Lecture 4: Exploratory Data Analysis

Intro to Data Science for Public Policy
Spring 2017

Jeff Chen + Dan Hammer

Roadmap

- **Revised Syllabus**
- Homework Review
- EDA
- Code-along

Syllabus [Before]

Section 1: Fundamentals

- Lecture 4: Exploratory Data Analysis

Section 2: Models

- Lecture 5: Intro to Supervised Learning
  - Homework #2 assigned
- Lecture 6: Simulation and Bias
- Lecture 7: Causal Inference
- Lecture 8: Classifiers

Syllabus [Revised]

Section 1: Fundamentals

- Lecture 4: Exploratory Data Analysis

Section 2: Models

- Lecture 5: OLS, Simulation and Bias (Dan Hammer)
  - Homework #2 assigned
- Lecture 6: Intro to Supervised Learning
- Lecture 7: Supervised Learning: Classifiers -- #1
- Lecture 8: Supervised Learning: Classifiers -- #2

Roadmap

- Syllabus Change
- **Homework Review**
- EDA
- Code-along

Two rationale for the homework:

- Replication based on instructions
  - Science is only credible if it can be reproduced
- Implement among the most common programming paradigms
  - Programming paradigms are often absent among new data scientists

Roadmap

- Syllabus Change
- Homework Review
- **EDA**
- Code-along

"**Exploratory Data Analysis** is really the process of understanding how you react to data. It is not about the graphs and statistics, but a matter of defining your quantitative identity and process."

- Star Ying

Bayesian Statistician

Concretely, your quantitative identity is a function of how you deal with *separability*, *correlation*, and *usability*. Understanding your biases and propensity to treat the data in a certain way is also part of that identity.

There is no way to understand EDA than to conduct it 10+ times in the wild.

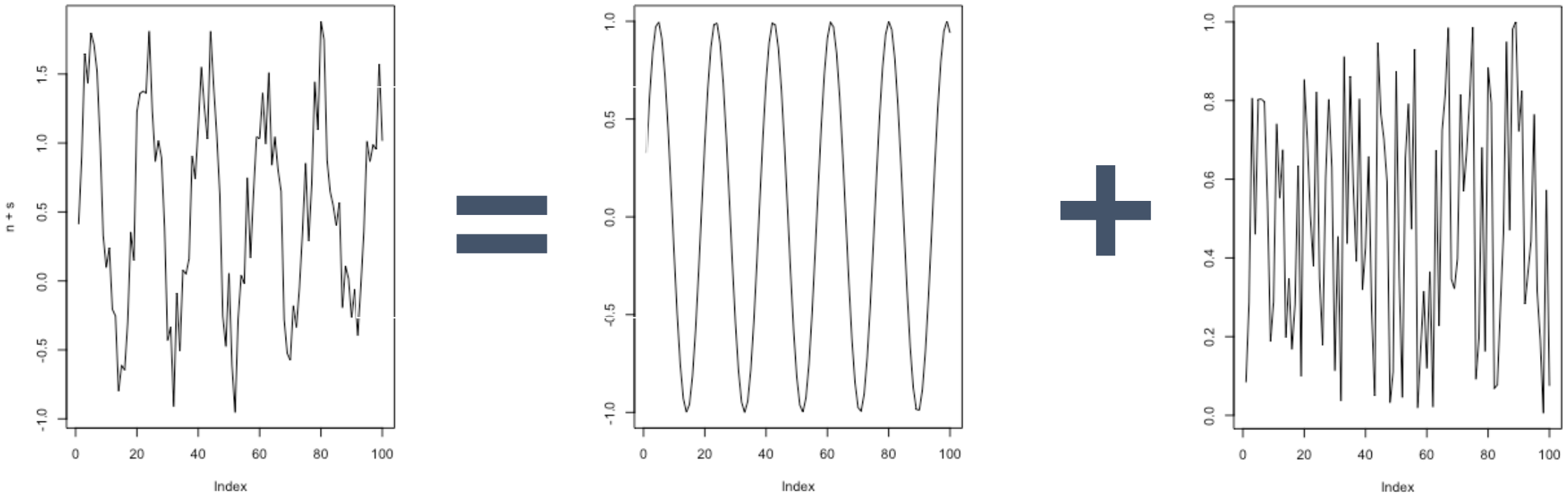
Dimension	Star Ying	Jeff Chen
Statistical School	Bayesian	Frequentist
First thing I do with text data	“What? What is the problem I’m trying to do?	I check to see if the target variable is available. Then, I read the first 5 lines, then proceed to write a vectorize function to parse words by spaces so that I can find most common terms.
First thing I do with numeric data		I check to see if the numeric is numeric, then run a kernel density graph by group and a Kolmogorov-Smirnov test

Usability

- Conversion of raw data into model-enabled information
  - Converting between data formats
  - Understanding missingness of data points
  - Extracting signal from raw data

Usability

- Extracting signal from raw data

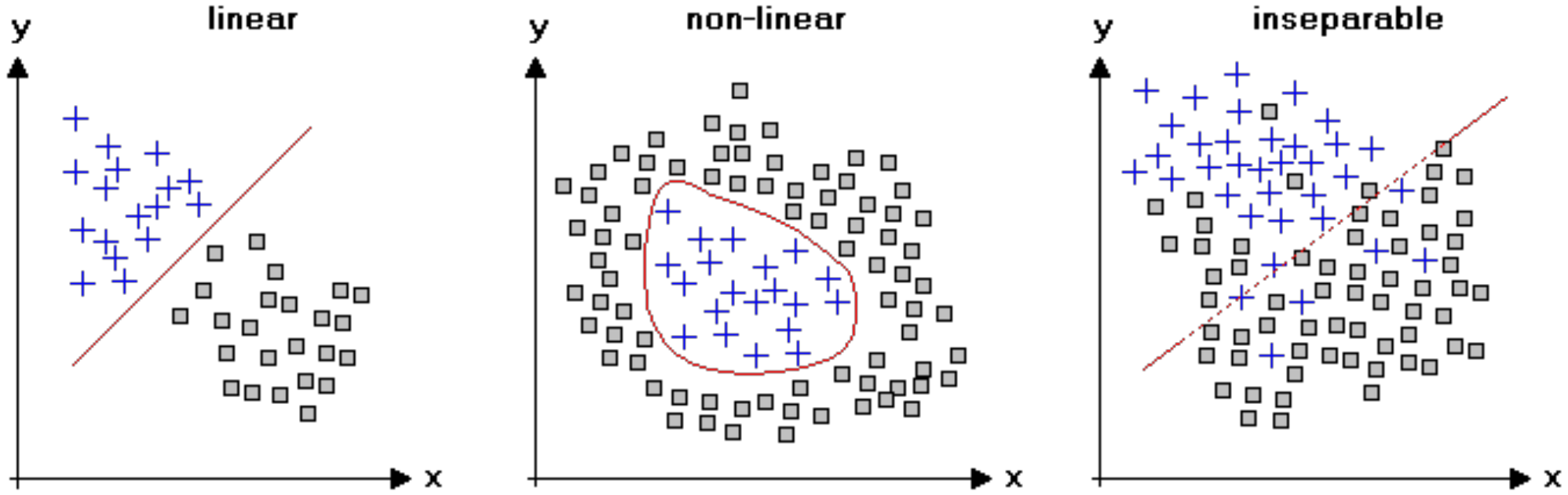


Separability

- The extent to which two or more groups can be distinguished using one or more features
- Idea: if you can see the differences with your eyes, there should be a mathematical solution.

