UCSF
University of California
San Francisco

MoonLAIT

MoonLAIT – Leng Lab

Interview for Research Data Analyst
Alice Albrecht
February 4th, 2025

Dataset Presentation

Dataset: Sleep Heart Health Study – Visit 1 (SHHS1)
Number of Subjects: Subsample of 50

For Each Subject:

- **EDF File:** Contains full **Polysomnography (PSG)** recordings, from which we will extract key biosignals:
  - **Electrocardiogram (ECG)** – Heart activity
  - **Electroencephalogram (EEG)** – Brain activity
  - **Electrooculogram (EOG):** Eye movement
  - **Electromyogram (EMG)** – Muscle activity

- **XML File:** Includes annotated sleep data, providing:
  - **Sleep stages** (e.g., Stage1, REM)
  - **Respiratory events** (e.g., apneas, hypopneas)
  - **Arousals** and other relevant physiological events

- **Subject Information:** Includes demographic and physiological data such as **age, gender, BMI**, and other relevant characteristics.

- **Outcome:** Contains follow-up data, including **vital status (alive or dead)** and **time since the most recent contact**, which can be used for **survival analysis**.

Assignment Task 1

HRV Metrics Calculation and Survival Analysis for Mortality Prediction

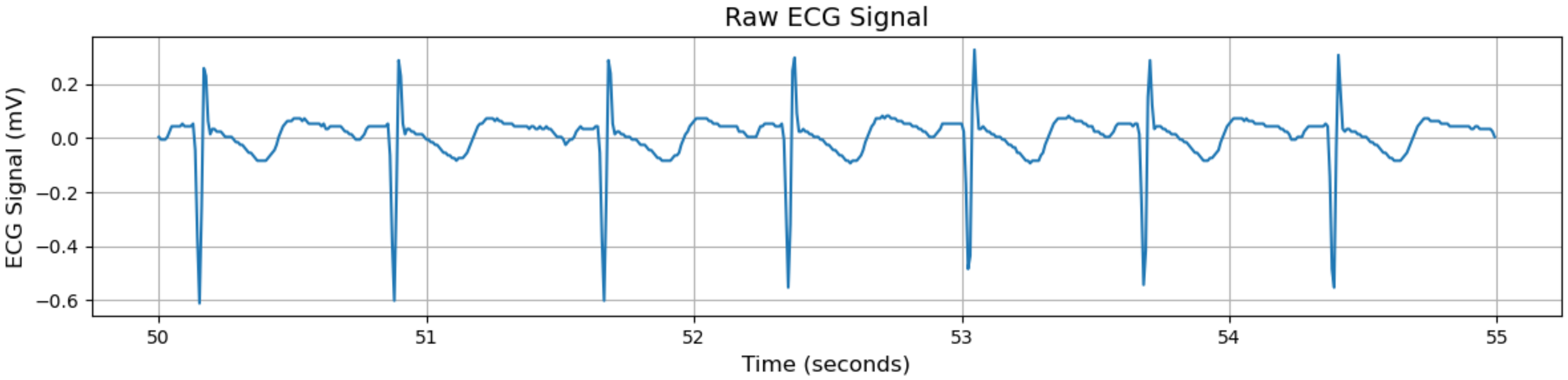
University of California
San Francisco

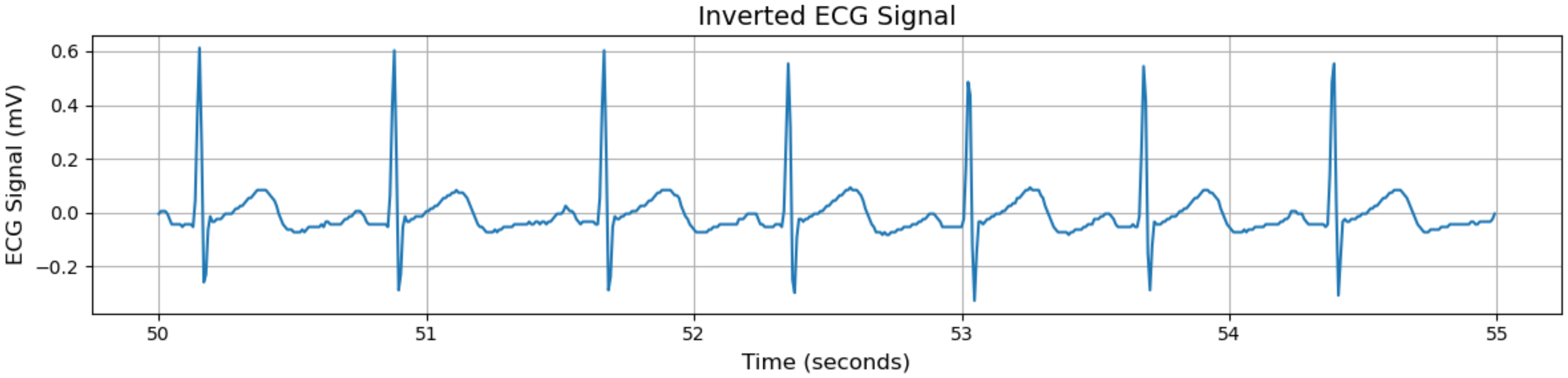
Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

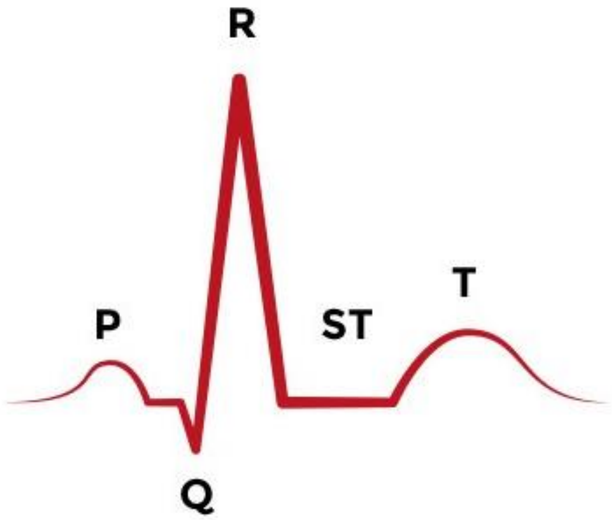
Task 1

Extraction of ECG signal

PSG signals inside EDF file





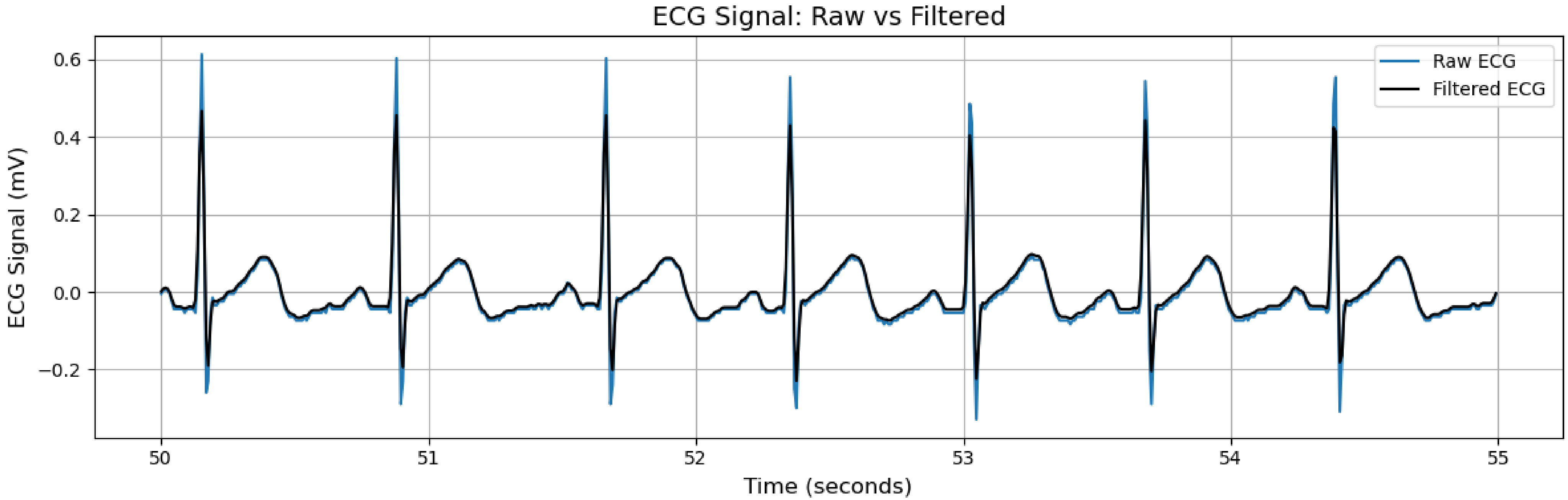


Conventional ECG waveform orientation (where the P-wave has an upward deflection)

Task 1

Filtering of ECG signal

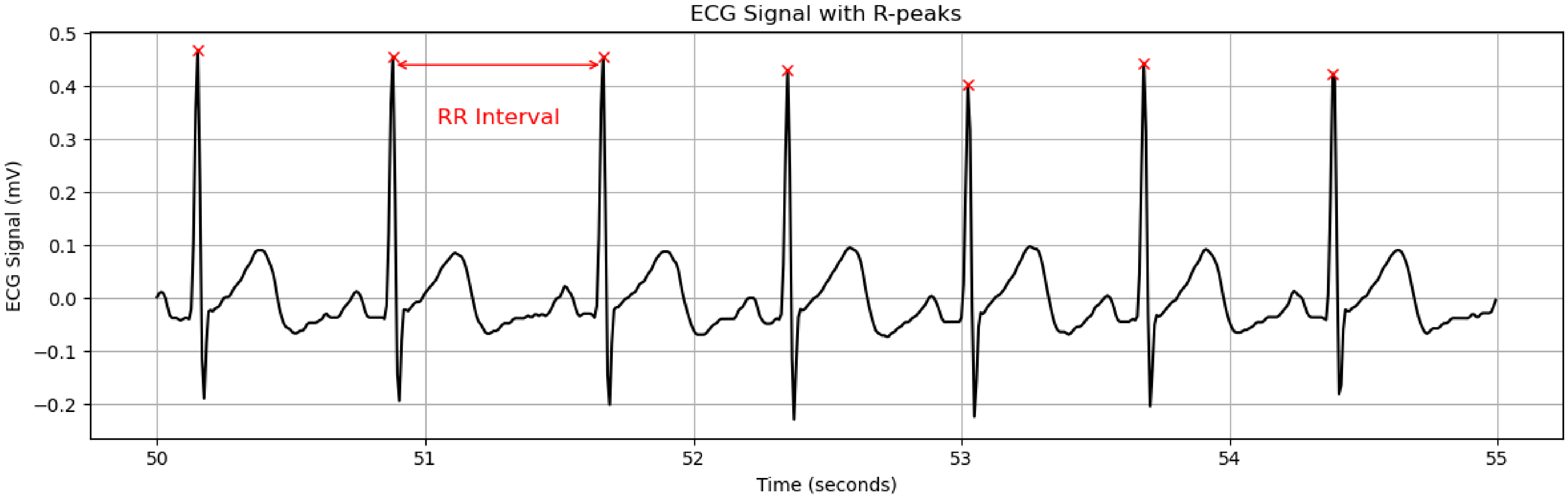
- **Cleaning applied:** NeuroKit *ecg_clean* function includes:
  - Muscle artifact noise removal
  - Baseline drift correction
  - General signal enhancement for analysis



