The Floow

# ANALYSES OF JOURNEY DATA

24 February 2020

Kulbhushan Shah

# Table of Contents

## 1. Abstract

Each journey data was gathered using sensors of different cell phones while traveling from one location to another. The process was relied on Accelerometer and GPS system as a source type to capture raw data. This report provides analyses of twenty-five journeys. The analysis was carried out to try to predict a person's emotions while driving behind the wheels. This study comprises an understanding of data collection, managing the data for analysis, understanding any further correlation exists, conveying insights about the data through visual aids, the technique used to identify 'event' that might be an accident and allocating a notional severity index and notional confidence.

## 2. Introduction

Mobile phones are a constant in our lives. In a country of around 66 million people, there are 83 million mobile phones in use. There are 38 million licensed vehicles in just Great Britain.

As per RAC motoring report [1], twenty-five percent of drivers admit that they have used their phones whilst driving. Twenty-five percent of people concede speaking on phone, thirty-nine percent admit to making or receiving a call whilst the car is stopped but the engine is running. Sixteen percent have posted on social media or sent texts or emails whilst driving. Distraction plays a vital role in causing road deaths. Numbers of steps can be taken to prevent causing disasters and impacting many lives around.

For insurance companies, assessment of risk is influenced more by how similar you are to a known risky population.  That's why young drivers tend to command higher premiums.  The cause might be due to their inexperience and immaturity, but that's almost beside the point. It's a numbers game.

Another problem of road accident can be due to driving while fatigued has contributed to more road accidents in Great Britain than drug-driving offense.

To overcome these problems, The Floow has taken initiative by offering a range of services for insurers, policyholders, passengers, and wellbeing of drivers. Using FloowDrive software, one can monitor drivers' driving skills, distraction, fatigue nature, to name a few. By collecting data from millions of journey data across a wide range of customers, an insurance company can apply machine learning to discern the patterns of drivers who make claims versus those who don't.

The purpose of the study is to elicit 'event' that might cause an accident in a driving setting. To provide safe and comfortable transportation, analyses of journey data was carried out to understand and examine a driver behaviour state in a driving context. Sensors play an important role in the study of human behaviour.

## 3. Method

I. Material Used

As per the raw data registered in twenty-five journeys' files, GPS and Accelerometer were reliable sources to gather the information.

What is GPS?

GPS (Global Positioning System) plays a very important role in all our lives. From allowing you to see where you are on the planet, to help you get to destinations quickly, GPS has evolved how we live our lives.

What is Accelerometer?

Accelerometers are devices that measure acceleration, which is the rate of change of the velocity of an object. Cell phones equipped with an accelerometer ('a sense of motion') sensors can assist in tracking and detecting the orientation of the phone. It only measures the linear acceleration of movement. By using Gyroscope sensors, rotation or twist information can be gathered.

Smartphone measures acceleration across three axes, traditionally labelled as x, y, and z. The measurements are sampled multiple times per second. Each measurement reflects acceleration across one of the axes. Taken together, you can get a sense for the phone's overall direction. Acceleration is measured in meters per second squared, or m/s2 or in G-forces (g).

