ML IN HEALTHCARE

SURVIVAL ANALYSIS

Մարիամ Սարգսյան

Survival analysis based on Roman Emperors dataset

I am going to perform a thorough analysis of emperor dataset, I'll look at **Assassination**, **Captivity** and **Execution** as traumatic ends to their reigns (event = 1), while Natural Causes, Died in Battle and Suicide are more "normal" ways for an emperor's reign to come to an end (event = 0).
Survival Analysis is used to analyse time until the roman emperor reign.
Typical research question is what is the probability that an emperor reign during specific time.

Survival analysis is a set of statistical approaches used to determine the time it takes for an event of interest to occur. We use survival analysis to study the **time** until some **event** of interest occurs. Time is usually measured in years, months, weeks, days, and other time measuring units. The event of interest could be anything of interest. It could be an actual death, a birth, a retirement, along with others.

Kaplan-Meier plots to visualize survival curves.

Here, we start by defining fundamental terms of survival analysis, including:

1. Survival time and event.
2. Censoring of data.
3. Survival function and hazard function.

Survival Function(S)
Hazard Function (H)

1. To find the survival probability of a subject, we will use the survival function S(t), the Kaplan-Meier Estimator. The survival function is defined as the probability that an individual (subject) survives from the time origin (diagnosis of a disease) to a specified future time t. Please note that the time can be in various forms like minutes, days, weeks, months, or years.
2. The hazard probability, denoted by H(t), is the probability that an individual (subject) who is under observation at a time t has an event (death) at that time.

The **Kaplan–Meier estimator** is a non-parametric statistic used to estimate the survival function (probability of a person surviving) from the lifetime data. The probability of survival at time ti, which is denoted by S(ti), is calculated as follow:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

Where,

$\hat{S}(t)$ = *The probability of being alive at time t*

n_i = *The number of subjects alive just before time* t_i

d_i = *The number of events at time* t_i

	Name	Birth	Succession	Reign	Time	Deaths
0	Augustus IMPERATOR CAESAR DIVI FILIVS AVGVST...	23 September 63 BC, Rome , Italia \n	Great-nephew and adopted son of Julius Caesa...	16 January 27 BC – 19 August 14 AD\n	40 years, 7 months and 3 days\n	19 August 14 AD (aged 75) Natural causes \n
1	Tiberius TIBERIVS CAESAR DIVI AVGVSTI FILIVS...	16 November 42 BC, Rome , Italia \n	Natural son of Livia Drusilla , Augustus' th...	18 September 14 AD – 16 March 37 AD\n	22 years, 5 months and 27 days\n	16 March 37 AD (aged 77) Probably natural cau...
2	Caligula GAIVS IVLIVS CAESAR AVGVSTVS GERMANI...	31 August 12 AD, Antium , Italia \n	Great-nephew and adoptive grandson of Tiberiu...	18 March 37 AD – 24 January 41 AD\n	3 years, 10 months and 6 days\n	24 January 41 AD (aged 28) Assassinated in a ...
3	Claudius TIBERIVS CLAVDIVS CAESAR AVGVSTVS ...	1 August 10 BC, Lugdunum , Gallia Lugdunens...	Uncle of Caligula; brother of Germanicus; nep...	25/26 January 41 AD – 13 October 54 AD\n	13 years, 8 months and 18/19 days\n	13 October 54 AD (aged 63) Probably poisoned ...
4	Nero NERO CLAVDIVS CAESAR AVGVSTVS GERMANICV...	15 December 37 AD, Antium , Italia \n	Great-nephew, stepson, son-in-law and adopted...	13 October 54 AD – 9 June 68 AD\n	13 years, 7 months and 27 days\n	9 June 68 AD (aged 30) Committed suicide afte...

```
Total number of emperors: 174
Total number of raw features: 6
Features available: ['Name', 'Birth', 'Succession', 'Reign', 'Time', 'Deaths']
```

We use regex to extract meaningful features from the raw data. It's a good alternative to manually iterating and extracting information from individual rows. Note that years before AD (year zero) are encoded as a negative integer. The process is not 100% robust yet, as the output still has a few outliers due to the messiness of the raw data

