

Replicate Intl
Michael Coggins, PhD

BIOMETRIC DATA SCIENCE HRV PROJECT

Current Progress



CONTENTS

1. Purpose
2. About Me
3. Data Science Life Cycle
 1. Initial Ideation / Problem Definition
 2. Data Acquisition and Exploration
 1. EDA (.csv and .xlsx files); Transformation
 2. Domain understanding and Hypothesis
 3. Outlier analysis and data cleaning concerns
 4. Composite metric development
 3. R&D / Model(s) Building
 1. Using $\langle \rangle$ to guess at missing age values
 2. Underlying contributors / key
 3. Using Gen AI to increase adult data
4. Delivery and Monitoring (not there yet, so...)
 1. Using EKG-based HRV data models to evaluate PPG-based proprietary data



PURPOSE

Why? But....why...(or why not)?

I want to understand us and data can help

- For years, I collected biometric data – mostly performance, but some relevant to basic medical health and wellness – as part of my own curiosity regarding client success (or not) as a coach and trainer
 - In 2016-2019, I created a small pilot study to look at wearable technology as a potentially unbiased data collection/analysis engine to predict improvements in a person's strength, lean muscle mass, bone density and fat loss based on different exercise and diet interventions
- During the pandemic, as gyms closed, I took online courses in Data Science to broaden my analytical techniques as most of what I used in previous research 'lives' were basic parametric and non-parametric methods
- In 2022, I thought...why not put these together?
 - But, a professor friend, suggested I use some publicly available data first, as way to have a comparison to other researchers' data analytics and conclusions.



ABOUT ME

Entrepreneur (Replicate International)
Business Analyst (Credit Suisse, Campbell Alliance)
Researcher (Yale, Cornell)
Military Engineer (USAF)



Data Scientist

- Started computational side projects after career change from equity research to athletic coach / trainer
- Started Coursera courses during Covid lockdown (IBM Data Science cert amongst other classes)



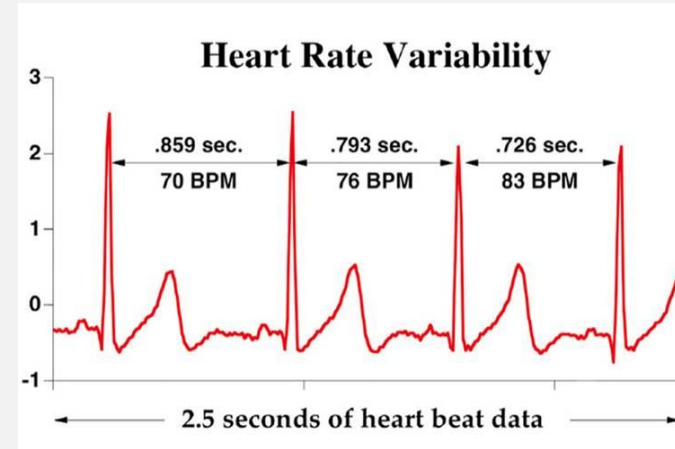


INITIAL IDEATION / PROBLEM DEFINE

*Initial idea and first
look at a public
database*

IDEATION, PROBLEM DEFINITION OR PURPOSE

Can we learn something new from old data



- There exist various publicly-accessible databases of human biometric data with potential for preventative health decisions, or performance-based tailoring of performance/recovery programs
- There appears to be untapped potential for data to be analyzed by alternative methods including ML programs
- Why not run some of the public data through alternative methods to look for possible uncovered predictors/relevant metrics?
- EKG data represent a simple test case with significant years of study (and possibly a number of databases), yet little consensus about predictive value

Can we identify new predictive metrics from EKG data?

INITIAL DATA VISUALIZATION - GLOSSARY

Quick glossary and terminology from PubMed research (Largest searchable academic physiological database)

- Interbeat interval (IBIs)
 - What is displayed (in milliseconds) on pre-processed EKG outputs (raw output are graphs with characteristic electrical current traces relating to nervous input and muscle contraction of heart)
- Instantaneous Heart Rate (iHRs)
 - If an interval between beats (eg 800 ms/beat) is converted to a rate of beats per minute ($60 \text{ sec/min} / 0.8 \text{ sec/beat} = \text{beats/min}$) then an instantaneous heart rate is calculated (75 bpm)
 - Standard measurement displayed on wearable tech, Holter displays, etc.
 - In reality, these change as the recorded IBIs are not all exactly the same, they're more like the speedometer in a car and vary around a mean. This variability is termed Heart Rate Variability (HRV)
- Heart Rate Variability
 - Can be reported by Time Domain (TD), Frequency Domain (FD), or other measurements (entropy) (Task Force, *Circulation*, 1996)
 - TD and FD are the most commonly reported and intuitive both mathematically and physiologically
 - TD and FD each have 2 characteristic measurements that have been related physiologically to different chemical processes with different time courses (Akselrod, et al. *Science*, 1981; Schwab et al. *Heart*, 2003)

INITIAL DATA VISUALIZATION

First, we sought to look at the data to get a feel for the potential for different classes based on age (sex)

Notes:

- Inter-Beat Intervals (IBIs, recorded in milliseconds per heart beat) are the reported data. We also converted IBIs to Instantaneous Heart Rates (iHRs, beats/min)
 - Due to relationship between IBIs and iHRs:
$$\text{iHR} = 60 \text{ (seconds per min)} / \text{IBI (in seconds)}$$
- Database included 24-hour Holter monitor recordings of different aged subects:
 - 56 Infants aged (1 month – 3 years 11 months)
 - 11 Adults aged (17 – 53)
 - 24 Adolescents (4 – 15)
 - 4 Unknown

	A	B
1	539	
2	539	
3	547	
4	539	
5	539	
6	531	
7	539	
8	539	
9	539	
10	539	
11	547	
12	547	
13	555	
14	547	
15	562	
16	563	
17	554	
18	563	
19	570	
20	563	
21	570	

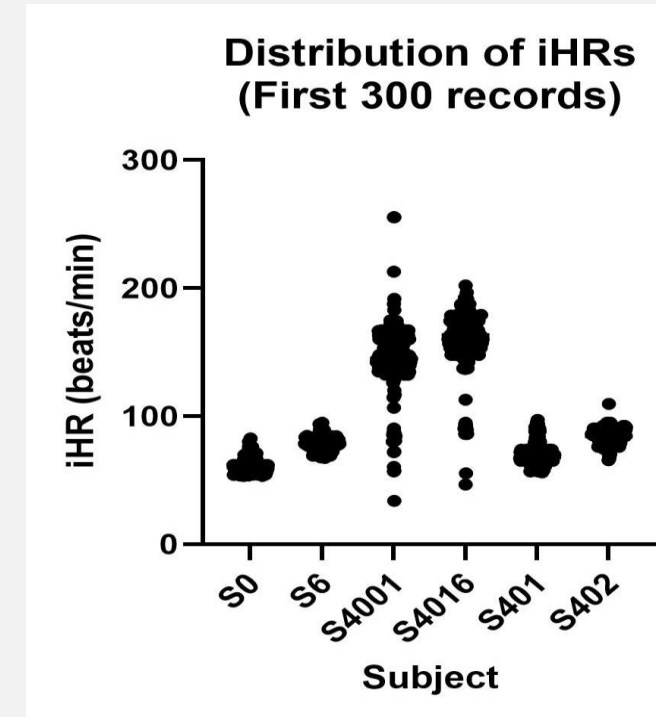
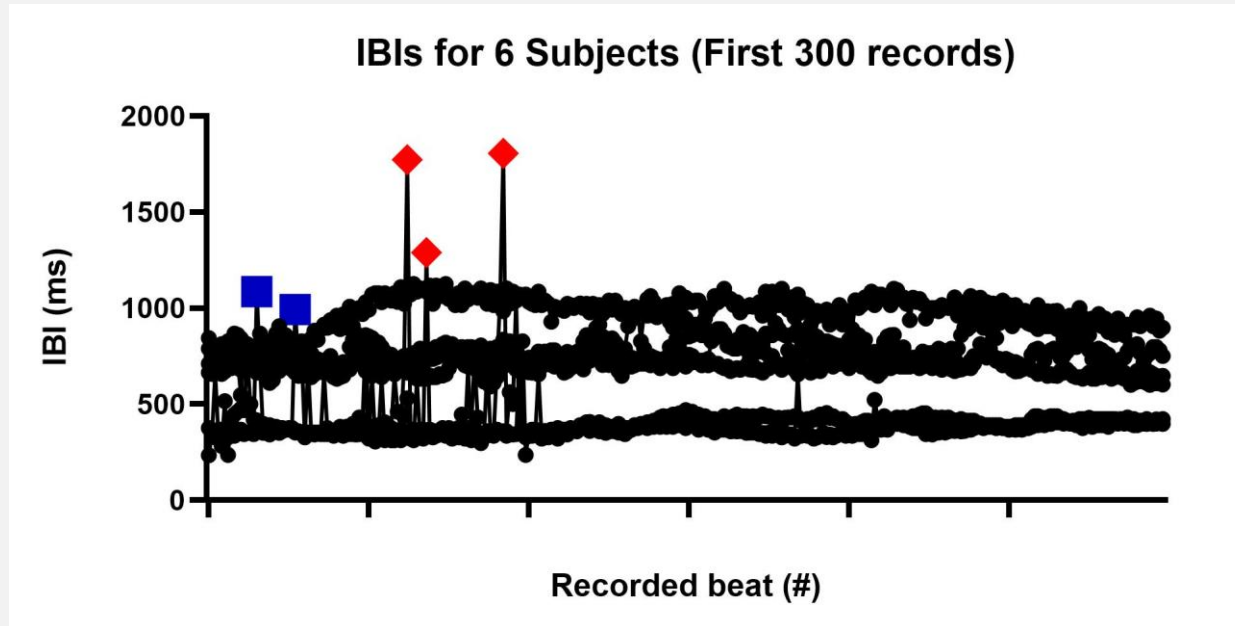
Single Subject IBIs

	A	B	C
1	File	Age (years)	Gender
2	0	53	M
3	2	17	F
4	3	46	F
5	5	38	F
6	6	32	M
7	7	51	F
8	8	39	M
9	9	24	F
10	10	55	M
11	11	17	M
12	12	20	F
13	13	39	F
14	401	12	
15	402	10	
16	403	13	
17	404	5	
18	405	15	
19	406	15	
20	407	6	
21	408	13	

Metadata

INITIAL DATA TRANSFORMATION

First, we sought to look at the data to get a feel for the potential for different classes based on age (sex)



Notes:

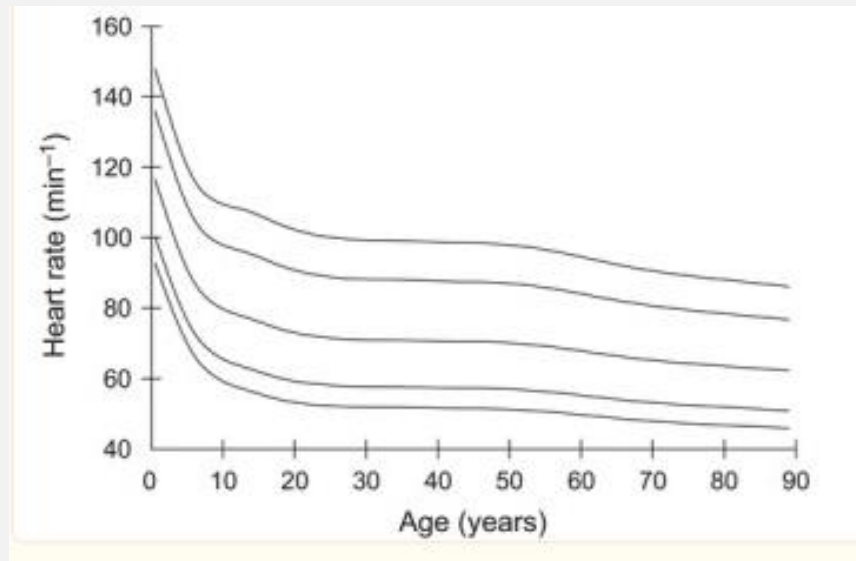
- 6 Traces shown:
 - S0 is 53, S6 is 32, S4001 is 6 months, S4016 is 1 year 2 months, S401 is 12 and S402 is 10 years old
- Infants had higher overall average iHRs (and corresponding smaller IBIs)
- Infants had recordings containing more potential 'missed' heart beats (multiples of previous/post recording, Red Diamonds) or large jumps between consecutive records (~10-30% change in recorded IBI, Blue Squares))

INITIAL DATA VISUALIZATION – DOMAIN CHECK

Do our observations square with previous observations and what is current domain understanding of differences between infants, adolescents and adults in EKG recordings

Domain Literature Review:

- Average IBIs for infants
 - Unfortunately, the accompanying paper ([Garavaglia, et al. 2021](#)) did not provide comparison of average heart rates (or distributions) for different ages/genders.
 - A marked change in two HRV metrics – discussed later – did have a large change breakpoint at ~12 years
 - A separate study and medical textbook ([Claiborne, et al. 2023](#), [Doherty, et al Neonatal Physiology](#)) detail that infants have higher average heart rates than adults
- Adolescent IBIs
 - Abstracts and at least 1 recent study in China ([Cattermole et al, 2019](#)), demonstrate age-dependent heart rates lowering and leveling off at ~10-15 years old





FIRST DATABASE ANALYSIS

*How data analysis has
been done in the past,
and our hypothesis*

TIME DOMAIN ANALYSIS – IN THE PAST

Two main measurements:

- SDNN – Standard Deviation (from mean over chosen time period)
 - Evaluates deviations from each beat to the mean
 - Corresponds to slower time course, lower frequency changes (*possibly* parasympathetic)
- rMSSD – Root Mean Square Standard Deviation (from immediately previous IBI)
 - Evaluates deviation from immediately previous recorded heart beat
 - Corresponds to faster time course, higher frequency changes (sympathetic nervous system and respiration, *possibly*)

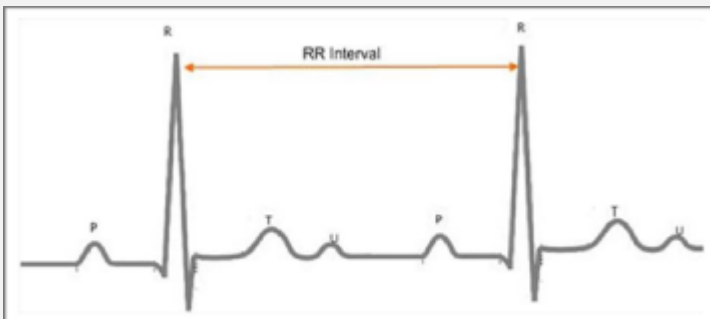
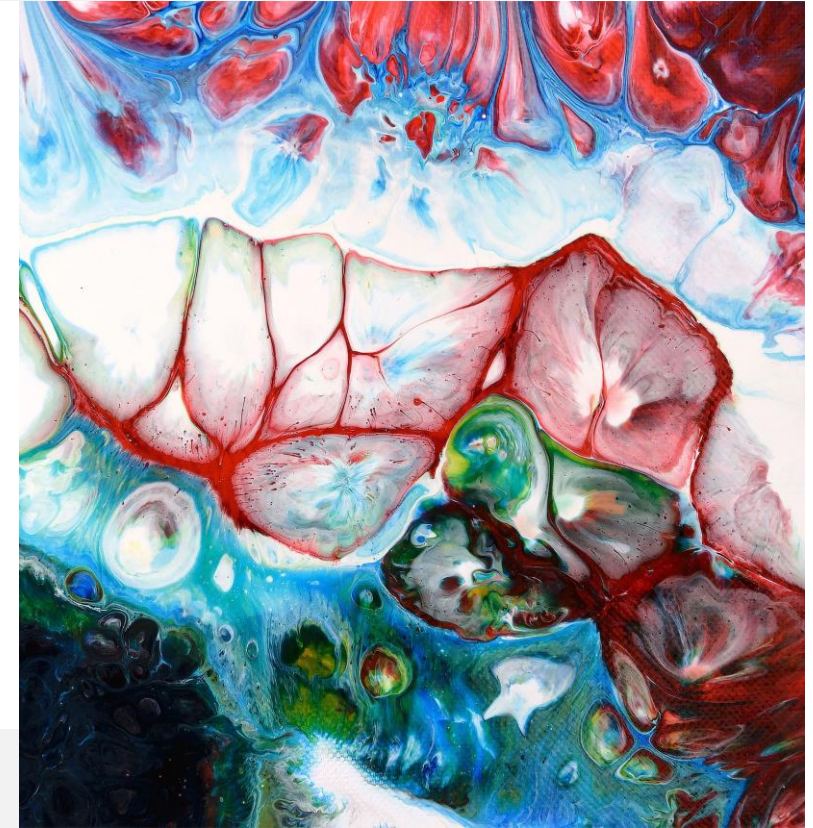


Fig. 1 RR interval used for HRV measurement

Murukesan et al., 2013



TIME DOMAIN ANALYSIS

A general display of formulas and calculated values for SDNN and rMSSD

Formulas:

$$SDNN = \sqrt{\frac{1}{N-1} \sum_{n=2}^N [I(n) - \bar{I}]^2},$$

$$RMSSD = \sqrt{\frac{1}{N-2} \sum_{n=3}^N [I(n) - I(n-1)]^2}.$$

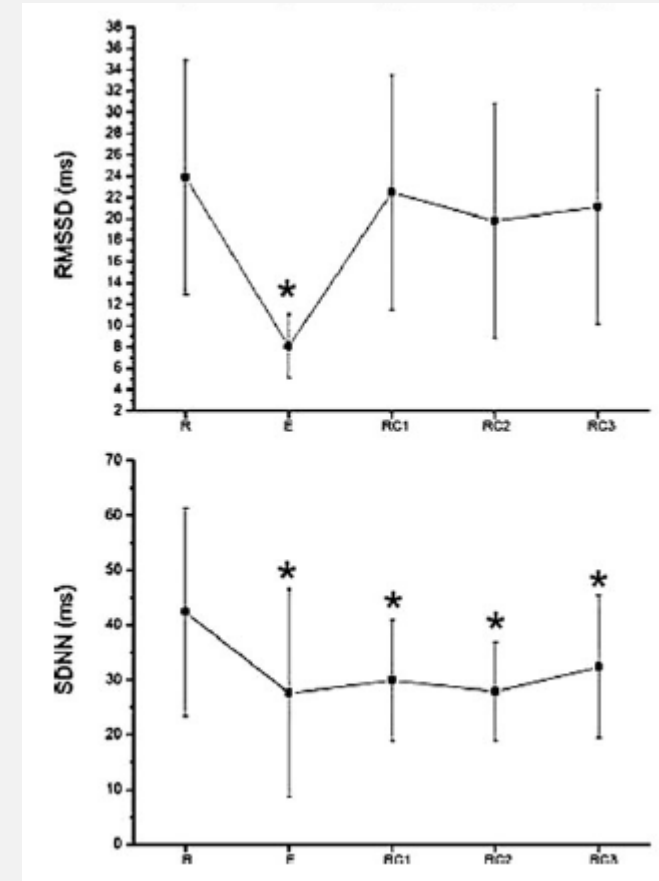
Wang and Huang, 2012

General values for SDNN:

- Vary (often) between 20 to 120 ms for adults
- In Irurzun, et al, values ~40-140 ms, (Claiborne et al showed infants with ~10-50 ms)

General values for rMSSD:

- Vary (often) between 20 to 100 ms for adults
- Irirzun et al, values ~5-50ms (Claiborne et al showed infants with ~10-25 ms)



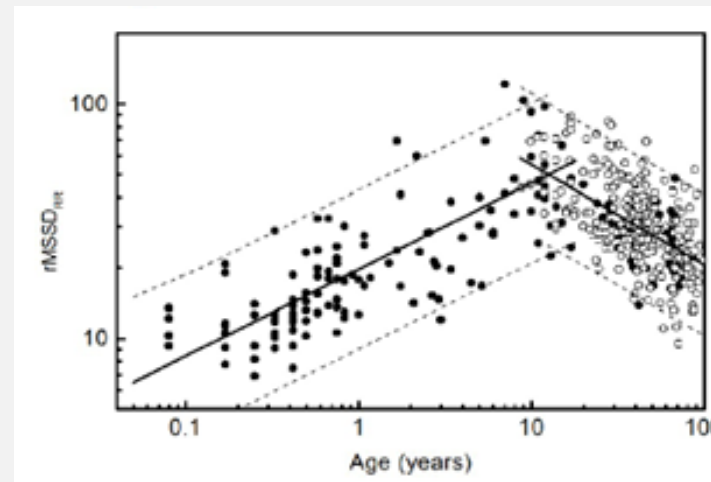
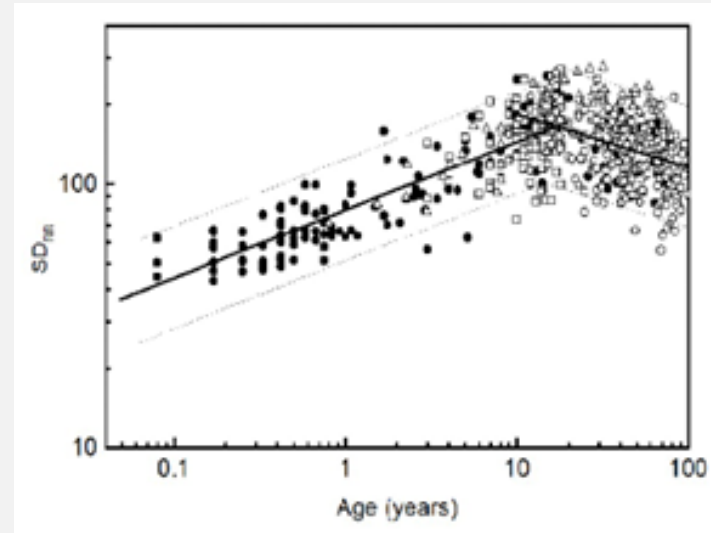
Raimundo, et al., 2013

INITIAL DATA VISUALIZATION – DOMAIN CHECK

Do our observations square with previous observations and what is current domain understanding of differences between infants, adolescents and adults in EKG recordings

Domain Literature Review:

- Time Domain HRV metrics relating to subject age.
 - Paper with online database ([Garavaglia, et al. 2021](#)) displayed an age-related log trend in SDNN and rMSSD increase until ~age 10-20 and then potentially a decline
 - Domain reviews indicate that SDNN is a longer time constant process likely related to parasympathetic nerve input while rMSSD is a shorter time constant process related to sympathetic and respiratory inputs ([Billman, 2011](#))
 - There is much disagreement over whether either/both of these metrics have simply too much noise to be of value. Also, if they are truly separable or convolved.
 - Coefficient of variations in populations have been recorded at even ~1.0



HYPOTHESIS

Versions of normalizing TD data to account for variable expected values given different average HRs

My idea for analysis creates values similar to Z Scores

- $Z = (x - \mu) / \sigma$ as a new version of a 'normalized' comparator given a pre-existing IBI (or HR)
 - I chose to use a simple version of an 'expected noise' or 'expected jitter' appropriate for SDNN (around the time period mean IBI) and rMSSD (immediately previous IBI)
- Shown is a single subject with created Z scores (for SDNN comparison metric)

Trace	aPR	IBI(ms)	((n) - (n-1))	E[SDNN]512	Z[SDNN]
461	130.1518	0	0	1.735612767	0
453	132.4503	8	64	1.705362594	13.62410465
438	136.9863	15	225	1.657180472	64.82695304
437	137.2998	1	1	1.631944205	0.149950134
461	130.1518	24	576	1.651921188	183.0213445
461	130.1518	0	0	1.665443658	1
469	127.9318	8	64	1.683254414	14.08274278
469	127.9318	0	0	1.69680065	1
469	127.9318	0	0	1.707449986	1
484	123.9669	15	225	1.726485454	59.10796569
469	127.9318	15	225	1.732668954	58.63216679
484	123.9669	15	225	1.746710209	57.57125161
469	127.9318	15	225	1.750456973	57.2926512
476	126.0504	7	49	1.757323967	8.900253548
477	125.7862	1	1	1.763789356	0.187522666
461	130.1518	16	256	1.762008285	65.29524855
453	132.4503	8	64	1.756761066	12.6297431
453	132.4503	0	0	1.752116523	1
453	132.4503	0	0	1.747976474	1
446	134.5291	7	49	1.741473361	9.117892873
453	132.4503	7	49	1.738264039	9.162766311
445	134.8315	8	64	1.732471725	13.08760844
453	132.4503	8	64	1.729952747	13.13630267

TIME DOMAIN ANALYSIS – DATASET CHARACTERISTICS

General concerns and limitations of the dataset

Time of Recording:

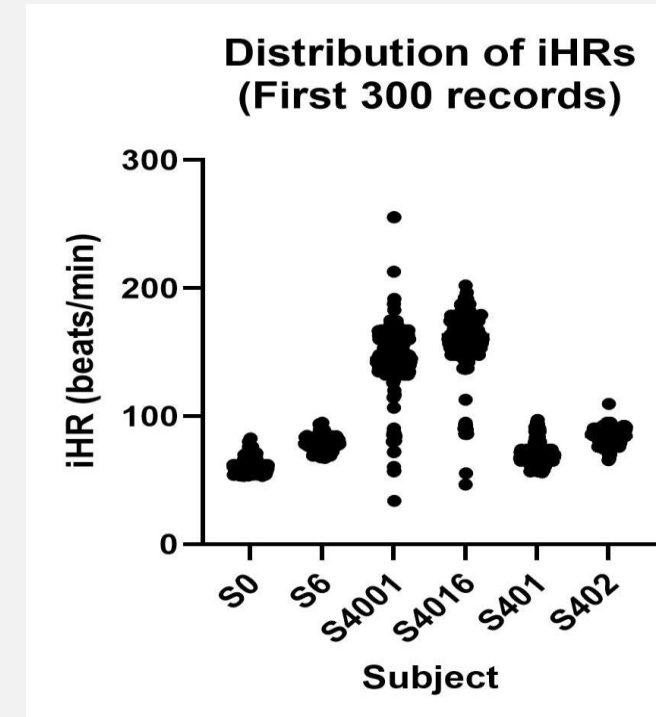
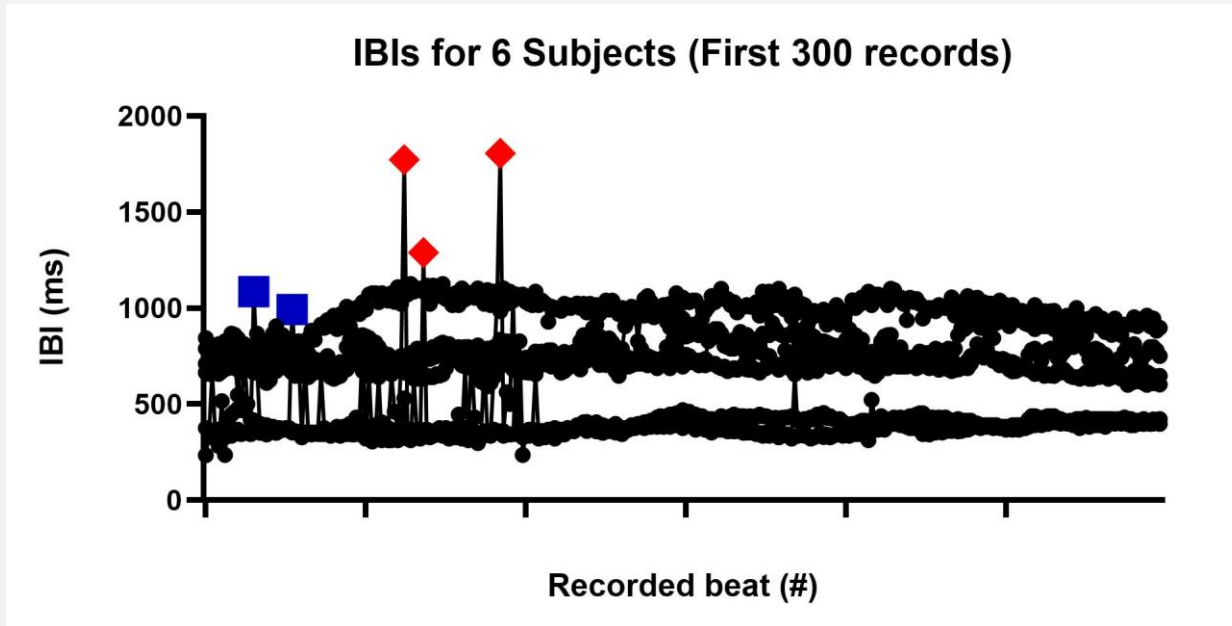
- Our available database has ~24-hour continuous recording but no characterization of when start time occurs (unknown if same for each subject)
- Unknown activities, and at what time and for what duration, occur during the recordings
 - Significant changes could occur during or after certain activities such as sleep, exercise, etc.

Time Epochs for Data Analysis:

- Domain literature is wide-ranging on amount of time used to calculate HRV values. Analysis has been performed on time epochs varying from 1 min to 24 hours
- EKG-based literature in particular has disagreements over what amount of time is sufficient – but not too large – to represent a single physiological state, while allowing enough data to potentially capture longer time course processes.
 - Akselrod and others using pharmacological blockers (and often using FD metrics) have described process time constants of ~2.5 to 40 s which implies needing at least 2-3 minutes to ‘see’ changes in longer term processes
 - Unclear if the processes can affect *both* SDNN and rMSSD (or equally affect). It remains possible that longer processes ‘only’ affect the ‘longer’ metric of SDNN
- What about the possibility that having *more heart beats* during a time epoch means more opportunity for modulation? Does analyzing same time as opposed to same *number* of beats make sense?

DATA OUTLIERS

First, we sought to look at the data to get a feel for the potential for different classes based on age (sex)



Notes:

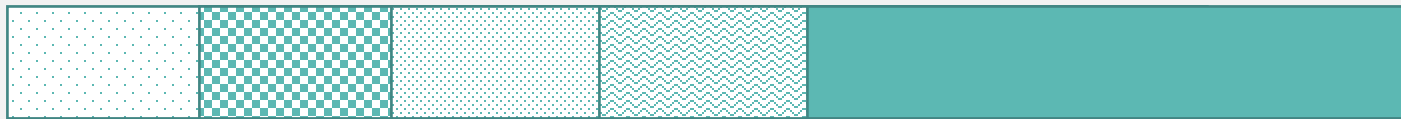
- Infants had recordings containing more potential ‘missed’ heart beats (multiples of previous/post recording, Red Diamonds) or large jumps between consecutive records (~10-30% change in recorded IBI, Blue Squares))

CREATED COMPOSITE DATA

Given different size datasets (row length) and disparities in known attributes, what might better represent individual data

Created a sliding windows of time periods (now many HRV ‘epochs’)

- I chose two time periods of 512 IBIs (513 heart beats) and 256 IBIs (257 IBIs)
 - Given infants have ~2x as many heart beats per time period, this allows for a comparison of similar #IBIs or similar periods of time
 - 512 IBIs is ~6-10 minutes for adults, ~3.5-5 minutes for infants
 - Now there are thousands of short time epochs (short windows) to potentially better capture periods of stabilized physiological state



T=0, IBI 1	T=x1, IBI 2	T=x2, IBI 3	T=x3, IBI 4	T=x4, IBI 5
---------------	----------------	----------------	----------------	----------------

CREATED COMPOSITE DATA

Given different size datasets (row length) and disparities in known attributes, what might better represent individual data

Created a group of simple statistical data as comparators

- Sample descriptors such as mean, coefficient of variation, count, etc. help minimize ectopic and missing beats
 - Some metrics not shown below

Subect	Age (years)	Gender	Avg IBI (ms, 512)	CoV Avg IBI	Avg iPR (bpm, 512)	CoV Avg iPR	No Change Count Z[SDNN, 512]	Low Count Z<2 [SDNN, 512]	High Count Z>4 [SDNN, 512]	Avg C[SDNN, 512]	CoV C[SDNN, 512]	Low Count <2 C[SDNN]	High Count >4 C[SDNN]
<u>0</u>	53	M	943.3	0.17	66.0	0.21	0.04	0.16	0.63	7.39	0.44	0.00	0.94
<u>2</u>	17	F	841.5	0.19	74.3	0.21	0.03	0.09	0.76	10.88	0.27	0.00	1.00
<u>3</u>	46	F	711.4	0.11	85.4	0.12	0.08	0.16	0.57	6.76	0.74	0.00	0.83
<u>5</u>	38	F	769.4	0.17	82.3	0.18	0.06	0.14	0.62	7.82	0.45	0.00	0.96
<u>6</u>	32	M	858.4	0.17	70.2	0.17	0.07	0.20	0.48	5.04	0.33	0.00	0.70

An abstract, textured background with a dark overlay. The background features a mix of colors including red, yellow, green, blue, and purple, with a cracked, marbled appearance. A semi-transparent dark rectangle is overlaid on the bottom left, containing the title text.

FIRST DATABASE ANALYSIS

*Initial (new) data
analysis*

VISUALIZATION OF COMPONENT DATA ANALYSIS

Looking for potentially irrelevant or redundant data for model building

Are there data components (column vectors) that are irrelevant?

- Column values across a row that do not differ (intra-subject variation)?

Subect	Age (years)	Gender	Avg IBI (ms, 512)	CoV Avg IBI	Avg iPR (bpm, 512)	CoV Avg iPR	No Change Count Z[SDNN, 512]	Low Count Z<2 [SDNN, 512]	High Count Z>4 [SDNN, 512]	Avg C[SDNN, 512]	CoV C[SDNN, 512]	Low Count <2 C[SDNN]	High Count >4 C[SDNN]
<u>0</u>	53	M	943.3	0.17	66.0	0.21	0.04	0.16	0.63	7.39	0.44	0.00	0.94
<u>2</u>	17	F	841.5	0.19	74.3	0.21	0.03	0.09	0.76	10.88	0.27	0.00	1.00
<u>3</u>	46	F	711.4	0.11	85.4	0.12	0.08	0.16	0.57	6.76	0.74	0.00	0.83
<u>5</u>	38	F	769.4	0.17	82.3	0.18	0.06	0.14	0.62	7.82	0.45	0.00	0.96
<u>6</u>	32	M	858.4	0.17	70.2	0.17	0.07	0.20	0.48	5.04	0.33	0.00	0.70

Are some data components likely redundant?

- Data values across columns that don't differ (little inter-subject variation)?

VISUALIZATION OF COMPONENT DATA ANALYSIS

Looking for potentially irrelevant or redundant data for model building

Are there data components (column vectors) that are irrelevant?

- Column values across a row that do not differ (intra-subject variation)?
- Data values across columns that don't differ (little inter-subject variation)?

Subect	Age (years)	Gender	Avg IBI (ms, 512)	CoV Avg IBI	Avg iPR (bpm, 512)	CoV Avg iPR	No Change Count Z[SDNN, 512]	Low Count Z<2 [SDNN, 512]	High Count Z>4 [SDNN, 512]	Avg C[SDNN, 512]	CoV C[SDNN, 512]	Low Count <2 C[SDNN]	High Count >4 C[SDNN]
<u>0</u>	53	M	943.3	0.17	66.0	0.21	0.04	0.16	0.63	7.39	0.44	0.00	0.94
<u>2</u>	17	F	841.5	0.19	74.3	0.21	0.03	0.09	0.76	10.88	0.27	0.00	1.00
<u>3</u>	46	F	711.4	0.11	85.4	0.12	0.08	0.16	0.57	6.76	0.74	0.00	0.83
<u>5</u>	38	F	769.4	0.17	82.3	0.18	0.06	0.14	0.62	7.82	0.45	0.00	0.96
<u>6</u>	32	M	858.4	0.17	70.2	0.17	0.07	0.20	0.48	5.04	0.33	0.00	0.70

Are some data components likely of little predictive value?

- Very little variation between subjects

MISSING DATA

Versions of normalizing TD data to account for variable expected values given different average HRs

Metadata file missing some key ages and genders

- For initial model building, I only used existing data records with complete age and gender
- Later used model to guess missing ages
 - Too many missing genders from initial public dataset for a reasonable model guess
 - Authors of dataset concluded no effects of gender on their ~700 subject sample set
- Also tried generative AI to create further adolescent and adult data samples
 - Put technique/website here
- Additionally, database had a duplicate subject (403/404)

Subect	File	Age (years)	Gender
0	0	53	M
2	2	17	F
3	3	46	F
5	5	38	F
6	6	32	M
7	7	51	F
8	8	39	M
9	9	24	F
10	10	55	M
11	11	17	M
12	12	20	F
13	13	39	F
401	401	12	
402	402	10	
403	403	13	
404	404	5	
405	405	15	
406	406	15	
407	407	6	
408	408	13	
409	409	10	M
410	410	12	F
411	411	8	F



