Capstone 2: Milestone Report 1

Problem: Can we detect someone's sobriety/drunkenness by their movement?

Client: Public health organizations, sociologists, universities, younger people. Potential to create app that warns users of heavy alcohol use.

Data: UCI Detecting Heavy Drinking Dataset
https://archive.ics.uci.edu/ml/datasets/Bar+Crawl%3A+Detecting+Heavy+Drinking

Data Description: Accelerometer data for 13 participants involved in a "bar crawl" event. CSV file includes position in 3 axes and time of measurement in milliseconds. Additionally, separate CSV files for each participant of TAC (transdermal alcohol content) readings, taken roughly every half hour. "Clean" readings are shifted back in time by 45 minutes to account for time it takes to release alcohol through the skin.

Data Wrangling: The following steps were taken:

1) The first two rows of the accelerometer data were dropped, as they contained all zero values and therefore appear to be a data entry error.

2) Time measurements for accelerometer data are taken in milliseconds since UNIX Epoch (01/01/1970). Values converted to correct datetime objects using the pandas .to_datetime function.

3) TAC readings were concatenated into a single dataframe by looping over the unique participant ID's. Here, time measurements are changed to datetime objects as well.

4) TAC readings for two participants are found to be identical, indicating a data entry error. One participant is dropped.

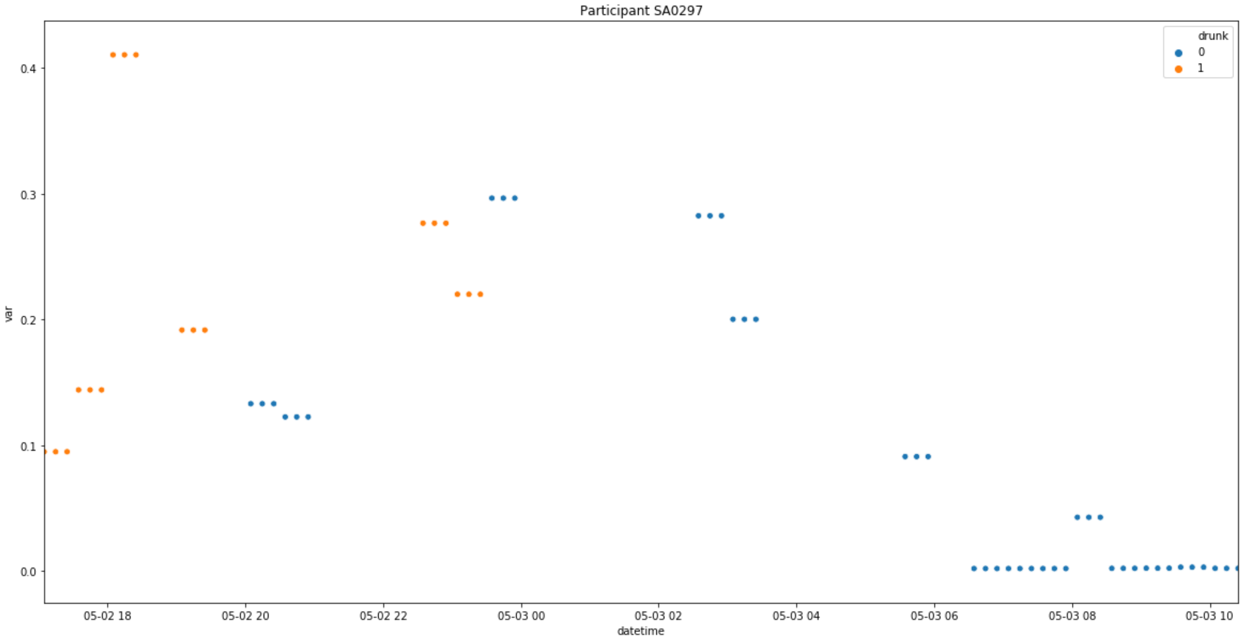
5) Variance is calculated from accelerometer data.
    a) Loop over readings for an individual participant.
    b) Take average of TAC readings within a thirty minute window. Because readings happen roughly every half hour, usually only one or two readings are aggregated.

c) For every one TAC window, there are 3 sub-windows, each representing the variance in movement over a ten minute period.
d) If the TAC reading is greater than or equal to 0.08 (legal limit), label row as “drunk” (1 in “drunk” column).
e) Loop over all participants.