Task 1

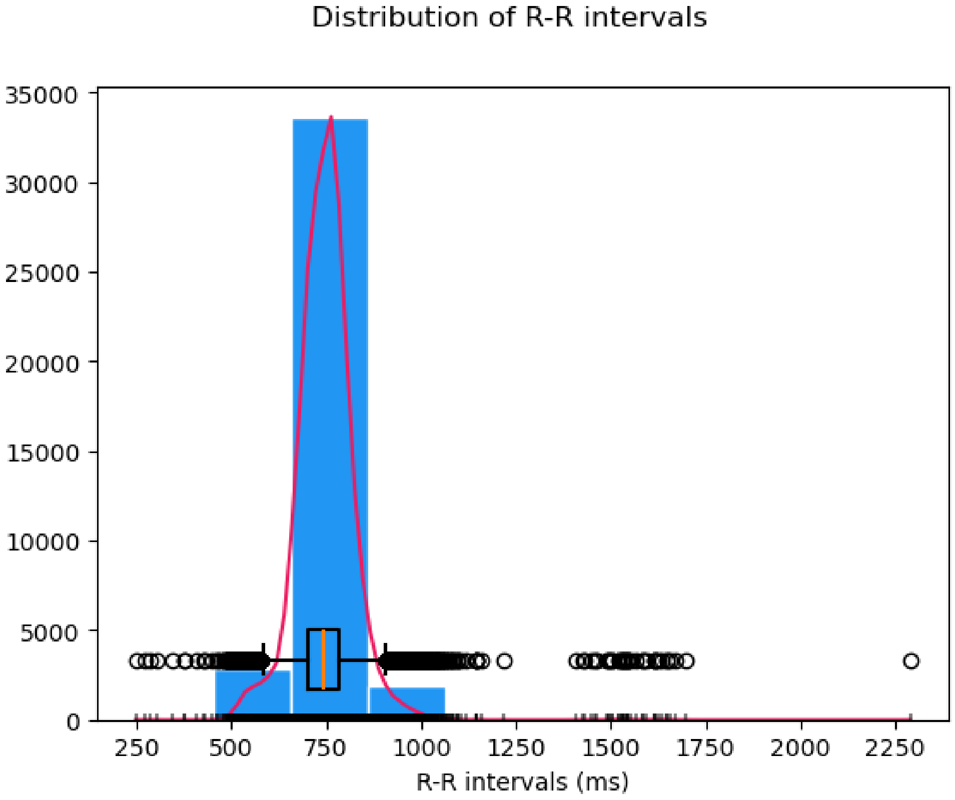
R-Peak Detection and RR-intervals

- **R-peak Detection :** NeuroKit *ecg_findpeak* function based on Probabilistic Methods-Agreement via Convolution (ProMAC)
- **RR-intervals:** Time differences between successive R-peaks.
- **NN-intervals:** after abnormal RR intervals (> 2.5s) removed



Task 1

Temporal Analysis for HRV Metrics



AVNN (Average NN Interval)
Mean time in ms between successive **normal** R-peaks (NN intervals)

SDNN(Standard Deviation of NN Intervals)
Measures overall HRV by quantifying **beat-to-beat variability** over a period.

RMSSD (Root Mean Square of Successive Differences)
Reflects short-term HRV by measuring rapid **fluctuations in heart rate**.

pNN50 (Percentage of NN50)
Percentage of consecutive NN intervals that differ by **more than 50 ms**, indicating **parasympathetic activity**

UCSF University of California San Francisco

MoonLAIT

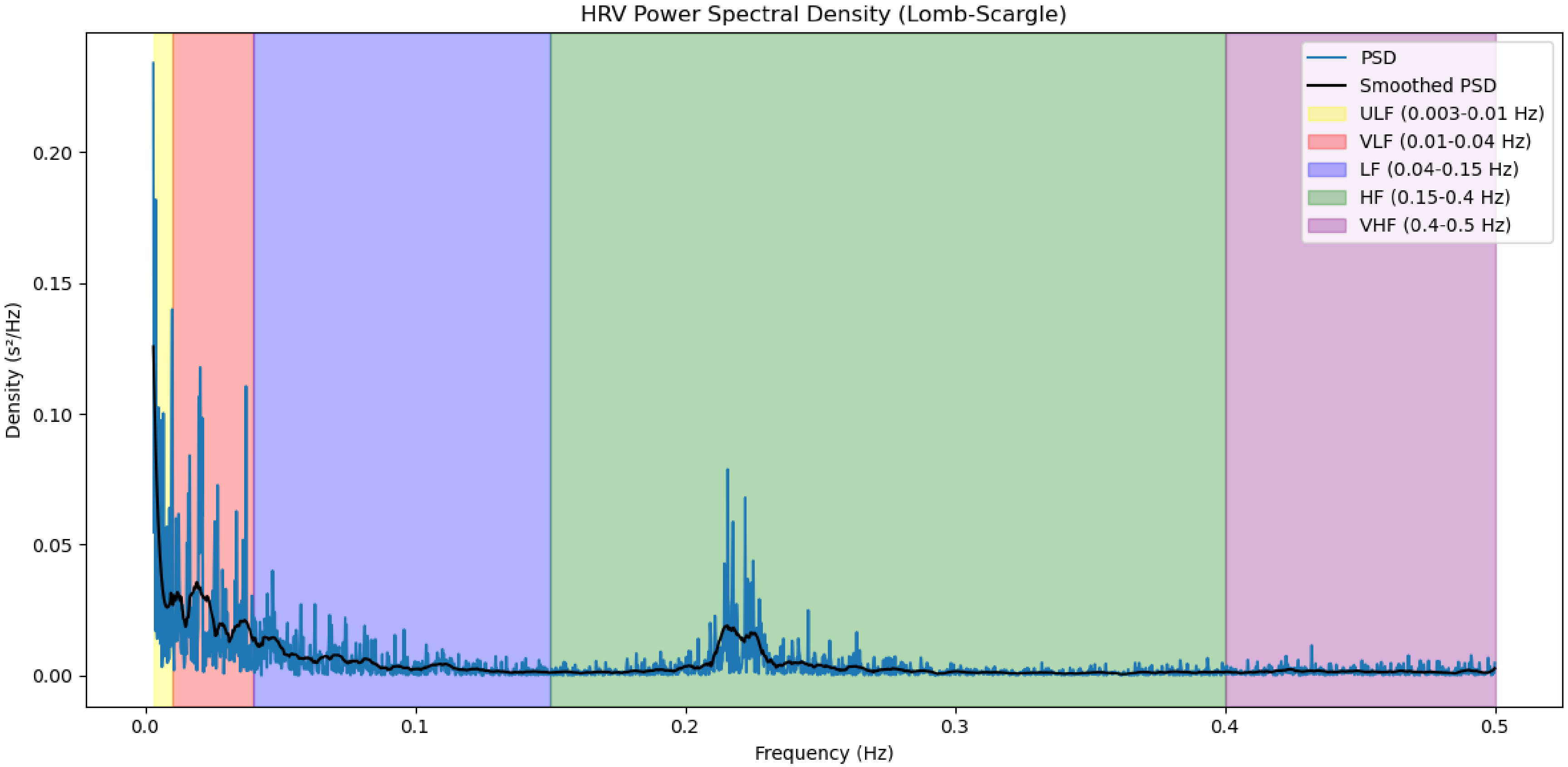
Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 1

Frequency Analysis for HRV Metrics

- **NN intervals** are transformed to the **frequency domain** using the **Lomb-Scargle Periodogram**, ideal for unevenly spaced data like RR intervals.

- The resulting **Power Spectral Density (PSD)** reveals how signal power is distributed, allowing calculation of **HRV metrics** to assess autonomic heart rate regulation.



University of California San Francisco

MoonLAIT

Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 1

Frequency Analysis for HRV Metrics

ULF (Ultra-Low Frequency, <0.003 Hz)
Very slow HRV changes, linked to long-term **circadian and metabolic regulation**.

VLF (Very-Low Frequency, 0.003–0.04 Hz)
Associated with **thermoregulation, hormonal activity, and vagal tone**.

LF (Low Frequency, 0.04–0.15 Hz)
Reflects a mix of **sympathetic and parasympathetic activity**; often linked to blood pressure regulation.

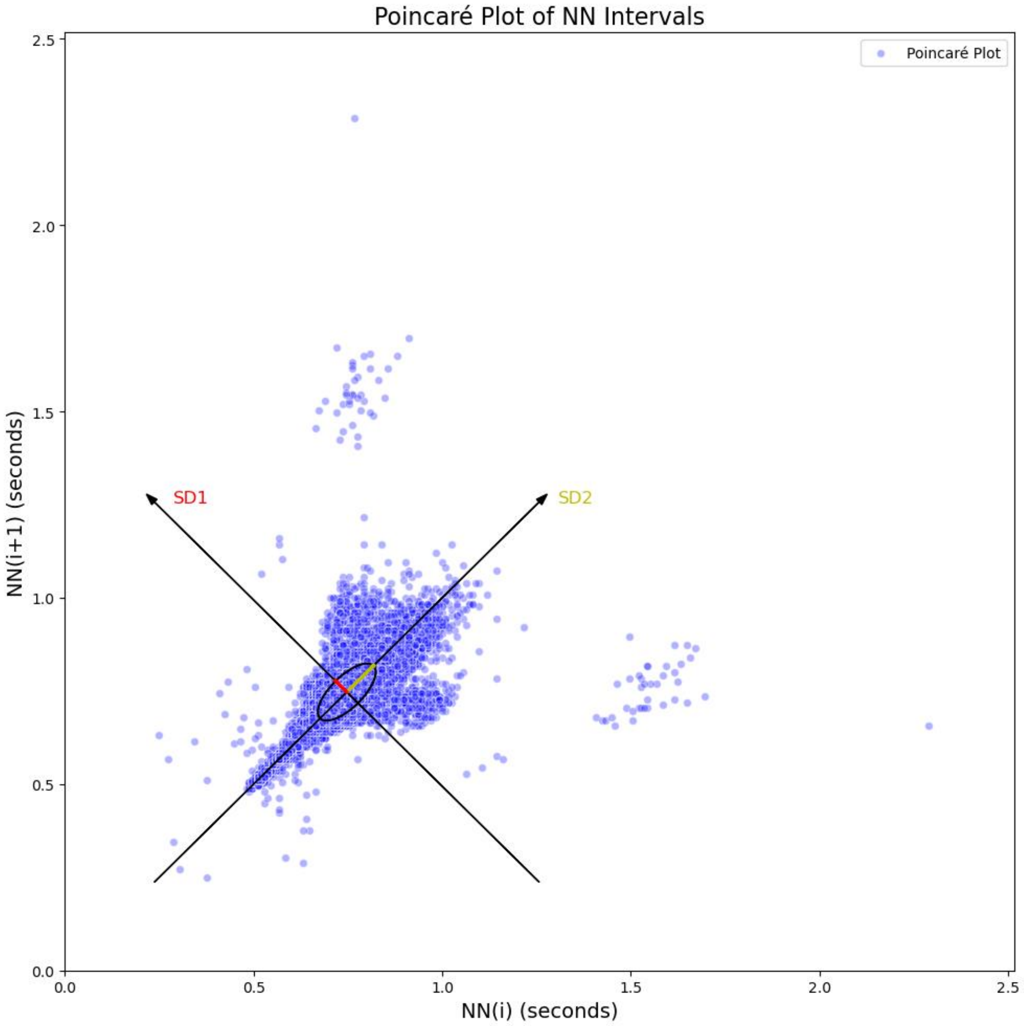
HF (High Frequency, 0.15–0.4 Hz)
Represents **parasympathetic (vagal) activity**, closely tied to breathing rate.

VHF (Very-High Frequency, >0.4 Hz)
Less commonly analyzed, potentially related to **mechanical cardiac processes**.

LF/HF Ratio
Balance between sympathetic and parasympathetic activity, used as indicator of autonomic nervous system modulation.

Task 1

Poincaré Plot and Non-Linear HRV Metrics



- ➢ The **Poincaré plot** visualizes the relationship between successive **NN intervals**, showing nonlinear HRV patterns not captured by linear methods.
- ➢ The ellipse illustrate the direction and strength of these HRV components.

SD1 (Short-Term HRV)
Reflects **short-term variability** and **parasympathetic activity**.

SD2 (Long-Term HRV)
Represents **both sympathetic and parasympathetic influences** over longer time scales.

Task 1

All metrics used for Survival Analysis

HRV METRCIS

ULF (Ultra-Low Frequency, <0.003 Hz)
Very slow HRV changes, linked to long-term **circadian and metabolic regulation**.