	AgeOfDeath	ReignYears	ReignMonths	ReignDays	BirthName	Title	CauseOfDeath	BirthYear	BirthDay	BirthMonth	ReignStartYear	ReignEndYear	ReignStartDay	ReignEndDay	ReignStartMonth	ReignEndMonth
0	75	40	7	3	Augustus	IMPERATOR	natural causes	-63	23	September	-27	14	16	19	January	August
1	77	22	5	27	Tiberius	TIBERIVS	probably natural causes, possibly assassinated...	-42	16	November	14	37	18	16	September	March
2	28	3	10	6	Caligula	GAIVS	assassinated in a conspiracy involving senator...	12	31	August	37	41	18	24	March	January
3	63	13	8	19	Claudius	TIBERIVS	probably poisoned by his wife agrippina the yo...	-10	1	August	41	54	26	13	January	October
4	30	13	7	27	Nero	NERO	committed suicide after being declared a publi...	37	15	December	54	68	13	9	October	June

Analyzing cause of death

Lastly we look at the most occurring cause of death amongst Roman emperors. It seems that, despite popular believe, that many in fact died from natural causes, but Constantinople is mentioned quite a lot as well. Also suicide wasn't that uncommon amongst emperors either.

```
Augustus died of natural causes
Caligula died of assassinated in a conspiracy involving senators and praetorian guards
Nero died of committed suicide after being declared a public enemy by the senate
Otho died of committed suicide after losing battle of bedriacum to vitellius
Vespasian died of natural causes
```