FIRST DATABASE ANALYSIS

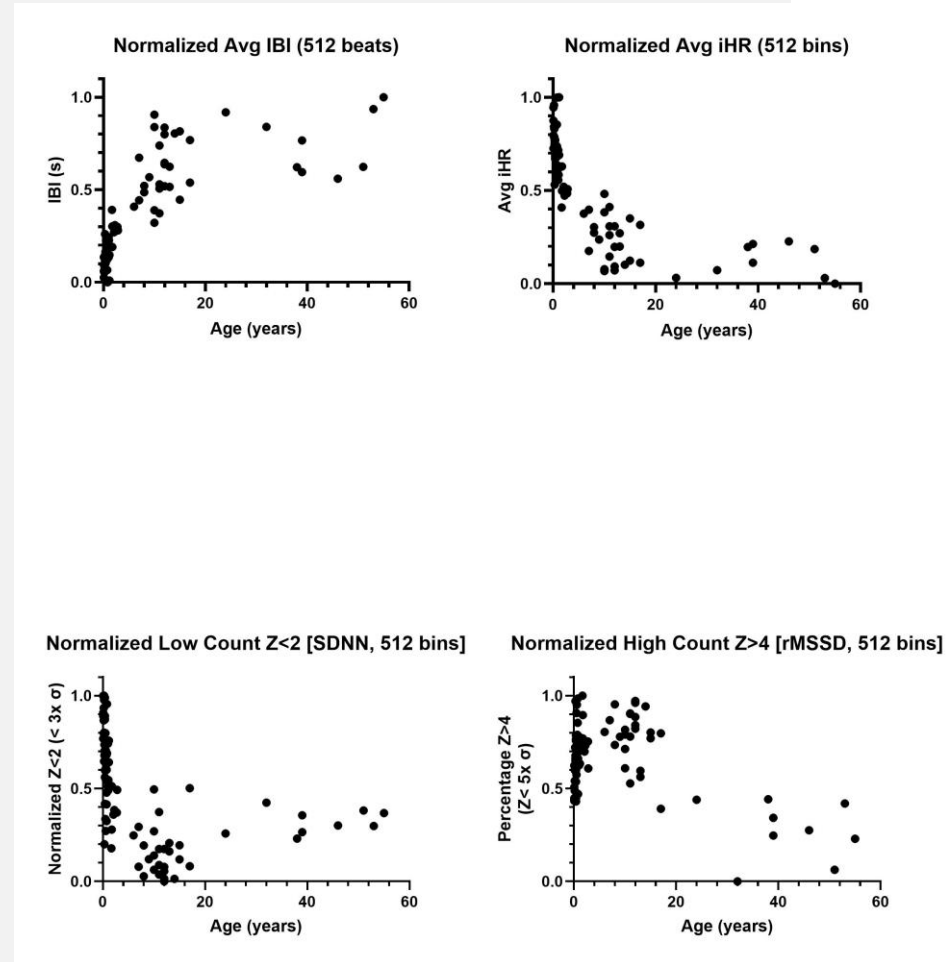
*Initial (new) data
analysis*

TIME DOMAIN COMPOSITE DATA VISUALIZATION

Do any of the composite metrics show variation by age or possible 2-3 groupings (infant, adolescent, adult)

Created metric data presented (used for AI/ML processing)

- Non-normalized data showed a number of expected trends (and are also visible after normalization)
 - Infants have significantly smaller IBIs / higher iHRs: top row
- Normalized data showed potential for some metrics to group ages into 2 or 3 groups by our created 'expectation' Z metric
 - Low Count represents the percentage of time bin windows that had IBIs $< X$ (initially we chose X to be 2 representing a change of < 3 standard deviations)
 - High Count represents the percentage of time bin windows that had IBIs $> X$ (initially we chose X to be 4 representing a change of > 5 standard deviations)



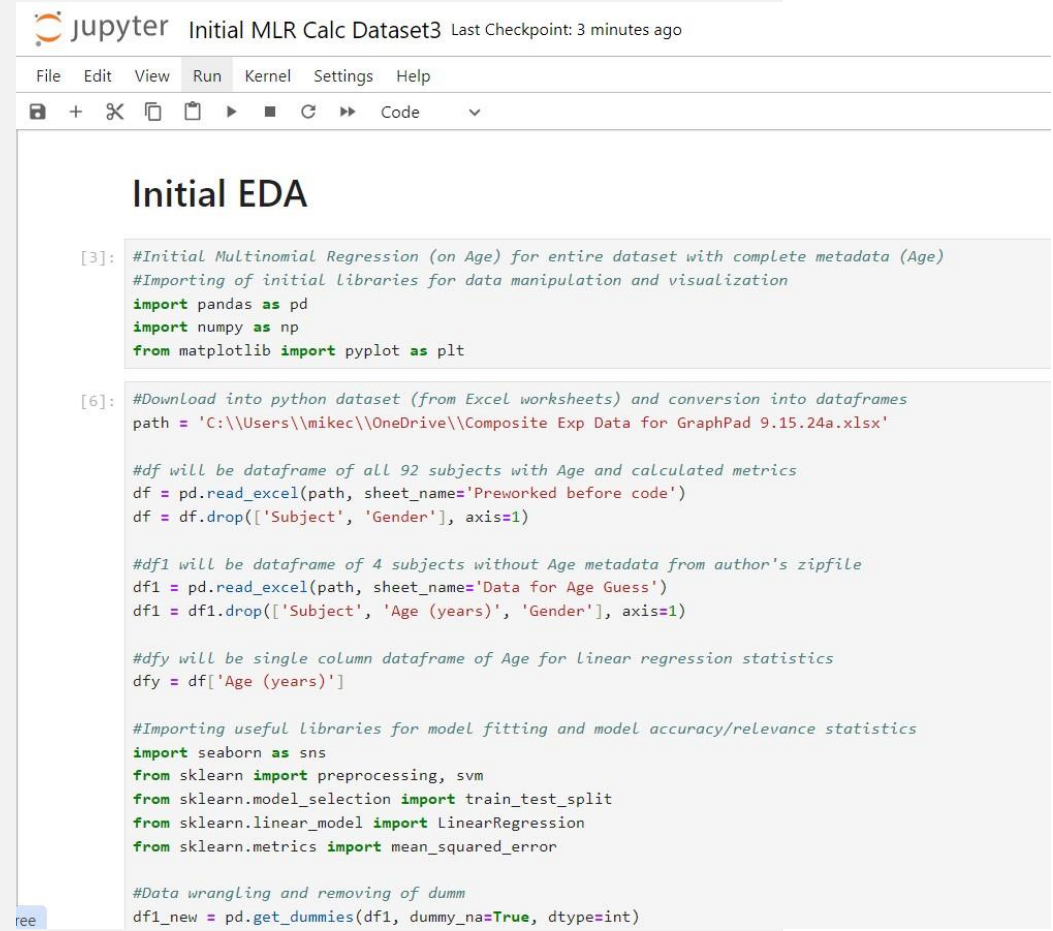
LINEAR REGRESSION ANALYSIS

Naive linear regression using 32 features of 91 Subjects with full metadata (age)

Selected Python Code

Initially I chose a supervised multiple linear regression model based on age as the dependent variable

- Initial model will use all 91 subjects and is easy to understand and has significant literature on model benefits and limitations
- Full code uploaded to GitHub here:
<https://github.com/mrc7mrc49/HRV-Public-DB-Irirzun>



The screenshot shows a Jupyter Notebook titled "Initial MLR Calc Dataset3" with a last checkpoint of 3 minutes ago. The interface includes a menu bar (File, Edit, View, Run, Kernel, Settings, Help) and a toolbar with icons for saving, adding, deleting, and running code. The notebook content is titled "Initial EDA" and contains two code cells. The first cell, labeled [3], imports pandas, numpy, and matplotlib. The second cell, labeled [6], downloads an Excel dataset, reads it into a DataFrame, drops unnecessary columns, reads a second dataset, and imports libraries for model fitting and statistics.

```
[3]: #Initial Multinomial Regression (on Age) for entire dataset with complete metadata (Age)
#Importing of initial libraries for data manipulation and visualization
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt

[6]: #Download into python dataset (from Excel worksheets) and conversion into dataframes
path = 'C:\\Users\\mikec\\OneDrive\\Composite Exp Data for GraphPad 9.15.24a.xlsx'

#df will be dataframe of all 92 subjects with Age and calculated metrics
df = pd.read_excel(path, sheet_name='Peworked before code')
df = df.drop(['Subject', 'Gender'], axis=1)

#df1 will be dataframe of 4 subjects without Age metadata from author's zipfile
df1 = pd.read_excel(path, sheet_name='Data for Age Guess')
df1 = df1.drop(['Subject', 'Age (years)', 'Gender'], axis=1)

#dfy will be single column dataframe of Age for linear regression statistics
dfy = df['Age (years)']

#Importing useful libraries for model fitting and model accuracy/relevance statistics
import seaborn as sns
from sklearn import preprocessing, svm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

#Data wrangling and removing of dumm
df1_new = pd.get_dummies(df1, dummy_na=True, dtype=int)
```

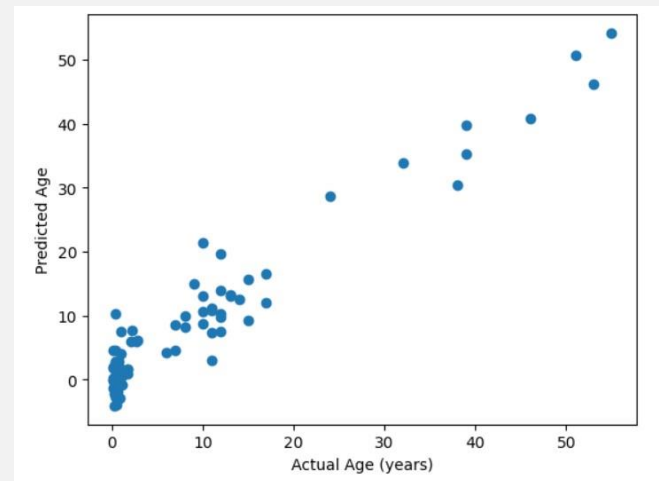
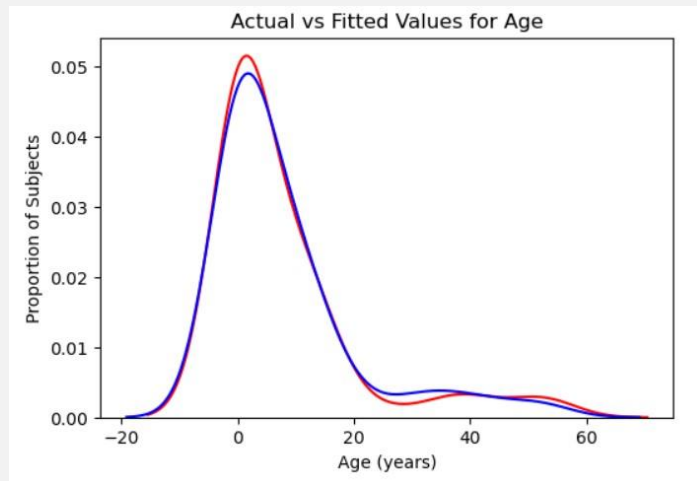

LINEAR REGRESSION ANALYSIS

Naive linear regression using 32 features of 91 Subjects with full metadata (age)

Initial Linear Regression Model

Good agreement between model and data for Ages over 1:

- $R^2 = 0.93$ (Predicted naïve ages were: 0.86, 4.59, -2.98 and -118)
- As code was not forced to y-intercept of 0, some ages predicted to be <0 (I chose to repeat the model with forced 0 y-intercept; that model has $R^2 = 0.93$)



```
yhat_t1 = lm.predict(df1_new1)
yhat_t2 = lm.predict(df1_new2)
yhat_t3 = lm.predict(df1_new3)
yhat_t4 = lm.predict(df1_new4)

print("Predicted naive dataset Ages from our model are: %.2f" % yhat_t1[0], " , %.2f" % yhat_t2[0], " , %.2f" % yhat_t3[0], " , %.2f" % yhat_t4[0])

Predicted naive dataset Ages from our model are: 0.86 , 4.59 , -2.98 , -118.37
```

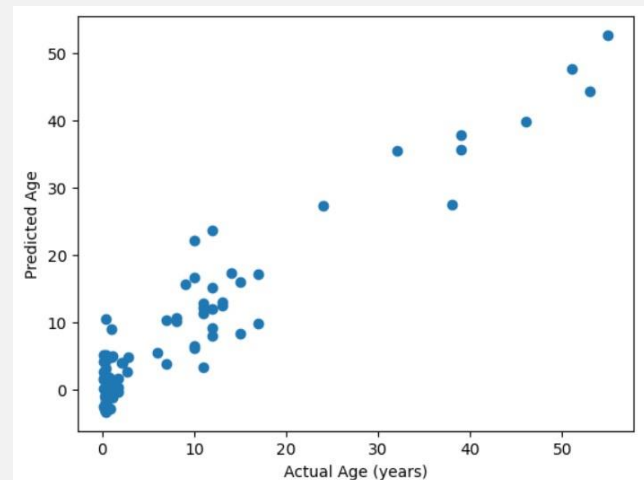
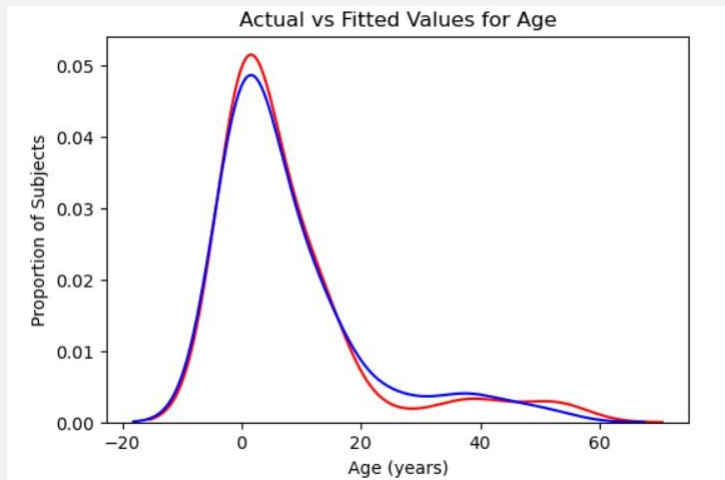
LINEAR REGRESSION ANALYSIS