HF (High Frequency, 0.15–0.4 Hz)
Represents **parasympathetic (vagal) activity**, closely tied to breathing rate.

AVNN (Average NN Interval)
Mean time in ms between successive **normal** R-peaks (NN intervals)

SDNN(Standard Deviation of NN Intervals)
Measures overall HRV by quantifying **beat-to-beat variability** over a period.

VLF (Very-Low Frequency, 0.003–0.04 Hz)
Associated with **thermoregulation, hormonal activity, and vagal tone**.

VHF (Very-High Frequency, >0.4 Hz)
Less commonly analyzed, potentially related to **mechanical cardiac processes**.

RMSSD (Root Mean Square of Successive Differences)
Reflects short-term HRV by measuring rapid **fluctuations in heart rate**.

pNN50 (Percentage of NN50)
Percentage of consecutive NN intervals that differ by **more than 50 ms**, indicating **parasympathetic activity**

LF (Low Frequency, 0.04–0.15 Hz)
Reflects a mix of **sympathetic and parasympathetic activity**; often linked to blood pressure regulation.

LF/HF Ratio
Balance between sympathetic and parasympathetic activity, used as indicator of autonomic nervous system modulation.

SD1 (Short-Term HRV)
Reflects **short-term variability** and **parasympathetic activity**.

SD2 (Long-Term HRV)
Represents **both sympathetic and parasympathetic influences** over longer time scales.

COVARIATES

AGE

GENDER

BODY MASS INDEX (BMI)

CHOLESTEROL

HYPERTENSION

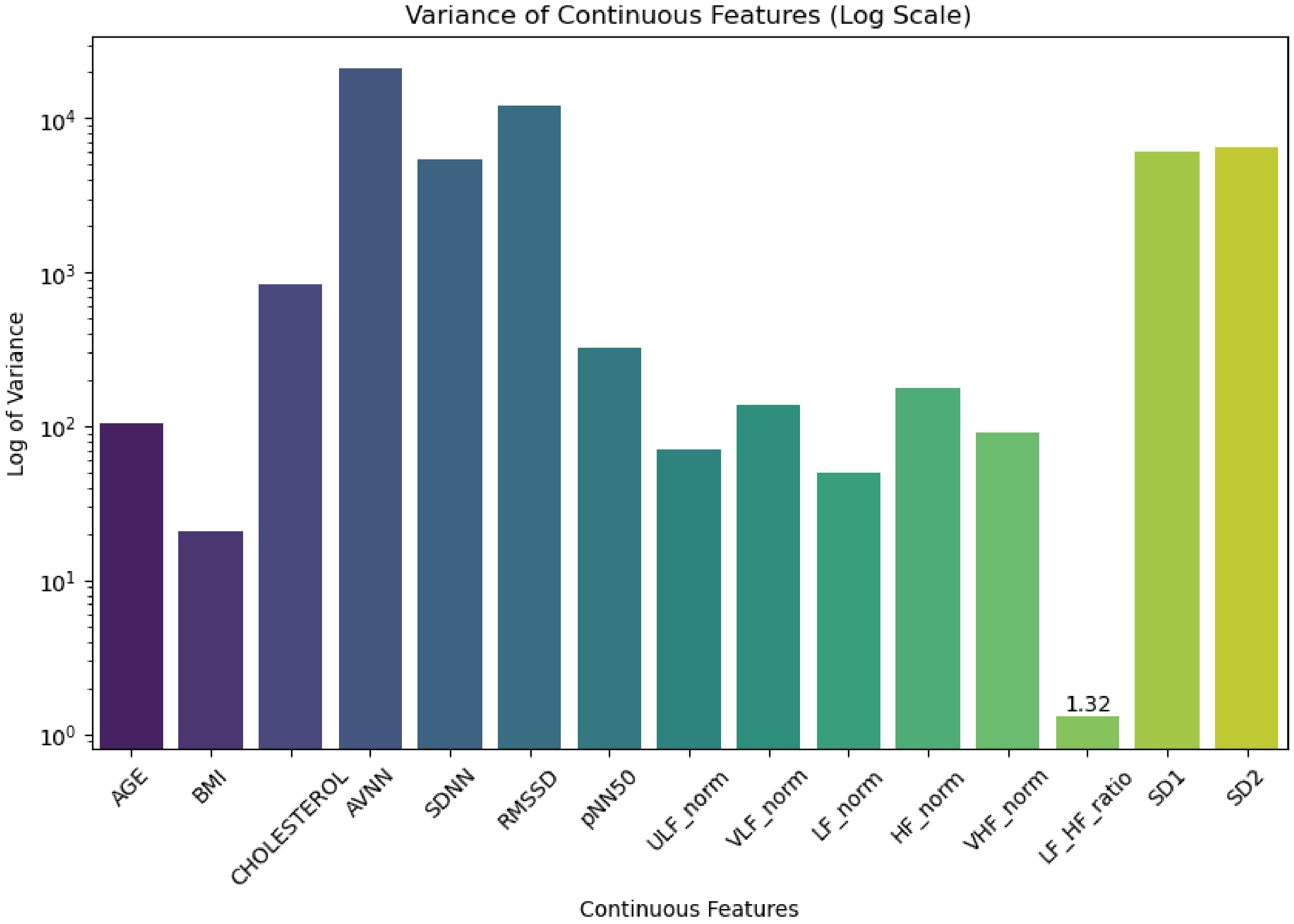
UCSF
University of California
San Francisco

MoonLAIT

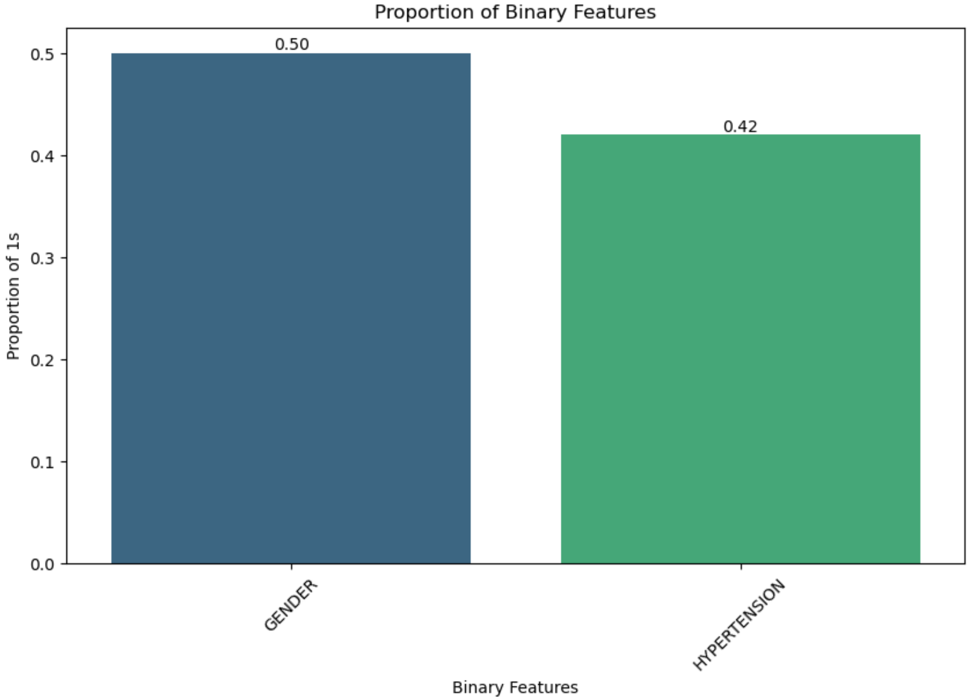
Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 1

Data Preprocessing: Variance and Balance



- High variance for the continuous variable
- Good balance for binary features



University of California
San Francisco

MoonLAIT

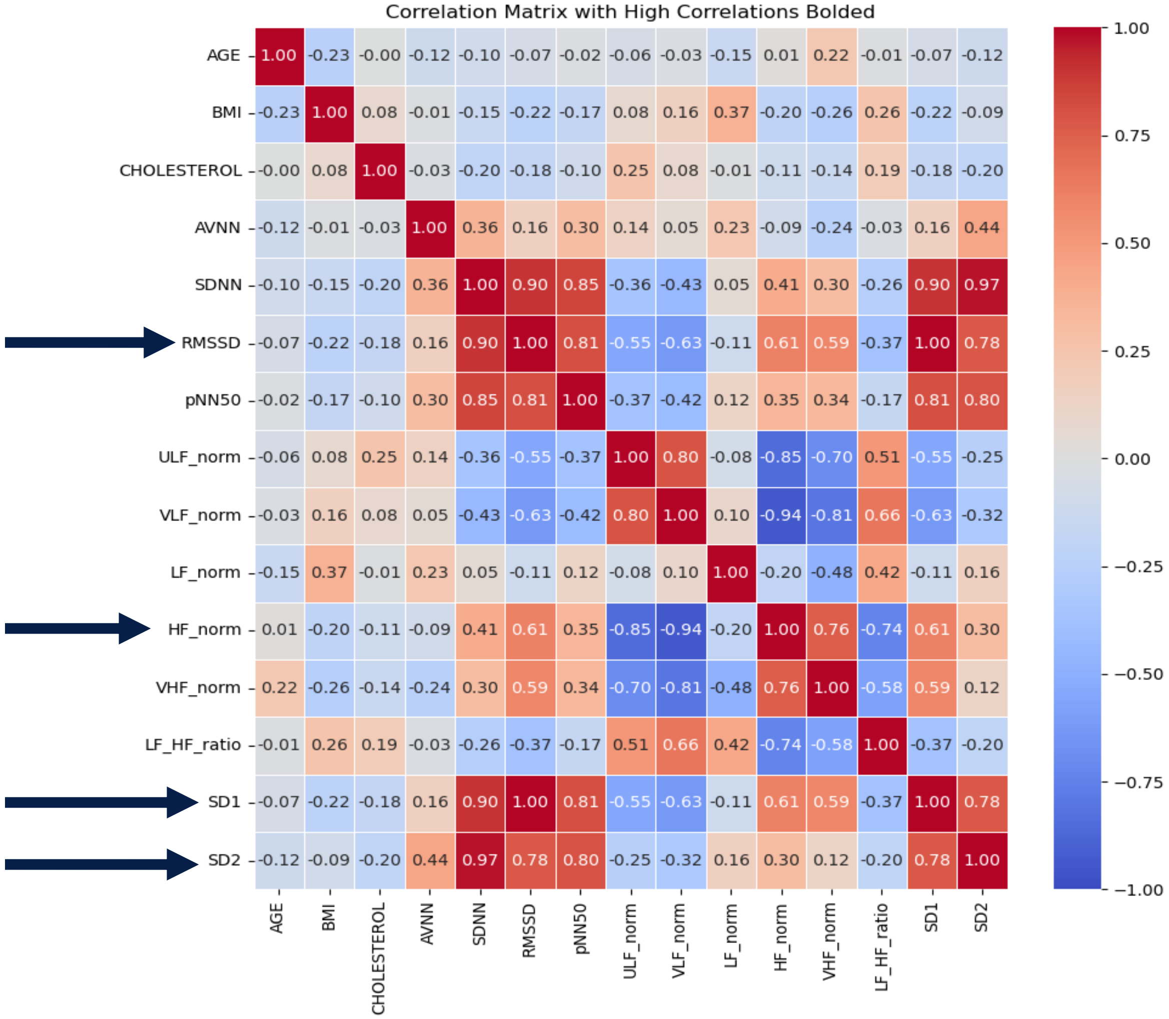
Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 1

