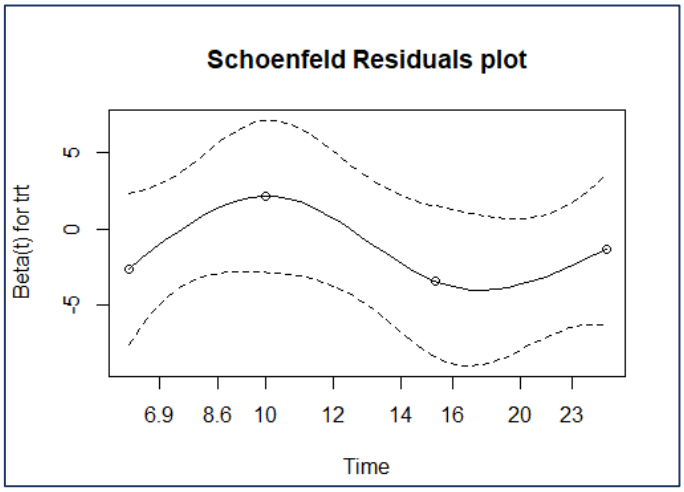

Figure 6: Schoenfeld Residuals Plot for Treatment Type

The points are centred at zero, so the proportional hazard assumption is met for this covariate.

4 Data Exploration

4.1 Background

Employee turnover is the rate at which staff leave an organisation. Employee turnover is a significant obstacle for some companies because employee loss affects the organisation financially. The loss can also be in terms of productivity and innovation if they are losing their talented hirers. This issue highlights the importance of using analytics to make predictions and identify the variables that most affect employee turnover. These predictions might aid companies with their hiring strategies and their development of the work environment. Some of the most relevant variables to consider and investigate are shown in **Figure 7.** [4]

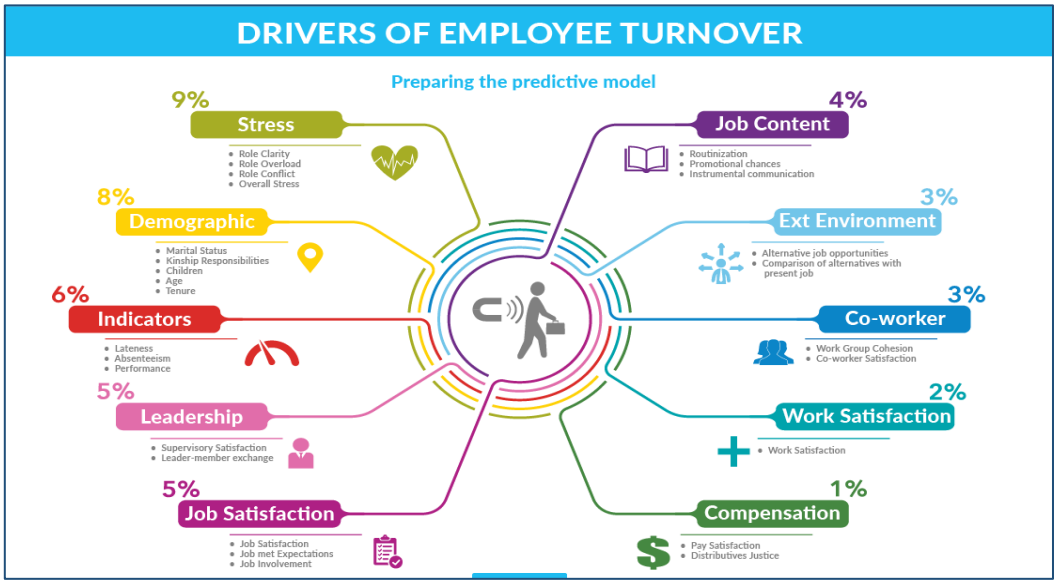

Figure 7: Popular Drivers of Employee Turnover [4]

LinkedIn survey data also showed that 25% of employees leave their organisation due to personal problems like family issues, interpersonal issues with co-workers or supervisors, or dependent care issues. [5]

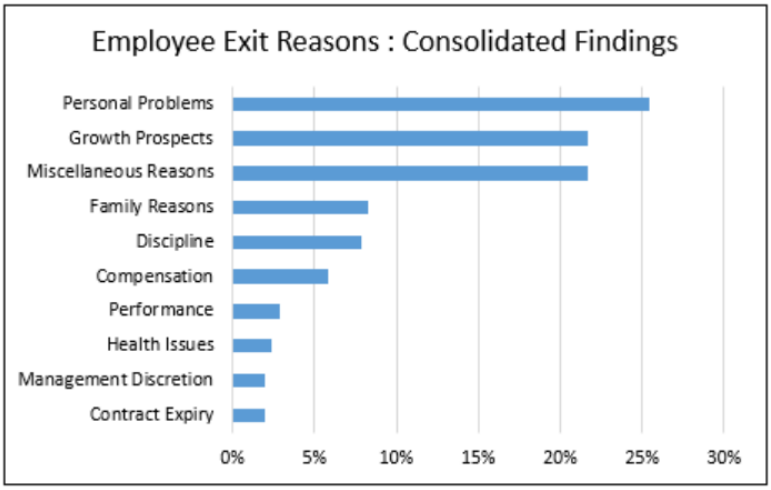

Figure 8: Causes of Employee Turnover [5]

4.2 The Data

The data contain information about 1129 real employees from Russia in various industries and professions and their turnover rate over fifteen years. It was obtained from Edward Babushkin Blog. Edward Babushkin is an HR Analyst from Moscow, Russia.

The data contain information on each employee, such as their age, gender, profession, and the industries they work in. it also contains information about their mental health, like their independent score and anxiety level. **Table 7** shows the variables in the dataset and their descriptions.

Variable name	Description	Variable name	Description
time	Time in a company (Months)	**head_gender**	Manager's gender
event	1= quit, 0 = Censored	**greywage**	Salary bracket (White, Grey)
Gender	Employee's gender	**way**	Method of commuting
Age	Employee's age	**extraversion**	Extraversion score (1-10)
industry	Employee's industry	**Independ**	Independent score (1-10)
profession	Employee's profession	**selfcontrol**	Self-control score (1-10)
traffic	The hiring agency	**anxiety**	Anxiety score (1-10)
coach	Presence of a coach at the start of the position	**novator**	Innovation score (1-10)

Table 7: Turnover Dataset Variables

4.3 Exploratory Analysis

The dataset is examined in R (using R program ***initial_analysis.R***) to check its attributes and ensure there are no missing values. The exploratory analysis revealed that the data contain information about employees with the age that ranges from 15- 58, with the mean age being 31 years. It can be seen in **Figure 9,** a histogram of the age distribution by gender. It can also be noticed that the number of female employees is higher than that of male employees.

