(11) **Task 1**

A (2) There is no truncation present in the study, according to the description.
One example of truncation could be if the study only includes employees who have already resigned or retired by the time of analysis, then employees who are still active at the company are never observed. Or, if the study only looks at employees with a minimum tenure, individuals who are only employed a very short period would not be included.

B (6)

i) The employee was still employed as of February 2025, the event thus hasn't been observed yet. This is a case of **right censoring**.

ii) The event (resignation or retirement) was observed during the study period, which means that there is **no censoring**.

iii) We know that the employee resigned sometime between October and December 2021, **interval censoring** is thus present here.

C (3) i) We have a right censored observation at 85 months.

ii) This is an event at 52 months.

iii) An interval censored observation at somewhere between 49 and 51 months.

$$\mathbf{L = \mathit{S}\,(85) \cdot \mathit{f}(52) \cdot \{\mathit{S}(49) - S(51)\}}$$

(26) **Task 2**

A (2) Since the event is defined as an employee voluntarily leaving their job, one obvious competing risk would be if an employee leaves their job, but not voluntarily (e.g. being fired). This would prevent us from observing the event of interest.

B (8) Small companies:
25th percentile = 4.7 years. This means that 25% of the employees in small companies have voluntarily left their jobs within 4.7 years from the start of their employment.

Median = 6.5 years. This means that 25% of the employees in small companies have voluntarily left their jobs within 6 years from the start of their employment.

75th percentile = 10.9 years. 75% of the employees in small companies have voluntarily left their jobs within 10.9 years from the start of their employment (actually 100% have left their jobs at this time).

.

Medium sized companies:
25th percentile = 5.4 years. This means that 25% of the employees in medium sized companies have voluntarily left their jobs within 5.4 years from the start of their employment.

Median and 75% percentile = at least 10.6 years. Neither the median nor the 75th percentile can be estimated since less than 50% of the employees in medium sized companies have voluntarily left their jobs during the study. Both the median and 75th percentile must however be at least 10.6 years, which is the longest observed time in this group.

These measures have larger values for employees in medium sized companies, which means that the employees tend to stay longer with medium sized than small companeis.

C (3) Approximate 6-year probability of still being employed can be seen from the Kaplan-Meier curve.
For employees in small companies: approx. **60%**
For employees in medium sized companies: approx. **70%**

The probability of still being employed after 6 years is a higher for employees in medium sized companies, almost 10 percentage points higher than for employees in small companies.

D (13) $H_0: \ h_{small}(t) = \ h_{medium}(t) \ = \ h_{large}(t)$ for all $t \leq \tau$
H_a: $h\,(t)$ differ between some of the groups for some $t \leq \tau$

Where τ = largest time at which all groups have at least one subject at risk.

Test: Should be discussed with field expert(s), if there is a larger interest in earlier differences Gehan's/Wilcoxon test should be chosen, and if all time points are equally interesting the Log-Rank test should be chosen.

Only one test should be chosen here, with a motivation.

Assumptions:
- Random sample – not specified in the task, which means that the results below should be interpreted with care (the inference only holds if the sample is random)
- Independent samples – reasonable to assume that career longevity in the three compared groups are independent
- Non-informative/random censoring (i.e. that the censoring times are not related to the later, unknown, retirement times) – not specified in the task. Should be discussed with field experts, but at least the censoring plot doesn't contradict this since the censoring pattern is similar in all three groups.
- Right censored data – OK (information provided in task)

- Survival probabilities are the same for subjects recruited early and late in the study - not specified in the task. Should be discussed with a field expert, but reasonable to assume.
- Large samples (both tests are based on large-sample approximations to the distribution of the chi-square statistics) – OK, 38 events in the smallest group.

Choice of significance level:

Wrongly rejecting the null hypothesis here would mean that we claim that there is a difference in time to retirement between the three company size groups, when in fact there is no difference. There are hardly any serious consequences from this, 5% is a reasonable choice.

Result:
P-value = 0.0502 (for the Log-rank test).
The *P*-value is not smaller than the chosen α, thus H_0 is not rejected.