Data Preprocessing: Correlation

Features to drop to avoid correlation *(threshold=0.9)*

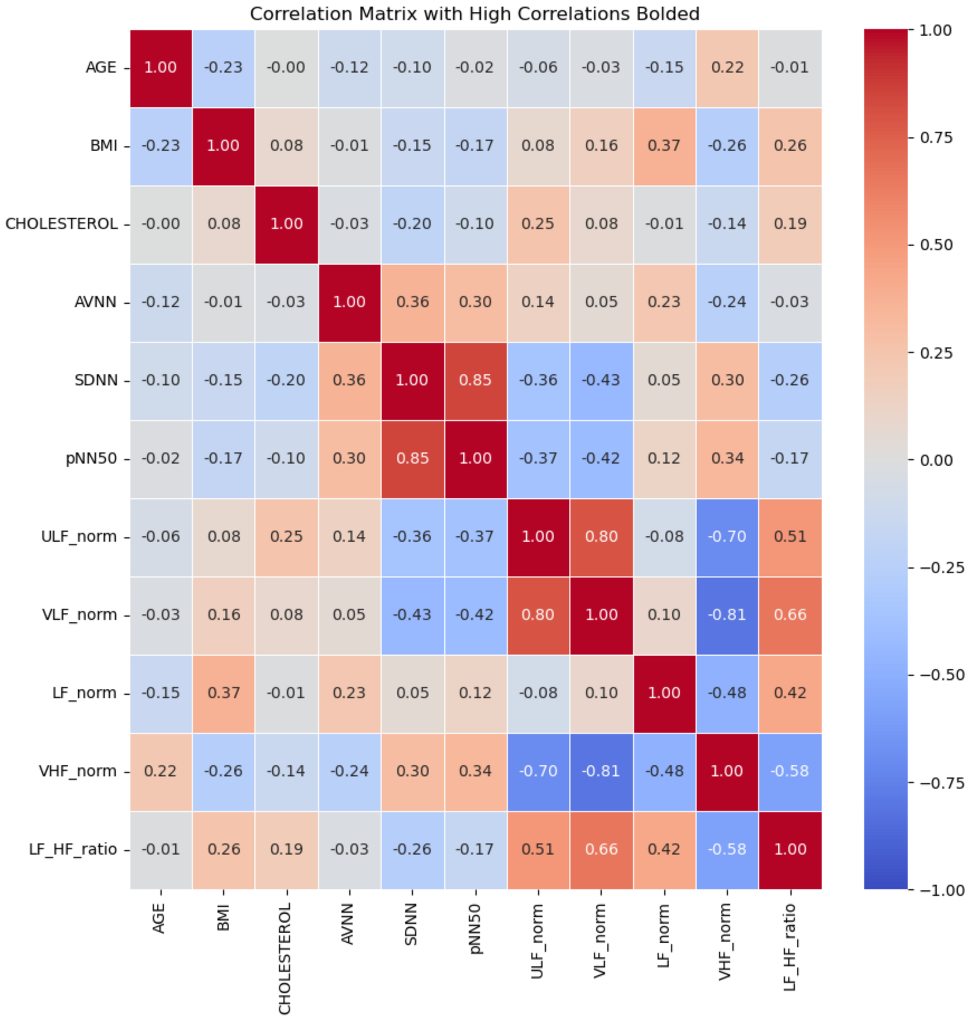
Note: Cox Proportional Hazards is highly sensitive to correlations between features



Correlation Matrix with High Correlations Bolded

	AGE	BMI	CHOLESTEROL	AVNN	SDNN	RMSSD	pNN50	ULF_norm	VLF_norm	LF_norm	HF_norm	VHF_norm	LF_HF_ratio	SD1	SD2
AGE	1.00	-0.23	-0.00	-0.12	-0.10	-0.07	-0.02	-0.06	-0.03	-0.15	0.01	0.22	-0.01	-0.07	-0.12
BMI	-0.23	1.00	0.08	-0.01	-0.15	-0.22	-0.17	0.08	0.16	0.37	-0.20	-0.26	0.26	-0.22	-0.09
CHOLESTEROL	-0.00	0.08	1.00	-0.03	-0.20	-0.18	-0.10	0.25	0.08	-0.01	-0.11	-0.14	0.19	-0.18	-0.20
AVNN	-0.12	-0.01	-0.03	1.00	0.36	0.16	0.30	0.14	0.05	0.23	-0.09	-0.24	-0.03	0.16	0.44
SDNN	-0.10	-0.15	-0.20	0.36	1.00	0.90	0.85	-0.36	-0.43	0.05	0.41	0.30	-0.26	0.90	0.97
RMSSD	-0.07	-0.22	-0.18	0.16	0.90	1.00	0.81	-0.55	-0.63	-0.11	0.61	0.59	-0.37	1.00	0.78
pNN50	-0.02	-0.17	-0.10	0.30	0.85	0.81	1.00	-0.37	-0.42	0.12	0.35	0.34	-0.17	0.81	0.80
ULF_norm	-0.06	0.08	0.25	0.14	-0.36	-0.55	-0.37	1.00	0.80	-0.08	-0.85	-0.70	0.51	-0.55	-0.25
VLF_norm	-0.03	0.16	0.08	0.05	-0.43	-0.63	-0.42	0.80	1.00	0.10	-0.94	-0.81	0.66	-0.63	-0.32
LF_norm	-0.15	0.37	-0.01	0.23	0.05	-0.11	0.12	-0.08	0.10	1.00	-0.20	-0.48	0.42	-0.11	0.16
HF_norm	0.01	-0.20	-0.11	-0.09	0.41	0.61	0.35	-0.85	-0.94	-0.20	1.00	0.76	-0.74	0.61	0.30
VHF_norm	0.22	-0.26	-0.14	-0.24	0.30	0.59	0.34	-0.70	-0.81	-0.48	0.76	1.00	-0.58	0.59	0.12
LF_HF_ratio	-0.01	0.26	0.19	-0.03	-0.26	-0.37	-0.17	0.51	0.66	0.42	-0.74	-0.58	1.00	-0.37	-0.20
SD1	-0.07	-0.22	-0.18	0.16	0.90	1.00	0.81	-0.55	-0.63	-0.11	0.61	0.59	-0.37	1.00	0.78
SD2	-0.12	-0.09	-0.20	0.44	0.97	0.78	0.80	-0.25	-0.32	0.16	0.30	0.12	-0.20	0.78	1.00

Data Preprocessing: Correlation



Correlation Matrix with High Correlations Bolded

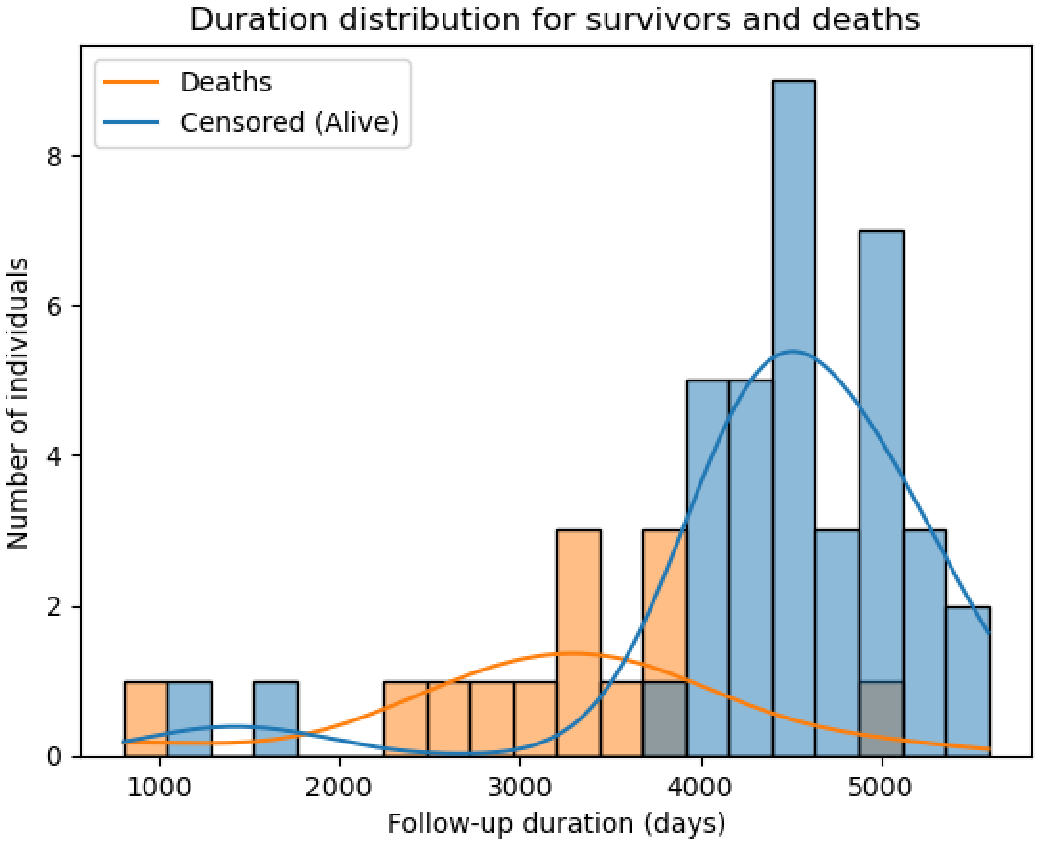
	AGE	BMI	CHOLESTEROL	AVNN	SDNN	pNN50	ULF_norm	VLF_norm	LF_norm	VHF_norm	LF_HF_ratio
AGE	1.00	-0.23	-0.00	-0.12	-0.10	-0.02	-0.06	-0.03	-0.15	0.22	-0.01
BMI	-0.23	1.00	0.08	-0.01	-0.15	-0.17	0.08	0.16	0.37	-0.26	0.26
CHOLESTEROL	-0.00	0.08	1.00	-0.03	-0.20	-0.10	0.25	0.08	-0.01	-0.14	0.19
AVNN	-0.12	-0.01	-0.03	1.00	0.36	0.30	0.14	0.05	0.23	-0.24	-0.03
SDNN	-0.10	-0.15	-0.20	0.36	1.00	0.85	-0.36	-0.43	0.05	0.30	-0.26
pNN50	-0.02	-0.17	-0.10	0.30	0.85	1.00	-0.37	-0.42	0.12	0.34	-0.17
ULF_norm	-0.06	0.08	0.25	0.14	-0.36	-0.37	1.00	0.80	-0.08	-0.70	0.51
VLF_norm	-0.03	0.16	0.08	0.05	-0.43	-0.42	0.80	1.00	0.10	-0.81	0.66
LF_norm	-0.15	0.37	-0.01	0.23	0.05	0.12	-0.08	0.10	1.00	-0.48	0.42
VHF_norm	0.22	-0.26	-0.14	-0.24	0.30	0.34	-0.70	-0.81	-0.48	1.00	-0.58
LF_HF_ratio	-0.01	0.26	0.19	-0.03	-0.26	-0.17	0.51	0.66	0.42	-0.58	1.00

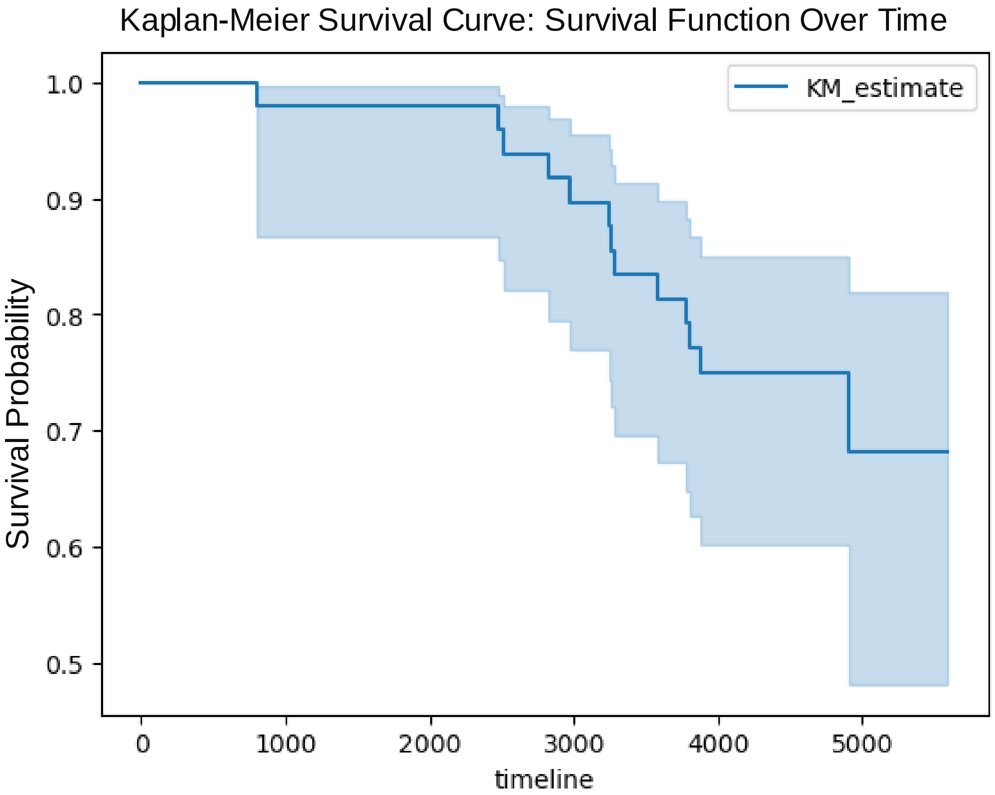
Less Correlated Dataset

University of California
San Francisco

Task 1

Survival Analysis Outcomes





University of California
San Francisco

MoonLAIT

Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Multivariate Survival Analysis – Statistical Metrics

- **C-index (Concordance Index)**: Measures the model's discriminatory power. A **higher** C-index indicates **better** ability to differentiate between individuals at different risk levels.

