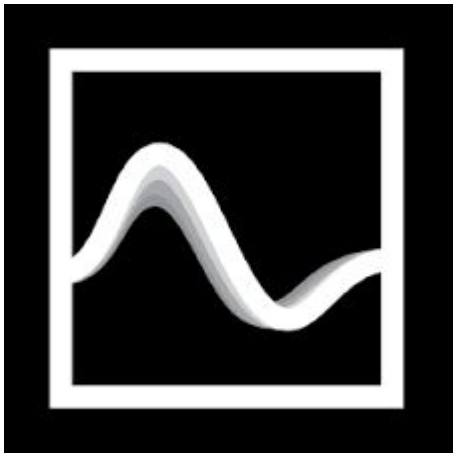


MAISI

CRA Week 5:

Survival Models + COX Proportional Hazards

Michigan Data Science Team

Fall 2025

Session 5 Agenda

01

Fun Icebreaker!!

Get to know your projectmates!

02

Intro to Survival Analysis

What is survival analysis?

03

COX Proportional Hazards

What is this survival model?

04

Interpreting COX PH Model

What do our findings tell us about our dataset?

05

Practice Time!

Work on the COX PH notebook!

Quick Icebreaker!!

What is/are your halloween costume(s)? Or, what is the best costume you've ever seen?

Share with the people around you :)

Survival Analysis

- Used for analyzing "**time-to-event**" data (in our case, recidivism)
- Traditionally used in healthcare: "What variables affect the risk of a patient in care dying, and how does this look across time?"
- Example in context:
    - Helps us answer questions like: How long does a person go without reoffending after release?
- Identify factors that influence how long individuals remain "event-free" (without reoffending) after release

Facets of Survival Analysis:

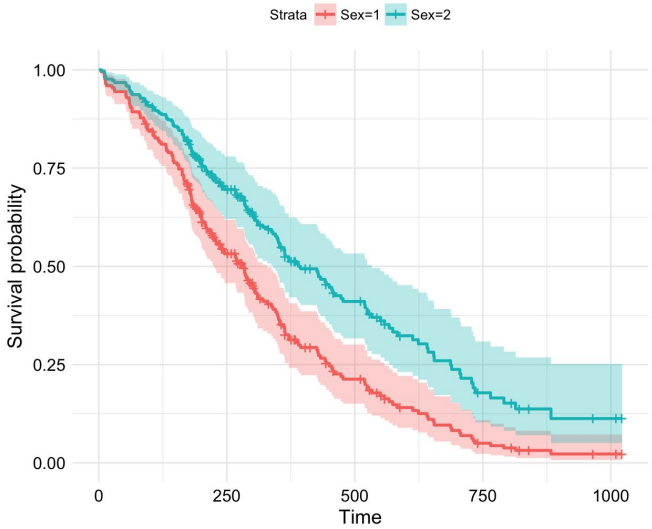
- **Survival Function S(t)**: Probability of "surviving" (no event) up to time t
- **Hazard Function h(t)**: Rate of event occurrence at time t

How does this relate to COMPAS?

- COMPAS predicts the likelihood of **recidivism**
- Survival analysis can help understand if COMPAS predictions are **biased** or **accurate** over **time**
- **Cox Proportional Hazards Model (Cox PH)**
  - Understand which factors (age, race, prior offenses) increase or decrease recidivism risk, and how that interacts with COMPAS scores
- **Kaplan-Meier Curve (next week)**
  - Understand descriptive trends by demographics (e.g., age, race)
  - How does recidivism probability change over time for different demographic groups?
  - How does COMPAS assign different “meanings” of scores to different groups?
- Today we will be going over **Cox Proportional Hazards (CPH)**

Cox Proportional Hazards Model

- **Cox PH** is a semi-parametric model that estimates the impact of covariates on the hazard rate (likelihood of reoffending)
  - **Semi-parametric**: model that has both parametric and non-parametric components
  - **Covariates**: variables that we believe might influence the “hazard”==”likelihood” of the event happening
- Helps us see which factors impact recidivism risk while accounting for **time**
- Cox PH doesn’t need to assume a **specific distribution** for time-to-event data, making it **flexible** (such as exponential, normal distributions)
  - i.e., the model is robust across different contexts where the underlying survival patterns are unknown or variable
- Example:
  - If “**age**” decreases the hazard, older individuals might have a **lower risk** of reoffending



Key Assumptions

- **Proportional Hazards Assumption**: Hazard ratios between groups are constant over time (e.g., the hazard ratio for older vs. younger individuals remains the same at different times)
- **Independence of Survival Times and Censoring**: The end of an observation period doesn't bias the survival times
  - Censoring: when we don't observe the event (like reoffending) for everyone in the dataset
  - Essentially assuming that censoring doesn't affect the survival distribution
- Example:
  - If prior offenses double the risk of reoffending, this effect stays constant over time
- If these assumptions **don't** hold, then the estimated hazard ratios **can't** be trusted as accurate representations of the relationships in the data, leading to flawed insights