Conclusion:
The test does not suggest that there is a significant difference in career longevity between small/medium/large companies in the investigated population of software engineers, where employees in larger companies in general retire later than the others. Note however that this conclusion is only valid if the sample was taken randomly.

(30) **Task 3**

A (3) The variables *company_size, education_level* (and *salary_group,* if used) have to be recoded to 0/1 variables, or denoted as "class" variables in proc phreg.

The Martingale plot suggests that *starting_salary* can be used as a continuous covariate (and *salary_group* is thus not needed).

B (15) **Assumptions** for the Cox model:
(Many of them are already discussed in Task 2, OK to refer to that)
• Random sample (for inference to be correct) - unknown.
• Non-informative censoring. Not contradicted.
• Right-censored or left truncated data. Right censoring can be seen from task.

New for this task:
• Large sample (common rule of thumb: ≥ 10 events per covariate). We have a total of 450 observations, and 142 events. This is sufficient for the estimated model which has 5 covariates.
• Proportional hazards (to be checked when building the model)

Check of **proportionality (PH) assumption**:

1) include time-dependent covariate in model
To test the assumption of proportional hazards you can include the time-dependent covariate ln(t)*covariate in the model (if significant, the PH assumption is rejected)

H_0: The hazards for different values of covariate i are proportional
(all i=1 to p covariates are to be examined)
H_a: The hazards are not proportional

Significance level α = 5% fine to use (no serious consequences if we claim that the hazards are not proportional when they in fact are)

The test above is rejected for the *starting_salary* and *education_level* covariates, but fine (non-significant) for *remote_work* and *salary_group*.

2) Arjas plots
You should always make use of a graphical method in addition to the test above. Arjas plots have been provided, which show that the PH assumption is violated for *company_size* (the green curve is nonlinear) and *salary_group*, but looks okay for *education_level* and *remote_work* (some of these curves follow the 45 degree line to start with, and there is some minor crossing, but the departures from linearity are not obvious).

This concludes that the PH assumption doesn't hold for *company_size*, *starting_salary*, and *education_level*.

Thus, the model should either include these covariates as time-dependent, or stratify on them. The presented model includes a time-dependent covariate for *starting_salary*, and stratifies on *education_level*, but there is no remedy for *company_size*.

Test of equality of strata for *education_level*:
For the stratification on *education_level* to be valid, we need to check that it is reasonable to assume that the regression coefficients are the same in each stratum.

H_0 : All β's are the same for all s strata
H_a : At least one of the β's is/are different

This can be tested, using the Likelihood ratio test.

Significance level α=5% fine to use (no serious consequences if we claim that the covariates are not the same when they in fact are, then we'll estimate separate models instead)

Test statistic:

$$-2\left[LL(\mathbf{b}) - \sum_{j=1}^{s} LL_j(\mathbf{b}_j)\right] \sim \chi^2_{(s-1)p}$$

where s= no. of strata and p=no. of covariates
s = 3, p = 5

$-2LL(\mathbf{b}) = 1260.920$
$LL(\mathbf{b}) = -630.460$

$-2LL(\mathbf{b}_{Bachelor}) = 653.427$
$-2LL(\mathbf{b}_{Master}) = 503.093$
$-2LL(\mathbf{b}_{Phd}) = 90.015$

Sum $-2LL_j(\mathbf{b}_j) = 1246.535$
Sum $2LLj(\mathbf{b}j) = -623.2675$

Test statistic = 1260.920-1246.535 =
= -2(-630.460+623.2675) =
= 14.385

df = (3-1)*5 = 10 df

According to Table c.2 the null hypothesis should be rejected if $\chi^2_{test} > \chi^2_{crit} = 18.3$ (the corresponding p-value is larger than 0.01).
The null hypothesis is not rejected, since $\chi^2_{test} = 14.385$ is smaller than 18.3.

Conclusion:
It is okay to assume that the regression coefficients are the same in each of the two strata and the stratified model can be used.

Cox-Snell plot
The Cox-snell plot shows that the model fit isn't that good.