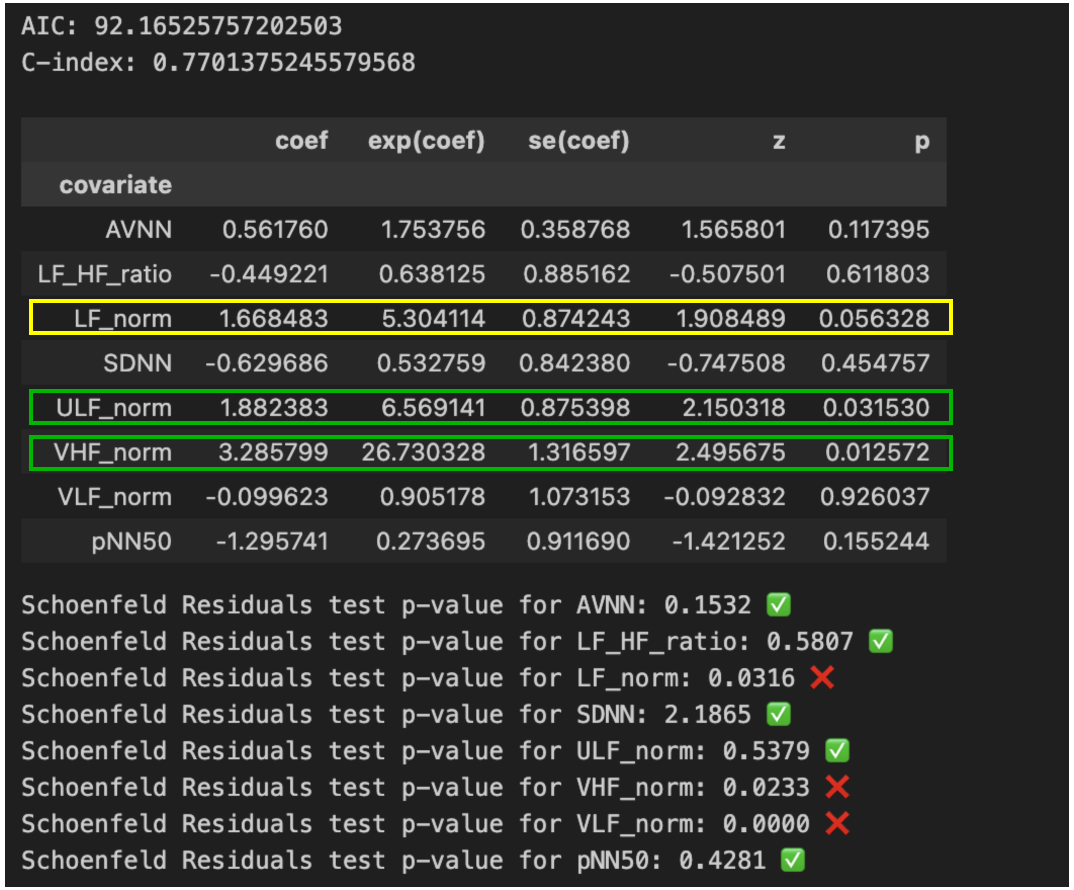
- **AIC (Akaike Information Criterion)**: A measure of the model's goodness of fit, balancing model fit and complexity. **Lower** AIC values indicate a **better**-fitting model.

- The **Cox Proportional Hazards model** estimates the relationship between features and the risk of an event.
  - **p-value > 0.05** → Feature is **not significant**, suggesting it doesn't impact survival.
  - **p-value <= 0.05** → Feature is **significant**, indicating it affects survival."

- The **Schoenfeld Residuals test** checks the proportional hazards assumption for each feature.
  - **p-value > 0.05** → No violation of the proportional hazards assumption, should be **constant over time**.
  - **p-value <= 0.05** → There is a potential violation of the proportional hazards assumption, meaning its effect on hazard **may not be constant over time.**

Task 1

Multivariate Survival Analysis – HRV metrics

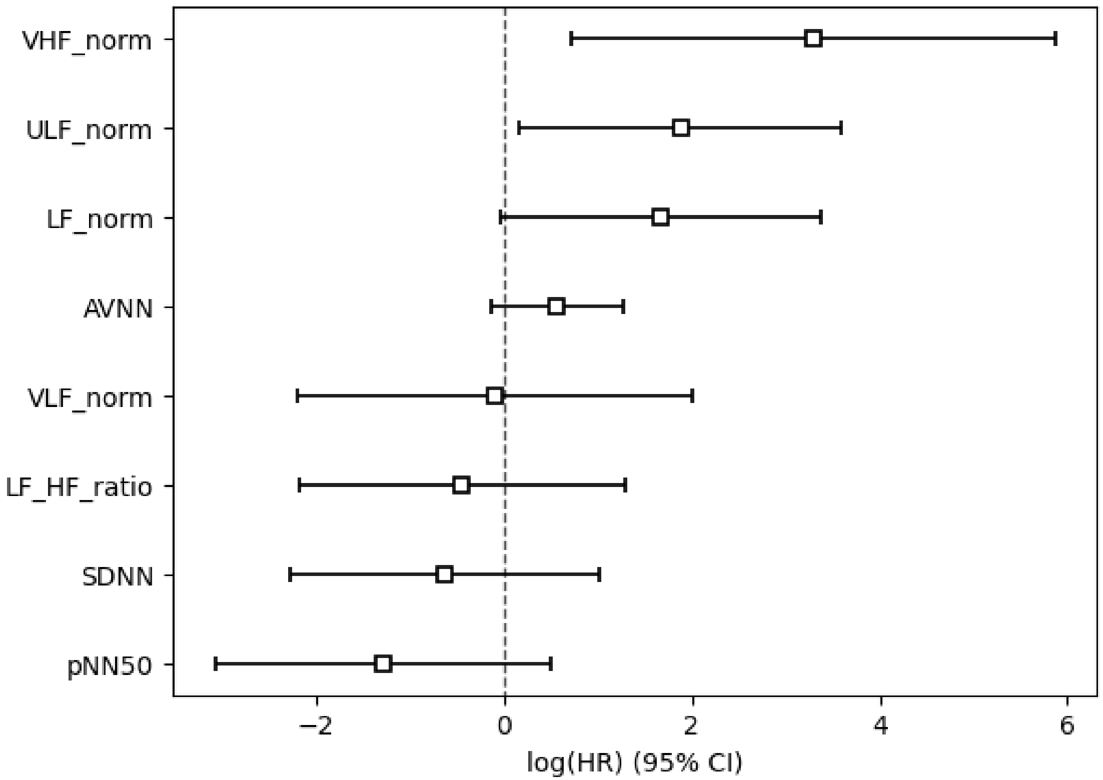


```
AIC: 92.16525757202503
C-index: 0.7701375245579568
```

covariate	coef	exp(coef)	se(coef)	z	p
AVNN	0.561760	1.753756	0.358768	1.565801	0.117395
LF_HF_ratio	-0.449221	0.638125	0.885162	-0.507501	0.611803
LF_norm	1.668483	5.304114	0.874243	1.908489	0.056328
SDNN	-0.629686	0.532759	0.842380	-0.747508	0.454757
ULF_norm	1.882383	6.569141	0.875398	2.150318	0.031530
VHF_norm	3.285799	26.730328	1.316597	2.495675	0.012572
VLF_norm	-0.099623	0.905178	1.073153	-0.092832	0.926037
pNN50	-1.295741	0.273695	0.911690	-1.421252	0.155244

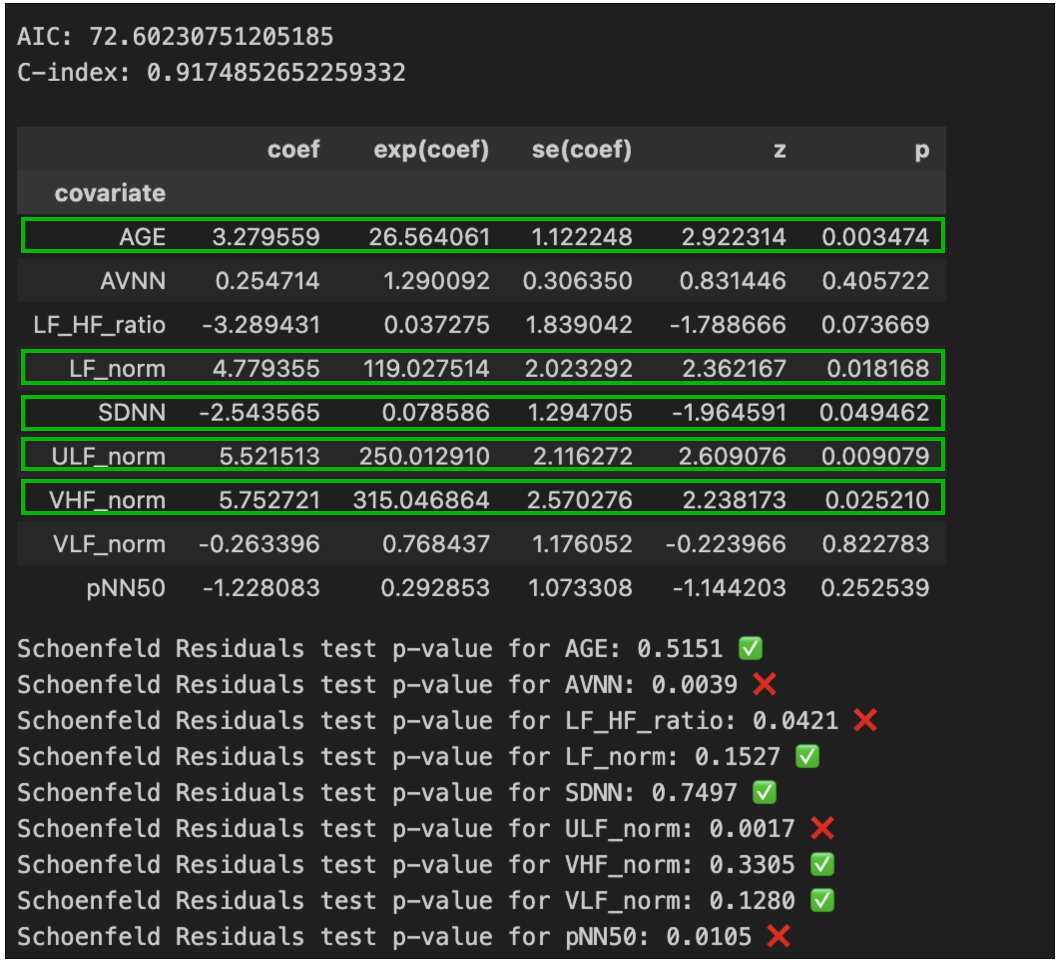
```
Schoenfeld Residuals test p-value for AVNN: 0.1532 ✅
Schoenfeld Residuals test p-value for LF_HF_ratio: 0.5807 ✅
Schoenfeld Residuals test p-value for LF_norm: 0.0316 ❌
Schoenfeld Residuals test p-value for SDNN: 2.1865 ✅
Schoenfeld Residuals test p-value for ULF_norm: 0.5379 ✅
Schoenfeld Residuals test p-value for VHF_norm: 0.0233 ❌
Schoenfeld Residuals test p-value for VLF_norm: 0.0000 ❌
Schoenfeld Residuals test p-value for pNN50: 0.4281 ✅
```