Visual Separability

Graph that shows clear differences in clustering.



Statistical Separability

Continuous

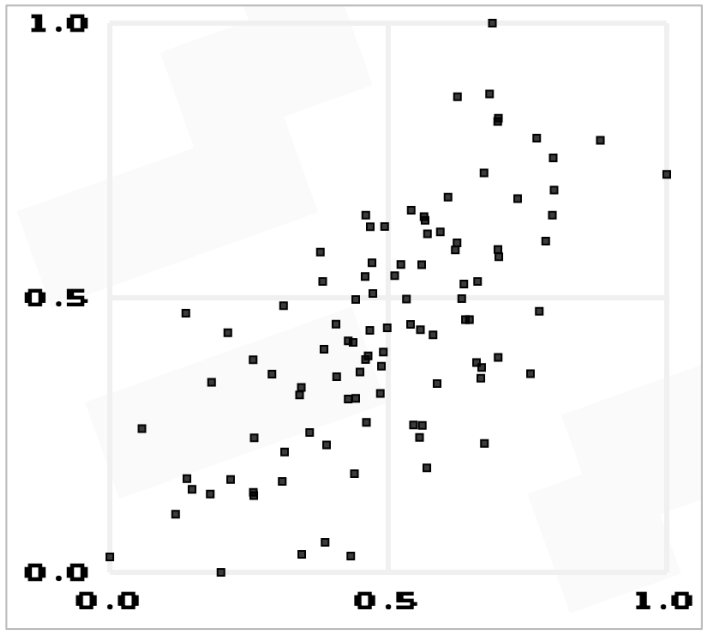
- T-test//Z-test
- Kolmogorov-Smirnov Test

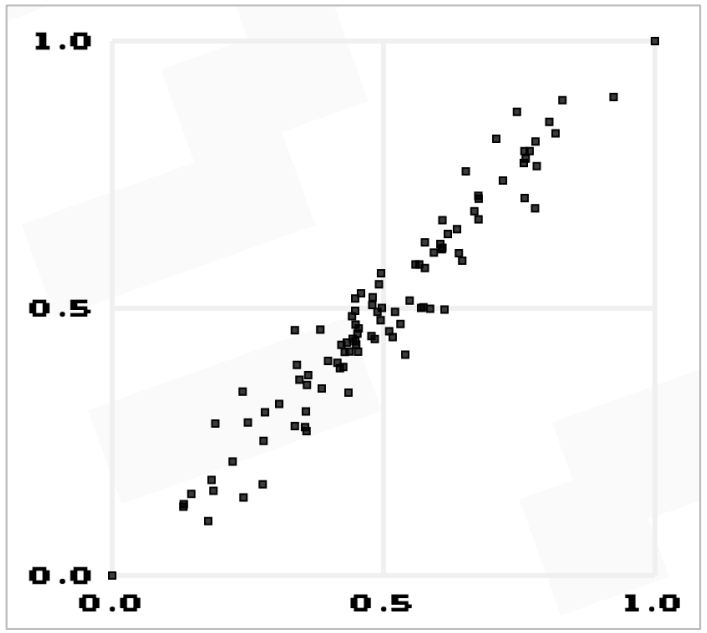
Discrete

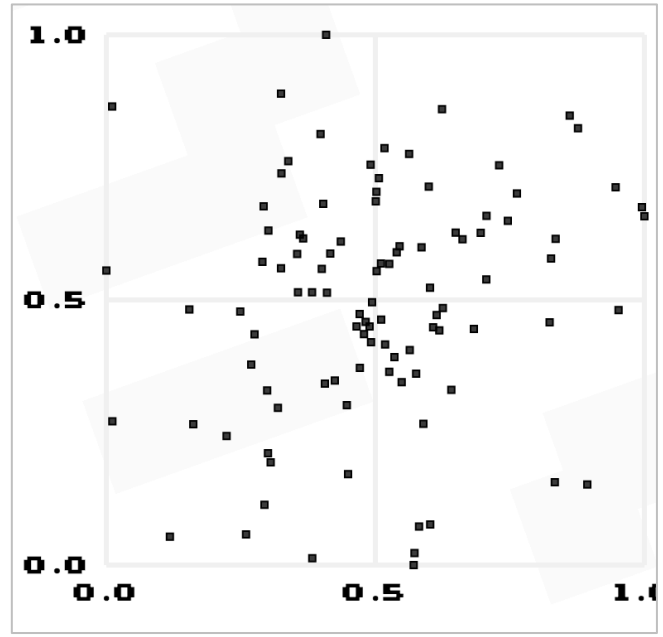
- Chi-squared
- Confusion Matrix

Correlation

- The extent to which two or more quantities are related or connected.







Statistical Correlation

Continuous

- Pearson's Correlation Coefficient
- Cosine Similarity

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Discrete

- Cosine Similarity
- Jaccard Similarity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Visual Correlation (Continuous)

Use scatter plots, train your eyes on:

guessthecorrelation.com

A Framework for EDA

Area	Goal	Common Questions
Usable	Assess the data types	Are the data categorical, numerical, factor, strings, other? What manipulations will you need to perform to get the data into usable shape?
Separable, Correlated	Understand the empirical distributions	Does the data fall into a commonly recognized shape? Is it unimodal, bimodal? Is there any indication of time-dependence?
Usable	Detect outliers, missingness and errors	Are there anomalous values? - Do records spike or occur during odd times? - How complete is the data? - Which variables need to be standardized and cleaned?
Usable	Check the assumptions	How exactly is the data collected? Does the data reflect what would be expected?
Correlated	Identify important variables	Which variables are correlated with one another?

Example

Smartphones have accelerometers that can measure an user's motion. Accelerometers are the basis of activity tracking devices.

Example

Goal is to understand how we can use acceleration readings to characterize physical activity.

Pre-Work: Assess

Goal	Common Questions
Assess the data types	Are the data categorical, numerical, factor, strings, other? What manipulations will you need to perform to get the data into usable shape?