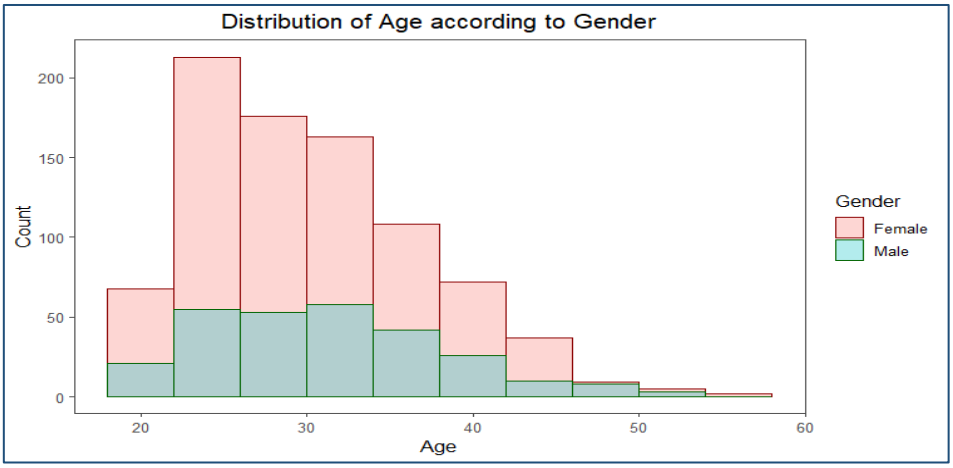

Figure 9: Distribution of Age by Gender

There are 16 various industries within the data. Retail, IT, Manufacturing, and banking are the most repeated ones, with retail having the highest attrition rate, as shown in **Figure 10.**

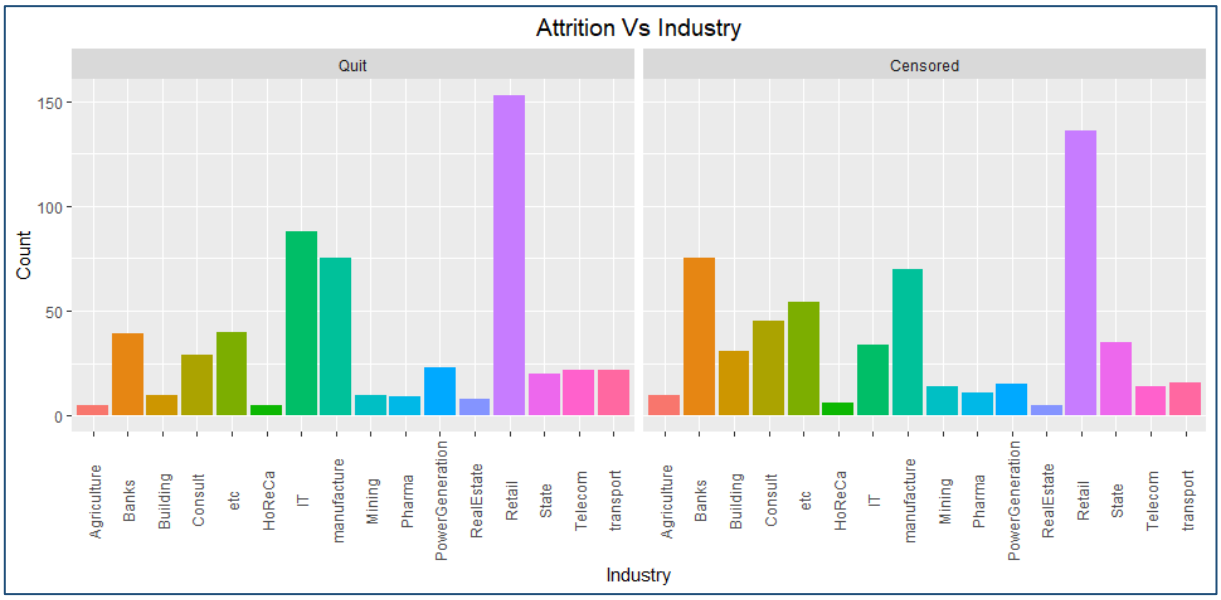

Figure 10: Attrition of Industry

The correlation between some variables and survival time is studied to identify the variables that could be important in predicting turnover. **Figure 11** shows a negative correlation between turnover time and age, innovation, and extraversion scores. In comparison, the turnover time has a positive correlation with self-control and anxiety scores. So, the more experienced, innovative employees would be more likely to leave a company.

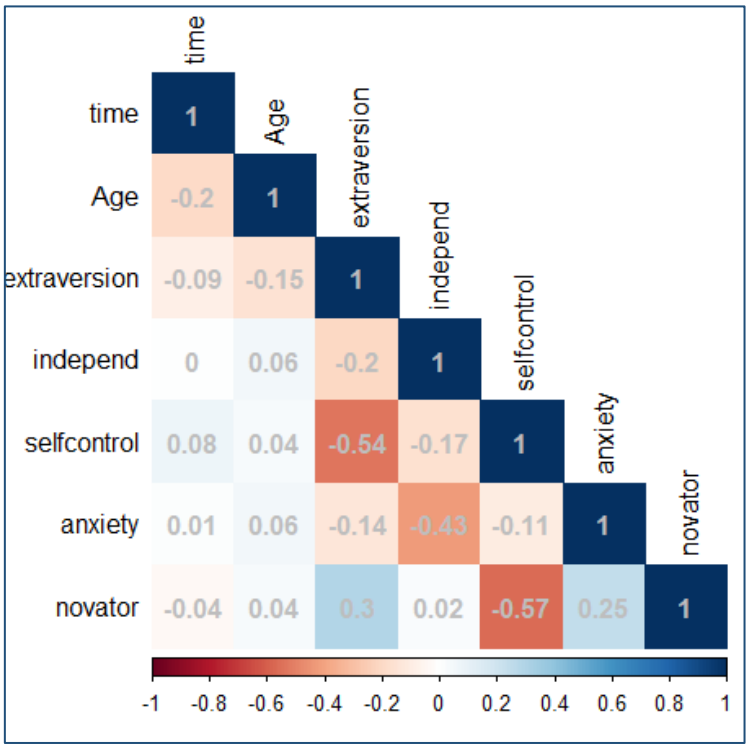

Figure 11: Correlation Matrix for Turnover Time and Other Variables

4.4 Objectives of the Analysis

This analysis will look at critical variables like age, gender, and self-control score (variable selfcontrol) and their effect on turnover rate in four industries: Retail, IT, Manufacturing, and Banking because they are most common in the dataset as highlighted in the exploratory analysis. So, there will be 670 employees only after filtering by these industries instead of the 1129 employees total mentioned before. Survival analysis methods will be implanted on the data to understand the turnover rate.

The main objectives of the analysis are:

- Evaluate the overall turnover rate
- Evaluate the effect of age and gender on the turnover rate
- Compare employees' turnover rate in the four industries
- Identify the factors that have the most effect on the turnover rate
- Evaluate and compare the turnover rate for each factor with a significant effect
- Make predictions based on the results obtained

The basic steps for the analysis are:

- Use the Kaplan-Meier method to estimate the survival rate and plot the survival curve
- Use the Log Rank method to test whether the difference between groups' survival is significant
- Use the Cox proportional hazard model to evaluate the factors' effects on the turnover rate and plot the survival curve for the model
- Use the Wald test and the Score test to determine if the model and the results obtained from it are significant
- Use the significant Cox model to make predictions

5 Survival analysis with Kaplan-Meier

The dataset is filtered to include employees from four industries: Retail, IT, Manufacturing, and Banking, as they are the most common in the dataset as highlighted in the exploratory analysis. The R program corresponding to the Kaplan-Meier model analysis is ***analysis_1.R.***

The assumption made when analysing the data with Kaplan- Meier are

- The censoring is taken as a right, non-informative censoring
- The censored employees were included in the analysis up to their censored time (assuming they survived until the censored time).

5.1 Overall Survival with Kaplan-Meier

A Kaplan-Meier estimator is used to calculate the overall survival for all the employees in the data. The results show that 315 employees quit out of the 670 by the end of the 15 years. So, there was a loss of 47% of employees. The mean survival was approximately four years.

Figure 12 shows the survival curve for this data. The vertical mark on the curves represents a censored employee at that time. We can see that the survival probability decreases the longer an employee works for a company.