Hazard Ratios for Predictors of Mortality



Positive coef expected: VHF
Negative coef expected : AVNN, SDNN, pNN50, ULF, LF/HF ratio
Context dependent: LF, VLF

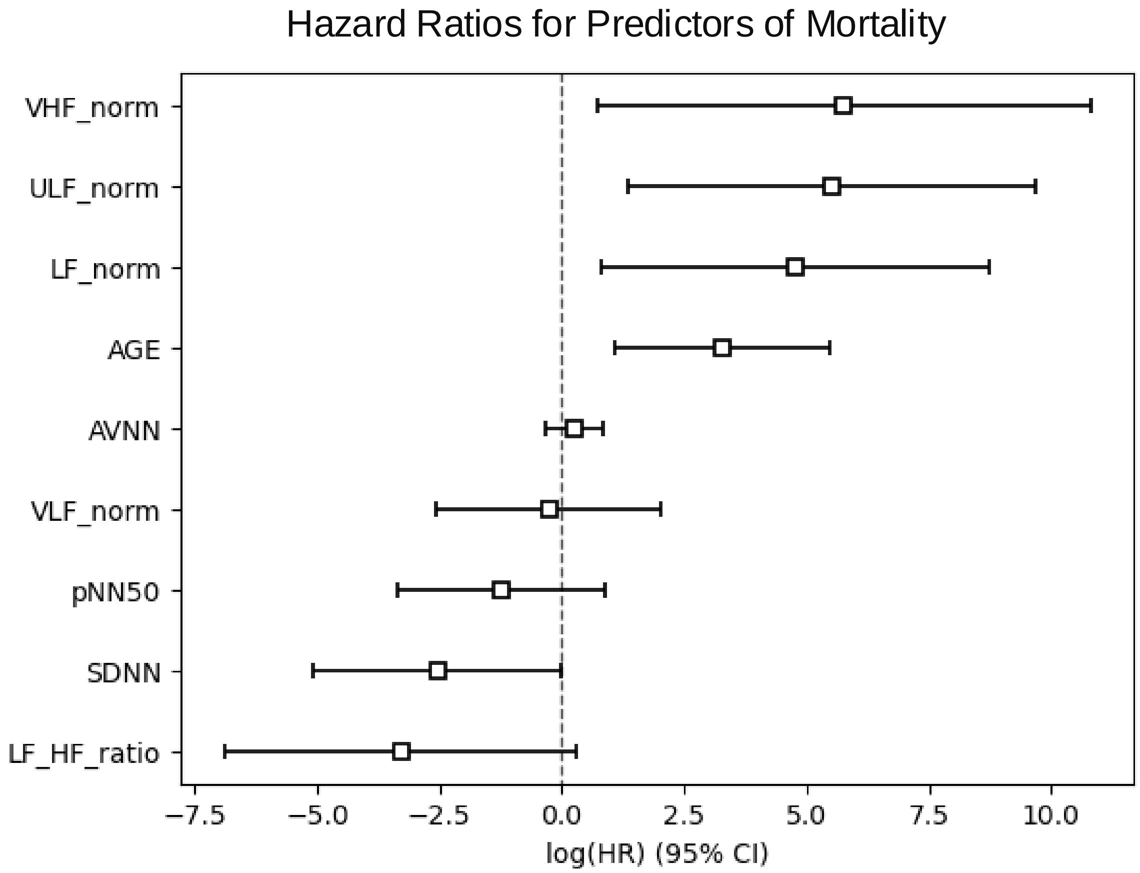
Task 1 Multivariate Survival Analysis – HRV metrics with AGE



```
AIC: 72.60230751205185
C-index: 0.9174852652259332
```

covariate	coef	exp(coef)	se(coef)	z	p
AGE	3.279559	26.564061	1.122248	2.922314	0.003474
AVNN	0.254714	1.290092	0.306350	0.831446	0.405722
LF_HF_ratio	-3.289431	0.037275	1.839042	-1.788666	0.073669
LF_norm	4.779355	119.027514	2.023292	2.362167	0.018168
SDNN	-2.543565	0.078586	1.294705	-1.964591	0.049462
ULF_norm	5.521513	250.012910	2.116272	2.609076	0.009079
VHF_norm	5.752721	315.046864	2.570276	2.238173	0.025210
VLF_norm	-0.263396	0.768437	1.176052	-0.223966	0.822783
pNN50	-1.228083	0.292853	1.073308	-1.144203	0.252539

```
Schoenfeld Residuals test p-value for AGE: 0.5151 ✅
Schoenfeld Residuals test p-value for AVNN: 0.0039 ❌
Schoenfeld Residuals test p-value for LF_HF_ratio: 0.0421 ❌
Schoenfeld Residuals test p-value for LF_norm: 0.1527 ✅
Schoenfeld Residuals test p-value for SDNN: 0.7497 ✅
Schoenfeld Residuals test p-value for ULF_norm: 0.0017 ❌
Schoenfeld Residuals test p-value for VHF_norm: 0.3305 ✅
Schoenfeld Residuals test p-value for VLF_norm: 0.1280 ✅
Schoenfeld Residuals test p-value for pNN50: 0.0105 ❌
```



Hazard Ratios for Predictors of Mortality

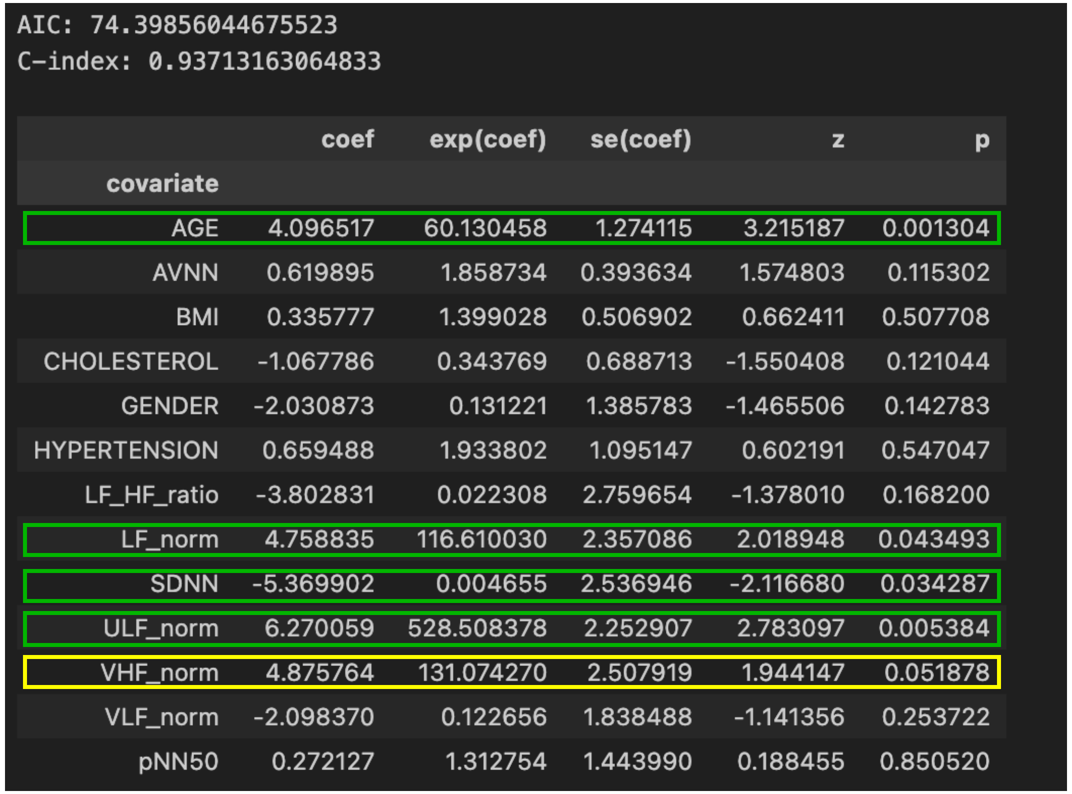
Positive coef expected: AGE, VHF
Negative coef expected : AVNN, SDNN, pNN50, ULF, LF/HF ratio
Context dependent: LF, VLF

University of California
San Francisco

MoonLAIT

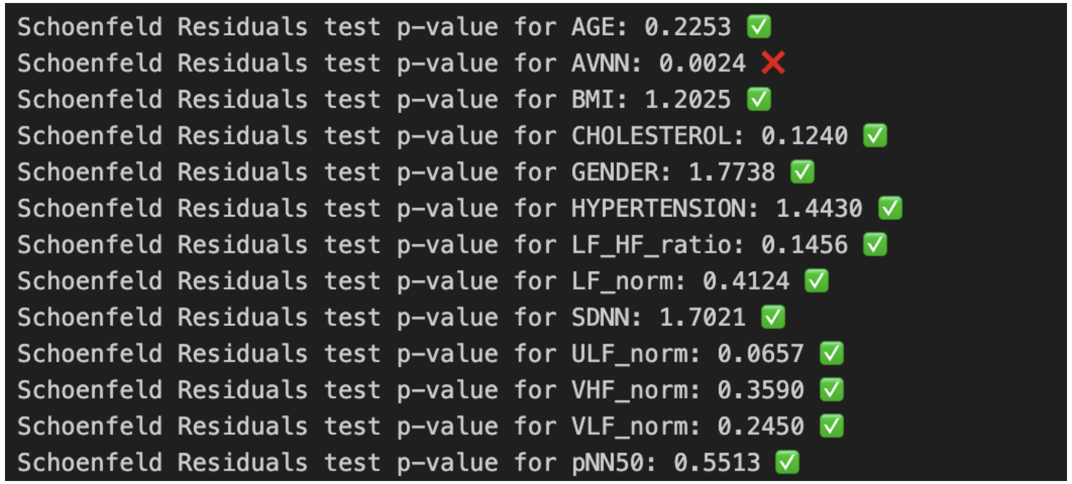
Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 1 Multivariate Survival Analysis – HRV metrics with all covariates