- Convert all variables into the right format
- Reshape data (transform to wide form)
- Filter out noise and/or transform data into derivative form

Assess Raw Dataset

Convert data from long form to analyzable wide form

Time	Type	Value
1	Accel.X	0.068
1	Accel.Y	0.357
1	Accel.Z	0.673
1	Label	Walk
2	Accel.X	0.403
2	Accel.Y	0.476
2	Accel.Z	0.359
2	Label	Walk

Assess Raw Dataset

Convert columns

- Time = as.Date()
- Accel.X = num
- Label = as.character()

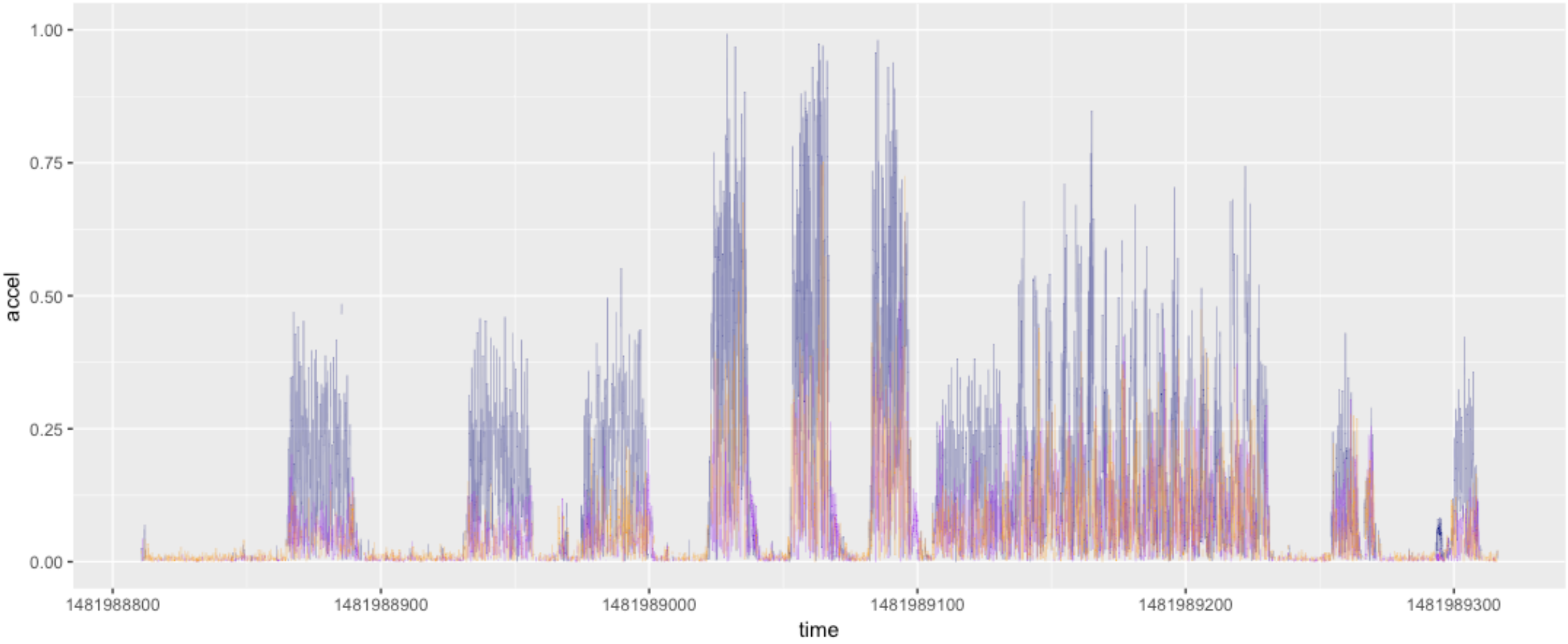
Time	Accel.X	Accel.Y	Accel.Z	Label
1	0.068	0.357	0.673	Walk
2	0.403	0.476	0.359	Walk
3	0.683	0.298	0.186	Walk
4	0.505	0.054	0.753	Walk
5	0.264	0.662	0.240	Walk
6	0.021	0.044	0.017	Stand
7	0.024	0.011	0.001	Stand
8	0.003	0.019	0.042	Stand

#2 -- visualize

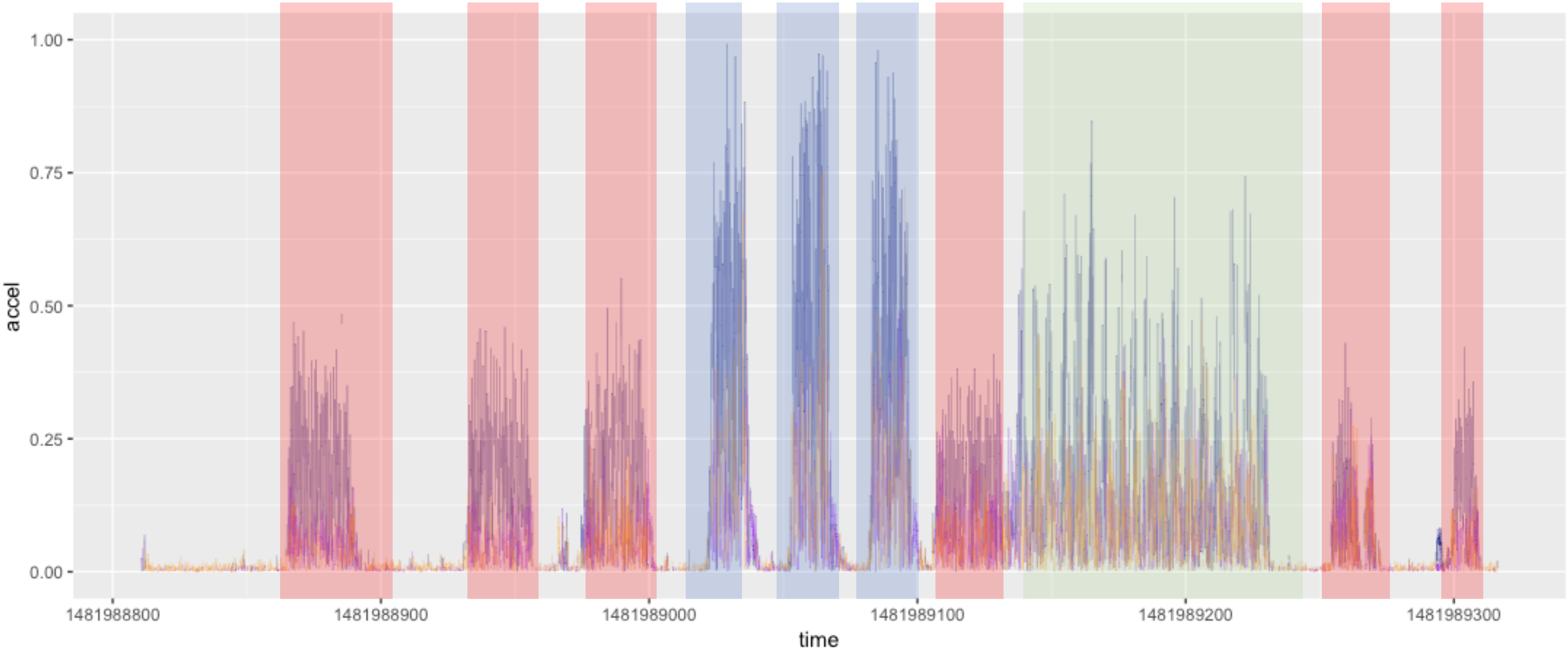
Goal	Common Questions
Understand the empirical distributions	Does the data fall into a commonly recognized shape? Is it unimodal, bimodal? Is there any indication of time-dependence?

```
#How JC approaches a time series: Goal - distill signal
##check structure for NAs
####Graph data as line graph and kernel density
####If multiple related variables
######Combine into an index to reduce dimensionality
##Graph the kernel densities globally and by group
```