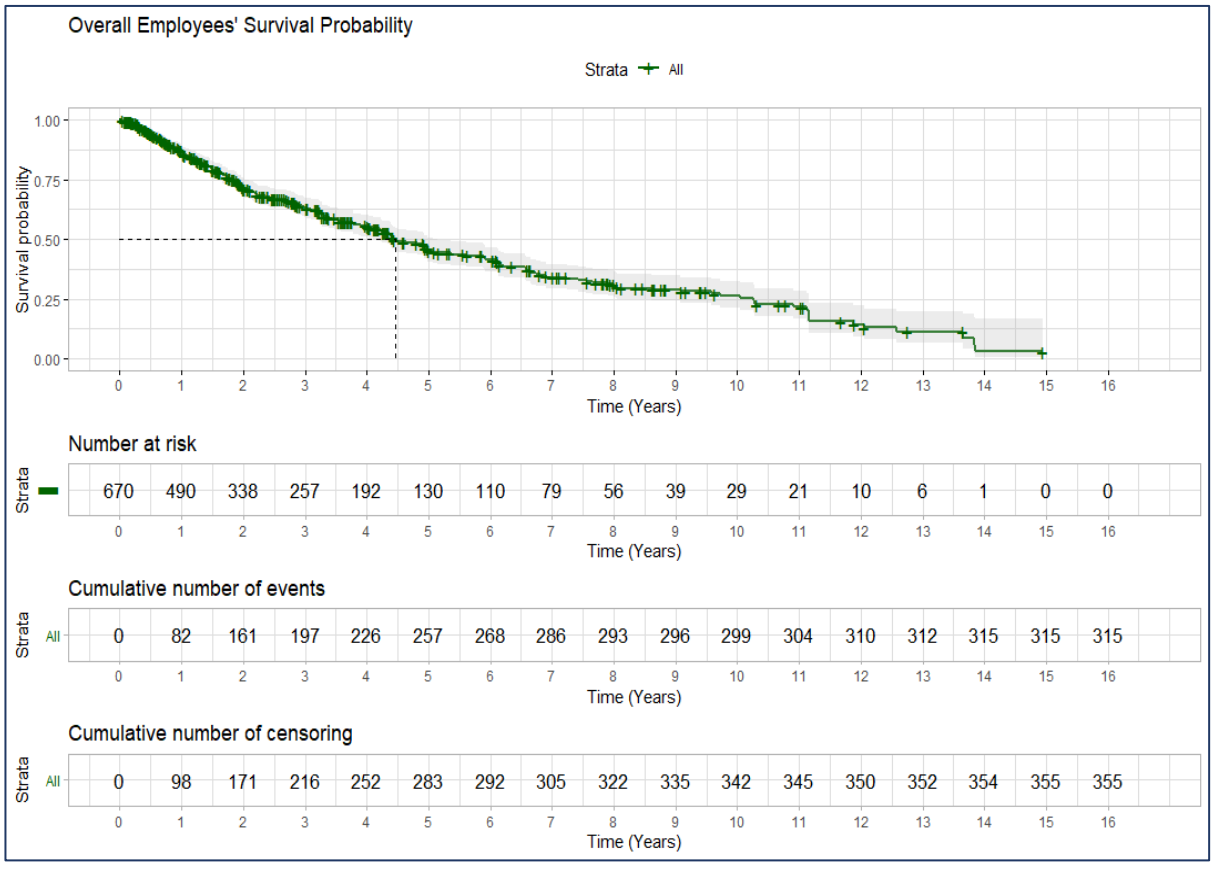

Figure 12: Kaplan-Meier Survival Curve for The Data

5.2 Gender Analysis

The survival for male and female employees in all industries is estimated using a Kaplan-Meier model. It found that 230 female employees quit over the observed period of 15 years with a median survival of 4 years. Only 85 male employees quit over the same period with a median survival of approximately five years.

The Log-rank test is used to determine whether the difference is statistically significant. The test is carried out at a 5% significant level. The calculations are shown in **Table 8**.

Gender	Total n	Total Observed	Total Expected	$\frac{(O-E)^2}{E}$
Female	489	230	213	1.29
Male	181	85	102	2.71

Table 8: Log-Rank Test by Gender

The chi-square test statistic with one degree of freedom is $\chi_1^2 = 4.1$, and the p-value is 0.04. Thus, the difference is statistically significant. In conclusion, the gender of an employee affects their turnover rate. The survival curve of employees by gender is shown in **Figure 13**.

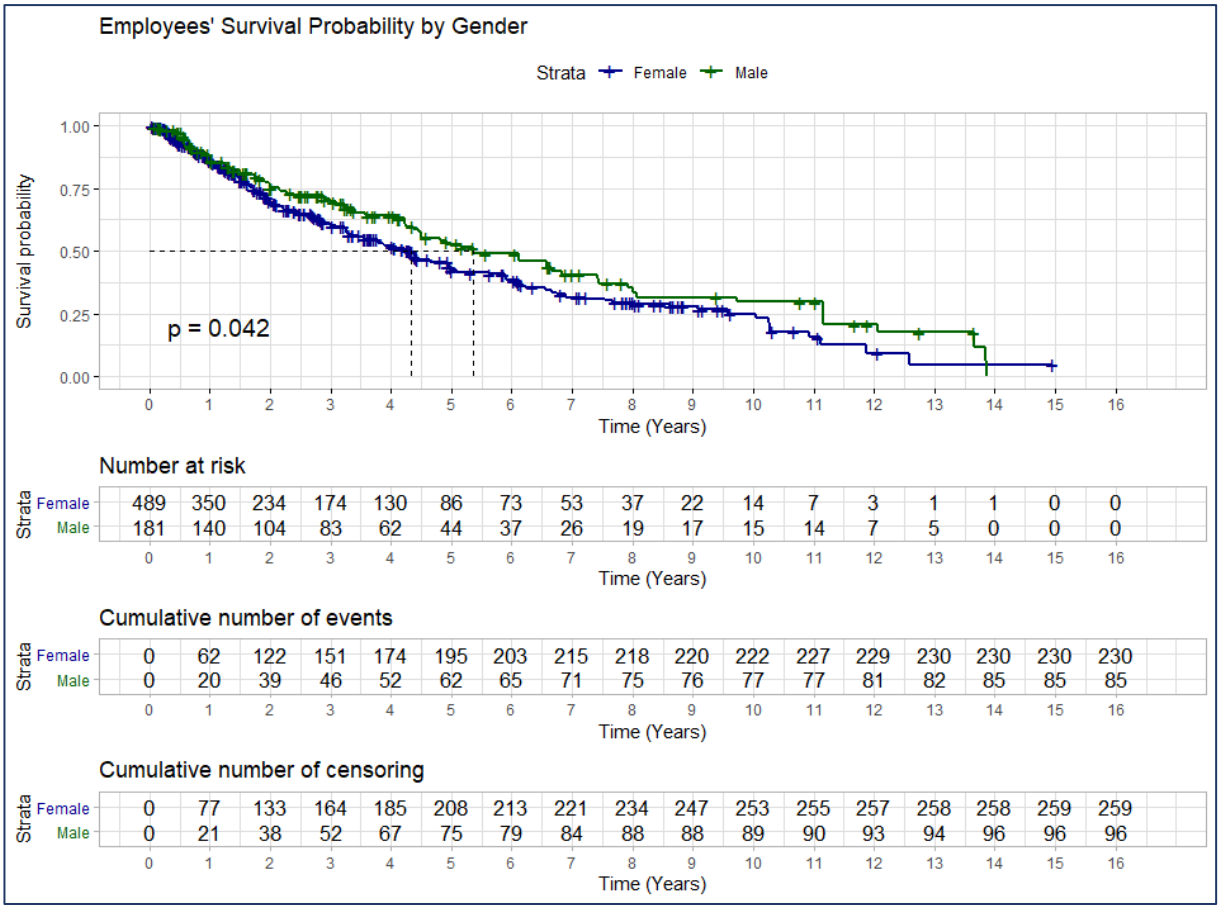

Figure 13: Kaplan-Meier survival Curve for Employees by Gender

5.3 Age Group Analysis

The employees have been divided into four age groups which are 18-30, 30-40, 40-50 and 50+. The survival is estimated using the Kaplan-Meier model. **Table 9** shows the results of this model from R.