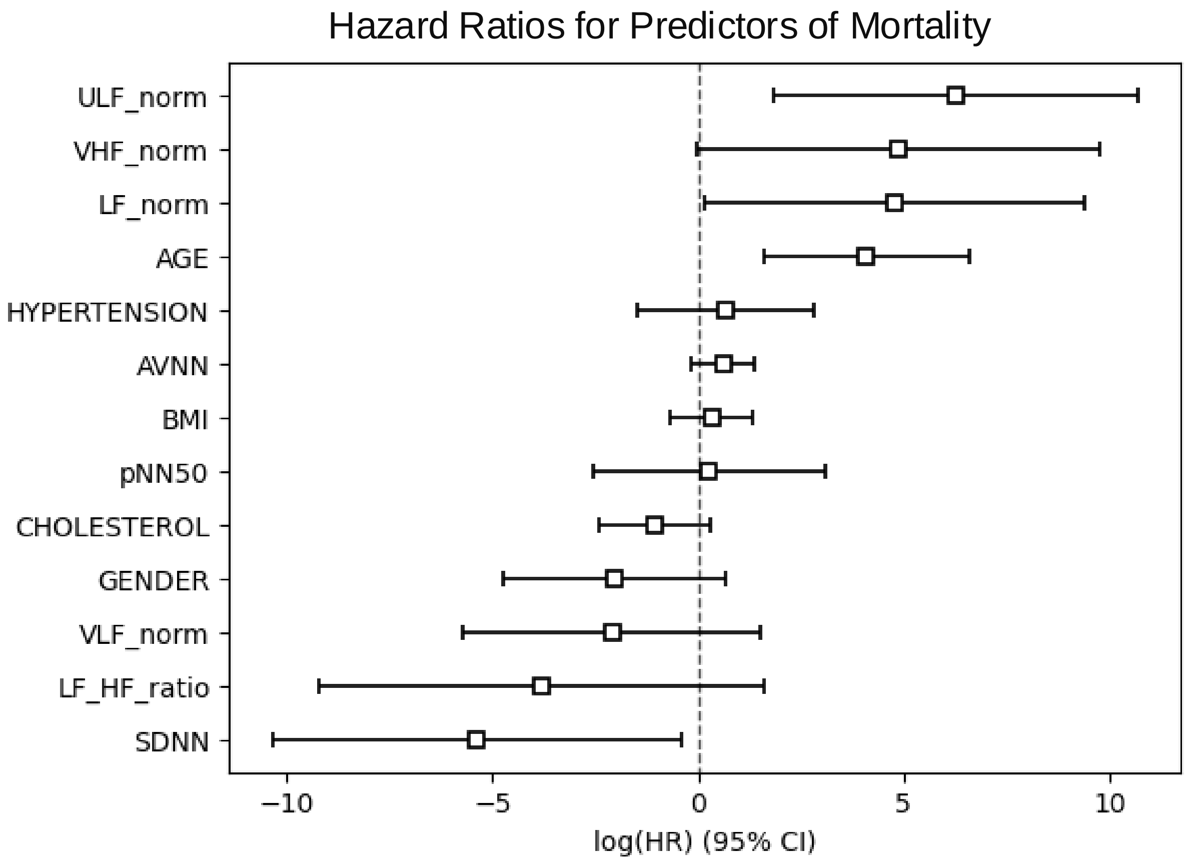
```
AIC: 74.39856044675523
C-index: 0.93713163064833
```

covariate	coef	exp(coef)	se(coef)	z	p
AGE	4.096517	60.130458	1.274115	3.215187	0.001304
AVNN	0.619895	1.858734	0.393634	1.574803	0.115302
BMI	0.335777	1.399028	0.506902	0.662411	0.507708
CHOLESTEROL	-1.067786	0.343769	0.688713	-1.550408	0.121044
GENDER	-2.030873	0.131221	1.385783	-1.465506	0.142783
HYPERTENSION	0.659488	1.933802	1.095147	0.602191	0.547047
LF_HF_ratio	-3.802831	0.022308	2.759654	-1.378010	0.168200
LF_norm	4.758835	116.610030	2.357086	2.018948	0.043493
SDNN	-5.369902	0.004655	2.536946	-2.116680	0.034287
ULF_norm	6.270059	528.508378	2.252907	2.783097	0.005384
VHF_norm	4.875764	131.074270	2.507919	1.944147	0.051878
VLF_norm	-2.098370	0.122656	1.838488	-1.141356	0.253722
pNN50	0.272127	1.312754	1.443990	0.188455	0.850520



```
Schoenfeld Residuals test p-value for AGE: 0.2253 ✅
Schoenfeld Residuals test p-value for AVNN: 0.0024 ❌
Schoenfeld Residuals test p-value for BMI: 1.2025 ✅
Schoenfeld Residuals test p-value for CHOLESTEROL: 0.1240 ✅
Schoenfeld Residuals test p-value for GENDER: 1.7738 ✅
Schoenfeld Residuals test p-value for HYPERTENSION: 1.4430 ✅
Schoenfeld Residuals test p-value for LF_HF_ratio: 0.1456 ✅
Schoenfeld Residuals test p-value for LF_norm: 0.4124 ✅
Schoenfeld Residuals test p-value for SDNN: 1.7021 ✅
Schoenfeld Residuals test p-value for ULF_norm: 0.0657 ✅
Schoenfeld Residuals test p-value for VHF_norm: 0.3590 ✅
Schoenfeld Residuals test p-value for VLF_norm: 0.2450 ✅
Schoenfeld Residuals test p-value for pNN50: 0.5513 ✅
```

Task 1 Multivariate Survival Analysis – HRV metrics with all covariates



Positive coef expected: AGE, GENDER (male), BMI, CHOLESTEROL, HYPERTENSION, VHF
Negative coef expected : AVNN, SDNN, pNN50, ULF, LF/HF ratio
Context dependent: LF, VLF

Model Performance:

- C-index = 0.937
- AIC: 74.34

Significant p-values that passing proportional hazards assumption:

- AGE, LF, ULF, SDNN, VHF

Key Coefficients:

- AGE, LF, VHF: Positive → Higher values = higher mortality risk
- SDNN: Negative → Higher SDNN = lower mortality risk

Conclusion

The model is promising, with key HRV metrics and Age as a confounder significantly improving survival prediction.

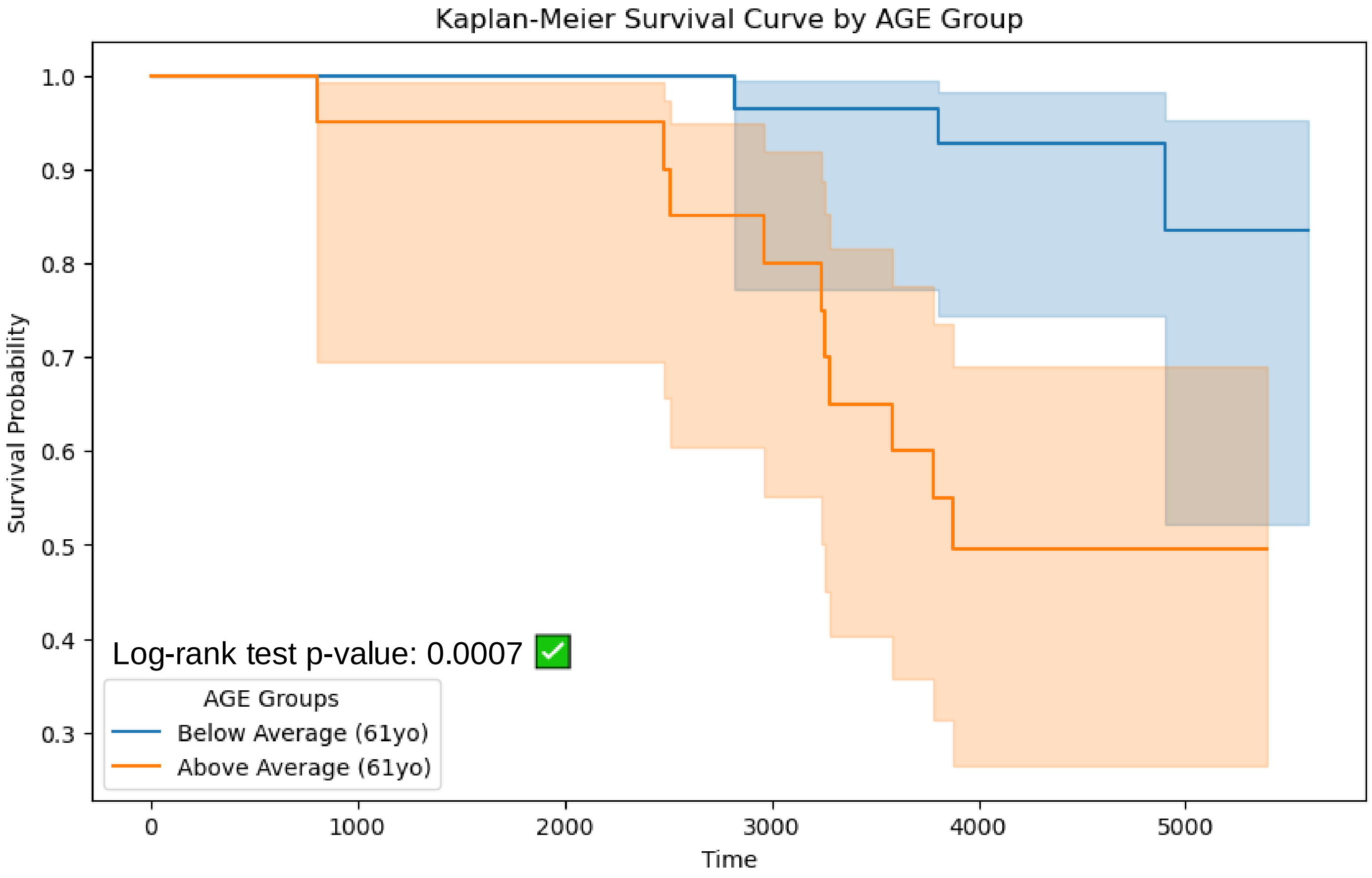
University of California San Francisco

MoonLAIT

Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 1

Univariate Survival Analysis – AGE



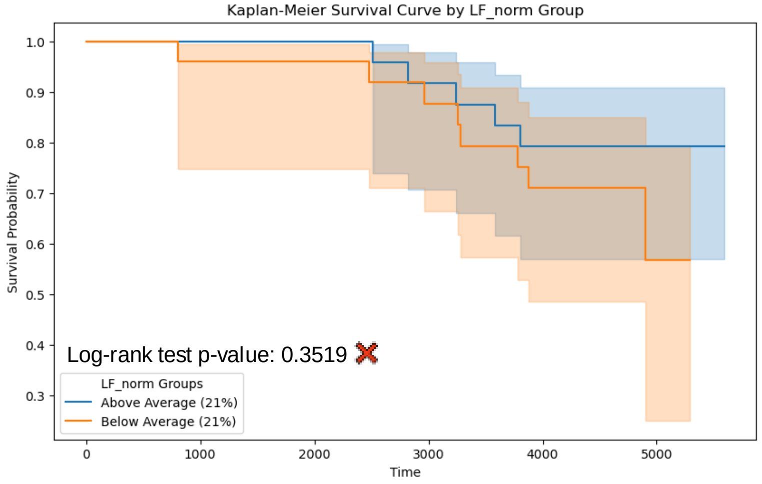
Log-Rank Test Interpretation

- **p-value > 0.05 → Feature is not significant**, indicating no clear difference in survival between groups.
- **p-value <= 0.05 → Feature is significant**, suggesting a statistically significant difference in survival between groups.

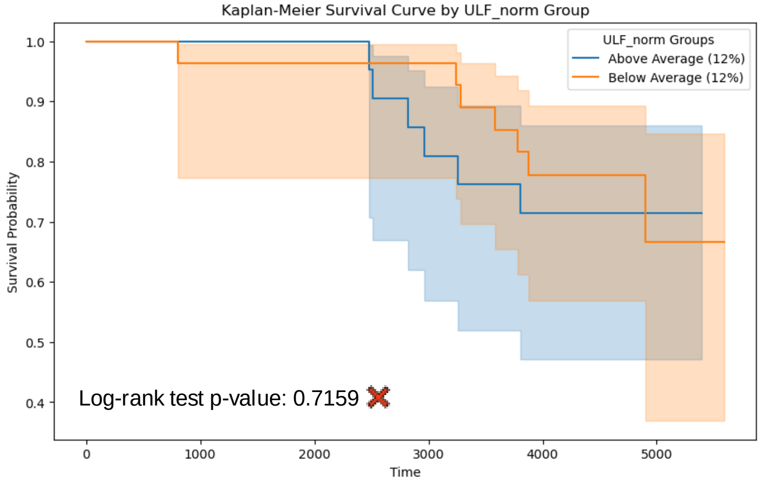
Age remains significant in both univariate and multivariate models, confirming its strong impact on survival. This aligns with the general understanding that older age is associated with higher mortality risk.