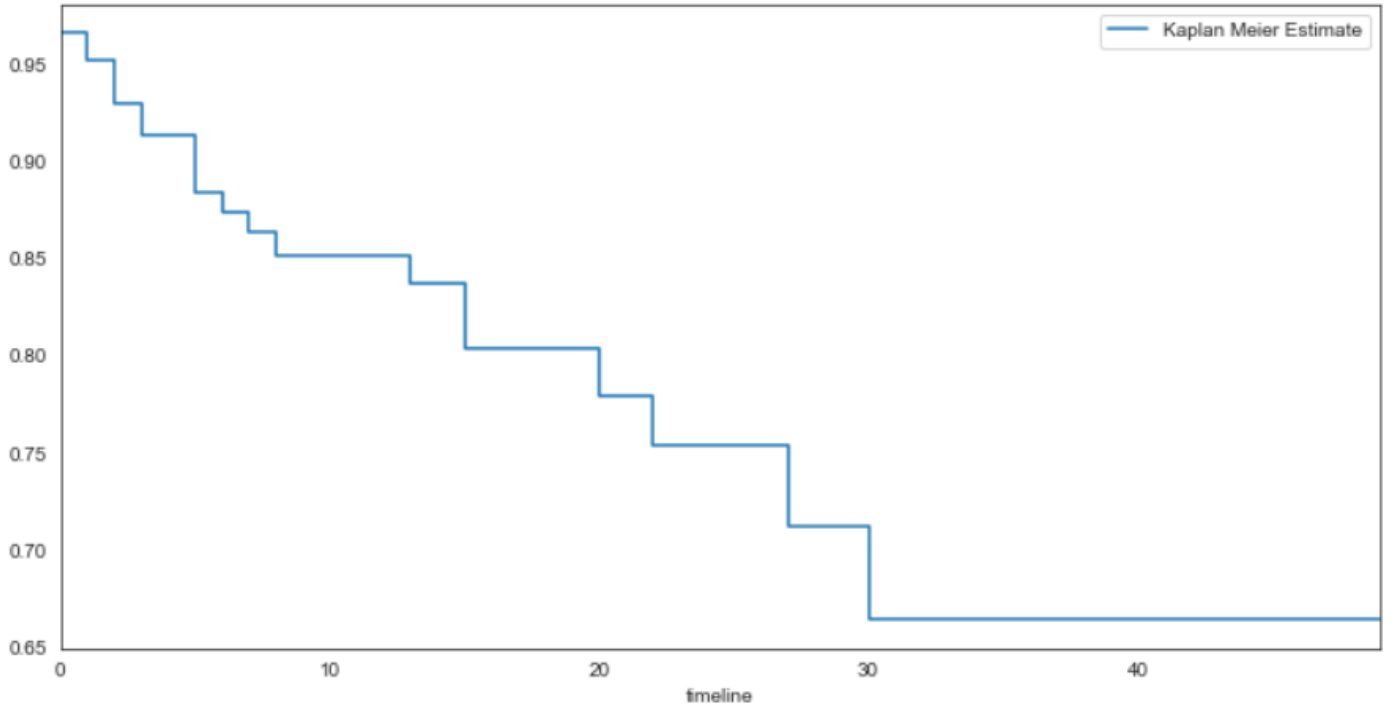
Definition of Survival

event=1 -> Assasiation, captivity, execution

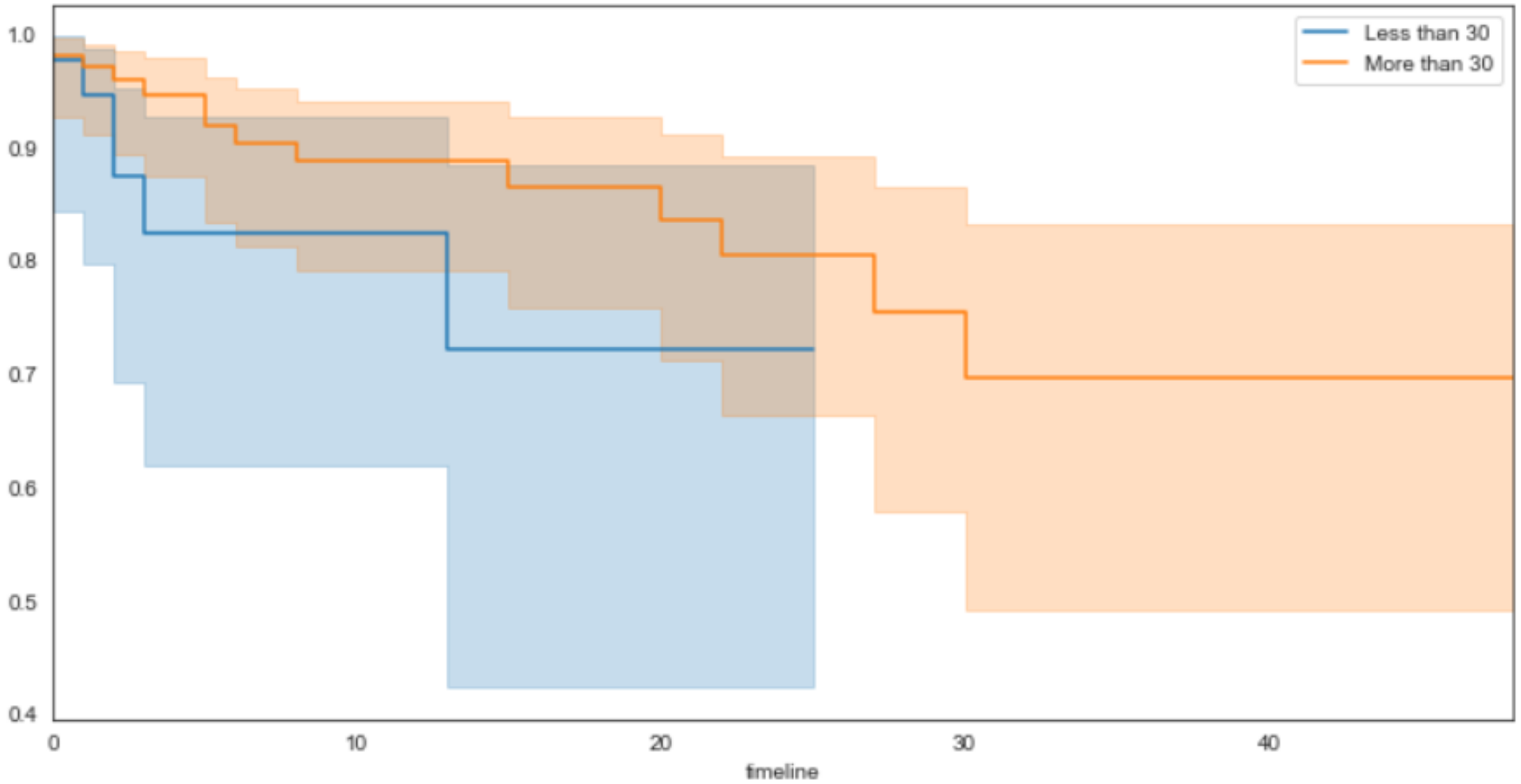
event=0 -> in battle, Suicide (more natural ways)