Age Group	Number in each group	Number of events	Median survival (Years)
18-30	349	163	5.9
30-40	249	116	3.9
40-50	68	34	4.2
50+	4	2	2.4

Table 9: Kaplan-Meier Model Results by Age Group

The Log-rank test is used to verify that the difference is statically significant. **Table 10** shows the calculations for this test.

Age Group	Total n	Total Observed	Total Expected	$\frac{(O-E)^2}{E}$
18-30	349	163	185.26	2.675
30-40	249	116	107.57	0.661
40-50	68	34	21.03	7.993
50+	4	2	1.14	0.655

Table 10: Log-Rank Test by Age Group

Chi-square test statistic with three degrees of freedom equals $\chi_3^2 = 12.3$, and the p-value is 0.006. Thus, the difference is statistically significant at a 5% significant level. So, the employees within the age group 50+ have the lowest survival and are more likely to quit than any younger employee. The same applies to employees in the 40-50 age category. This shows that older employees have a higher turnover rate. The survival curve for employees by age group is shown in **Figure 14.**

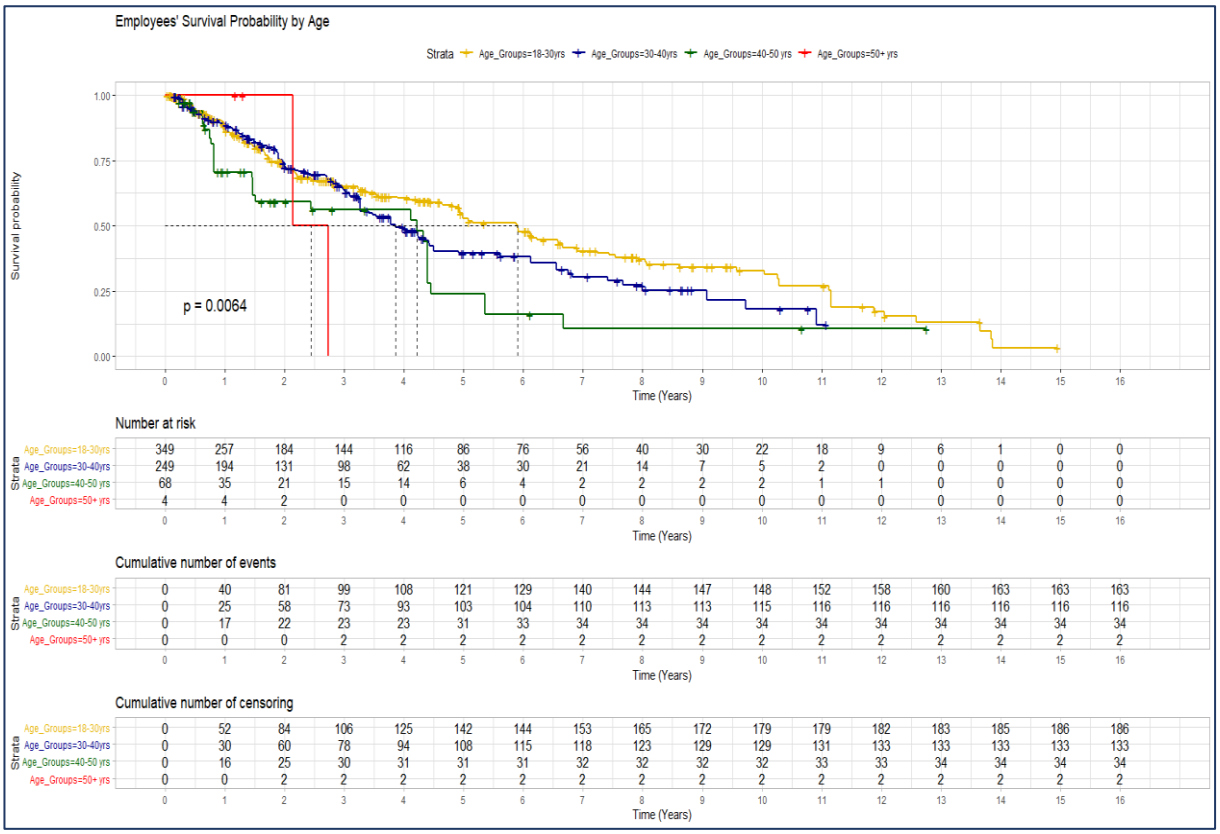

Figure 14: Kaplan-Meier Survival Curve for Employees by Age Group

6 Analysis with Cox Proportional Hazard

The dataset is filtered to include employees from four industries: Retail, IT, Manufacturing, and Banking because they are most common in the dataset as highlighted in the exploratory analysis. The R program used to analyse the data using the Cox model is ***analysis_2.R.***

The assumptions made when analysing the data with the Cox Proportional Hazard model are:

- The censoring is taken as a right, non-informative censoring
- Constant hazard ratio over time
- The variables are assumed to be time-independent
- Survival times are independent for each employee
- The censored employees were excluded from the analysis

6.1 Covariates Selection

A Cox model was applied in R on a combination of the covariates in the dataset. Then, an Akaike Information Criterion quantity is used to determine the covariates that have a statistically significant effect on survival. The Akaike Information Criterion, or AIC. This quantity is given by [1]

$$AIC = -2 \cdot l(\hat{\beta}) + 2 \cdot k \quad (37)$$

Where, $l(\hat{\beta})$ denotes the value of the partial log-likelihood for a particular model, and k is the number of parameters in the model. The use of the AIC in combination with a stepwise procedure in R enabled the choice of the covariates with the most significant effect on survival. The covariates are:

- Industries
- Self-control score (variable selfcontrol)
- Wage (variable greywage)
- Age

Age here is taken as a continuous variable not separated into age groups. The reason for that is, during the variable's selection process, the age of employees was presented as a continuous covariate (as the variable Age in the dataset) and presented in four age groups mentioned in the previous section (as the variable Age_Groups). It was found that age as a continuous covariate has a more significant effect.

A forest plot was created to display the hazard ratio and the $p - value$ to show the statistically significant difference in the effect of each category in each covariate, as shown in **Figure 15.**

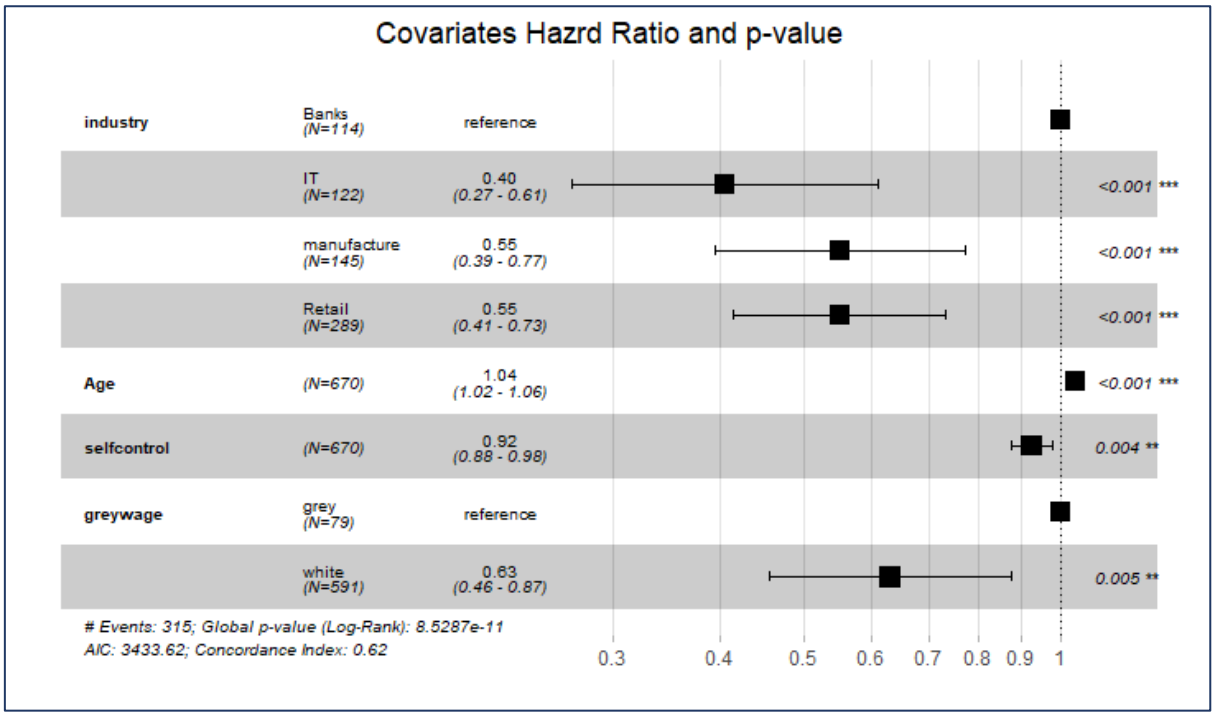

Figure 15: Forest Plot for Covariates' Hazard Ratio and The p-value for Significant Model