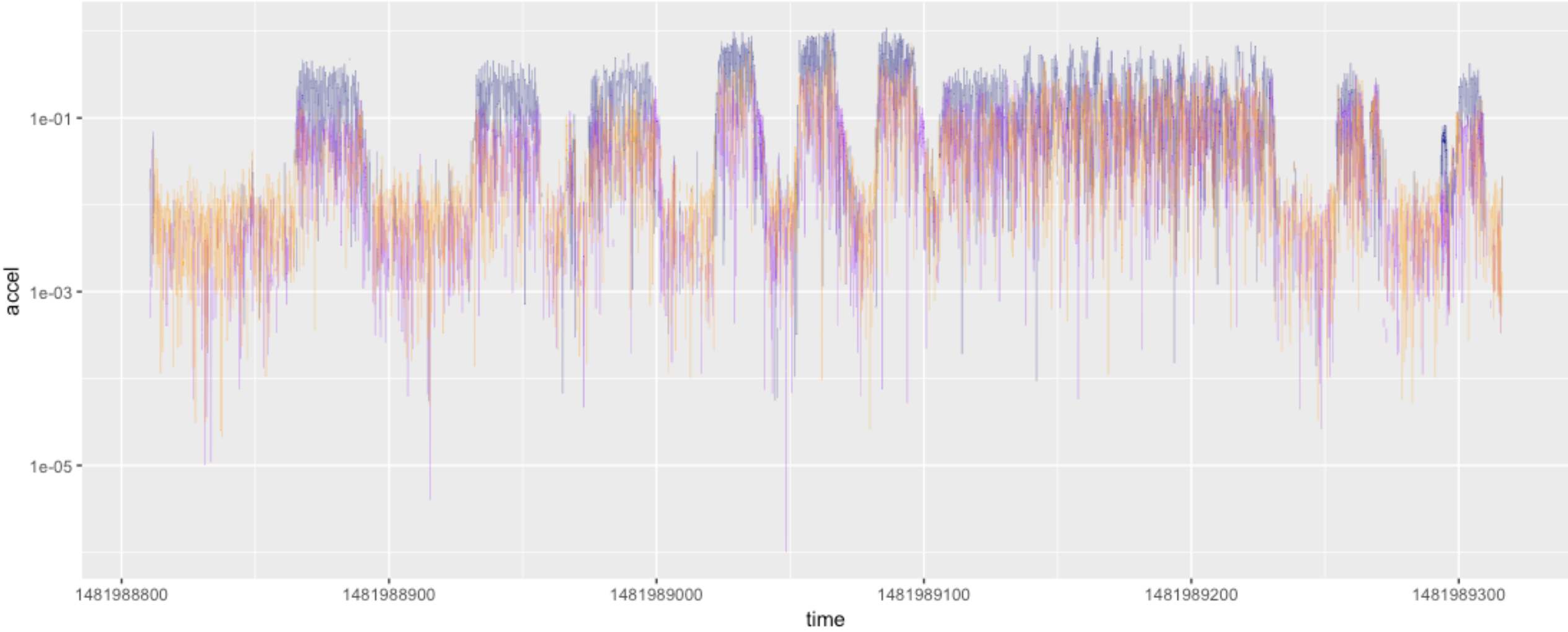
Three-Series Graph of Raw (20 hertz)



Three-Series Graph of Raw (20 hertz)



Log_{10}(Raw)

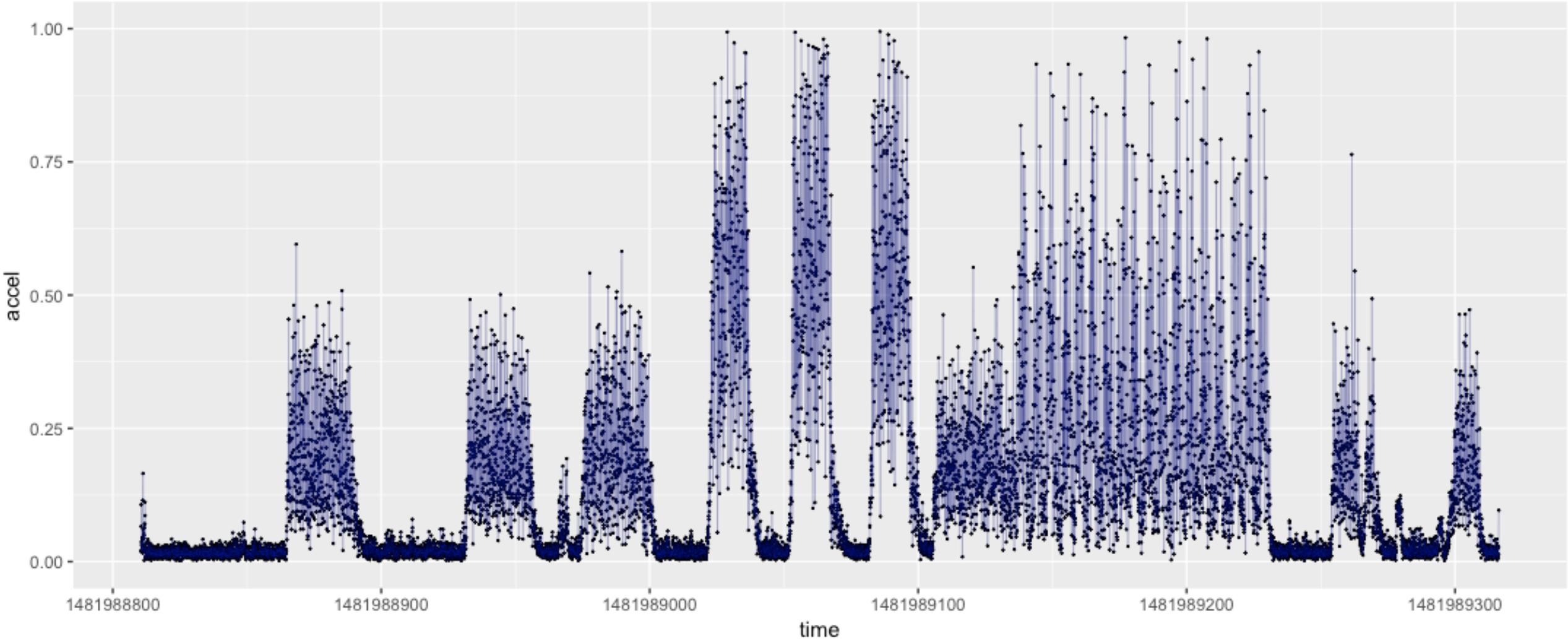

1e-01
1e-03
1e-05
accel
1481988800
1481988900
1481989000
1481989100
1481989200
1481989300
time

