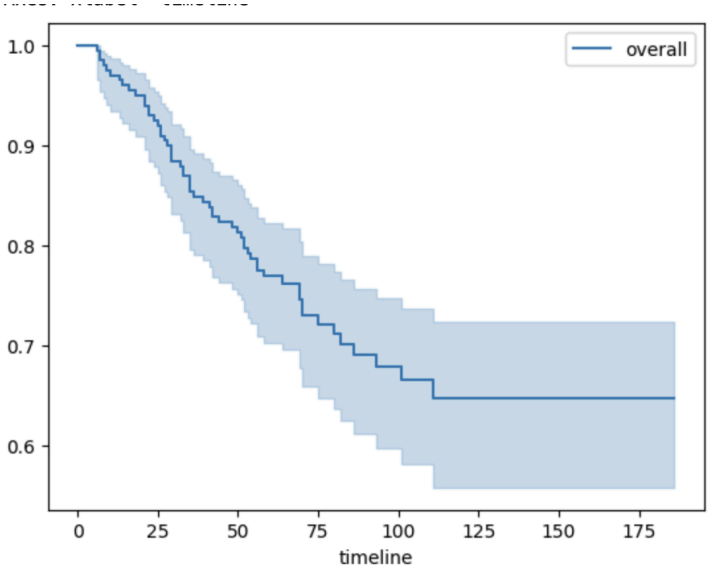
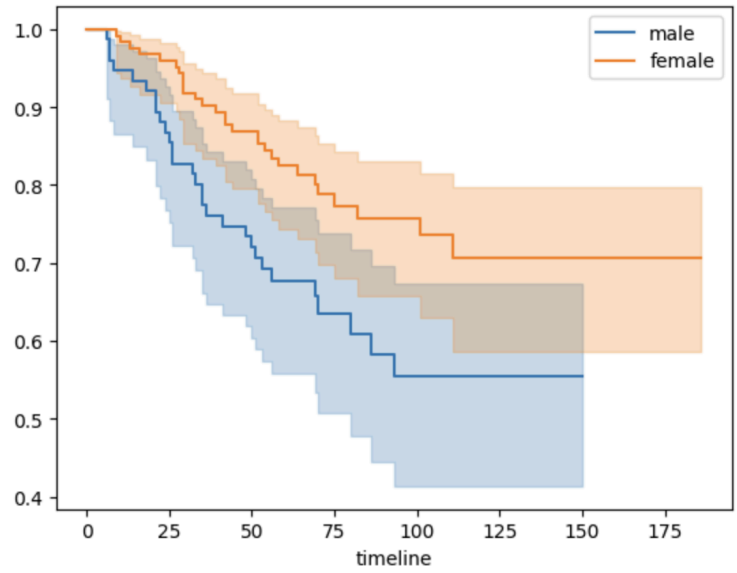
(i) Plot the Kaplan Meier curve for survival. Do the same for male vs female [5]





(ii) Estimatea Cox regression to estimate the impact of the covariates. Write the equation and then estimate the model. How will you interpret these estimates? [10]

$$h(t|X) = h_0(t)\exp(\beta_1 \times \text{gender} + \beta_2 \times \text{age} + \beta_3 \times \text{year} + \beta_4 \times \text{thickness} + \beta_5 \times \text{ulcer})$$

	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)
Gender	0.60	1.83	0.27	0.07	1.14	1.07	3.13	0.00	2.20	0.03	5.18
age	0.02	1.02	0.01	-0.00	0.03	1.00	1.03	0.00	1.77	0.08	3.69
year	-0.08	0.92	0.06	-0.20	0.03	0.82	1.03	0.00	-1.42	0.16	2.68
thickness	0.15	1.16	0.03	0.08	0.21	1.08	1.24	0.00	4.29	<0.005	15.76
ulcer	0.33	1.39	0.27	-0.21	0.86	0.81	2.37	0.00	1.19	0.23	2.10

Holding other constants, on average:
Being male will increase the log of baseline hazard by 83%
On year increase in age of operation will increase the log of baseline hazard by 2%
One year increase of the operation time will decrease the log of baseline hazard by 8%
One mm in tumor thickness will increase the log of baseline hazard by 16%
The presence of ulcer will increase the log of baseline hazard by 39%

(iii) Can you provide some rationale for testing the proportionality assumption? How would you do it in Python? Now provide the results of the test and comment on what you find. [7]
The proportionality assumption says that the independent variable will shift the hazard up or down uniformly across time. But the assumption is very easily violated. Male partients could have more risk than female partients when they are old rather than young. Increase in age could expose the patients at higher risk right after the time of operation rather than many years after. Operation time could be due to advances in surgical methods, which could also only decrease the immediate hazard right after surgery rather than long term hazard. The assumption is very easy to break, so we must do a proportionality test.

```
from lifelines.statistics import proportional_hazard_test

results = proportional_hazard_test(cph, df)
results.print_summary(decimals=3, model="untransformed variables")
```

time_transform	rank
null_distribution	chi squared
degrees_of_freedom	1
model	<lifelines.CoxPHFitter: fitted with 205 total ...
test_name	proportional_hazard_test

	test_statistic	p	-log2(p)
Gender	1.97	0.16	2.64
age	2.55	0.11	3.18
thickness	2.80	0.09	3.41
ulcer	0.26	0.61	0.71
year	0.13	0.72	0.47

```
cph.check_assumptions(df, p_value_threshold=0.05,  show_plots = False)
```

```
Proportional hazard assumption looks okay.
[]
```

Although none of the features statistically significantly break the assumption, thickness and age, year still have a low p-value, and we need to take those into considerations.

(iv) In this model, you assume that the shape of the hazard is the same and the tumor thickness shifts it proportionally. However, it is possible that the shape of the hazard is different for different tumor thickness. You plan to stratify based on thickness. Since tumor thickness is a continuous measure, your first create a dummy variable which is 1 when tumor thickness is above median and 0 otherwise.

You now stratify your model based on thickness dummy. Operationalize this in your python code. Provide the output.[10] Why do you think the estimates have changed? [5]

	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)
Gender	0.44	1.55	0.28	-0.10	0.98	0.90	2.65	0.00	1.59	0.11	3.15
age	0.01	1.01	0.01	-0.01	0.03	0.99	1.03	0.00	1.34	0.18	2.48
year	-0.05	0.95	0.06	-0.16	0.06	0.85	1.06	0.00	-0.90	0.37	1.44
thickness	0.06	1.06	0.05	-0.03	0.15	0.97	1.16	0.00	1.36	0.17	2.53
ulcer	0.15	1.16	0.28	-0.39	0.69	0.68	2.00	0.00	0.55	0.58	0.78

$$h(t|X) = h_{0thickness>median}(t)\exp(\beta_1 \times \text{gender} + \beta_2 \times \text{age} + \beta_3 \times \text{year} + \beta_4 \times \text{thickness} + \beta_5 \times \text{ulcer})$$

$$h(t|X) = h_{0thickness<=median}(t)\exp(\beta_1 \times \text{gender} + \beta_2 \times \text{age} + \beta_3 \times \text{year} + \beta_4 \times \text{thickness} + \beta_5 \times \text{ulcer})$$

Before stratify, we assumed tumors of different thickness have the same baseline hazard and the tumor thickness shift it up proportionally. But after stratifying, all the coefficient or effect of features reduced. The different results says that the basine hazards are essentially different in the two groups, and also the redistribution of the coefficient says that the model now more accurately attributes some of the risk to the stratification variable itself. The assumption of proportional hazards may hold more accurately within each group than across the entire population.