6.2 Checking the Proportional Hazard Assumption

The Schoenfeld Residuals plots are created for the selected covariates to confirm that the proportional hazard assumption is met. The plots are generated using program ***analysis_2.R,*** as shown in **Figure 16**.

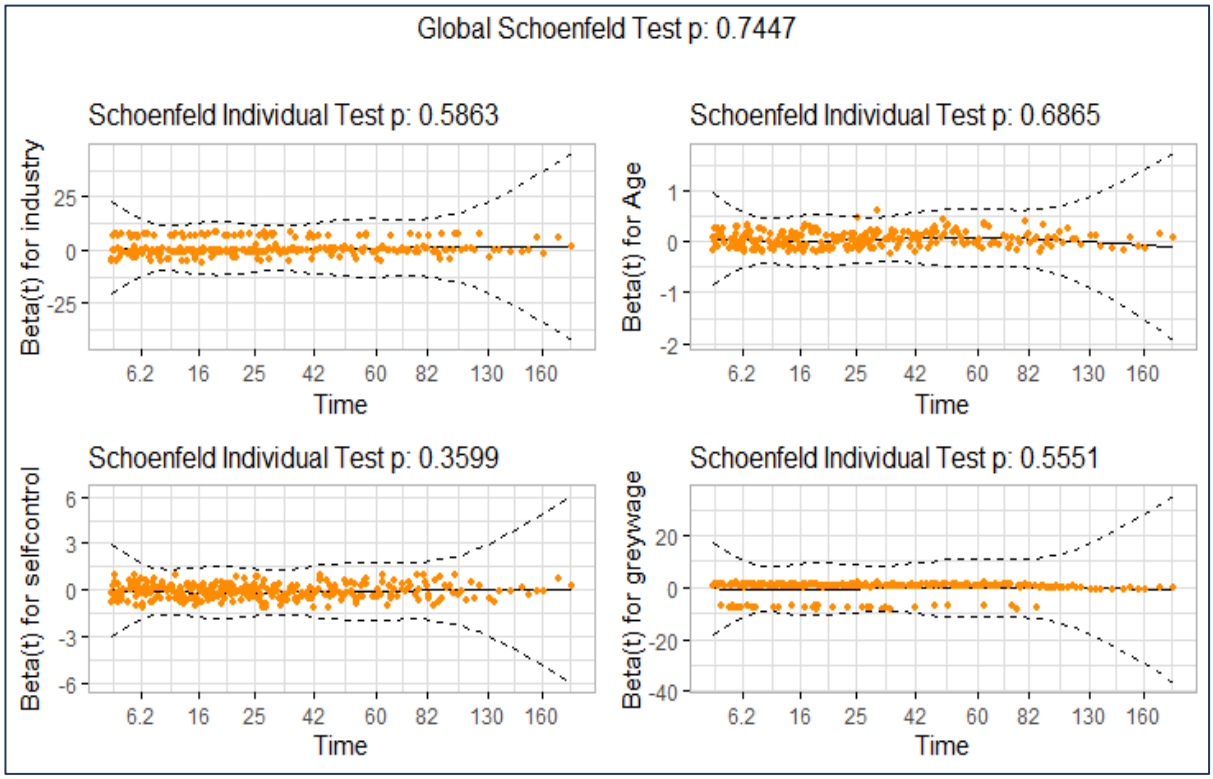

Figure 16: Schoenfeld Residuals Plots for Each Covariate

It can be seen in **Figure 16** that most of the points are centred at zero, so we can confirm that these covariates meet the assumption of constant HR. *R* also displays the p-value above each graph for a statistical test with the null hypothesis $\boldsymbol{H_0}$: *Covariates have proportional hazards*. It can be seen that all the p-values are high so that we can accept the null hypothesis. The hypothesis for checking the proportional hazard assumption can be obtained by fitting a straight line in the residual plot. The hypothesis test uses a Chi-square test statistic. The manual solving method for the statistical test will not be explained here. Instead, the R function $cox.zph()$ from the survival package is used to test this hypothesis.

6.3 Univariate Cox Model Analysis

6.3.1 Industry Analysis

The dataset is filtered to include the four industries: Retail, IT, Manufacturing, and Banking, as highlighted before. A Cox model is applied to the dataset in R. The results are as shown in **Table 11**. The category Banking is taken as the baseline to compare against the rest of the industry categories.

Variable Category	$\widehat{\beta}$	HR	Test Statistic	p-value
IT	−0.8396	0.4319	−4.054	0.0000503
Manufacturing	−0.4722	0.6236	−2.833	0.00461
Retail	−0.5744	0.5630	−3.982	0.0000683
N = 670, number of events = 315				
Likelihood ratio test with 3 degrees of freedom			20.99	0.0001059
Wald test with 3 degrees of freedom			22.4	0.00005
Score test with 3 degrees of freedom			23.13	0.00004

Table 11: Cox Model Outputs for Industries

Analysis results are interpreted as follows:

- The coefficients for these industries are negative, so they have a positive effect on survival. For example, if an employee belongs to the IT industry, there will be a decrease in the relative risk that they will quit their job
- The hazard ratio (HR) shows if an employee changed their career from the banking industry to IT. It would reduce the hazard of quitting by a factor of 0.43, or 57%. There is a 38% reduction in hazard when changed to Manufacturing and a 44% reduction when changed to retail.
- The test statistic column for each variable category displays the values for Wald test statistic, that tests if each coefficient β value has a statistically significant difference from 0

- The p-values for all these industries are very small, so we reject the null hypothesis $H_0: \beta = 0$. Thus, there is a statistically significant effect from this covariate on the hazard
- There are 670 employees, 315 of which had an observed event (quitting) occur for them. 355 employees were censored.
- The Likelihood ratio, Wald, and Score tests values at the bottom of the table provide insight into the overall significance of the model. The p-values for all these tests are very low, which shows that the model is statistically significant

The comparison between these industries shows that the IT industry has the lowest effect on hazard, so the turnover rate for an employee in IT is less than that in the other three industries mentioned above. The survival curve for four industries is displayed in **Figure 17**.