Net Accel

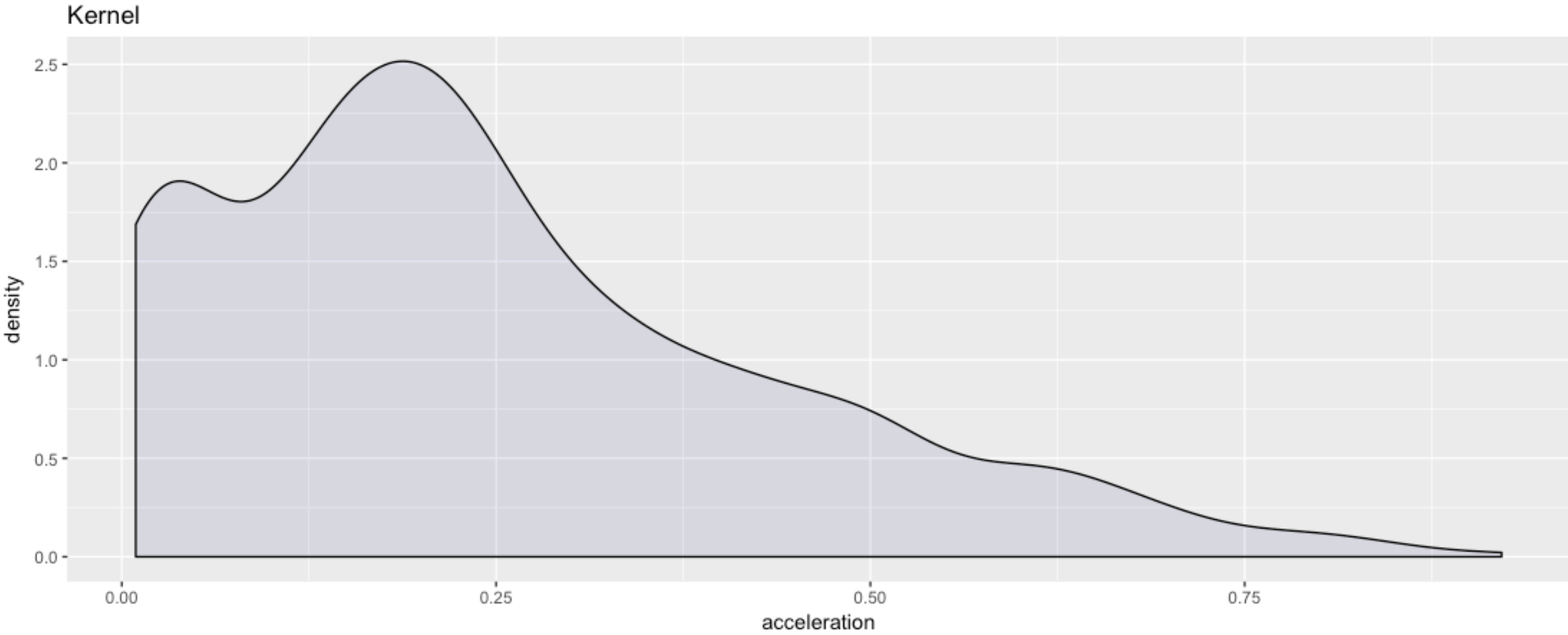
$$\text{Accel} = \sqrt{x^2 + y^2 + z^2}$$

Time	Accel.X	Accel.Y	Accel.Z	Label	Net.Accel
1	0.068	0.357	0.673	Walk	0.764854234
2	0.403	0.476	0.359	Walk	0.719629071
3	0.683	0.298	0.186	Walk	0.768042317
4	0.505	0.054	0.753	Walk	0.908267582
5	0.264	0.662	0.24	Walk	0.752023936
6	0.021	0.044	0.017	Stand	0.051633323
7	0.024	0.011	0.001	Stand	0.02641969
8	0.003	0.019	0.042	Stand	0.046195238