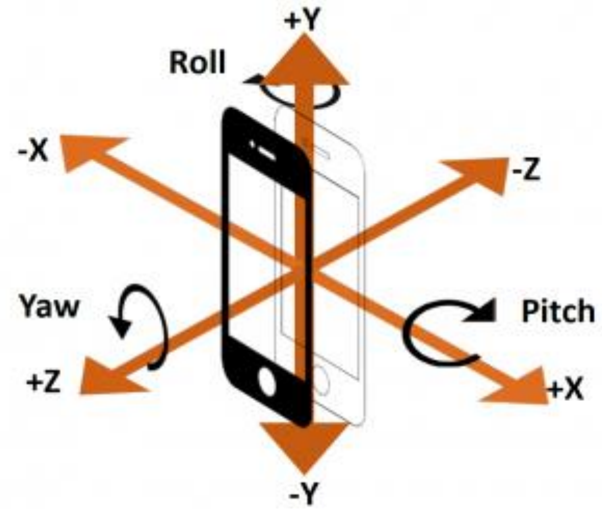


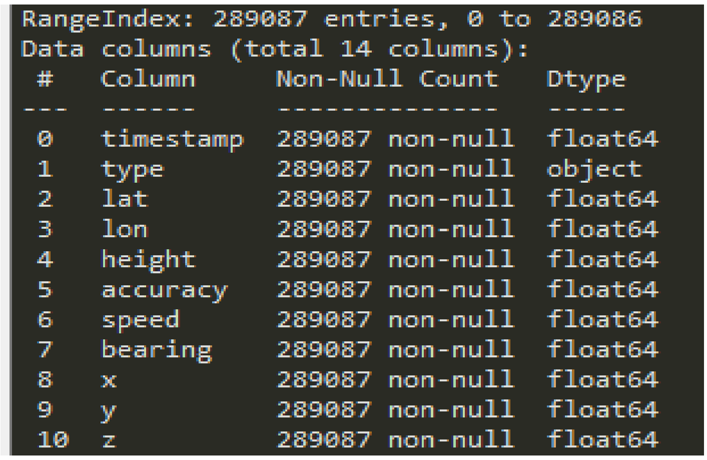
**Figure 1.0:** Represent Accelerometer's axes

Accelerometer can be used to track and detect the orientation of the phone. Whereas, GPS can determine their current location, time and velocity.

## II. Data Analysis

Descriptive analysis is used to give summary measures and statistic measures about the data without further interpretation.

- The data set contain 22 days (covering 25 journeys) files for year 2015. For all journey file a unique JID is created to represent a file.
- There are eleven variables in each journey file.



```
RangeIndex: 289087 entries, 0 to 289086
Data columns (total 14 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   timestamp   289087 non-null  float64
 1   type        289087 non-null  object
 2   lat         289087 non-null  float64
 3   lon         289087 non-null  float64
 4   height      289087 non-null  float64
 5   accuracy    289087 non-null  float64
 6   speed       289087 non-null  float64
 7   bearing     289087 non-null  float64
 8   x           289087 non-null  float64
 9   y           289087 non-null  float64
 10  z           289087 non-null  float64
```

Figure 2: Variable

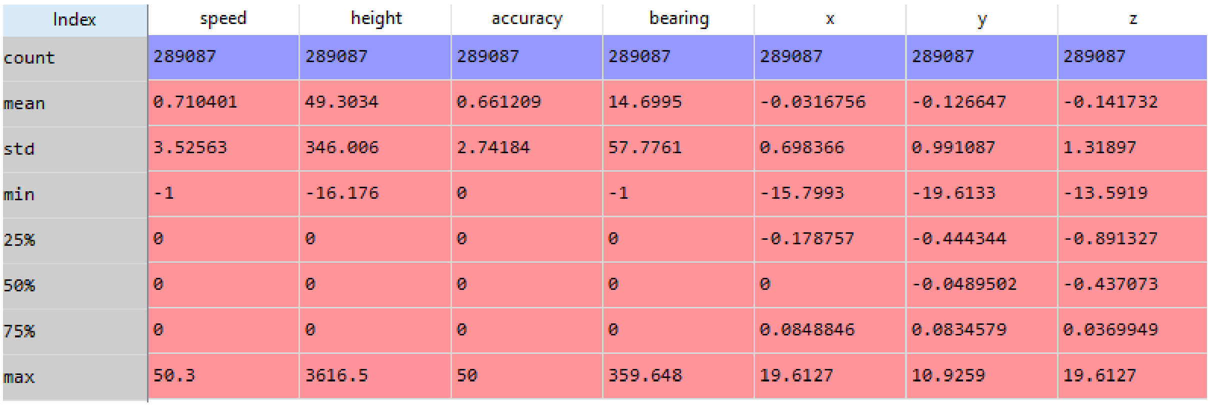
- Two journeys (with journey ID: JID6, JID23 respectively) were covered on 16/03/2015, two were covered on 20/04/2015 (JID4, JID3) and two were covered on 01/12/2015 (JID9 & JID21).

Outliers

- JID1(file name 00DAC437-FF8B-4DA3-9E24-4EE1B1AA12EC.csv) has registered with same time 20/06/2015 22:40 throughout the file. Values that are outliers of the common range for a single variable may lead to apparently large relationships in summary statistics like correlations or regression coefficients.

- Accuracy has potential errored values. The accepted value should be between 5 -10, beyond these are potential error. Accuracy values are registered up to 50.