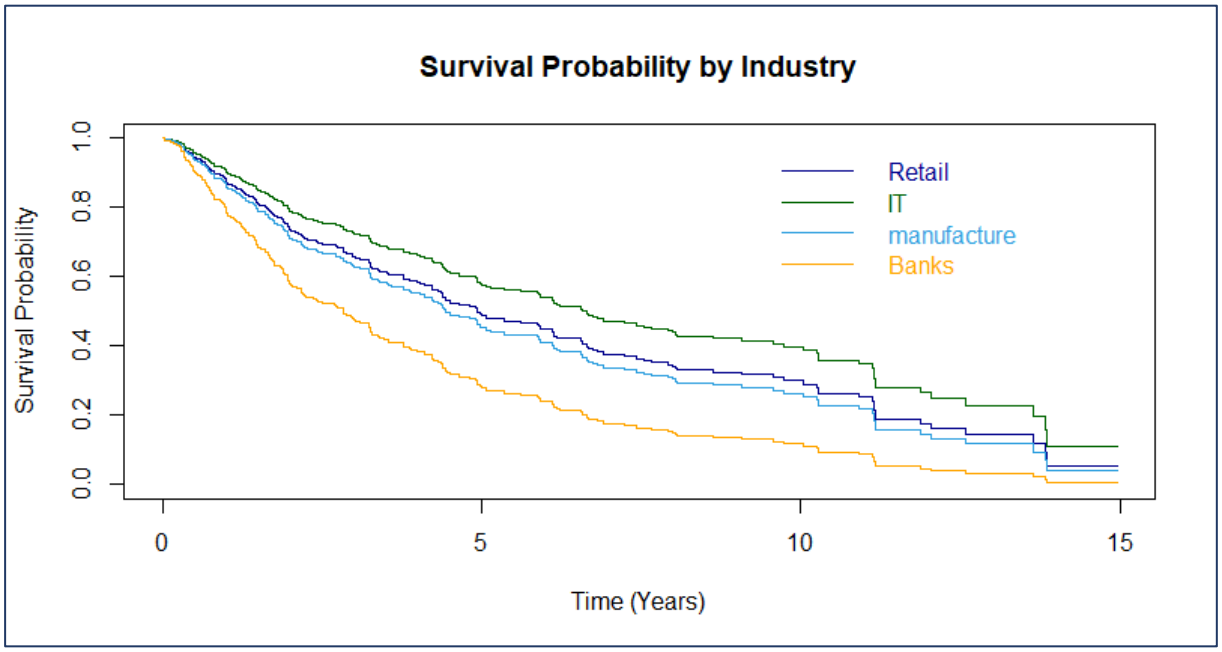

Figure 17: Survival Curves per Industry

6.3.2 Analysis of Age

A Cox model with age as a covariate is applied to investigate the effect on the turnover rate. **Table 12** shows the results of the analysis.

Variable Category	$\widehat{\beta}$	HR	Test Statistic	p-value
Age	0.034553	1.035157	4.248	0.0000216
N = 670, number of events = 315				
Likelihood ratio test with 1 degree of freedom			17.22	0.00003
Wald test with 1 degree of freedom			18.04	0.00002
Score test with 1 degree of freedom			18.1	0.00002

Table 12: Cox Model Outputs for Age

Age analysis results are interpreted as follows:

- The coefficient for age is positive, so they have a negative effect on survival. Thus, older employees have lower survival (higher risk of quitting)
- The hazard ratio (HR) shows that each additional year of age at the start of employment is associated with a 1.035 increase in the hazard of quitting
- The p-value for the coefficient's Wald test is very small, so we reject the null hypothesis $H_0: \beta = 0$. Thus, there is a statistically significant effect from this covariate on the hazard
- The p-values for The Likelihood ratio, Wald and Score tests for the model are very low, showing that the model is statistically significant

This shows that age is associated with a higher hazard. An employees' age at the start of their employment effect might increase their turnover rate.

6.3.3 Analysis of Self-control

The self-control variable represents the capacity to override an impulse in order to respond appropriately. A self-control score can provide much information about an employee's capabilities and mental health as self-control can be affected by factors such as mental exhaustion and stress. [7]

Self-control in psychology has three main parts:

- **Monitoring** involves keeping track of thoughts, feelings, and actions (**Emotional Control)**
- **Standards** are guidelines that steer us toward desirable responses
- **Strength** refers to the energy needed to control impulses (**Impulse Control)** [7]

A Cox model was applied to test the effect of self-control score on the turnover rate. **Table 13** shows the results.

Variable Category	$\widehat{\beta}$	HR	Test Statistic	p-value
Self-control score	-0.08317	0.92019	-3.008	0.00263
N = 670, number of events = 315				
Likelihood ratio test with 1 degree of freedom			9.10	0.003
Wald test with 1 degree of freedom			9.05	0.003
Score test with 1 degree of freedom			9.08	0.003

Table 13: Cox Model Outputs for Self-Control

Self-control score analysis results are interpreted as follows:

- The hazard ratio (HR) shows that each point increase in self-control corresponds to a 0.92 decrease in the hazard of quitting

- The coefficient for self-control is negative, so they have a positive effect on survival. Thus, employees with high self-control score have higher survival (lower risk of quitting)
- The p-value for the coefficient's Wald test is very small, so we reject the null hypothesis $H_0: \beta = 0$. Thus, there is a statistically significant effect from this covariate on the hazard
- The p-values for The Likelihood ratio, Wald and Score tests for the model are very low, showing that the model is statistically significant

This shows that a higher self-control score is associated with a lower hazard. An employees' self-control score can decrease their chance of quitting.

6.3.4 Analysis of Wage

Greywage variable represents the salary bracket for an employee in Russia or Ukraine. There are two categories present within the greywage variable in the dataset. They are white and grey and are interpreted as follows:

- **White**: are the salaries declared in an employee's contract, and tax is paid by both the employee and the employer
- **Grey**: it is a salary scheme divided into two parts, one of which is stated in the employment contract, and the other is issued in cash. So, lower tax is paid by both the employer and the employee [8]

The category Grey is taken as the baseline to compare against the other category **Table 14** shows the analysis result.

Variable Category	$\widehat{\beta}$	HR	Test Statistic	p-value
White	−0.5220	0.5933	−3.195	0.0014
N = 670, number of events = 315				
Likelihood ratio test with 1 degree of freedom			9.02	0.003
Wald test with 1 degree of freedom			10.21	0.001
Score test with 1 degree of freedom			10.44	0.001

Table 14: Cox Model Outputs for Wage

The analysis results are interpreted as follows:

- The coefficient for a White wage is negative, so it has a positive effect on survival. So, if an employee has their full wage declared in their employment contract. There will be a decrease in the relative risk that they will quit their job
- The hazard ratio (HR) shows if an employee's wage is changed from Grey to White. It would reduce the hazard of quitting by a factor of 0.59, or 41%

- The p-value for the coefficient's Wald test is very small, so we reject the null hypothesis $H_0: \beta = 0$. Thus, there is a statistically significant effect from this covariate on the hazard
- The p-values for the Likelihood ratio, Wald, and Score tests values at the bottom of the table for the overall significance of the model are very low, which shows that the model is statistically significant

The comparison between the wage types shows that employees with their full wage stated in their contract have the lowest effect on hazard, so the turnover rate for these employees is less than that for employees with Grey wage. The survival curve for the two wage types is displayed in **Figure 18.**