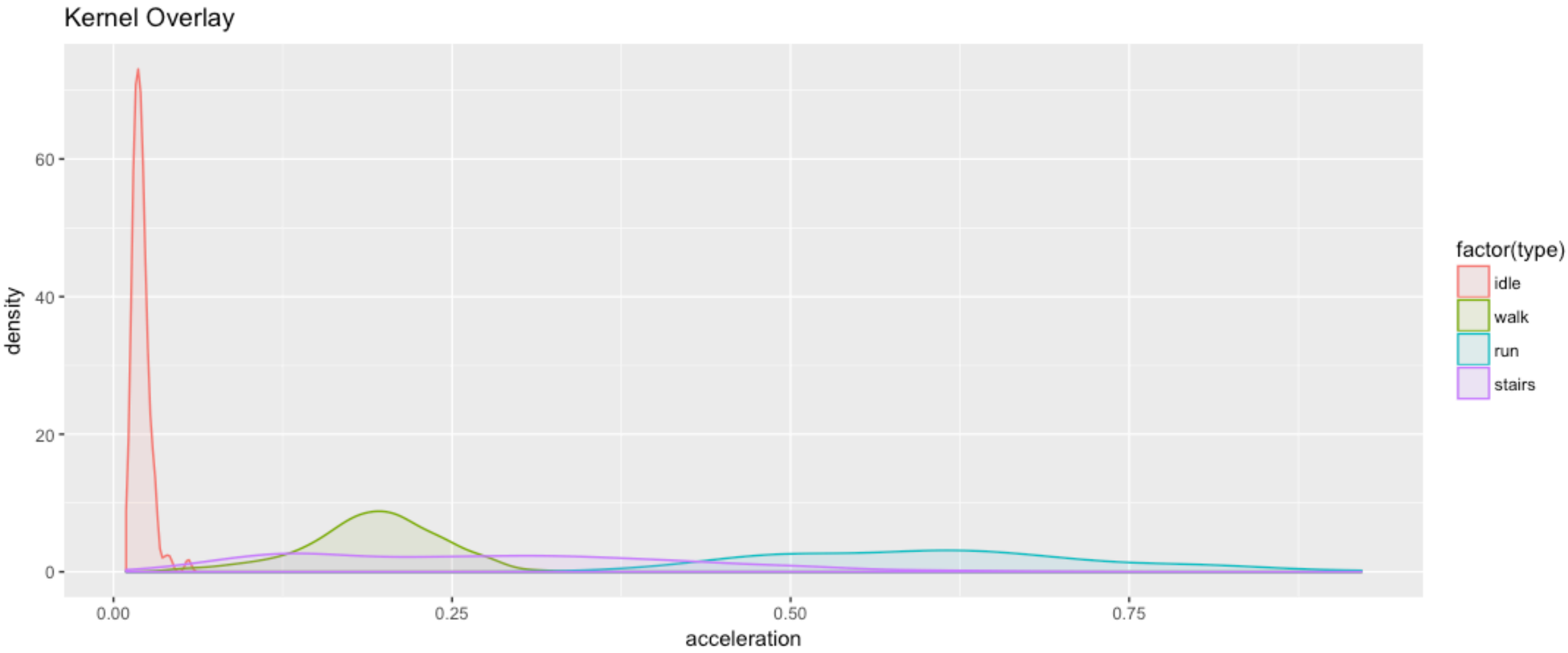
Transformation: Net acceleration



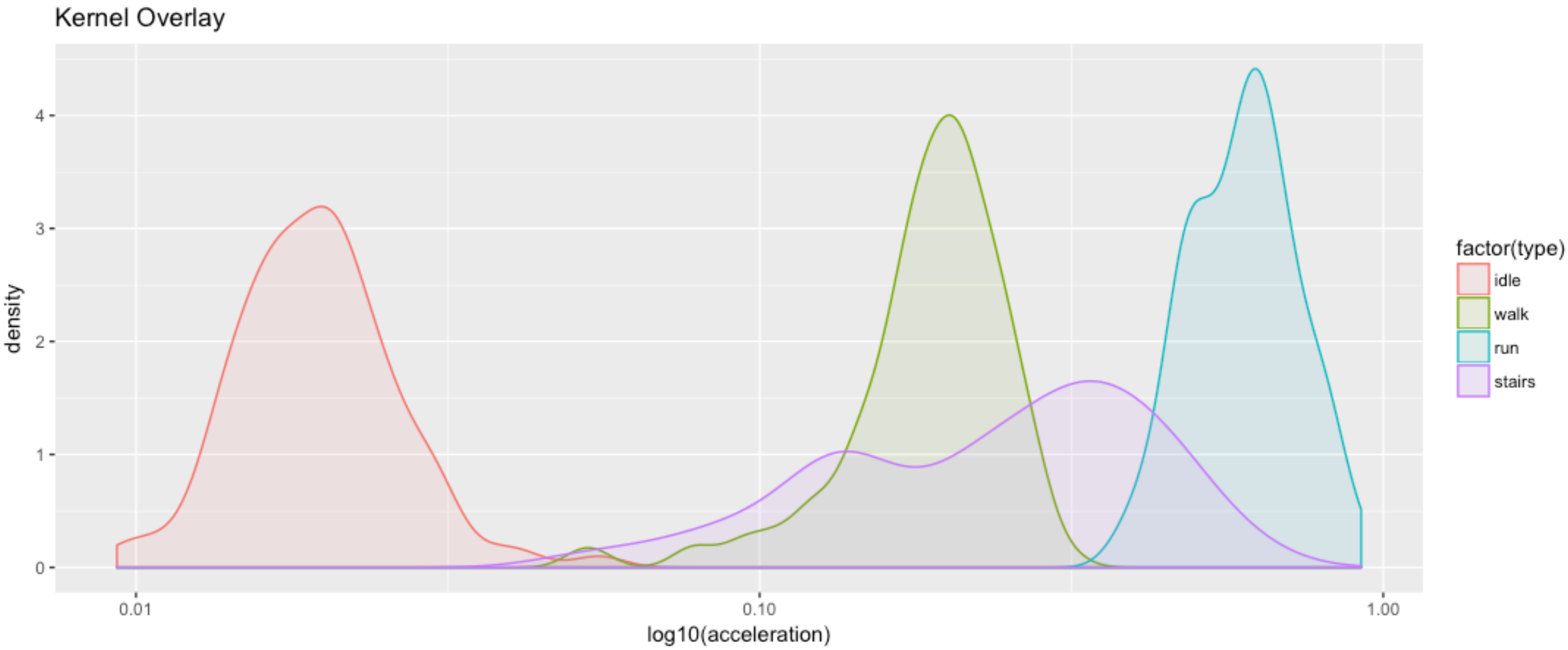
Distribution



Distribution



Distribution: Kernel Density

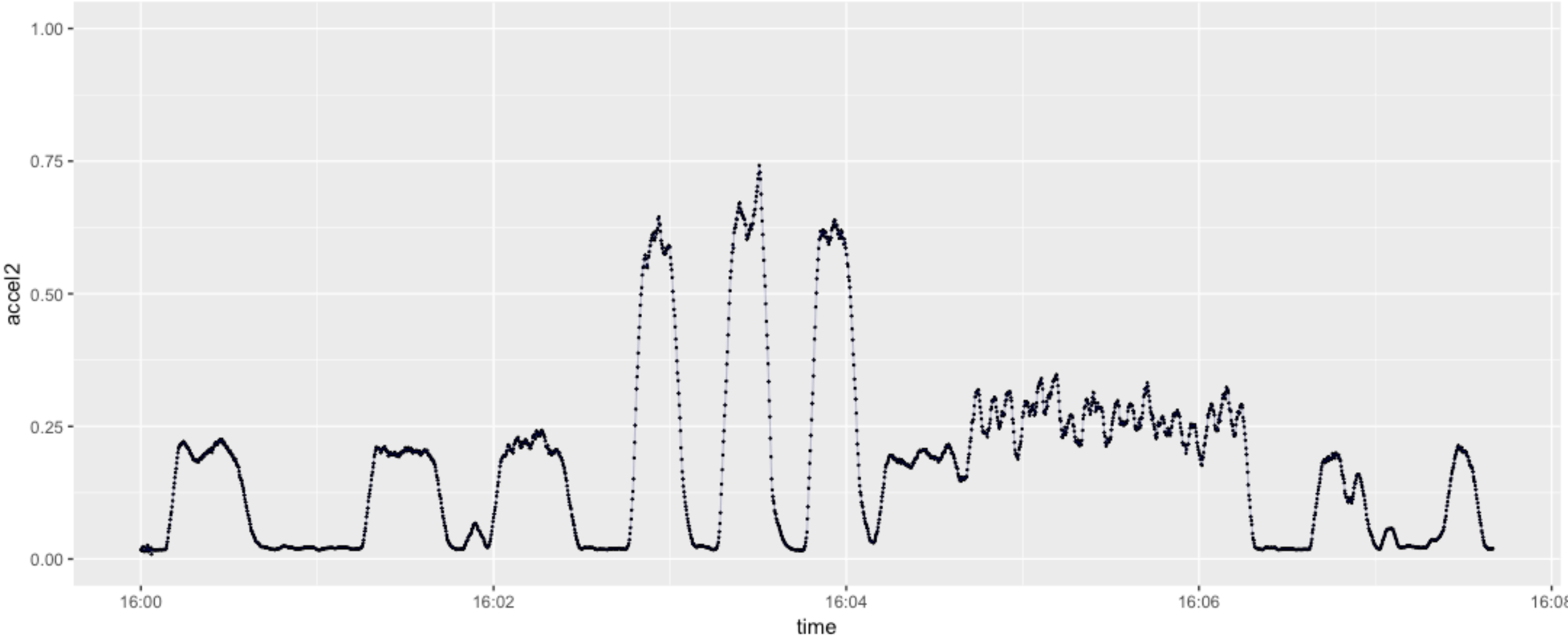


2 – Data Quality

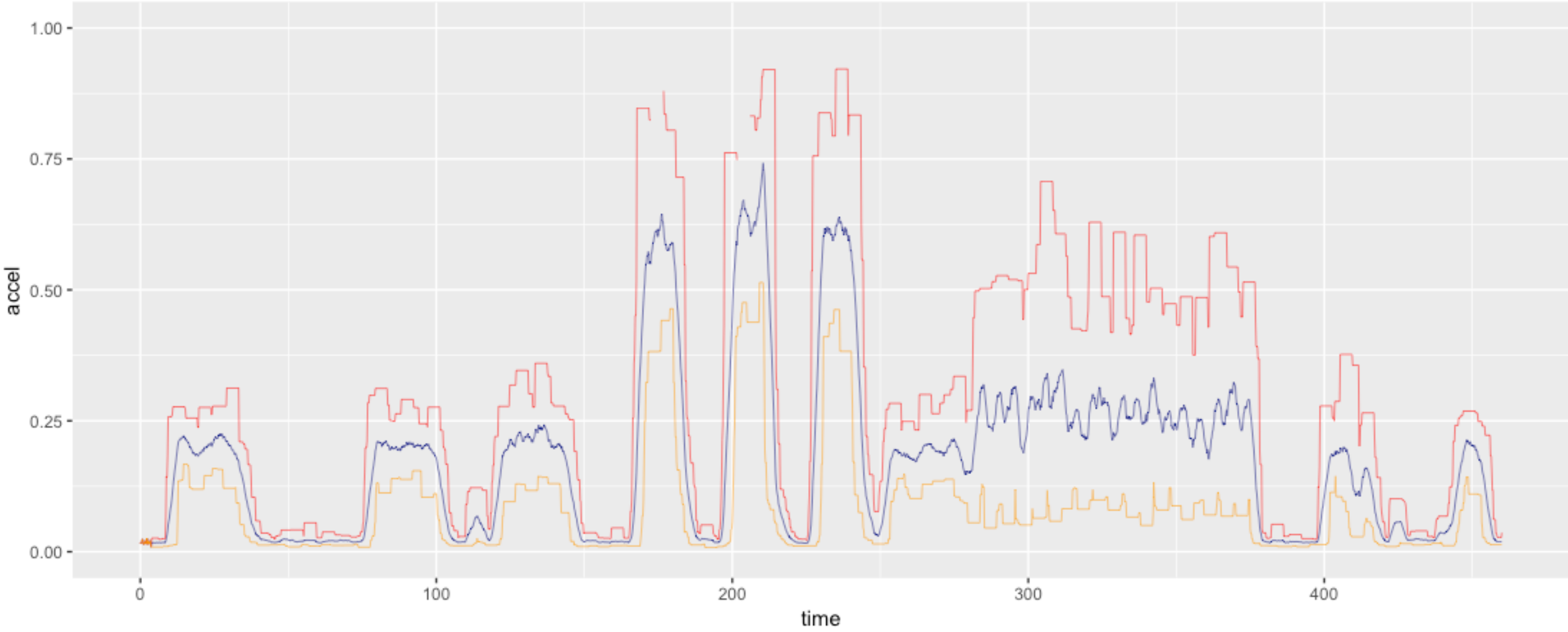
Goal	Common Questions
Detect outliers, missingness and errors	Are there anomalous values? Do records spike or occur during odd times? How complete is the data? Which variables need to be standardized and cleaned?