Hazard Function + Formula: $h(t|X) = h_0(t) \times e^{(X \cdot \beta)}$

- Hazard Rate h(t): Rate at which recidivism occurs at time t, given covariates
- Baseline Hazard $h_0(t)$: Hazard rate when all covariates are zero
- Effect of Covariates $e^{(X \cdot \beta)}$: Exponentiated term where each β reflects a covariate's effect
- Example:
  - If prior offenses (β = 0.5) increases the hazard, $e^{(0.5)} \approx 1.65$, meaning 65% higher hazard
- Measures the **immediate risk** (or rate) of the event occurring at **time** t, given a set of **covariates** X
- For example, if the hazard ratio for individuals with prior convictions compared to those without is **2,** this means that people with prior convictions have **twice** the risk of reoffending at **any point** in time

Interpreting Coefficients

- **Hazard Ratios**: e^{β} values represent multiplicative change in hazard per unit increase in the covariate
- Example:
  - If age has a β of -0.3, a one-year increase in age reduces the hazard by approximately 26%
  - $e^{-0.3} \approx 0.74$ hazard (or risk) is 74% if what it was before the increase in age
  - To see the relative decrease: 1 - 0.74 = 0.26, or 26%
  - This means older age is associated with a lower risk of reoffending by that amount per year

Concordance

For each possible pair of individuals, concordance checks whether the model correctly predicts which individual is more likely to experience the event first, based on their predicted risk scores (hazard ratios).

A pair is considered **concordant** if the model's prediction *aligns with the actual survival times*:

- If individual A has a higher predicted risk (i.e., higher hazard ratio) than individual B, and A indeed experiences the event before B, the pair is concordant.

Concordance is another way we can evaluate the accuracy of COMPAS, and the Cox model we'll use today will calculate concordance for us.

Practice Time!

Let's check out the survival rates across groups in the context of COMPAS!

Hands-On Data Science!! :O

Next Steps:

1. Find the F25 CRA repo in the MDST GitHub and download all the relevant files in the Week 5 directory (cox_proportional_hazards_blank.ipynb AND the new dataset, cox-parsed.csv)
2. Split into teams of 2-3 (new ones or same as last week)
    a. If you're working with new people, introduce yourselves!!
3. Work on the exercises in the notebooks!
    a. Ask us if you need help!

Pandas Cheat Sheet Seaborn Cheat Sheet

Reminders

- Don’t share colab notebooks with teammates if you are working at the same time

- Where to put csv and data files
  - Google Drive
    - Need to include:

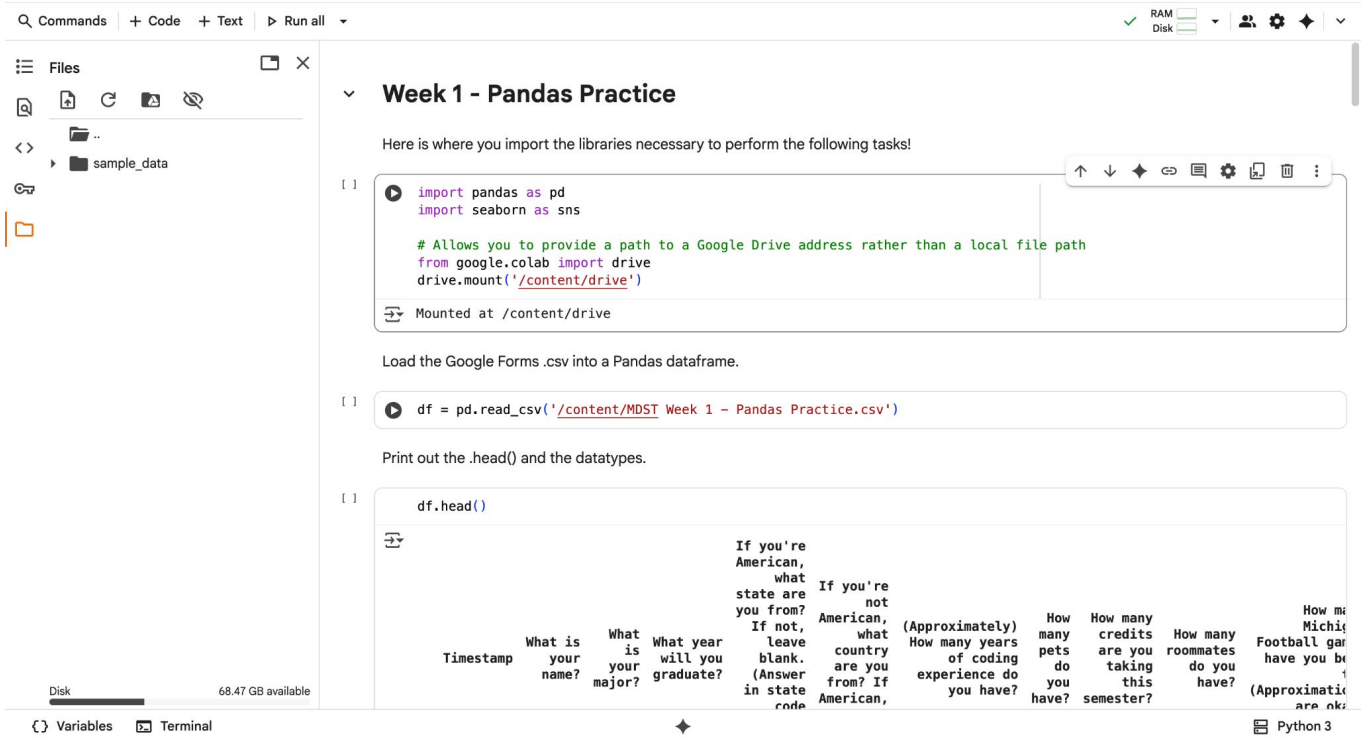
```
from google.colab import drive
drive.mount('/content/drive')
```

```
pd.read_csv('/content/drive/MyDrive/[FILE NAME]')
```