Task 1

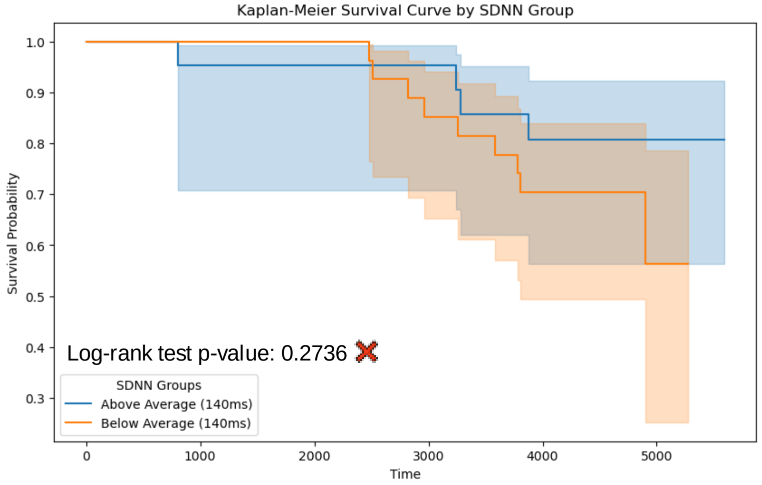
Univariate Survival Analysis – HRV metrics



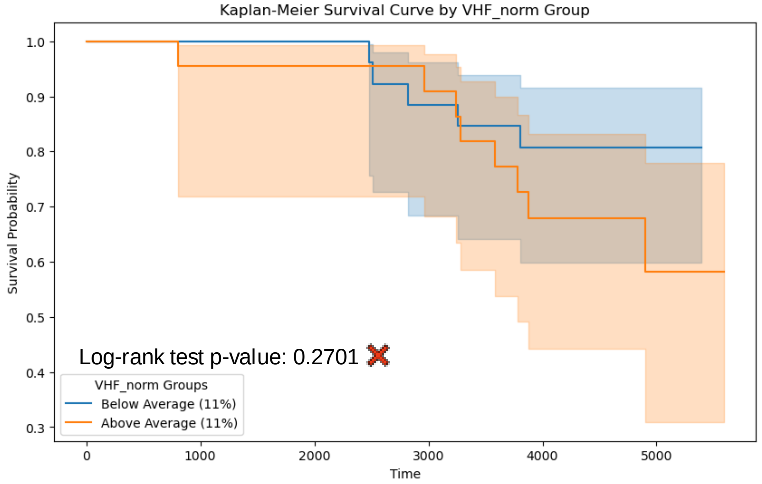
Log-rank test p-value: 0.3519 ❌



Log-rank test p-value: 0.7159 ❌



Log-rank test p-value: 0.2736 ❌



Log-rank test p-value: 0.2701 ❌

Log-Rank Test Interpretation

- **p-value > 0.05 → Feature is not significant**, indicating no clear difference in survival between groups.
- **p-value <= 0.05 → Feature is significant**, suggesting a statistically significant difference in survival between groups.

HRV metrics (LF, ULF, SDNN, VHF) are not significant univariately but become significant in the multivariate model, suggesting their combined effect with age improves survival prediction.

University of California
San Francisco

Assignment Task 2

Automatic Sleep Staging and Performance Evaluation

UCSF
University of California
San Francisco

MoonLAIT

Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025

Task 2

Sleep Stages Annotations

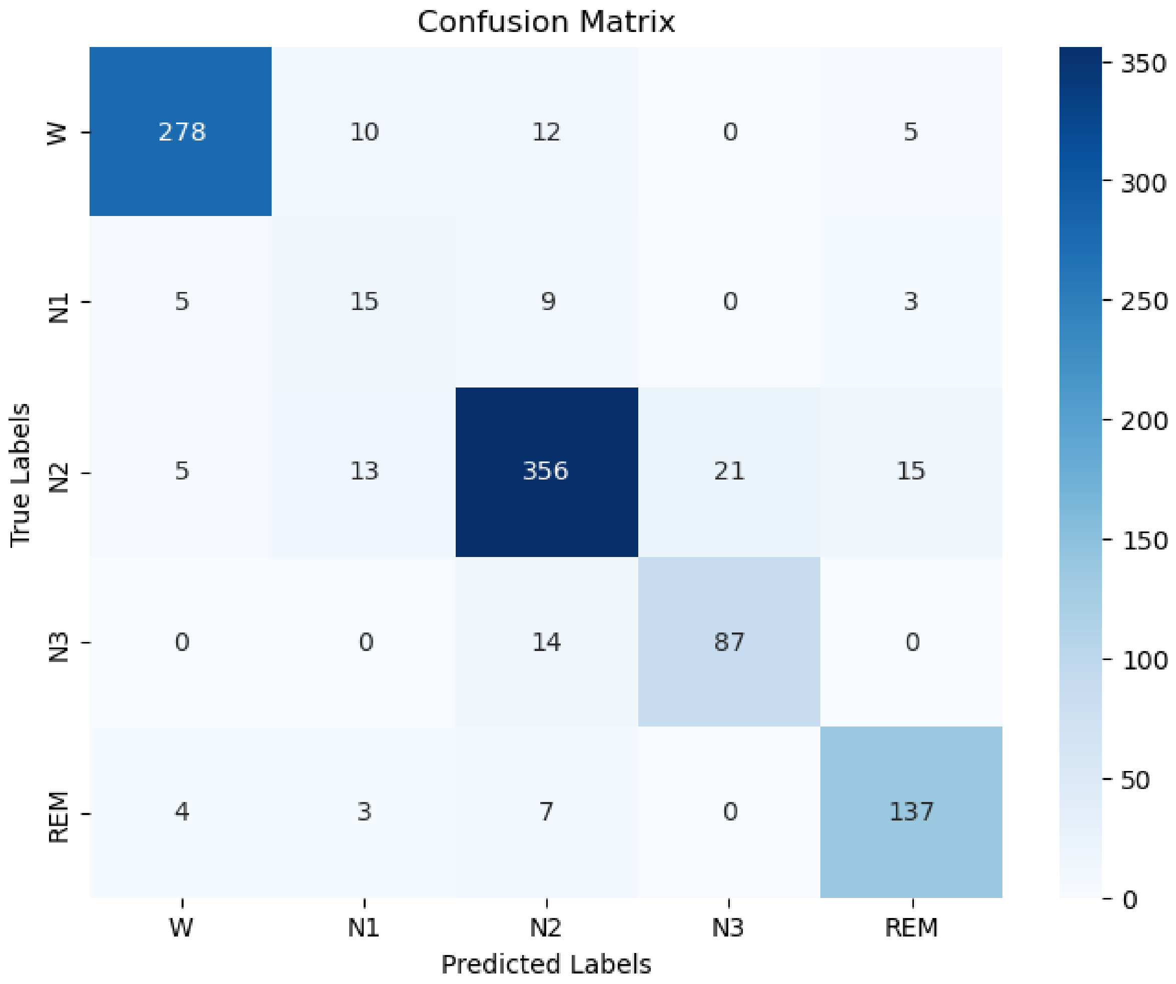
Experts

- Sleep stages annotated manually with **30-second epochs**.
- The **SHHS dataset** uses the **Rechtschaffen & Kales** guidelines (6 stages). For comparison with modern standards, the stages are adapted to **ASMM Guidelines**, combining **S3 and S4** into **N3**.

YASA Model

- Predicts sleep stages using **EEG (C4), EOG (left), and EMG (chin)** signals.
- Provides automated sleep stage labeling based on these signals, offering a more efficient and consistent approach.

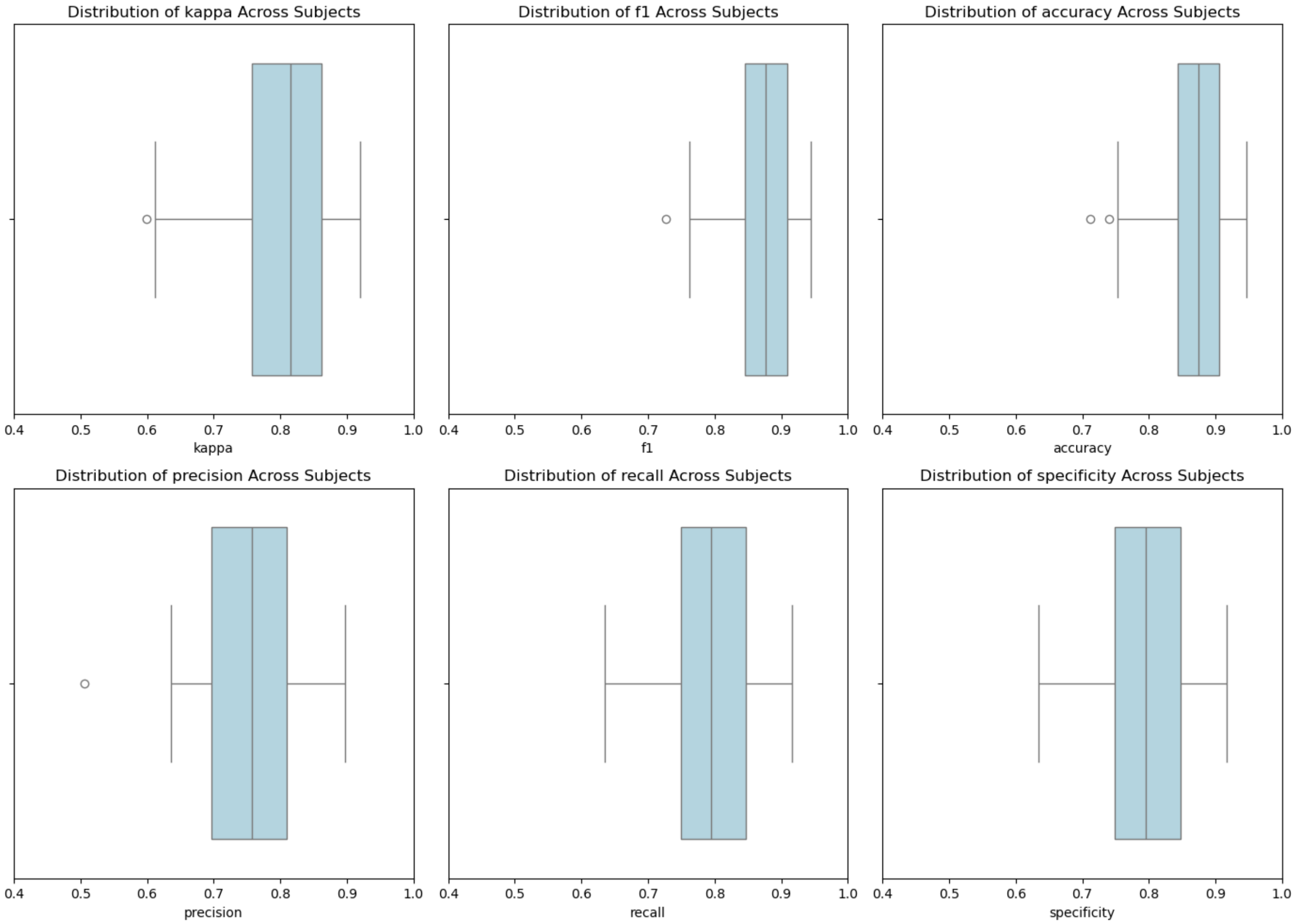
Performance Evaluation: Confusion Matrix



- ➢ The confusion matrix shows **strong agreement** between expert-labeled sleep stages in the SHHS dataset and the YASA model's predictions.
- ➢ The only **challenge is with the N1 stage**, which is often ambiguous due to its transitional nature between wakefulness and light sleep.

Task 2

Performance Evaluation: Other Metrics



The YASA model performs well in predicting sleep stages:

- **Kappa (0.81)**: High agreement with expert labels, minor inconsistencies.
- **F1 Score (0.87)**: Balanced precision and recall.
- **Accuracy (0.87)**: High classification accuracy.
- **Precision (0.75)**: Room to reduce false positives.
- **Recall (0.80)**: Good stage identification, room for improvement.
- **Specificity (0.80)**: Strong at identifying non-sleep stages.

Standard deviations show stable performance, with the most variation in **kappa** and **precision**. Overall, the model is reliable but can improve in distinguishing certain stages like N1 stage.

UCSF
University of California
San Francisco

MoonLAIT

Thank you for you attention !

Any Questions ?

Alice ALBRECHT
Interview for Research Data Analyst
February 4th, 2025