- Accuracy gives us information on whereabout our location is at certain point of time on the earth. It is calculated with the help of 3 to 4 satellite signals.



| Index | speed | height | accuracy | bearing | x | y | z |
|---|---|---|---|---|---|---|---|
| count | 289087 | 289087 | 289087 | 289087 | 289087 | 289087 | 289087 |
| mean | 0.710401 | 49.3034 | 0.661209 | 14.6995 | -0.0316756 | -0.126647 | -0.141732 |
| std | 3.52563 | 346.006 | 2.74184 | 57.7761 | 0.698366 | 0.991087 | 1.31897 |
| min | -1 | -16.176 | 0 | -1 | -15.7993 | -19.6133 | -13.5919 |
| 25% | 0 | 0 | 0 | 0 | -0.178757 | -0.444344 | -0.891327 |
| 50% | 0 | 0 | 0 | 0 | 0 | -0.0489502 | -0.437073 |
| 75% | 0 | 0 | 0 | 0 | 0.0848846 | 0.0834579 | 0.0369949 |
| max | 50.3 | 3616.5 | 50 | 359.648 | 19.6127 | 10.9259 | 19.6127 |

Figure 3: Descriptive Statistics on major variables

### Missing Values

- As there are multiple source type (GPS, Accelerometer), only the variable belonging to each type registered values in certain observation while in other observation they were recorded as missing entries. Missing values are handled (by replacing it with 0) using pandas function to ease the further analyses as by keeping it might result in bias results.

### Pros (Missing Values):

- Easy to understand and implement
- Can be applied to any model (decision trees, logistic regression, linear regression, etc)
- Addresses training and prediction time
- More accurate predictions

### Cons (Missing Values):

- Removing data points and features may remove important information from data
- Unclear when it’s better to remove data points versus features
- Doesn’t help if data is missing at prediction time.

An exploratory data analysis builds on a descriptive analysis by searching for discoveries, trends, correlations, or relationships between the measurements of multiple variables to generate ideas or hypotheses.

### Measures of variability

- Variability in dataset can be quantified by using measures of central tendency and measures of dispersion.
- Measures of central tendency are mean, median and mode. Measure of central tendency is performed on accelerometer axes (x, y, z) and speed, height, accuracy.

- The mean value varies for these coordiantes as it depends on acceleration, and altitude experienced in the journey over time.
- Measures of dispersion, on the other hand measures the dispersion of a set of data point around their mean.

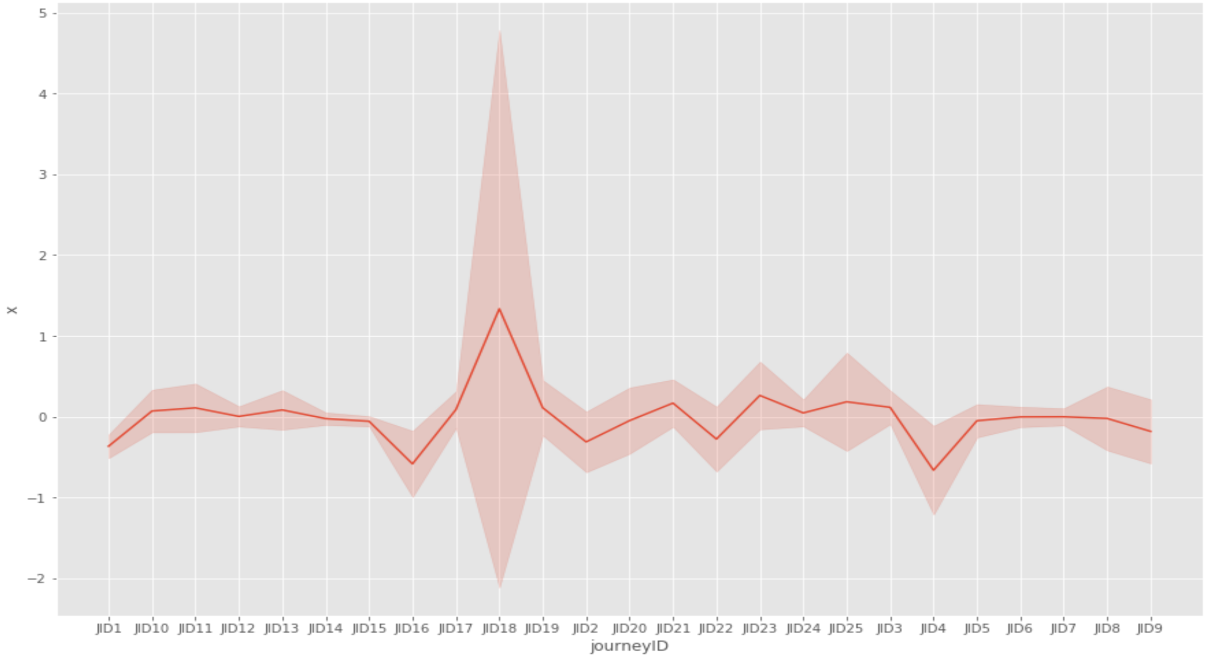


Figure 4: Standard Deviation of variable X for all journey files

- This figure elicit information about how the data is away from the average (dark line) value in each data file. The standard deviation is marked by the shaded part in the above graph. Journey JID18 has spike and dip that might represent more dramatic motion due to Roll, Yaw, or Pitch.

