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.

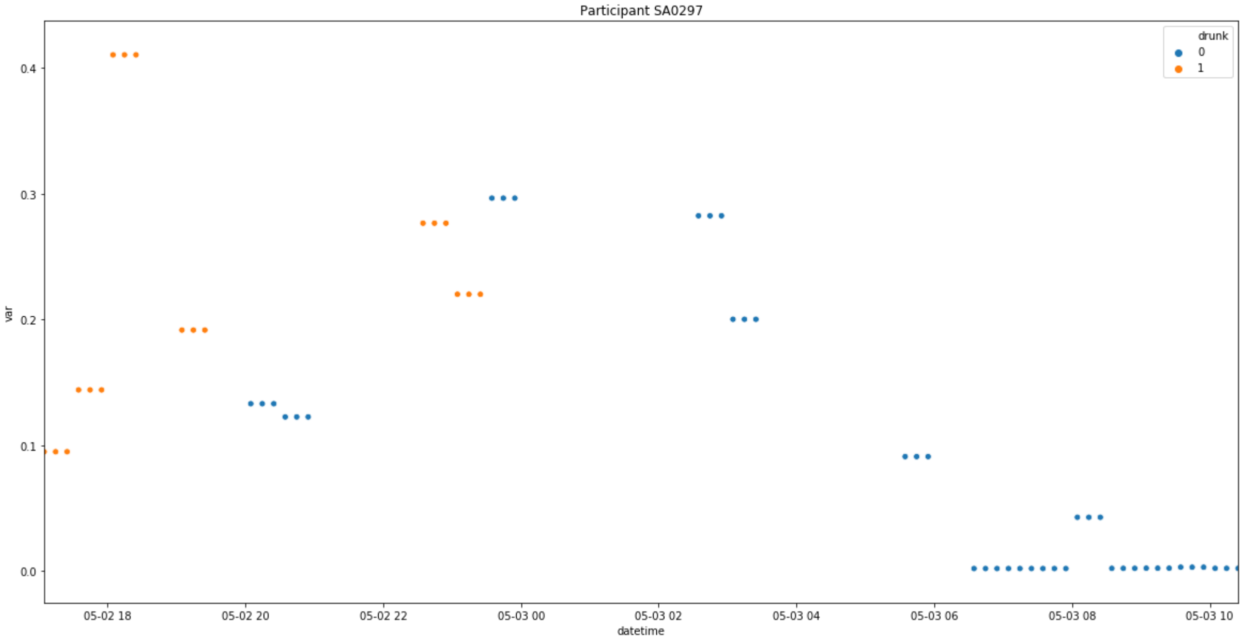
5) Variance is calculated from accelerometer data.
    a) Loop over readings for an individual participant.
    b) Take average of TAC readings within a thirty minute window. Because readings happen roughly every half hour, usually only one or two readings are aggregated.

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.

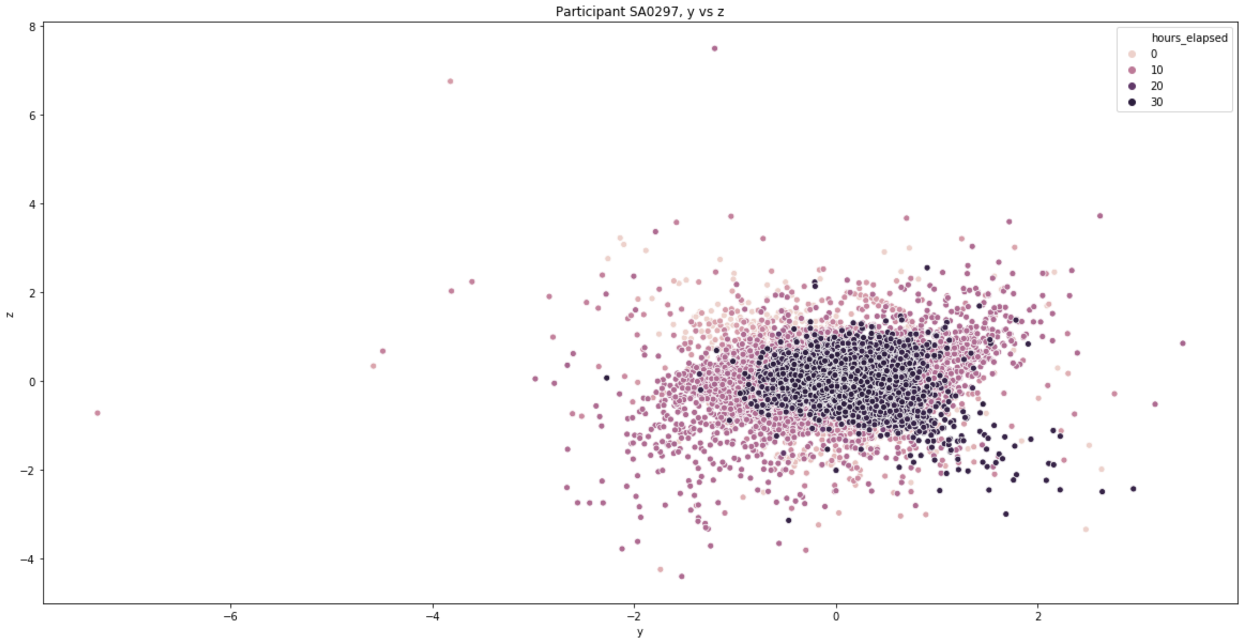
6) Resulting dataframe should have observations with datetime, variance in three axes, TAC reading, and “drunk” label.