```
#How JC approaches a time series: Goal – distill signal
##check structure for NAs
##If NA's less than 10%
#### Impute records
##If continuous
####Graph data as line graph and kernel density
####If noisy
######smooth records, downsample or conduct extraction of features
```

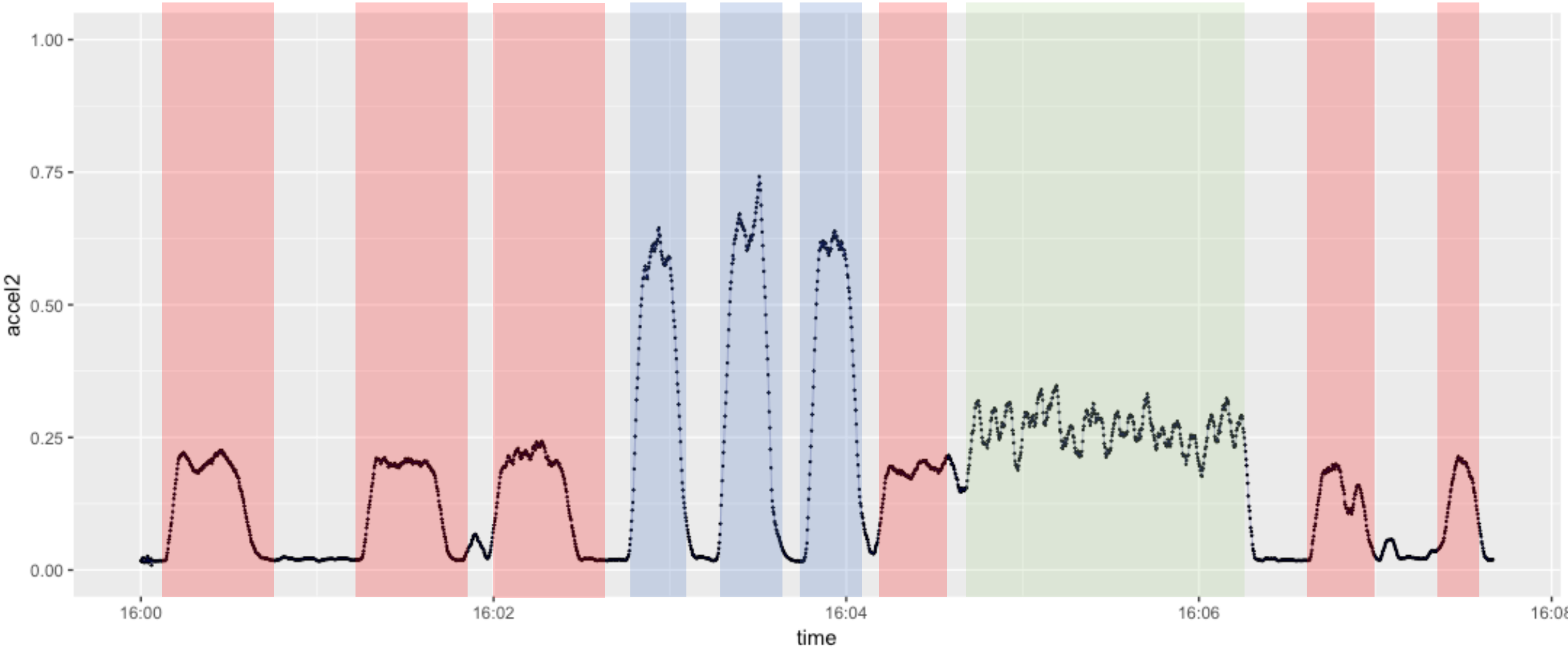
Transformation: Smoothed (1 hertz)



Transformation: Local min/max/mean extracted



Comparison with Labels: Smoothed (1 hertz)