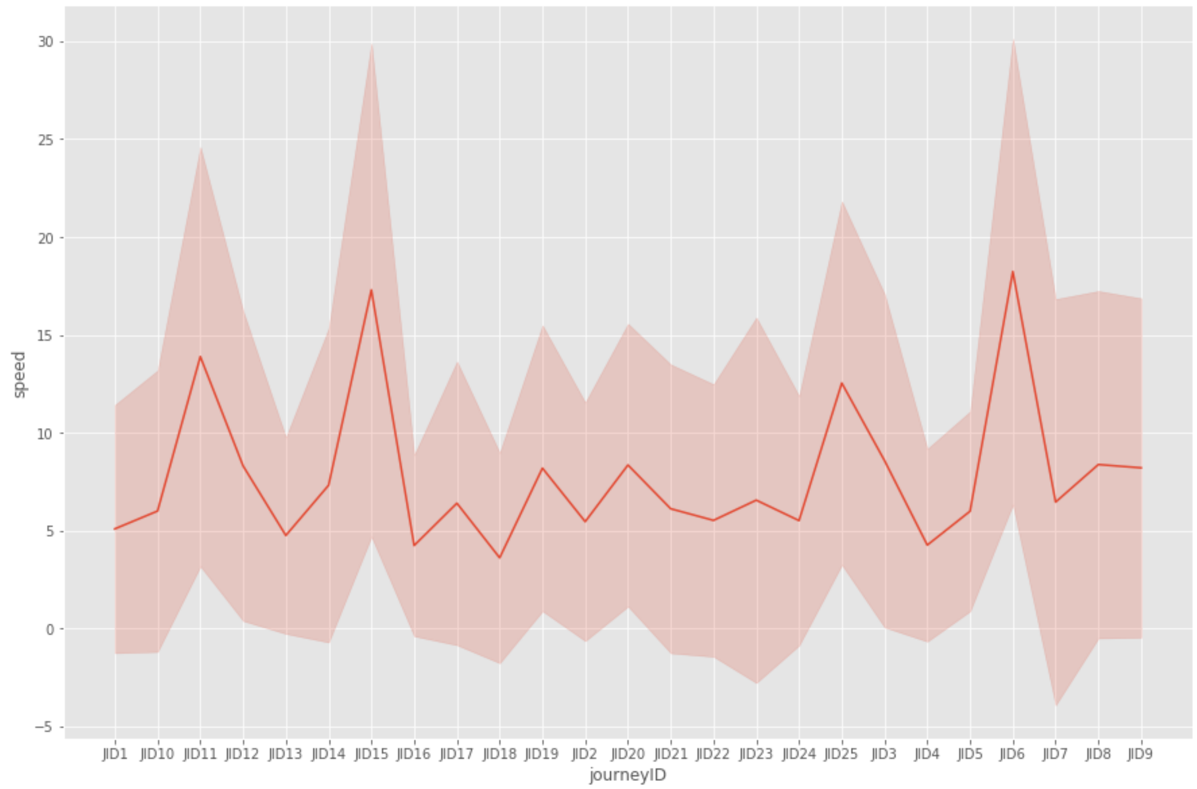


Figure 5: Standard Deviation of variable Speed for all journey files

- On the other hand, speed as well has high spikes and dips informing that from the average value there are observation which are far away than usual implying sudden brakes or accelerating or cornering has happened. JID14 & JID6 journeys have captured dramatic swift in the driving speed.

Pros (Measures of variability):

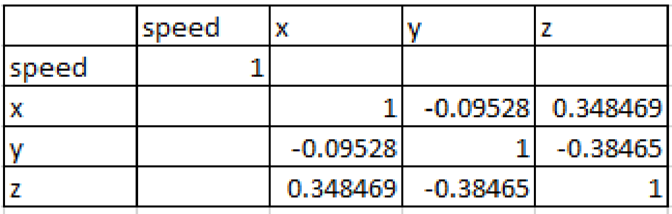
- It makes use of every element in the data set.

Cons (Measures of variability):

- Mean is a common measure, but it has a huge downside, it is easily affected by outliers.