Naive linear regression using 32 features of 91 Subjects with full metadata (age)

Fixed Linear Regression Model

Good agreement between model and data for Ages over 1:

- $R^2 = 0.91$ (Predicted naïve ages were: 0.61, 6.03, 1.96 and -103)
- As code was not forced to y-intercept of 0, some ages predicted to be <0 (I chose to repeat the model with forced 0 y-intercept; that model has $R^2 = 0.93$)



```
#Created single subject predictions from naive dataset
yhat_t1 = lmf.predict(df1_new1)
yhat_t2 = lmf.predict(df1_new2)
yhat_t3 = lmf.predict(df1_new3)
yhat_t4 = lmf.predict(df1_new4)

print("Predicted naive dataset Ages from our model are: %.2f" % yhat_t1[0], " , %.2f" % yhat_t2[0], " , %.2f" % yhat_t3[0], " , %.2f" % yhat_t4[0])

Predicted naive dataset Ages from our model are: 0.61 , 6.03 , 1.96 , -103.03
```

LINEAR REGRESSION ANALYSIS - COEFFICIENTS

Coefficients of our 2 models show conserved 'high importance' coefficients

Free Linear Regression Model

Relative ranking of most important features (Top 6):

- “No Change Count Z[SDNN, 512]”, “No Change Count Z[rMSSD, 512]”/” No Change Count Z[rMSSD]256”, “No Change Count Z[SDNN]256”, “Low Count Z[SDNN]256”, “Low Count Z<2 [SDNN, 512]”, “CoV Avg IBI”
- Avg IBI and Avg HR not among Top 6 is somewhat of a surprise given the strong, consistent relationship of infants possessing Avg HRs or ~2x adults

Fixed Linear Regression Model

Relative ranking of most important features (Top 6):

- “No Change Count Z[SDNN, 512]”, “No Change Count Z[rMSSD, 512]”/” No Change Count Z[rMSSD]256”, “No Change Count Z[SDNN]256”, “Low Count Z[SDNN]256”, “Low Count Z<2 [SDNN, 512]”, “CoV IBI256”

Removing IBI/HR for both models:

$R^2 = 0.91$ and $R^2 = 0.86$. Predicted Ages: 2.73, 1.89, 0.74, -72 and -4, 9.17, 1.86, -174.

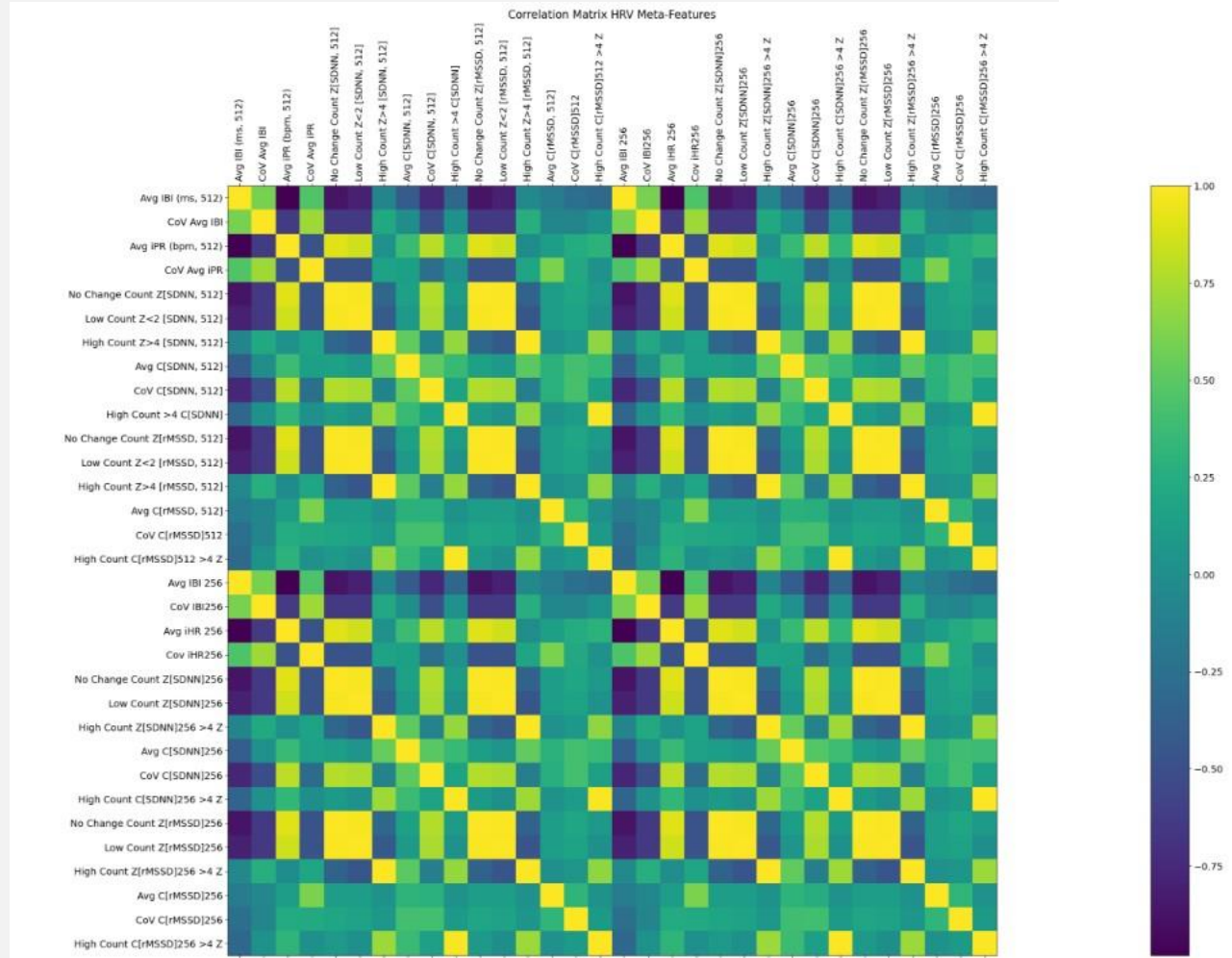
- Relative ranking of most important features did not introduce new features in Top 6

TIME DOMAIN COMPOSITE DATA CORRELATIONS

Which features show strong correlation to perhaps fine tune the model

Many Expected Correlations (post-hoc)

- 512 and 256 binned feature often correlated close to 1.0
 - Mostly expected, but worth double checking given long time scale, infant $\sim 2\times$ HR and potentially shorter time constants of physiological processes
- Most of Top 6 highly correlated
 - “No Change Count Z[rMSSD, 512]”/”No Change Count Z[rMSSD]256” correlated 1.0
 - “No Change Count Z[SDNN, 512]” and “No Change Count Z[rMSSD, 512]”/”No Change Count Z[rMSSD]256” correlated ~ 1.0
 - “No Change Count Z[SDNN, 512]” and “Low Count Z[SDNN]256” correlated ~ 0.98
 - “No Change Count Z[SDNN, 512]” and “Low Count Z<2 [SDNN, 512]” correlated to 0.98
 - “No Change Count Z[SDNN, 512]” and “CoV Avg IBI” correlated to -0.64



TIME DOMAIN COMPOSITE DATA CORRELATIONS

Which features show strong correlation to perhaps fine tune the model

What about less correlated Top Features?

- Keeping Two:
 - “No Change Count Z[SDNN, 512]”
 - “CoV Avg IBI”
- Others not strongly ($< |0.7|$) correlated with Top Two:
 - “High Count Z[SDNN]256 >4 Z” (-0.32 / 0.22)
 - “CoV C[rMSSD]512” (0.19 / 0.21 / 0.08)
 - “High Count C[rMSSD]512 >4 Z” (0.08 / 0.04 / 0.67 / 0.1)

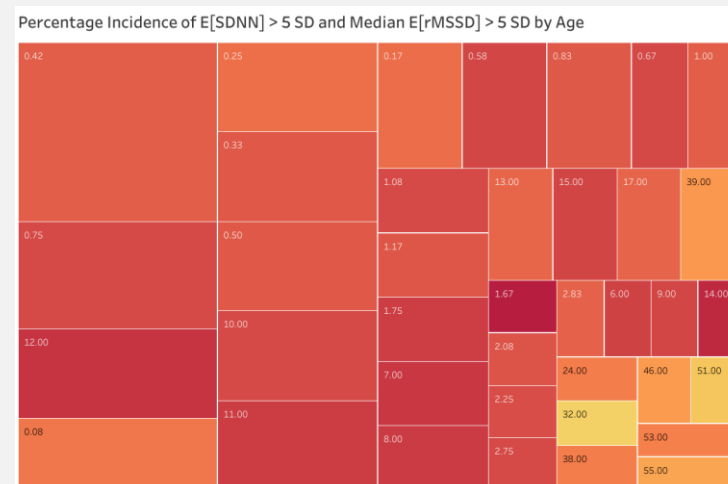
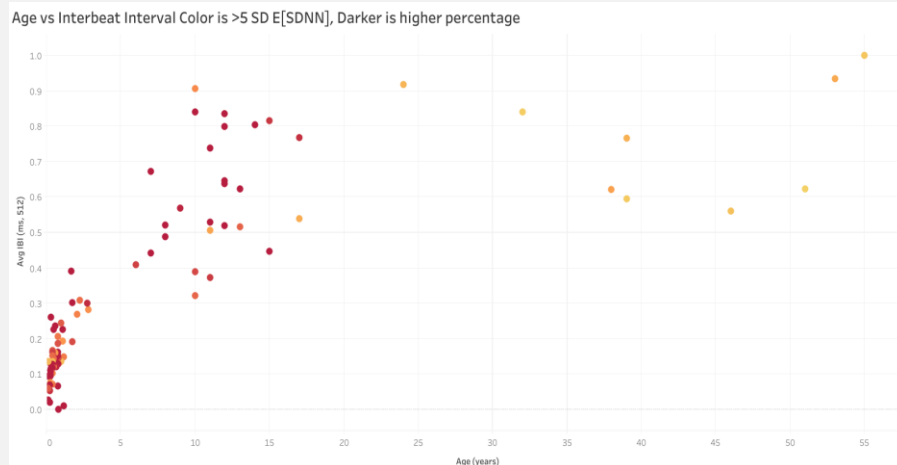
	No Change Count Z[SDNN, 512]	CoV Avg IBI	High Count Z[SDNN]256 >4 Z	CoV C[rMSSD]512	High Count C[rMSSD]512 >4 Z
No Change Count Z[SDNN, 512]	1.00	-0.64	-0.32	0.19	0.08
CoV Avg IBI	-0.64	1.00	0.22	-0.08	0.04
High Count Z[SDNN]256 >4 Z	-0.32	0.22	1.00	0.08	0.67
CoV C[rMSSD]512	0.19	-0.08	0.08	1.00	0.10
High Count C[rMSSD]512 >4 Z	0.08	0.04	0.67	0.10	1.00

INTERIM CONCLUSION AND CONCERNS

High R values for 3 different models suggest a high probability of overfitting

24-hour recordings may have significant challenges when comparing much shorter recording times

- As most of our most informative features involved ratios of percentages, this could strongly be influenced by recording time
- We have no knowledge of the amount or type of activities performed during the recording periods and thus, shorter time periods around specific activities or even positions of the body, could compromise the value of these models as predictors of age
- Is Age really an interesting marker at all
 - Teaser for later PPG-data is: Age might be an interesting marker of a type of 'health'...more to come



The background is an abstract, textured surface with a variety of colors including red, orange, yellow, green, blue, and purple. It has a cracked, marbled appearance. A dark, semi-transparent rectangular overlay covers the bottom half of the image, providing a background for the text.