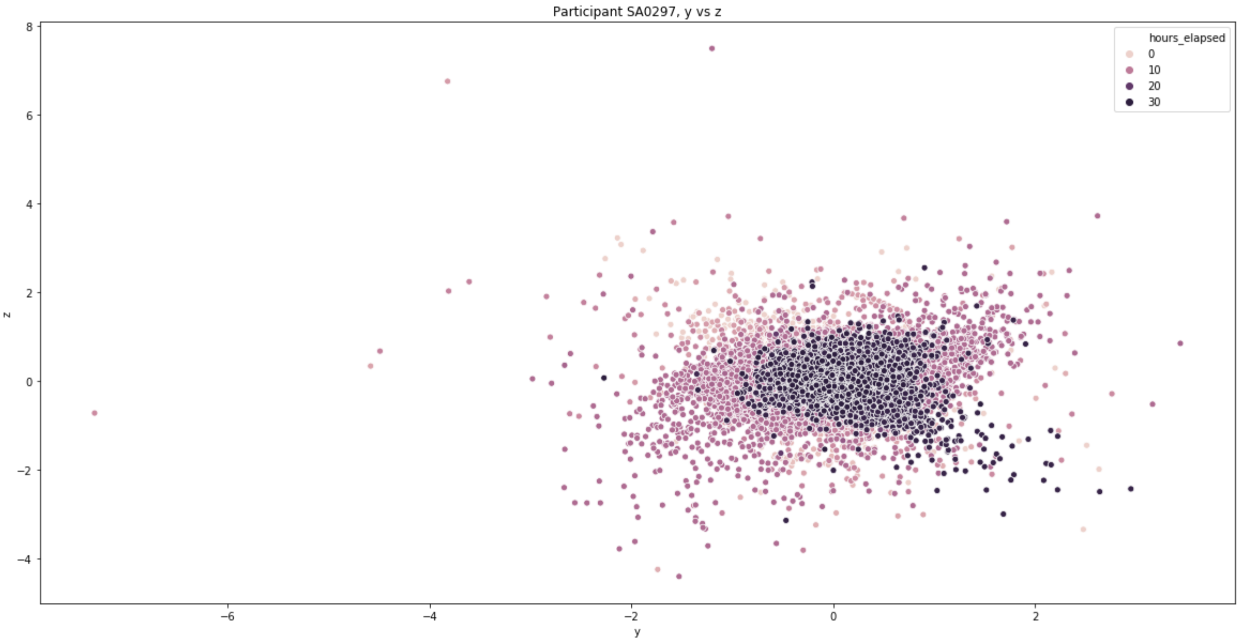
6) Resulting dataframe should have observations with datetime, variance in three axes, TAC reading, and “drunk” label.

Exploratory Data Analysis: Visual Analysis:

1) Visual examination of variance over time clearly shows higher average variance for a drunk participant than a sober one:



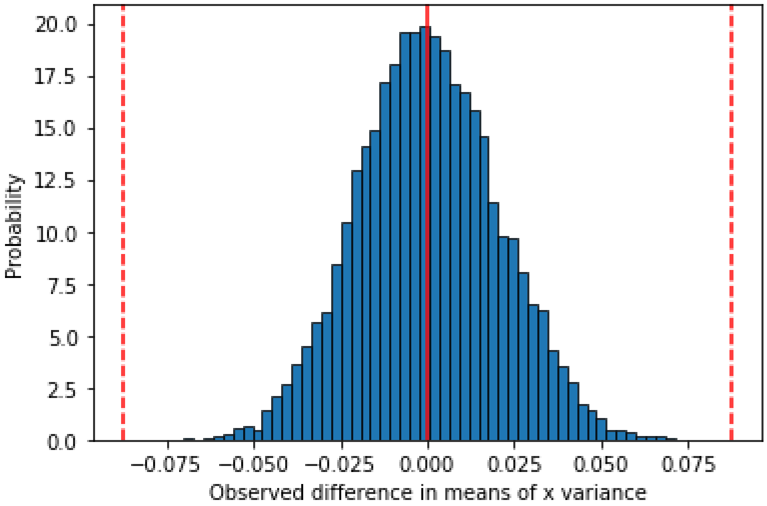
2) If the x-axis, which has lower variance than the other axes, is believed to be the vertical axis, then plotting the y- and z-axes against each other should show an “overhead” view of the participant’s movement:



We see the greatest variance for points around 10 hours into the event. This corresponds to midnight and later for this participant, the time when they would be most heavily drinking.

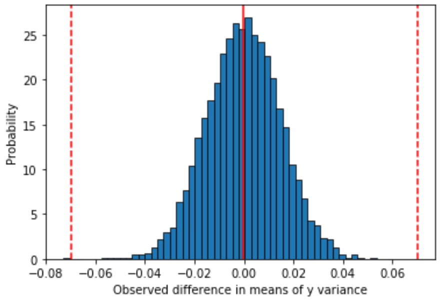
Inferential Statistics:

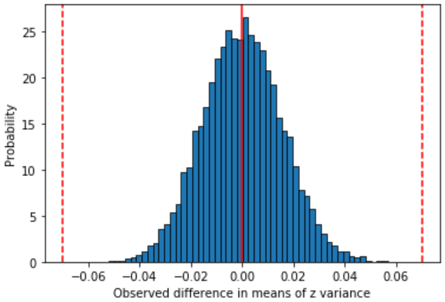
1) Bootstrapping is used to test whether there is a statistically significant difference between variance in sober and drunk data:



The dotted lines on the side represent the observed difference. The distribution shown is the simulated difference, made by shifting the two groups to have the same mean and then drawing 10000 samples. The p-value here is 0, which you can see on the graph -

no simulated values had the magnitude of the observed difference. The null hypothesis is rejected, meaning there is a difference between drunk and sober movement. The same was true for the y- and z-axes:





Intermediate conclusion:

Visual exploratory analysis and inferential statistics indicate that there is a difference in variance of movement while drunk. In other words, there is a relationship between variance and whether a person is drunk. A machine learning model should be able to analyze this relationship and predict a “drunk” label for unseen data.