Correlation Coefficient

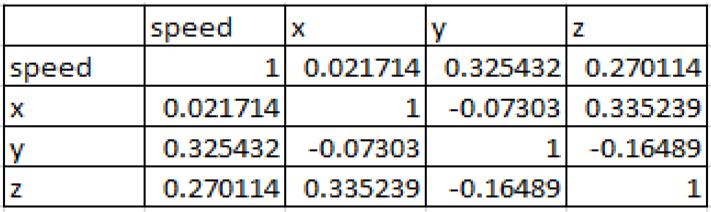
- A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
- In general, the correlation coefficient *r* measures (Asuero, 2006) the strength and direction of a linear relationship between two variables with a range of -1 to +1.



| | speed | x | y | z |
|---|---|---|---|---|
| speed | 1 | | | |
| x | | 1 | -0.09528 | 0.348469 |
| y | | -0.09528 | 1 | -0.38465 |
| z | | 0.348469 | -0.38465 | 1 |

Table 1: Correlation before handling missing data

- Since source are different for speed and x, y & z the data is captured with null entries. Due to which it is not ideal to see any correlation between them. Though, it would have been a good relationship to examine if there were no null entries.
- Negative relationship can be seen in X & Y, which cell phone Rolls to -ve side, Pitch value increase. Whereas there is a positive relationship between X and Z.



| | speed | x | y | z |
|---|---|---|---|---|
| speed | 1 | 0.021714 | 0.325432 | 0.270114 |
| x | 0.021714 | 1 | -0.07303 | 0.335239 |
| y | 0.325432 | -0.07303 | 1 | -0.16489 |
| z | 0.270114 | 0.335239 | -0.16489 | 1 |

Table 2: Correlation after handling missing data

- Table 2 shows an unaccepted correlation that is not ideal to interpret further (due to null entries). Speed would have been a good predictor for x, y and z, and vice versa.

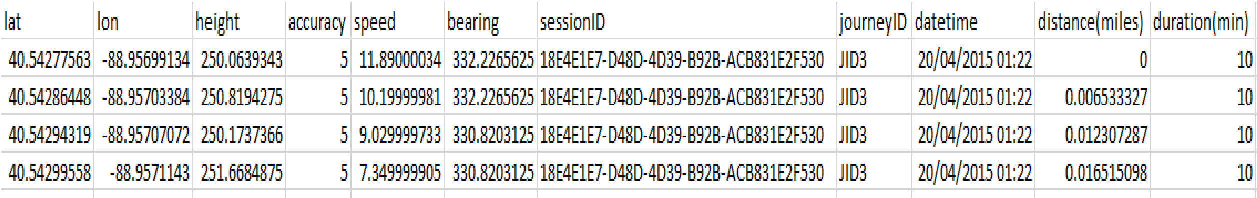
Pros (Correlation Coefficient):

- Relationship between axes (x, y, z) can be easily understood and interpreted.
- Sensor can capture dramatic motions.

Cons (Correlation Coefficient):

- Using current data set, speed or accelerometer coordinates can not be used as predictor to predict each other.

Tidy Data



| lat | lon | height | accuracy | speed | bearing | sessionID | journeyID | datetime | distance(miles) | duration(min) |
|---|---|---|---|---|---|---|---|---|---|---|
| 40.54277563 | -88.95699134 | 250.0639343 | 5 | 11.89000034 | 332.2265625 | 18E4E1E7-D48D-4D39-B92B-ACB831E2F530 | JID3 | 20/04/2015 01:22 | 0 | 10 |
| 40.54286448 | -88.95703384 | 250.8194275 | 5 | 10.19999981 | 332.2265625 | 18E4E1E7-D48D-4D39-B92B-ACB831E2F530 | JID3 | 20/04/2015 01:22 | 0.006533327 | 10 |
| 40.54294319 | -88.95707072 | 250.1737366 | 5 | 9.029999733 | 330.8203125 | 18E4E1E7-D48D-4D39-B92B-ACB831E2F530 | JID3 | 20/04/2015 01:22 | 0.012307287 | 10 |
| 40.54299558 | -88.9571143 | 251.6684875 | 5 | 7.349999905 | 330.8203125 | 18E4E1E7-D48D-4D39-B92B-ACB831E2F530 | JID3 | 20/04/2015 01:22 | 0.016515098 | 10 |

Table 3: Tidy data (Single Journey)

- A tidy data was created with additional feature implements (Duration, Distance) for further analyses.