UNSUPERVISED MODELS

Initial Public Dataset

K-MEANS ANALYSIS OF HRV DATA

Unsupervised models for comparison with described data provided unexpected potential underlying features

Using k-means modelling the public dataset (Irirzun, et al, n = 91)

Initial clustering used the entire dataset to see if infants formed a distinct cluster (later created model with only Age > 3 for comparison with other datasets using mobile subjects)

- Elbow Method showed 3 or 4 groups provided best information gain vs overfitting concerns (I show $k = 4$)
- Ages ≥ 7 were in 2 distinct groups (with one exception Age 10) from Ages < 7 (2 distinct groups)
 - Within those age ranges Adults ≥ 24 didn't show defined difference in grouping compared to Adolescents (Ages ≥ 7 and < 24)

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler

X = np.nan_to_num(X)
Xscale = MinMaxScaler().fit_transform(X)
Xscale

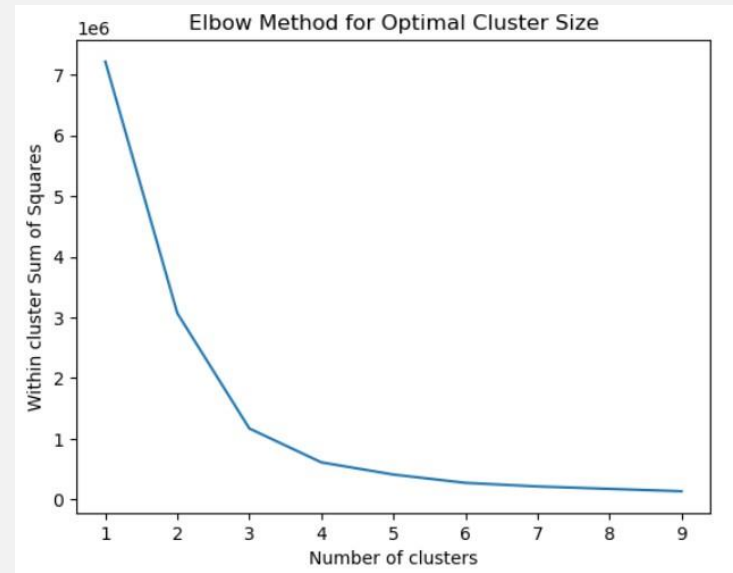
array([[0.93532013, 0.3366115 , 0.03058763, ..., 0.00653661, 0.02633328,
        0.86692963],
       [0.7679667 , 0.39551689, 0.11313647, ..., 0.00801747, 0.00624515,
        0.99699853],
       [0.55941865, 0.11955678, 0.22717258, ..., 0.00421398, 0.05732277,
        0.52718031],
       ...,
       [0.09463731, 0.04814777, 0.78624648, ..., 0.00609089, 0.03497125,
        0.99971296],
       [0.1105282 , 0.16824332, 0.7656001 , ..., 0.0109056 , 0.03855272,
        0.99861058],
       [0.13587704, 0.27528736, 0.72549346, ..., 0.01346948, 0.04546221,
        0.82262066]])

clusterNum = 4
k_means = KMeans(init = "k-means++", n_clusters = clusterNum, n_init = 12)
k_means.fit(Xscale)
labels = k_means.labels_
print(labels)

C:\Users\mikec\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1446: UserWarning: There are less chunks than available threads. You can avoid it by setting the
warnings.warn(

[[3 2 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 0 2 2 2 2 2 1 0
 0 1 1 1 1 1 0 1 0 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 0 1 1 0 0 0 1 0 1 1
 1 0 2 1 0 0 0 0 0 1 0 1 0 0 1 1 1]]

df1r["Clus_km"] = labels
df1r.head(5)
```



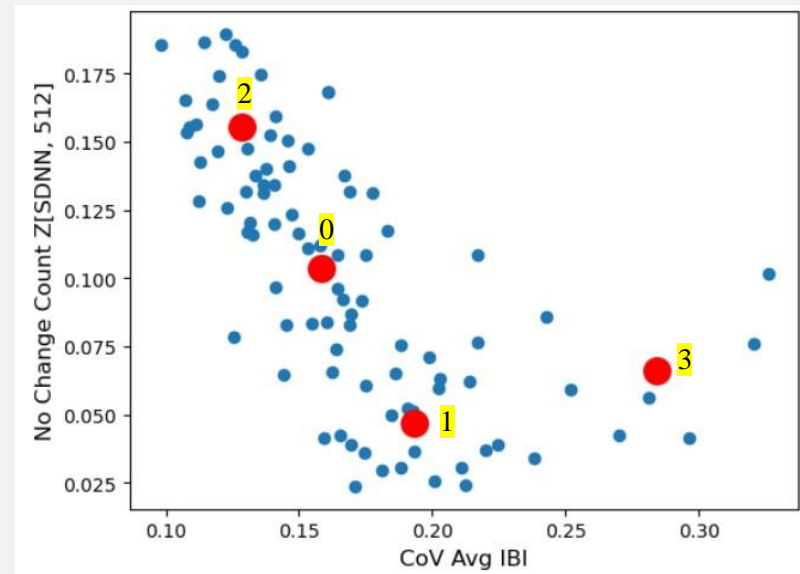
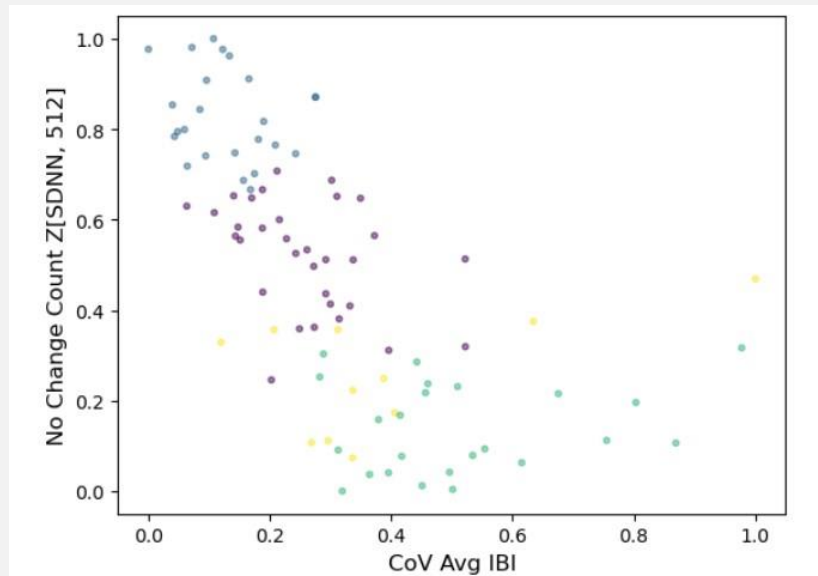
K-MEANS ANALYSIS OF HRV DATA

Unsupervised models for comparison with described data provided unexpected potential underlying features

Using k-means modelling the public dataset (Irirzun, et al)

We evaluated modelling of dataset with $k = 4$ clusters ($n = 35$ subjects):

- Older subjects consisting of Adolescents (Age 4 – 16, $n = 24$) and Adults (Age 17 – 55, $n = 11$) were grouped into 2 groups
 - Using StandardScaler normalization (MinMaxScaler produced similar distinct groups)
 - Adults and Adolescents were clustered around 3, 0 and Infants clustered around 1, 2
 - Using $k = 3$ Infants remained a mostly (87-95%) distinct cluster



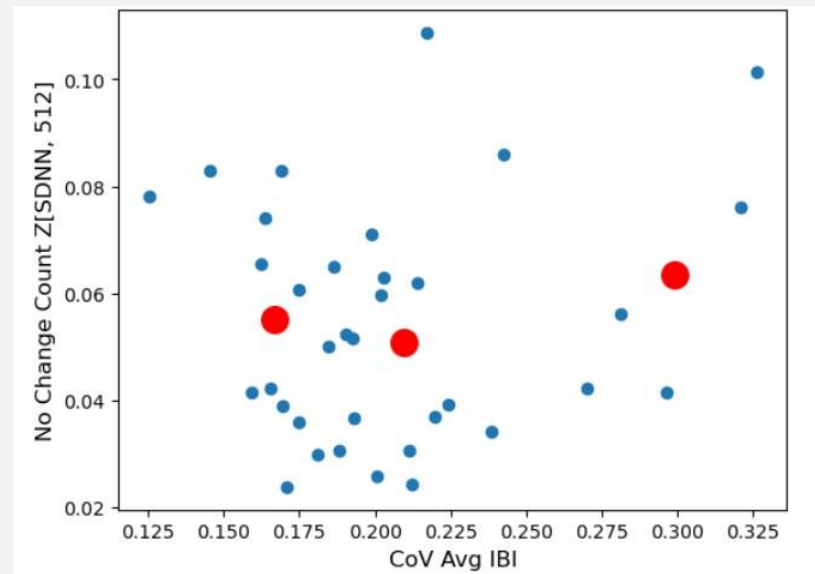
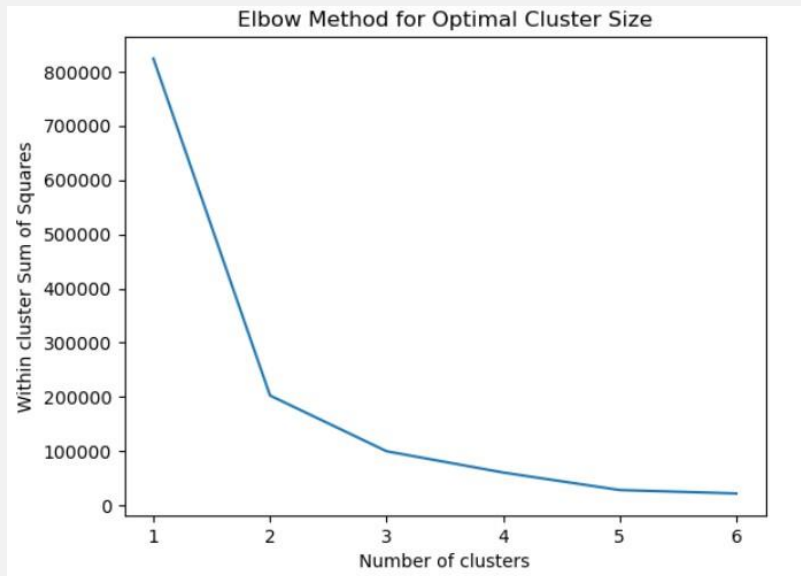
K-MEANS ANALYSIS OF HRV DATA – NO INFANTS

Unsupervised models for comparison with described data provided unexpected potential underlying features

Using k-means modelling the public dataset (Irizun, et al, n = 35)

Reasonable arguments about the inability of infants to perform certain tasks that increase/decrease heart rate rapidly, led to us choosing to exclude for this analysis

- Most datasets do not include infants and our models built off of an adolescent and older age range allows for comparison with other data
- Three clusters provided some unexpected support for supervised models to determine features that impact an Age prediction although less compelling than the cluster analysis for infant, adolescent and adults



The background is an abstract, textured surface with a variety of colors including red, yellow, green, blue, and purple. A dark, semi-transparent overlay covers the entire image, creating a moody atmosphere. The main title is positioned in the lower-left quadrant of this overlay.

MODEL TRANSFERABILITY

*Using PPG-derived
data in our model*

LINEAR REGRESSION ANALYSIS - APPLIED

PPG data often considered as exactly equivalent (magnitude, time course, effect) as HRV data

Using our Linear Regression Model trained on HRV data from EKGs

PPG-data consists of episodic epochs (~10-15 min for a recording period, 4-6 in a day):

- Predicted ages from two subjects with epochs 'stitched together' with linear interpolation showed low accuracy: predicted ages > 200% inaccurate
- Native dataset (outreach to authors ongoing) shows greater potential for reasonable age predictions
 - Ages not included in online file, potential to discover by author communication

```
[39]: #Created single subject predictions from naive dataset
yhat_t1 = lm_ni.predict(df1_td1)
yhat_t2 = lm_ni.predict(df1_td2)

print("Predicted naive dataset Ages from our model are: %.2f" % yhat_t1[0], " , %.2f" % yhat_t2[0])

Predicted naive dataset Ages from our model are: 152.62 , -482.01

[27]: #Created single subject predictions from naive dataset
yhat_t1ni = lm_ni.predict(df1_new1)
yhat_t2ni = lm_ni.predict(df1_new2)
yhat_t3ni = lm_ni.predict(df1_new3)
yhat_t4ni = lm_ni.predict(df1_new4)

print("Predicted naive dataset Ages from our model are: %.2f" % yhat_t1ni[0], " , %.2f" % yhat_t2ni[0], " , %.2f" % yhat_t3ni[0], " , %.2f" % yhat_t4ni[0])