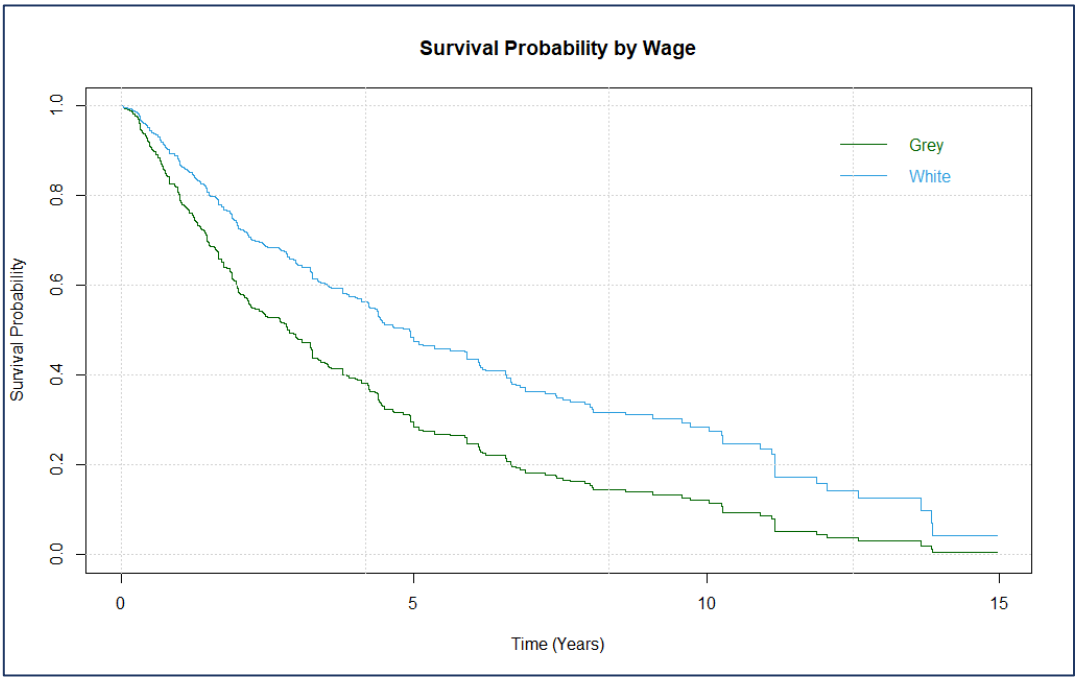

Figure 18: Survival Curves by Wage

6.4 Multivariate Cox Model Analysis

The four selected covariates are applied to the data to check all these covariates' combined effect. The results are shown in **Table 15.**

Variable Category	$\widehat{\beta}$	HR	Test Statistic	p-value
IT	-0.90440	0.40479	-4.305	0.0000167
Manufacturing	-0.59514	0.55148	-3.476	0.000509
Retail	-0.59556	0.55127	-4.091	0.0000429
Age	0.03994	1.04075	4.721	0.0000429
Self-control	-0.07868	0.92434	-2.872	0.004083
Wage: White	-0.45983	0.63139	-2.782	0.005406
N = 670, number of events = 315				
Likelihood ratio test with 3 degrees of freedom			20.99	0.0001059
Wald test with 3 degrees of freedom			22.4	0.00005
Score test with 3 degrees of freedom			23.13	0.00004

Table 15: Cox Model Outputs for Selected Covariates

The results of the multivariate analysis are as follows:

- The overall model is still significant as the p-values for the Likelihood ratio, Wald, and Score tests are small. Also, the p-values for the Wald test for the coefficients of the covariates are small, so we reject the null hypothesis $H_0: \beta = 0$ at the 5% significant level
- The various categories of the industry still correspond to a decrease in hazard, as can from the hazard ratio (HR). Similarly, for self-control and White wage. While age still corresponds to an increase in hazard

The survival curves for the combined effect of these covariates are plotted by industry, as shown in **Figure 19.**

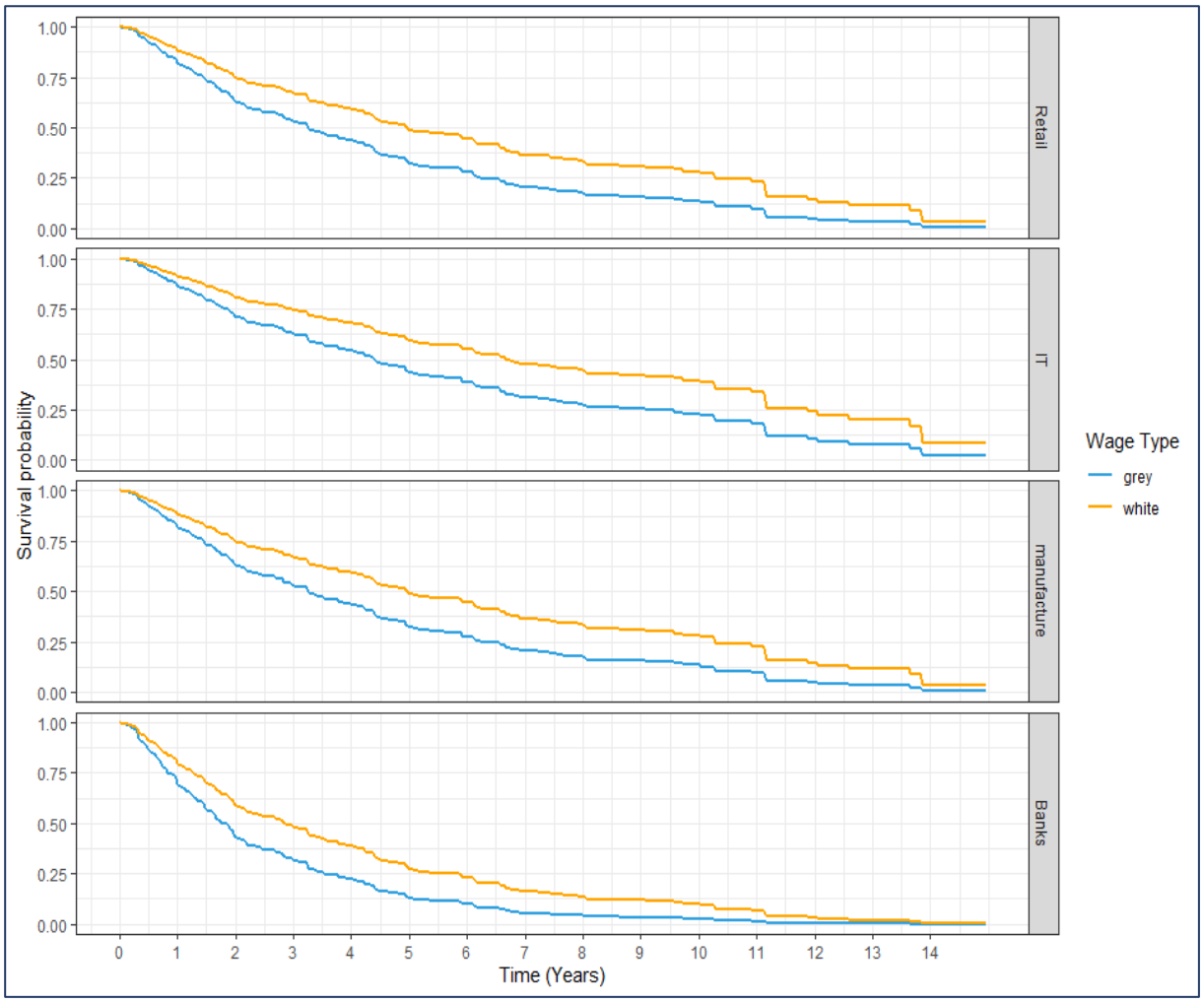

Figure 19: Survival Curves for The Combined Effect of The Covariates

6.5 Predictions

The model in section 6.4 can be used to make predictions. For example, if an employee has the attributes mentioned in **Table 16** and an employer need to evaluate their turnover rate after five years.