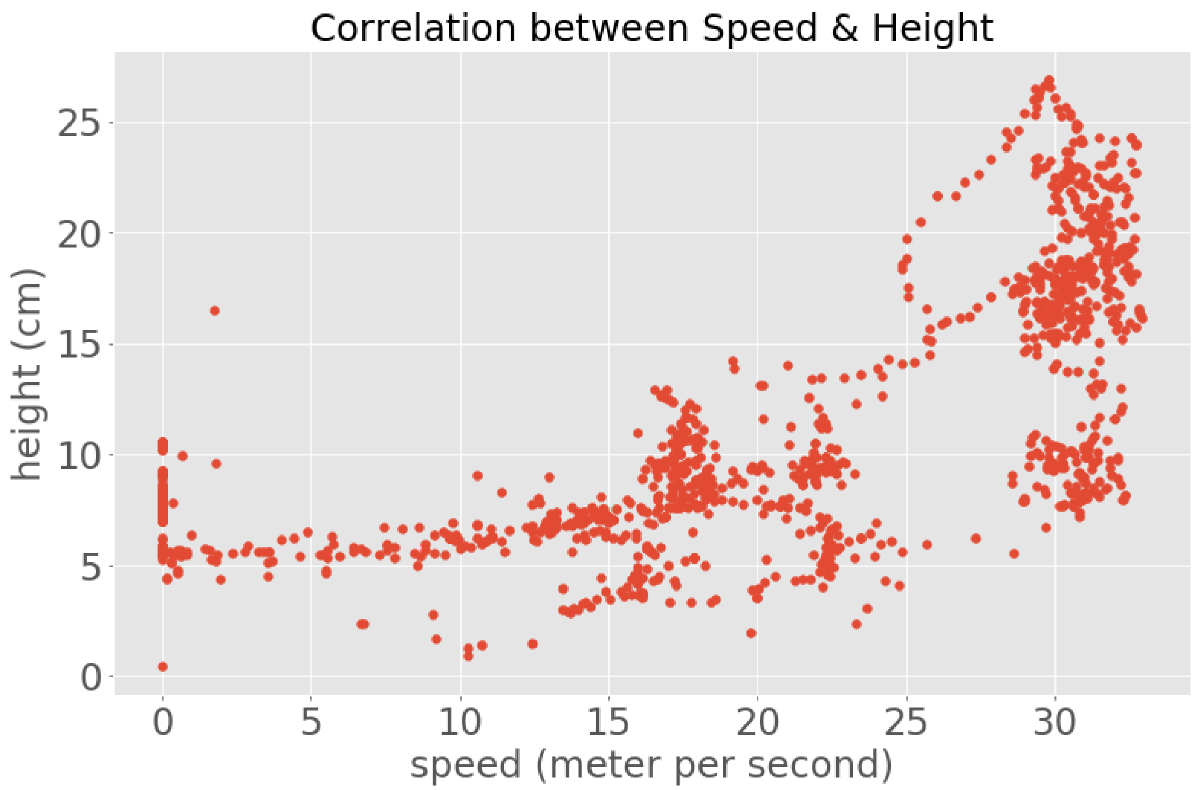


Figure 6: Correlation between speed and height

- Figure 6 is to understand if there is a correlation exists between these variables. The data elicit a positive relationship, though this correlation is not ideal to interpret.

## III. Sensor

What is Sensor?

It is a device that detects and responds to certain input from the physical environment. The specific input could be location, motion, time, etc which can be measured. The output is digital information that can be recorded and converted to a human-readable form and can also be transmitted electronically over a network for further processing.

### Leveraging smartphone to build safer transportation

Here's a challenge. You're a passenger in an automobile, and you've been asked to evaluate whether the driver's habits behind the wheel are "safe" or "risky." But there's a catch: you must collect all your information with your eyes closed.

Imagine your eyes shut, you are denied important information such as your location, traffic conditions, speed limits and traffic signals, and weather conditions. Sightless, your only source of data comes from your sense of motion as the vehicle accelerates, slows down, and turns.

For a quick physics refresher, let's review the difference between speed, velocity, and acceleration:

Speed: How fast an object is moving, usually expressed as distance over time (example: 10 meters per second)

Accelerometer: The rate of change in the velocity of an object. Since it's a rate of change, it's expressed as distance over time (speed), per unit of time. For example, to change speed from 0 to 60 miles-per-hour in 10 seconds, an object must accelerate at 2.682 meters per second per second.