Predicted naive dataset Ages from our model are: -18.08 , 384.54 , 0.21 , 162.14
```

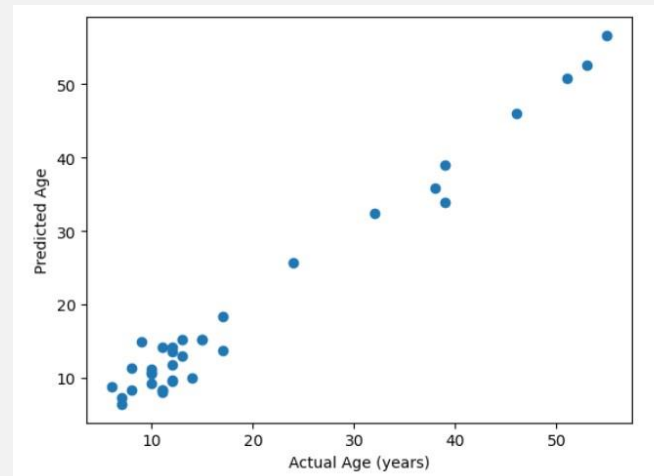
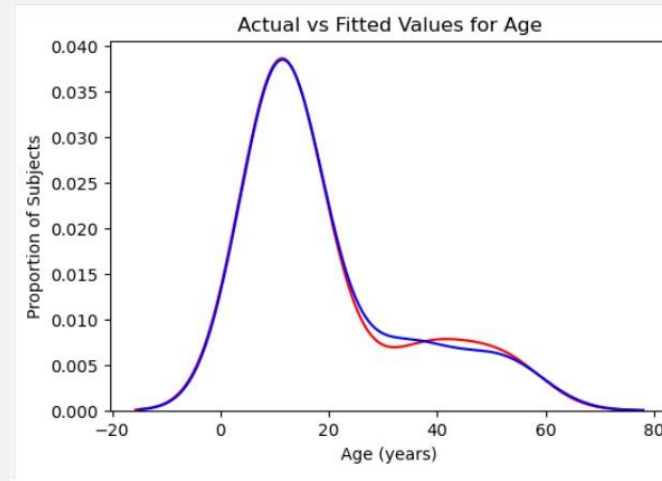

LINEAR REGRESSION ANALYSIS – AGE RESTRICTED

A critique of including infant HRV (61% of dataset) centers on lack of potential rapid high magnitude changes

Removing infants (Age < 3)
from data allows for comparison
of subjects able to stand, walk,
exercise (and recover)

Infants may be unable to have rapid
changes to HR given physical limitations
until upright standing is possible

- Infant data trace shows
- Irirzun described the recording for all subjects as (check Methods)
- Adolescent and Adults are a small dataset (n = 35)



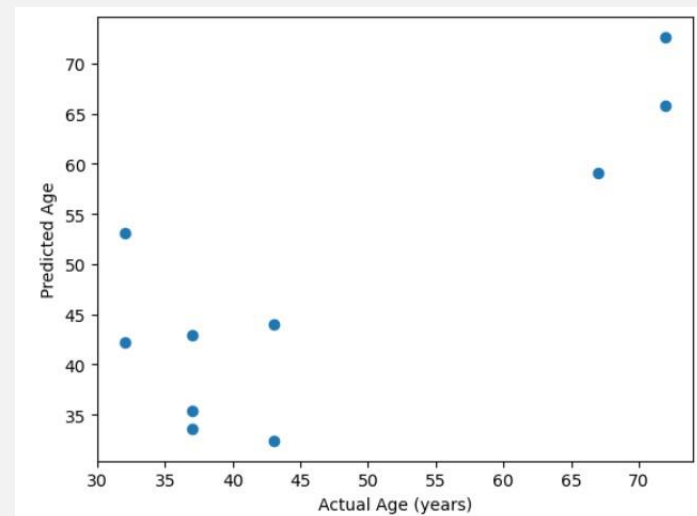
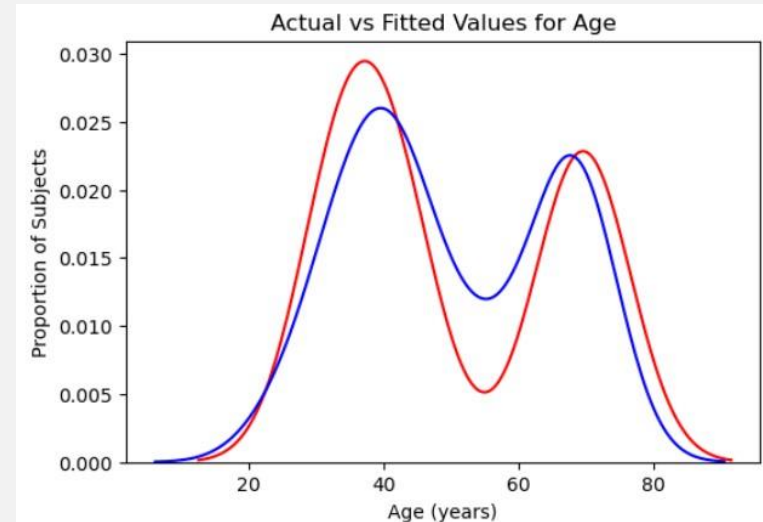
LINEAR REGRESSION ANALYSIS – PPG DATA

Five subjects performed repeated (10 days) measurements in the morning prior to any daily activities

Subjects performed recordings of ~10-12 minutes involving both seated and lying down positions

Order of position chosen semi-random such that 5 days began in seated and 5 began in lying down positions

- Previous experiments with HRV and PPG have utilized different positions with arguments that the position affects recording
- Seated positions associated with higher HR, shorter IBI, and changes to time constant of underlying physiological processes



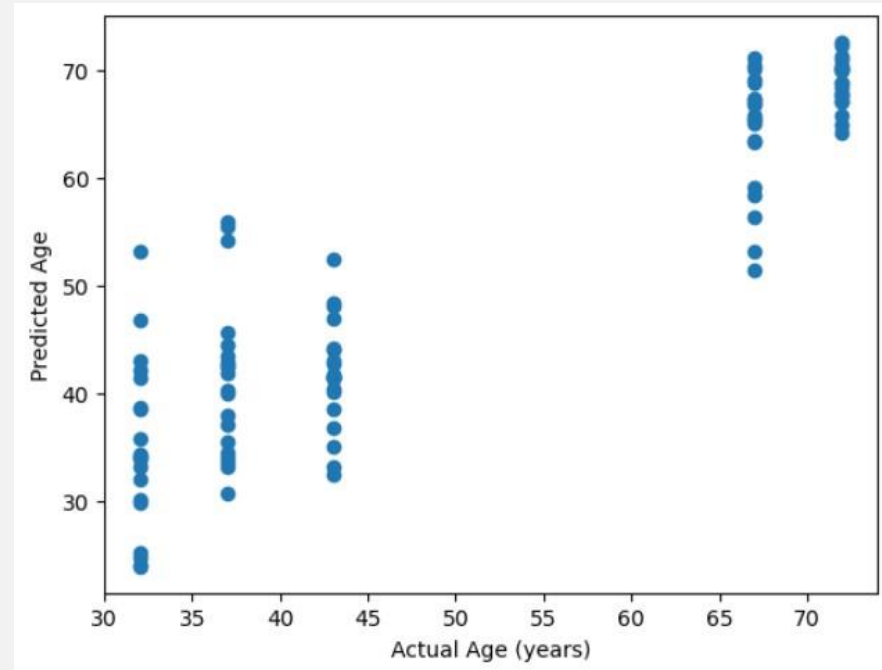
LINEAR REGRESSION ANALYSIS – PPG DATA

Repeated measurements showed variation that may reflect different underlying states (and certainly some noise)

Over relatively short time (< 5 months) subjects showed variation in morning PRV that appears age-dependent

Age range predicted for individual subjects shows more variation at younger ages

- Younger ages also show lower predicted possible 'ages'
- Potential usage as a 'Cardiovascular Age Marker' do define health states and response to treatment interventions
 - We apply next to days of different types of exercise, recovery, travel, stimulant consumption and activities of daily living (ADLs)



KNN ANALYSIS OF RECORDING POSITION

PPG recorded multiple times by 5 subjects in a seated (5 per subject) or lying down (5 per subject) position

Using KNN we first explored different K values for best accuracy

Test set of 0.10 to reduce chance of only including a single subject

- K of 6 was considered best accuracy (0.8)

```
[123]: yhat = neigh.predict(X_test_norm)
       yhat

[123]: array(['Lying', 'Lying', 'Lying', 'Lying', 'Sitting', 'Lying', 'Lying',
       'Lying', 'Lying', 'Sitting'], dtype=object)

[125]: comparison = [y_test, yhat]
       comparison

[125]: [array(['Lying', 'Lying', 'Sitting', 'Sitting', 'Sitting', 'Lying',
       'Lying', 'Lying', 'Lying', 'Sitting'], dtype=object),
       array(['Lying', 'Lying', 'Lying', 'Lying', 'Sitting', 'Lying', 'Lying',
       'Lying', 'Lying', 'Sitting'], dtype=object)]

[127]: from sklearn import metrics
       print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh.predict(X_train_norm)))
       print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))

Train set Accuracy: 0.6777777777777778
Test set Accuracy: 0.8

[113]: Ks = 10
       mean_acc = np.zeros((Ks-1))
       std_acc = np.zeros((Ks-1))

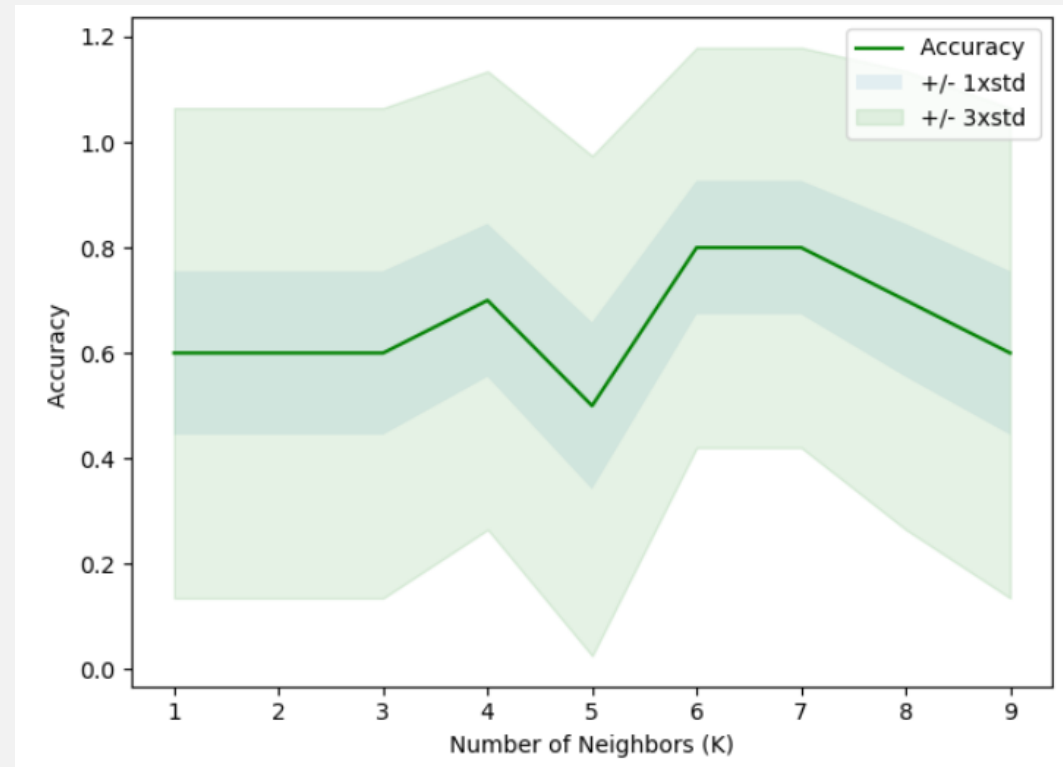
       for n in range(1,Ks):

           #Train Model and Predict
           neigh = KNeighborsClassifier(n_neighbors = n).fit(X_train_norm,y_train)
           yhat=neigh.predict(X_test_norm)
           mean_acc[n-1] = metrics.accuracy_score(y_test, yhat)

           std_acc[n-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])