Subject	Age	Industry	Self Control	Wage	Survival Probability	Hazard
1	25	IT	7.5	White	0.7004272	0.006495765
2	40	Manufacturing	6.5	Grey	0.2202198	0.02760565

Table 16: Example Employees Attributes and Predicted Survival

The hazard can be predicted using equation (23) in section 3.3, and also, using the values of the coefficients mentioned in **Table 15,** the base hazard is estimated using R programme ***analysis_2.R***. The base hazard is found to be

$$h_0(60) = 0.01689437$$

The hazard equation for the **first subject** will be as following:

$$h(60) = 0.01689437exp((0.03994 \times 25) + (-0.90440 \times 1) + (-0.07868 \times 7.5) + (-0.45983 \times 1))$$

$$h(60) = 0.006495765$$

Similarly, the hazard for the **second subject** can be calculated.

$$h(60) = 0.01689437exp((0.03994 \times 40) + (-0.59514 \times 1) + (-0.07868 \times 6.5) + (-0.45983 \times 0))$$

$$h(60) = 0.02760565$$

The survival probability for these two subjects is estimated using this model in R using programme ***analysis_2.R***. The results are shown in **Table 16.** The predicted survival curves for these subjects over 15 years is shown in **Figure 20.**

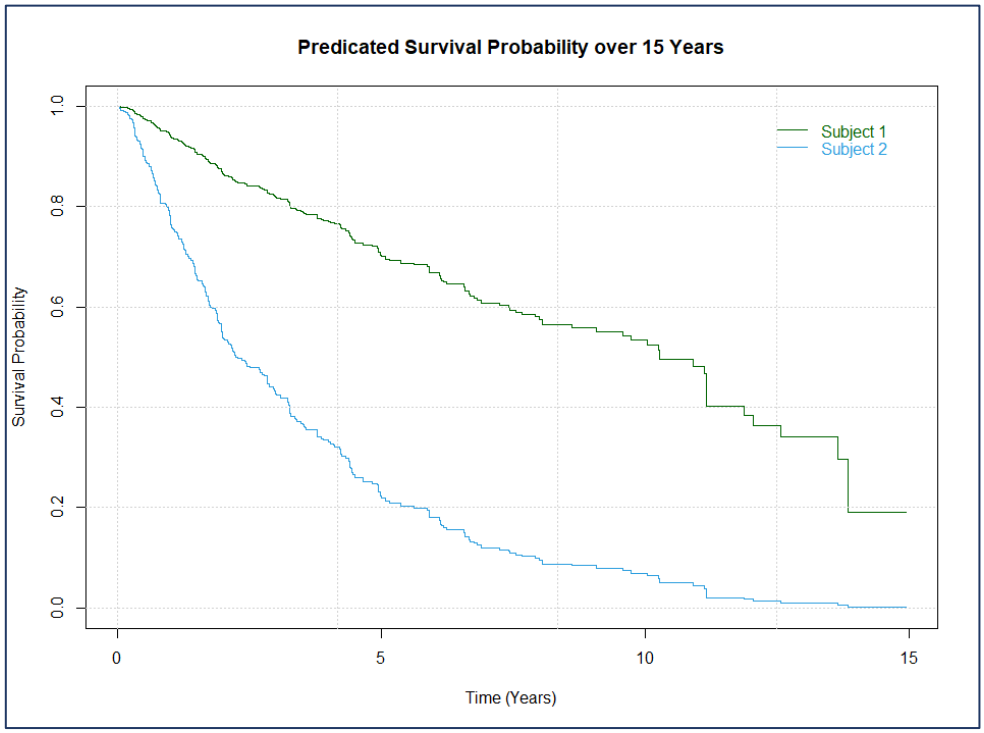

Figure 20: Predicted Survival Curves for The Two Subjects

7 Summary of Results

This document introduced survival analysis and the various models used to estimate survival probability and Hazard for datasets that contain information about time to an event. The dataset used in this document is concerned with employees' turnover rate (time till an employee quits). The data represents actual employees from Russia in various industries and professions.

The analysis focused only on four industries: Retail, IT, Manufacturing, and Banking, as they are the most common in the dataset as highlighted in the exploratory analysis. Kaplan-Meier and Cox regression model were the only models used to estimate overall survival and identify covariates effect on survival. The Cox model being semi-parametric, has the advantage of preserving the variable in its original quantitative form and using a maximum of information as fewer assumptions than typical parametric methods.

Partial likelihood estimation was implemented to estimate the regression coefficient β so that predictions can be made. One of the main objectives was to highlight the covariates that had the most effect on an employee turnover rate and use the results to estimate the survival probability for employees after a specific period. It was found that these covariates are:

- Industries
- Self-control score (variable selfcontrol)
- Wage (variable greywage)
- Age

Kaplan-Meier method was used to evaluate the overall survival of the employees in the dataset and evaluate the difference in survival within age groups and gender. The results show a loss of 47% of the employees as 315 employees quit by the end of the observed 15 years, and the median turnover was approximately four years. The survival probability decreased the longer an employee worked for the company.

It was also found that age and gender had a significant effect on survival. Females had a higher turnover rate than males, and employees in older age groups had a lower survival (higher turnover rate) than employees in younger age groups. For example, employees in the age group 50+ had a lower median survival time than employees in the 40-50 age group. Those had a lower median survival time than employees in the 30-40 age group.

The Cox model used with the selected covariates mentioned above to evaluate their effect. It was found that industries: Retail, IT, and Manufacturing have a positive effect on the hazard (reduce the hazard of quitting) when compared against Banking, with IT having the lowest turnover rate between these four industries with a 57% reduction in hazard if an employee changed their industry from banking to IT.

It was also found that age has a negative effect on hazard, where each year increase in an employee's age is associated with a 1.035 increase in the hazard of quitting. These results collaborate with the finding from the Kaplan-Meier method. The contrary effect for self-control, where each additional point in self-control associated with a 0.92 decrease in the hazard of quitting. Self-control scores can provide much information about an employee's mental health as self-control can be affected by mental exhaustion and stress. Employers can use this information to improve the work environment and provide better support for employees to reduce turnover.

The wage comparison shows that employees with contracts that states their full wage (White wage) would have a 41% reduction in hazard than employees who have a salary scheme divided into two parts, one of which is stated in the employment contract. The other is issued in cash (Grey wage), which shows that employees feel more secure in their role and less likely to quit if they have a contract stating their full salary and benefits clearly.

The Cox model is a regression model and can be used to make predictions for future employees and allow employers to plan their hiring strategy and implement changes to the work environment or other aspects to retain their most valuable employees.

8 Limitations and Recommendations

It can be seen that all the analysis objectives were met. However, these models had some limitations.

- The Kaplan-Meier method would only allow a univariate analysis of variables
- Cox model can be used only with variables that follow the assumptions of the model

The recommendations for future analysis of data are:

- Treat age and self-control as a time-dependent covariate as these variables will change with time
- Apply a piecewise analysis approach to the variables that do not meet the proportional hazard assumption
- Apply a parametric distribution to the dataset like Weibull distribution and examine the results
- Compare the results from the parametric distribution to the Cox model results

9 Bibliography

[1] Moore, Dirk F., Applied Survival Analysis Using R, Springer International Publishing AG, 2016

[2] Kleinbaum, David G and Klein, Mitchel, Survival Analysis: A Self-Learning Text Third Edition, Springer New York, 2012

[3] Baginska, Oleksandra, Applying Survival Analysis in HR Analytics on Real-Life Data. Viewed on 23/11/2020, *https://www.analyticsinhr.com/blog/applying-survival-analysis-reduce-employee-turnover-practical-case/*

[4] Van Vulpen, Erik, What Drives Employee Turnover? Viewed on 10/02/2021, *https://www.analyticsinhr.com/blog/what-drives-employee-turnover/*

[5] Handrick, Laura, 5 Reasons Your Retention Rate Is So Low. Viewed on 21/02/2021, *https://gethppy.com/employee-turnover/5-reasons-your-retention-rate-is-so-low*

[6] Sullivan, Lisa, Survival Analysis. Viewed on 17/03/2021, *https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival6.html*

[7] DeWall, Nathan, Self-control: Teaching students about their greatest inner strength, Viewed on 29/03/2021, *https://www.apa.org/ed/precollege/ptn/2014/12/self-control*

[8] Ostapenko, Raisa, Wages in envelopes: grey salaries in Russia and Ukraine, Viewed on 29/03/2021, *https://www.riskscreen.com/kyc360/article/wages-envelopes-russia-ukraine/#:~:text=White%20payments%20are%20the%20salaries%20declared%20in%20an%20employee's%20contract.&text=Another%20grey%20salary%20scheme%20involves,eligible%20for%20lower%20taxes%20rates*.