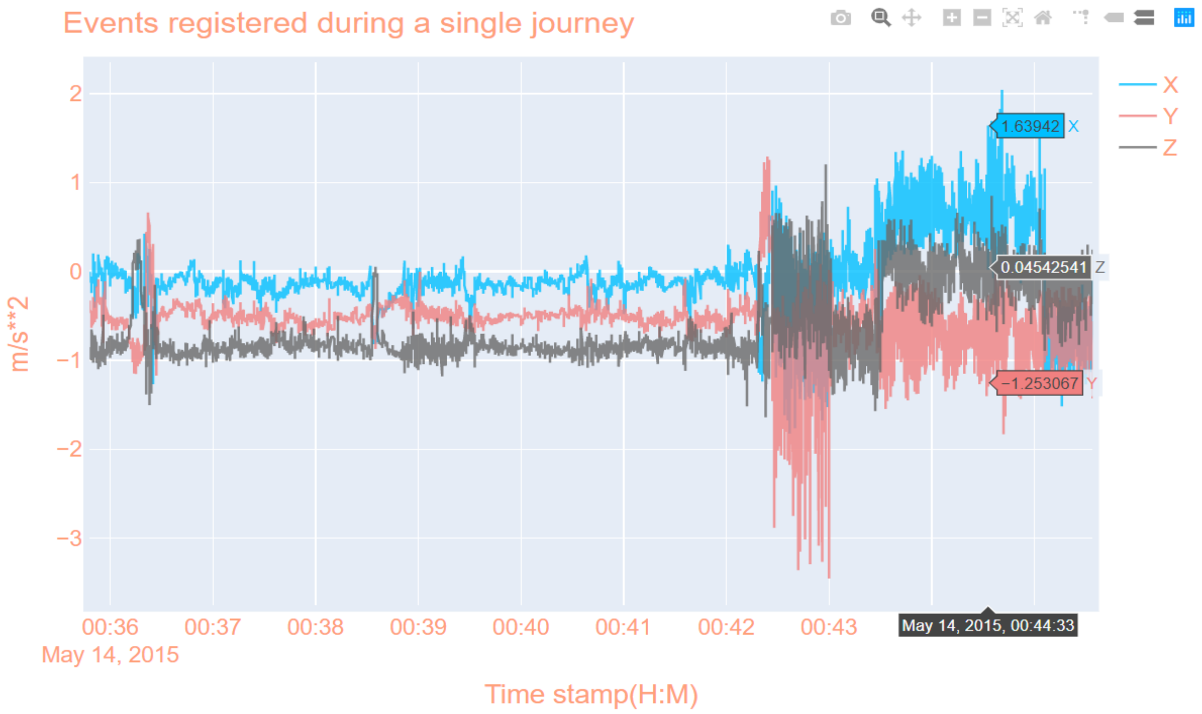


Figure 7: Time against accelerometer for single journey

This figure shows accelerometer data (x, y, z) over time. This journey seems simple enough to locate the mundane events of accelerating or braking after 00:44 hour (midnight). There are a few spikes and dips that might represent more dramatic braking events. Let's use histogram to see what it conveys.