  - Colab Files
    - See next slides

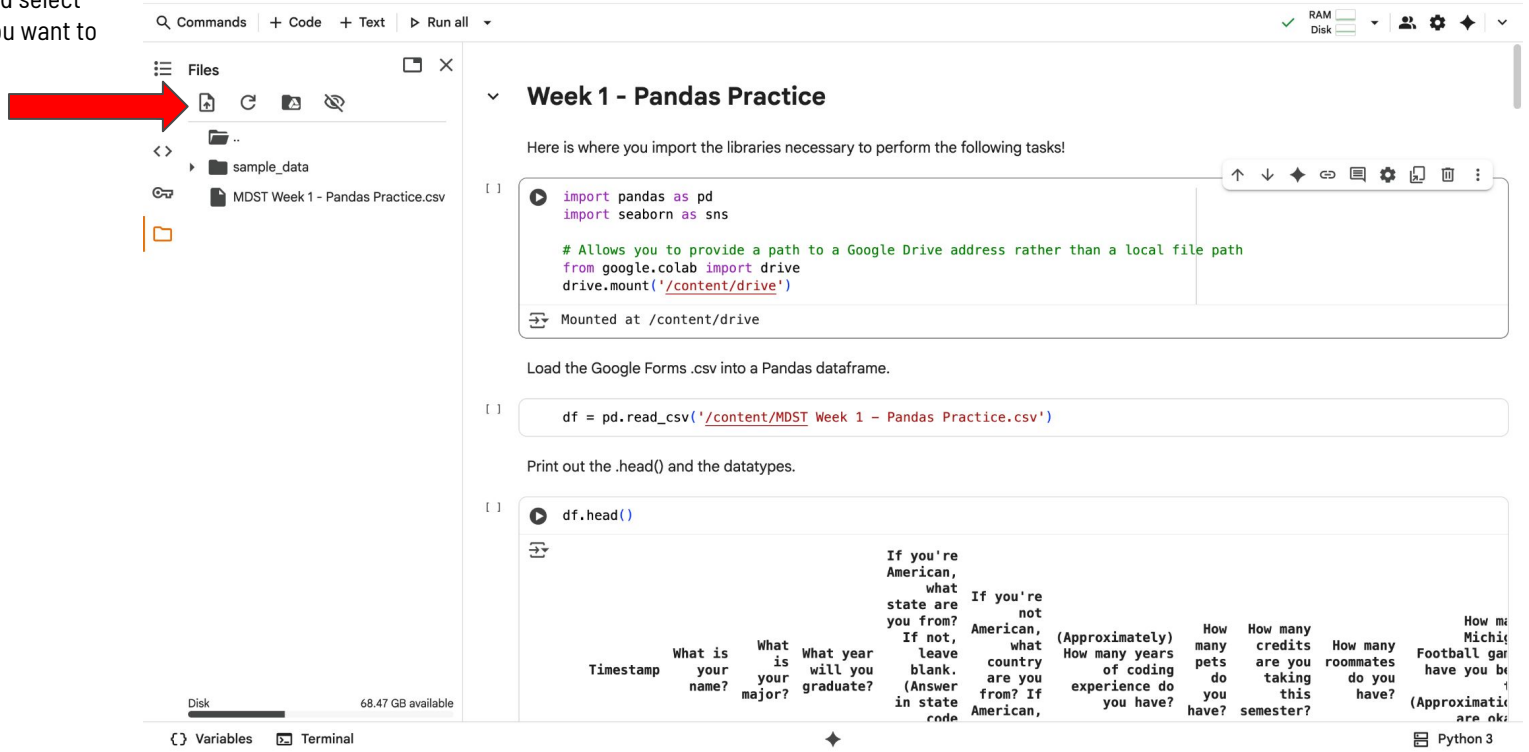
Reminders

Click on the folder in the sidebar



Reminders

Click the upload button and select the file you want to upload



Reminders

Click the three dots and copy the path. Put this in your read function