Kaplan-Meier

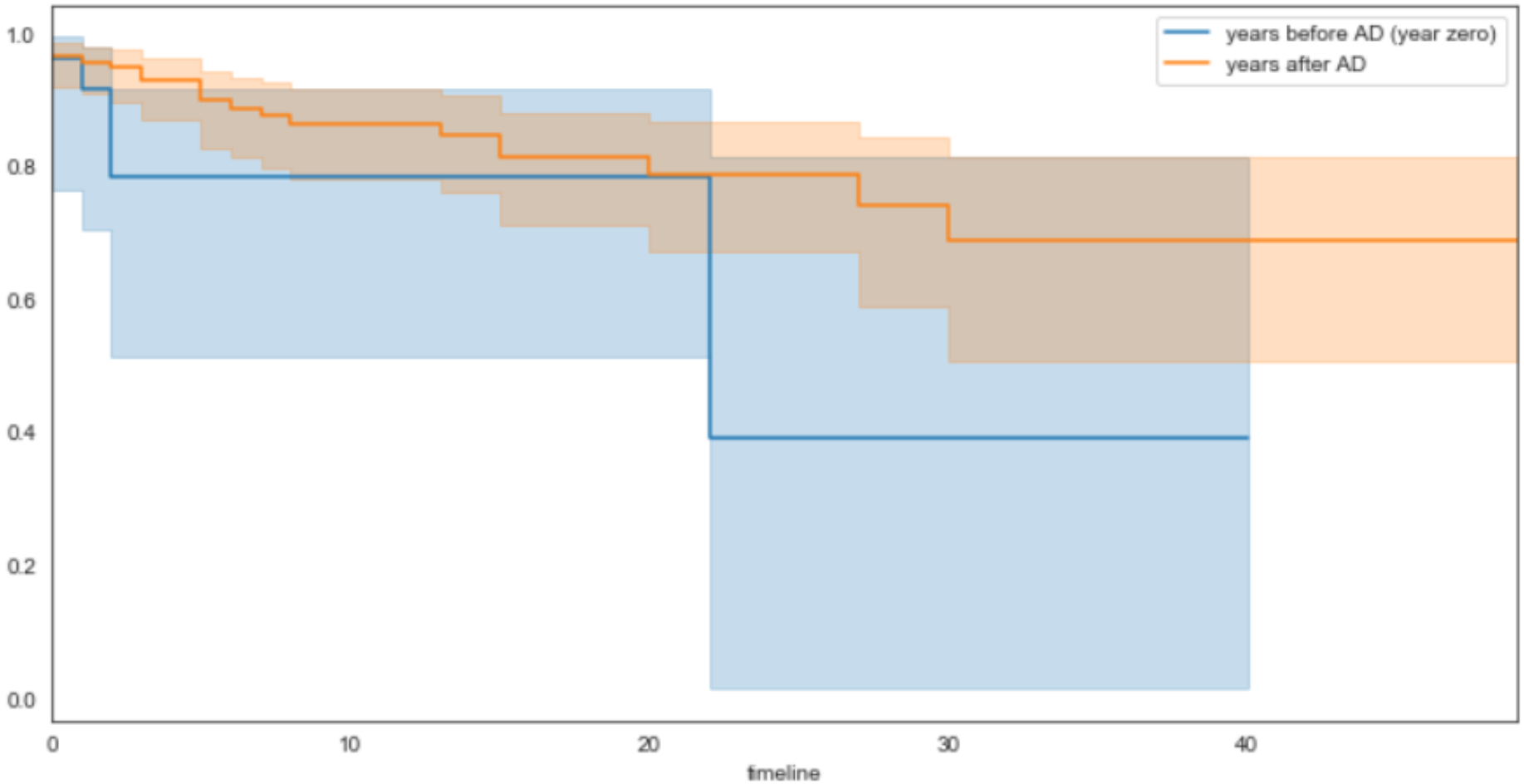
```
<matplotlib.axes._subplots.AxesSubplot at 0x25294299c48>
```



This plot shows the survival rate of emperors based on their length of reign



This plot shows the survival rate of emperors based on their length of reign and Age of Death



This plot shows the survival rate of emperors based on their length of reign and reign years (before and after AD).

Thank You