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:



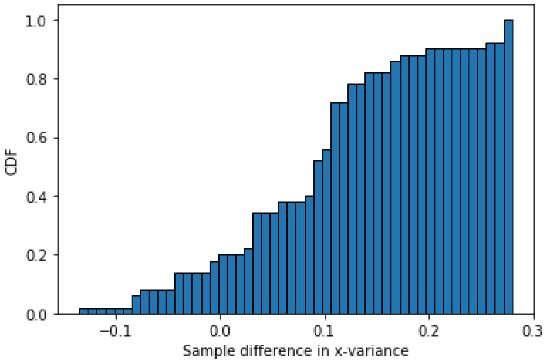
We see the greatest variance for points around 10 hours into the event. This corresponds to midnight and later for this participant, the time when they would be most heavily drinking.

Inferential Statistics:

1) A z-test is used to determine if there is a statistically significant difference between drunk and sober movement. The formula for computing z is as follows:

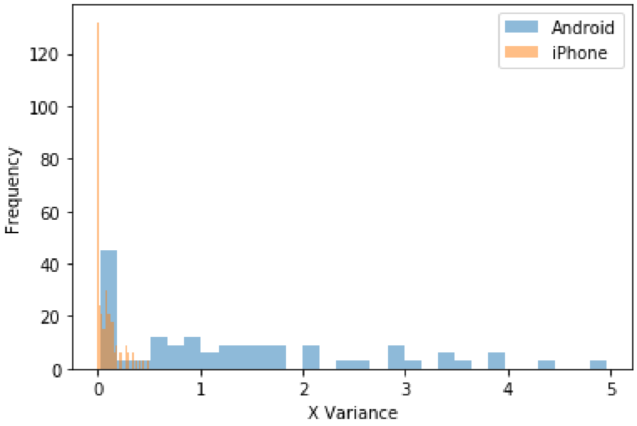
$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The standard deviation of both distributions can be computed from the observed data. To get sample means, I randomly sampled with replacement the observed variance with a sample size of 50. Comparing drunk and sober variance in the x-axis for one user, I found a z-score of 6.13, which equates to a two-sided p-value of 5.43e-5. With a significance level of 0.05, this disproves the null hypothesis, that there is no difference between the groups. The CDF for the sample distribution:



The p-values for the other axes were even smaller: 3e-12 for y-variance, and 7.61e-17 for z-variance.

2) Later, another z-test is used to examine the difference between Android and iPhone users. A quick look at the histograms for both groups shows a clear difference:

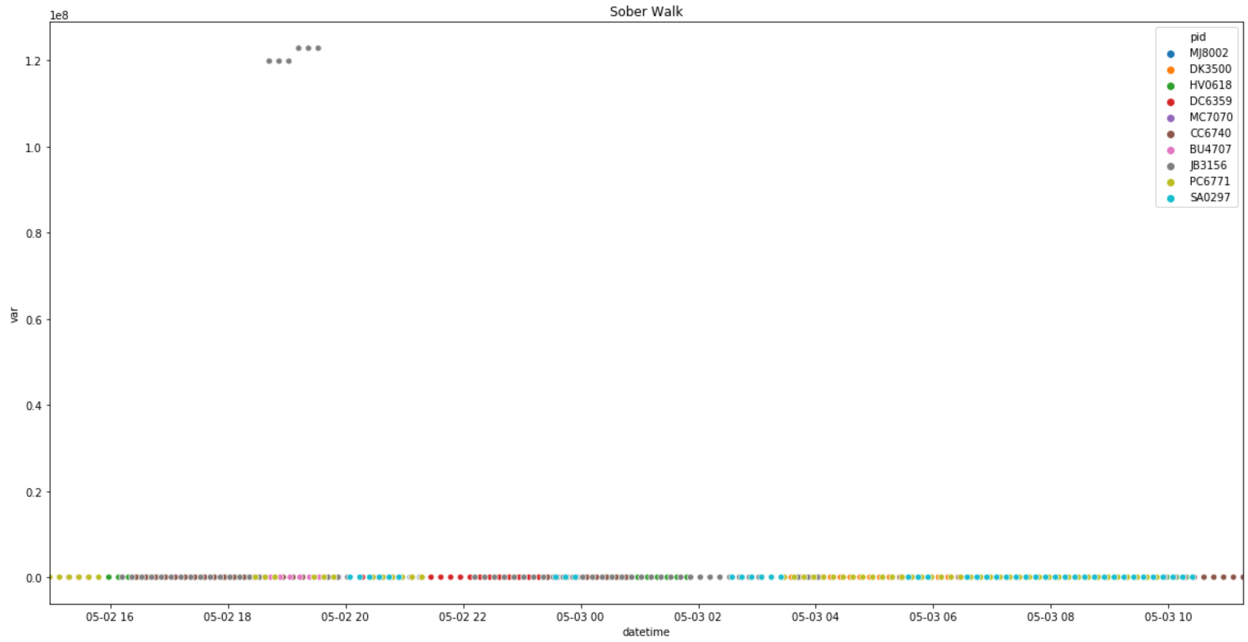


The p-value for the difference in means of x-variance between the groups is found to be 4.9e-10, so we reject the null hypothesis.

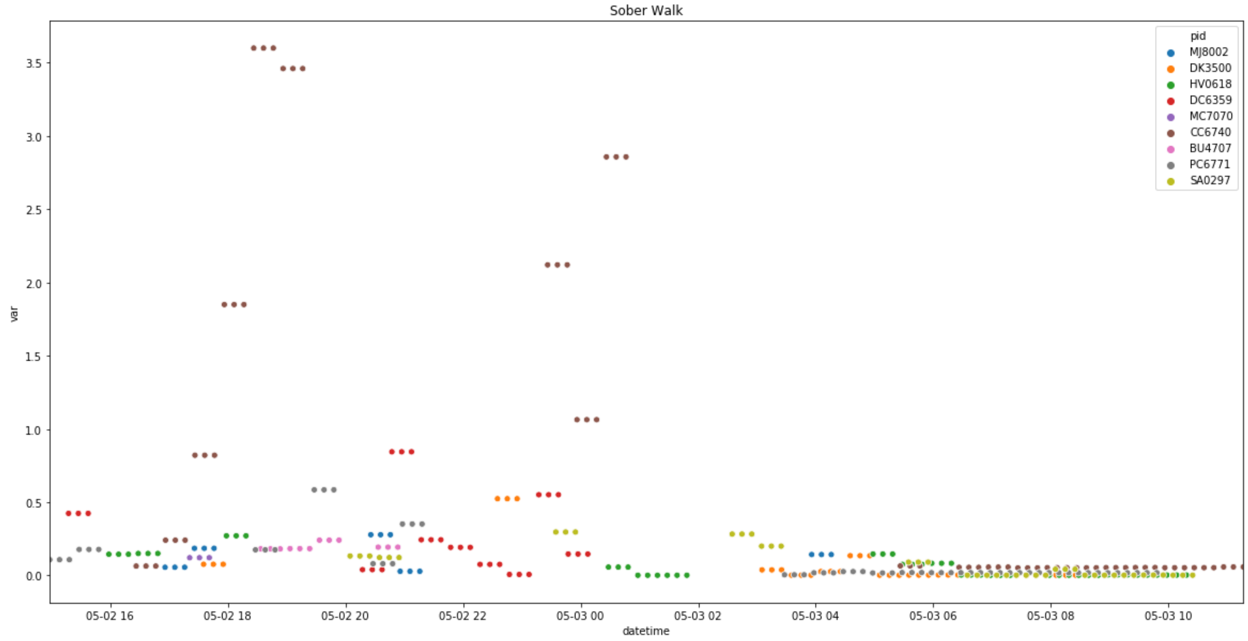
Intermediate conclusion:

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.

We also see clear evidence of a difference between Android and iPhone users. This could be potentially explained by a technical fault in how the android application in the original experiment recorded its data. In fact, one Android user in particular is clearly at least an outlier. Look at this following plot of variance over time for sober users:



This one Android user, JB3156, has variance measurements so much higher than all the other participants, the other measurements can't even be shown on the same scale. Qualitatively, we can also observe the impact this one participant has on the dataset. Counterintuitively, with all the users considered, the variance of sober movement is actually significantly higher than for drunk movement; the standard deviations of drunk and sober movement respectively are 1.33 and ~2,000,000. However, when this one participant is excluded, we see more reasonable output:



(Note that even here, the user with the highest variance is the other Android user). The new drunk variance is .78, compared to .25 for sober variance, which corresponds to our expectations. As a result of this investigation, one model will be trained without this outlier user, another will use phone type as a feature, and a final will also use “crazy_user” as a binary categorical feature.