       mean_acc

[113]: array([0.6, 0.6, 0.6, 0.7, 0.5, 0.8, 0.8, 0.7, 0.6])
```



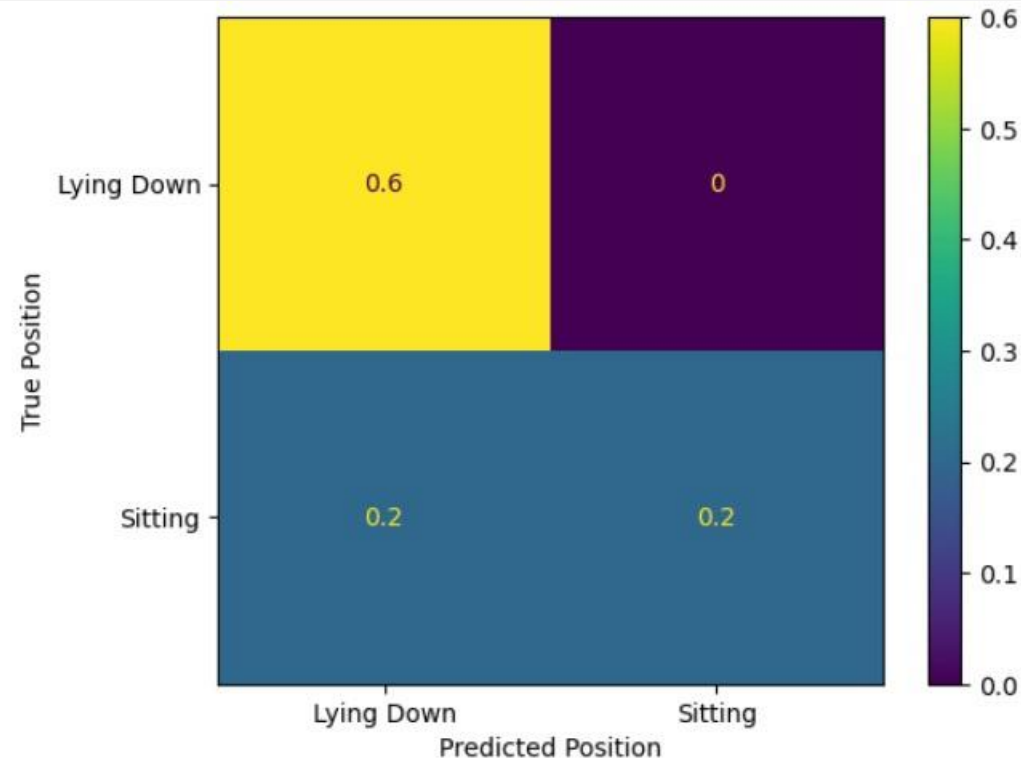
KNN ANALYSIS OF RECORDING POSITION

PPG recorded multiple times by 5 subjects in a seated (5 per subject) or lying down (5 per subject) position

Using KNN we first explored different K values for best accuracy

Accuracy was better for the Lying Down position compared to the Seated position

- May relate to underlying variation in Avg iHR where 38% of the recorded pairs (days) had unexpected higher Avg iHR for the seated position than the lying down position
- May relate to lack of high frequency data from the technological recording (used a low-cost PPG-based system that has the benefit of .csv output, but only to single digit precision of Avg iHR)
 - At low HRs this is a large source of error for short time constant / high frequency changes



DECISION TREE ANALYSIS PPG-DATA

Can we find a model to predict physical position (sitting vs lying down) with PPG-data

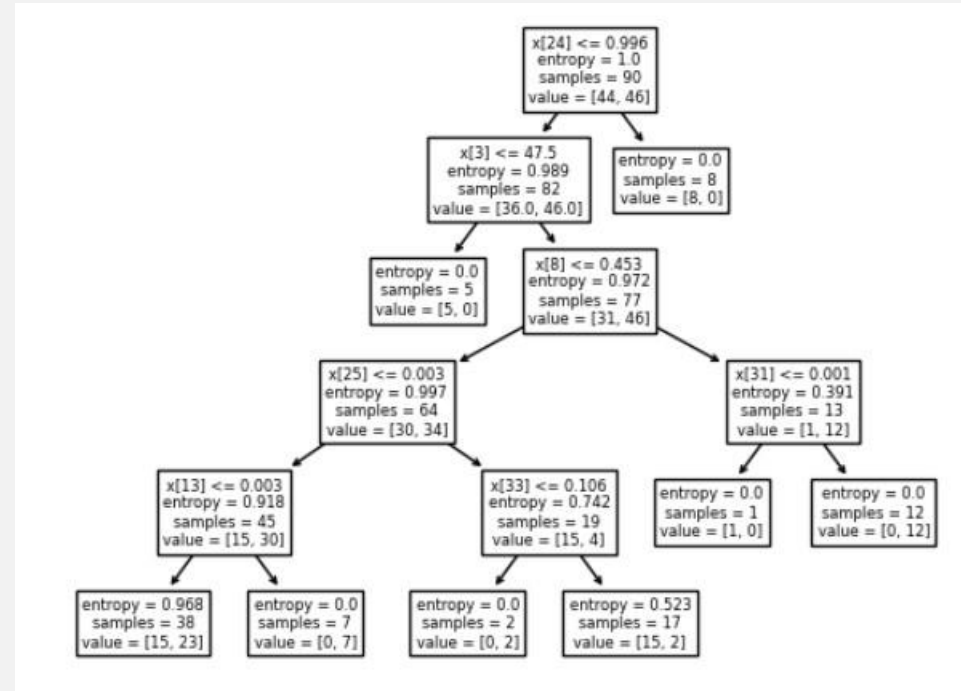
Short answer: No.

Various combinations of test size and layers (max depth)

- Best model found had accuracy of 0.6

Data collected noisy and missing high frequency components

- Avg HR difference ~1-2 beats higher for seated but had enough variation to create challenging prediction noise
- Low Avg HRs will miss more high frequency changes in sequential beats that might be necessary to fine tune model
- Seeking HRV public data with physical position categorization and changes under steady-state conditions to evaluate a general model



The background is an abstract, textured surface with a variety of colors including red, orange, yellow, green, blue, and purple. The texture appears to be like cracked paint or a marbled surface. A dark, semi-transparent overlay covers the entire image, making the colors appear more muted and creating a sense of depth.

MODEL TRANSFERABILITY

*Initial Gen AI
experiments*

INITIAL GEN AI FOR INCREASING CODE IDEATION

I used Bard and ChatGPT to derive code samples to check utility of polynomial and Ridge regression

Removing infants (Age < 3) from data allows

Dilemma of a now very small dataset ($n = 35$) with large gaps in ages untested originally

- Middle age very underrepresented (ie 40s)
- Mostly appears to use a method of creating synthetic data by finding feature coefficients with minimal correlation
 - Our chosen features were found to have some strong correlations between bin times (512 vs 256) and a few other higher order summary data

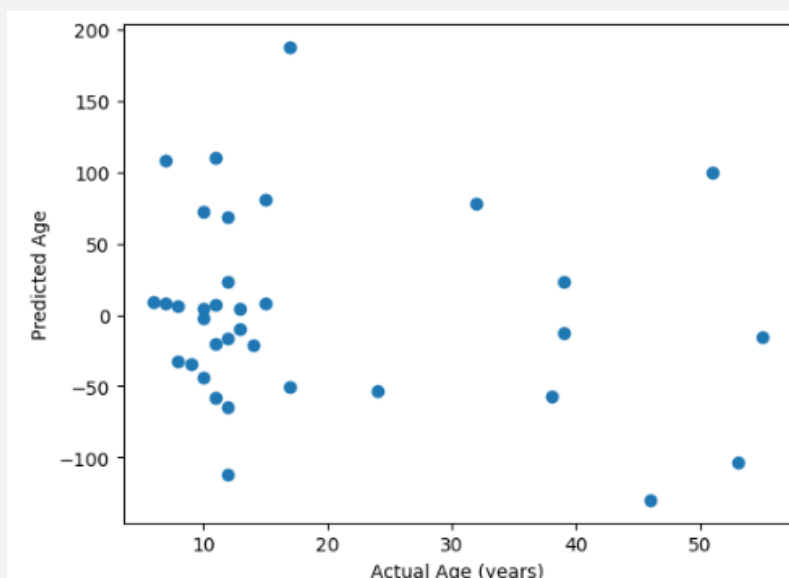
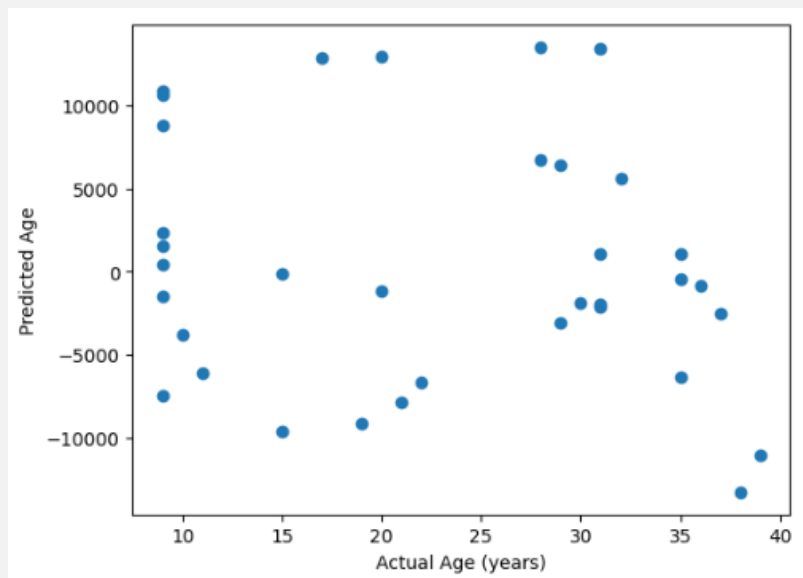
INITIAL GEN AI FOR INCREASING DATASET SIZE

Used Mostly.AI as a hub for initial attempts at increasing dataset

Removing infants (Age < 3) from data allows

Dilemma of a now very small dataset (n = 35) with large gaps in ages untested originally

- Middle age very underrepresented (ie 40s)
- Mostly appears to use a method of creating synthetic data by finding feature coefficients with minimal correlation
 - Our chosen features were found to have some strong correlations between bin times (512 vs 256) and a few other higher order summary data



	Avg IBI (ms, 512)	CoV Avg IBI	Avg IPR (bpm, 512)	CoV IPR	No Change Count Z(SDNN, 512)	Low Count Z=2 (SDNN, 512)	High Count Z=4 (SDNN, 512)	Avg C(SDNN, 512)	CoV C(SDNN, 512)	High Count Z=4 C(SDNN)	No Change Count Z(MSSD, 512)	Low Count Z=2 (RMSSD, 512)	High Count Z=4 (RMSSD, 512)	Avg C(RMSSD, 512)	CoV C(RMSSD)512
Avg IBI (ms, 512)	1.00	-0.02	-0.38	-0.12	-0.03	-0.11	0.09	0.10	-0.04	0.19	-0.12	0.15	0.05	0.17	0.16
CoV Avg IBI	-0.02	1.00	-0.08	-0.12	-0.06	0.02	-0.32	0.08	0.03	-0.06	0.16	-0.21	0.16	-0.01	0.25
Avg IPR (bpm, 512)	-0.38	-0.08	1.00	-0.17	0.04	0.23	-0.12	-0.26	0.07	-0.06	0.10	0.07	0.30	-0.35	-0.03
CoV IPR	-0.12	-0.12	-0.17	1.00	-0.37	-0.06	0.07	0.13	-0.07	-0.28	-0.03	-0.16	-0.31	0.03	0.10
No Change Count Z(SDNN, 512)	-0.03	-0.06	0.04	-0.37	1.00	0.18	-0.04	-0.17	0.16	0.25	0.15	0.28	0.12	-0.12	-0.09
Low Count Z=2 (SDNN, 512)	-0.11	0.02	0.23	0.06	0.18	1.00	-0.08	-0.31	-0.09	0.08	-0.05	-0.24	0.06	0.07	-0.26
High Count Z=4 (SDNN, 512)	0.09	-0.32	-0.12	0.07	-0.04	-0.08	1.00	-0.04	-0.18	0.19	-0.01	-0.06	-0.16	-0.00	-0.06
Avg C(SDNN, 512)	0.10	0.08	-0.26	0.13	-0.17	-0.31	-0.04	1.00	-0.09	0.07	-0.28	-0.15	-0.11	-0.00	0.29
CoV C(SDNN, 512)	-0.04	0.03	0.07	0.07	0.16	-0.05	0.18	-0.09	1.00	0.16	-0.07	0.25	0.07	-0.10	-0.37
High Count Z=4 C(SDNN)	0.19	-0.06	-0.05	-0.28	0.25	0.08	0.19	0.07	0.16	1.00	-0.10	0.20	0.29	0.18	-0.24
No Change Count Z(MSSD, 512)	-0.12	0.16	0.10	-0.03	0.15	-0.05	-0.01	-0.28	-0.07	-0.10	1.00	0.05	0.08	0.03	-0.32
Low Count Z=2 (RMSSD, 512)	0.15	-0.21	0.07	-0.10	0.28	-0.24	-0.06	-0.15	0.25	0.20	0.05	1.00	0.18	-0.40	-0.11
High Count Z=4 (RMSSD, 512)	0.05	0.16	0.30	-0.31	0.12	0.06	-0.16	-0.11	0.07	0.29	0.08	0.18	1.00	-0.31	0.06
Avg C(RMSSD, 512)	0.17	-0.01	-0.35	0.03	-0.12	0.07	-0.00	-0.00	-0.10	0.18	0.03	-0.40	-0.31	1.00	-0.16

INITIAL GEN AI FOR INCREASING DATASET SIZE

Used Mostly.AI as a hub for initial attempts at increasing dataset based on raw EKG data (and normalized)

Removing infants (Age < 3) from data allows

Dilemma of a now very small dataset ($n = 35$) with large gaps in ages untested originally

- Middle age very underrepresented (ie 40s)
- Mostly appears to use a method of creating synthetic data by finding feature coefficients with minimal correlation
 - Our chosen features were found to have some strong correlations between bin times (512 vs 256) and a few other higher order summary data

TO DO LIST

Not in order

- KNN of seated positions with graphs – Completed 10.24
- Decision trees of seated positions with graphs – Completed 10.24
- Log Regression
- Top 5 Features of PPG linear regression with correlation matrices
- Gen AI with HRV – Completed 10.24
- Gen AI with PPG
- More unsupervised models
- Show run AI code from Bard/ChatGPT for running through multinomial, polynomial (with optimization) and Ridge Regression as double check for previous work
- Change early images to Tableau if possible
- Show IBM DataRobot data and synthetic data from Mostly.ai – 10.24 Mostly.ai done
- Discuss model overfitting in early attempts based on Gen AI code – 10.29 done