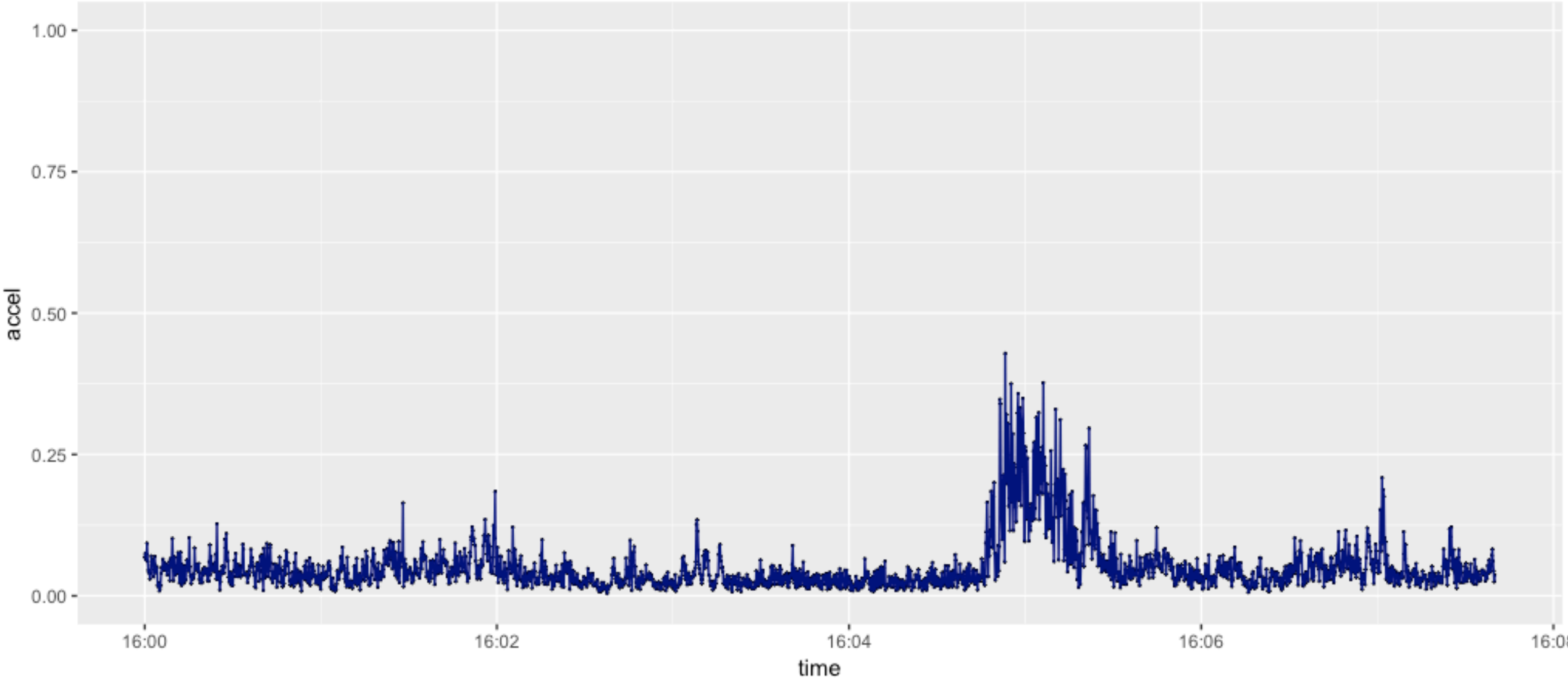


Follow-Up: Assumptions

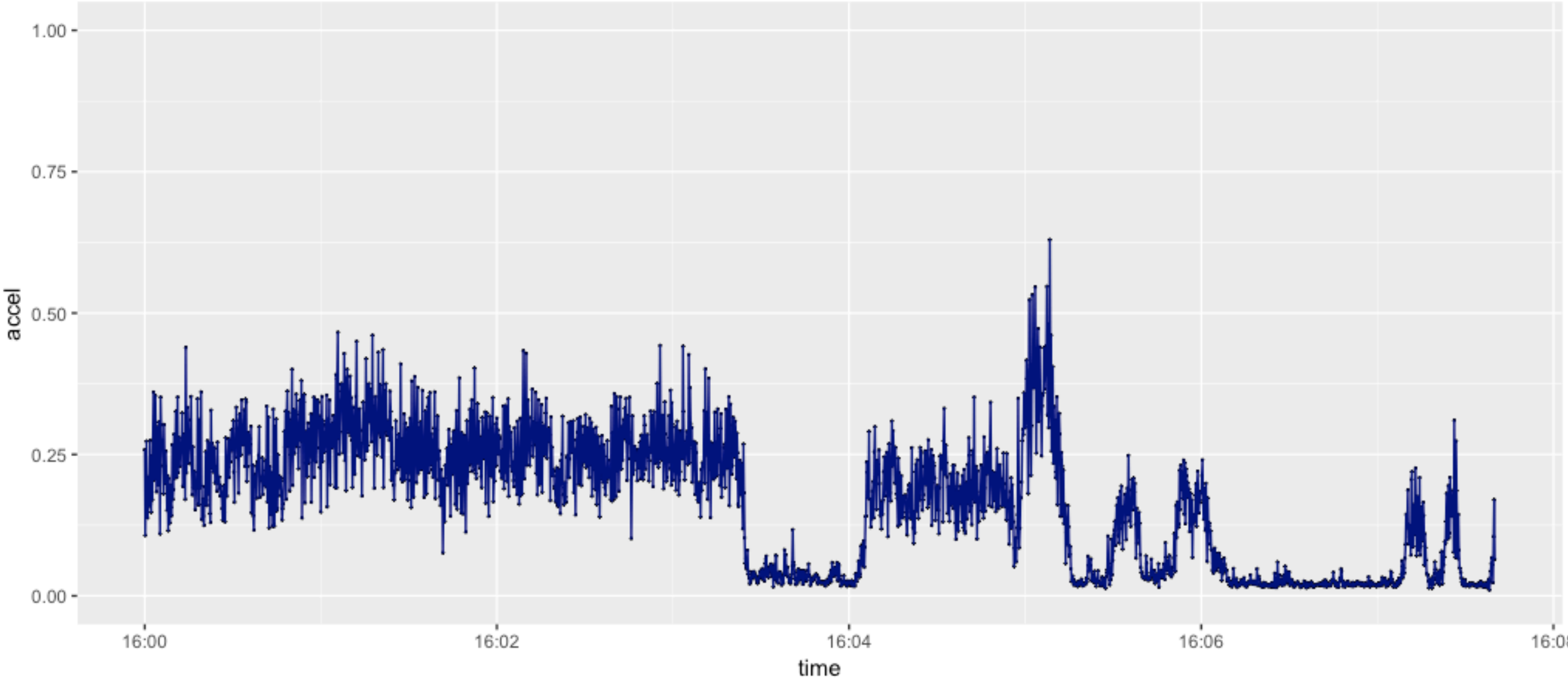
Goal	Common Questions
Check the assumptions	How exactly is the data collected? Does the data reflect what would be expected?

- Qualitative – discussion and review documents
- If possible,
  - Replicate data collection to quantify tolerances

Plane landing



Walk around a city



Roadmap

- Syllabus Change
- Homework Review
- EDA
- Code-along

For next time

- Read Chapter 5