Conclusion:
The suggested model should not be used, since it doesn't take into account that *company_size* violates the PH assumption. *Starting_salary* and *education_level* are however handled correctly in the model.

C (8) The marginal effects of the covariates are presented below, i.e. the effect of each covariate holding the other covariates constant.

Compared to employees at small companies, employees at large companies have on average a 45.5% lower risk of retirement, (hazard ratio 0.545, 95% confidence interval 0.356 to 0.826), and employees at medium sized companies have a 37.2% lower risk of retirement on average (hazard ratio 0.628, 95% confidence interval 0.423 to 0.927). Both of these risks are significantly different from the small company risk.

Employees that primarily work remotely have a 63.6% higher risk of retirement on average, compared to employees that don't primarily work remotely (hazard ratio 1.636, 95% confidence interval 1.168 to 2.297). This effect on the risk of retirement is also significant.

Since *starting_salary* is time-dependent, the effect on the risk of retirement is different at different time points.

The hazard ratio at 2, 5, and 8 years, respectively, for and increase of *starting_salary* by 1000 USD :

At 2 years: HR = exp(-0.01289-0.00790*(ln(2)) = 0.982
At 5 years: HR = exp(-0.01289-0.00790*(ln(5)) = 0.975
At 8 years: HR = exp(-0.01289-0.00790*(ln(8)) = 0.971

The risk of retirement seems to decrease over time, at 2 years there is a 1.8% risk reduction for every 1000 USD increase, at 5% the same risk reduction is 2.5%, and at 8 years it is 2.9%.

Since the model is stratified by education_level, there are no hazard ratios to interpret, but the plot of estimated survival shows that employees with a Bachelor's degree have a lower risk of resignation than employees with a Master's degree. The risk of resignment for employees with a PhD degree is difficult to separate from the others (it varies over time).

D (2) Relative risk = $\exp(\beta_{large} - \beta_{medium}) = \exp(\beta_{large})/\exp(\beta_{medium}) = 0.545 / 0.628 = 0.868$

This means that the risk of resigning for employees at large sized companies is on average 13.2% lower than the risk of resigning for employees at medium sized companies (all other covariates held constant).

E (2) Generalized $R^2 = 1 - e^{-(LRT/n)}$

where $LRT = -2\log L(0)-[-2\log L(p)] = 1297.853 - 1260.920 = 36.933$

$R^2 = 1 - \exp(-36.933/450) = 1 - 0.9212 = 0.0788$

According to the generalized R^2 the model shows some association between the covariates and time to resignment.

(9) **Task 4**

(2) **A** The model being estimated (an accelerated failure time model):

$$Y = \ln(X) = \mu + \gamma_1\, team\_performance_{high} + \gamma_2\, team\_performance_{medium} + \gamma_3\, nationality_{international} + \gamma_4\, age + \sigma W$$

where W follows the extreme value distribution (which yields a Weibull regression model).

(2) **B** Time to resignment is assumed to follow the Weibull distribution (i.e., the baseline hazard follows the Weibull distribution).

(5) **C** Since a Weibull regression model is used, the transformed parameter estimates can be interpreted as proportional hazards estimates.

The marginal effects of the covariates are presented below, i.e. the effect of each covariate holding the other covariates constant.

Compared to players in teams with low performance, players in medium performance teams have on average a 14.3% lower risk of retirement, (hazard ratio 0.857), and players in high performance teams have a 39.6% lower risk of retirement on average (hazard ratio 0.604). Both of these risks are significantly different from the risk for players in low performance teams.

Players in international teams have on average a 0.4% lower risk of retirement than players in domestic teams, (hazard ratio 0.996), this is however not a significant risk reduction.

For every year older a player is, the risk of retirement increases on average by 6.1%, (hazard ratio 1.061), and this is a significant effect.

(4) **Task 5**

As always, it is a good thing to explain to your employer that you are following ethics codes for statisticians.

As long as you have given a few examples of part(s) of one or more of the code of ethics document(s) that are applicable in this situation, and told which document you are referring to, you'll get points for this task.