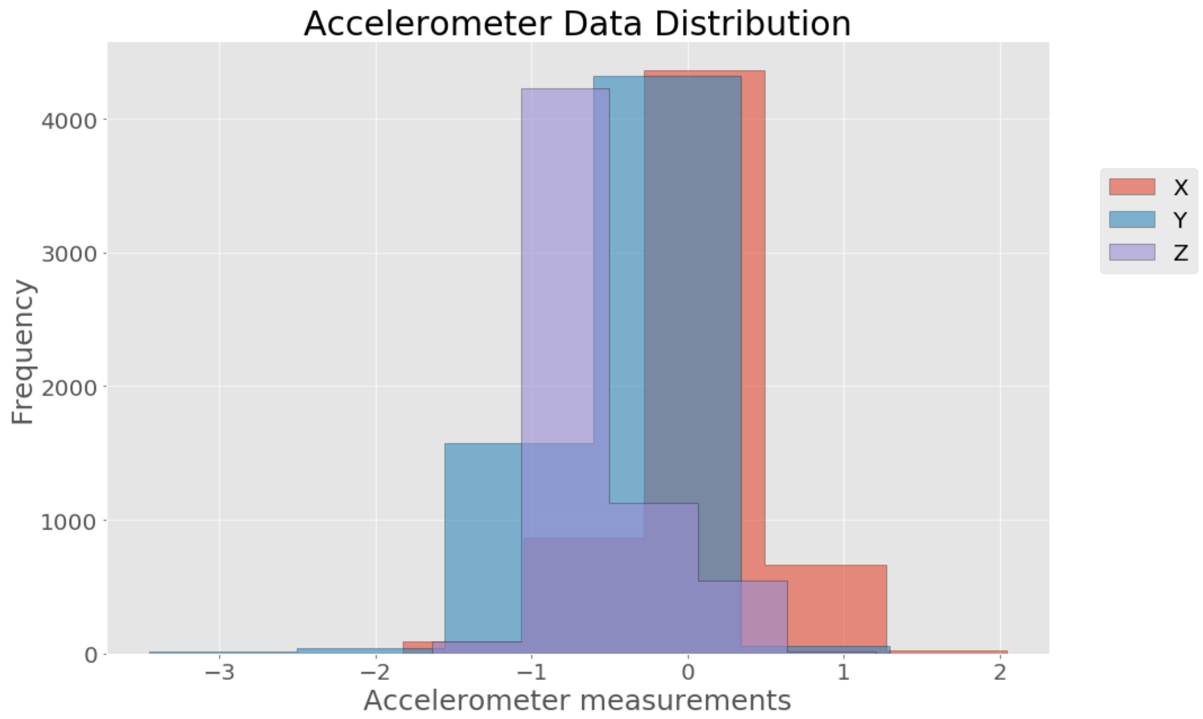


Figure 8: Distribution of X, Y, Z coordiantes

The histogram shows the x axis measurements are centred around 0. During this journey phone appeared to be positioned vertical.

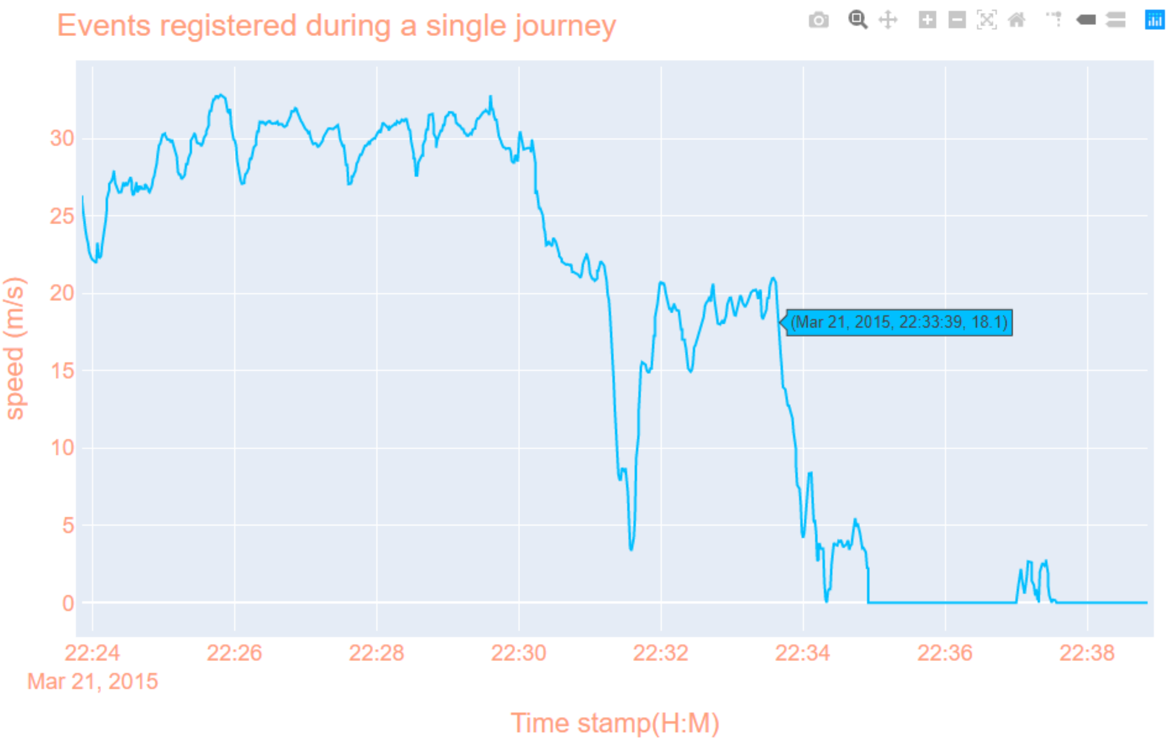


Figure 9: Speed over time for single journey

- Speed seems to not have registered at the start of the journey. As it shows to be travelling with 26 m/s at the beginning of the time. After 7-8 minutes, there has been hard braking or concerning following increase in speed. Later the journey seems to have lower in acceleration.

"Events" from the data

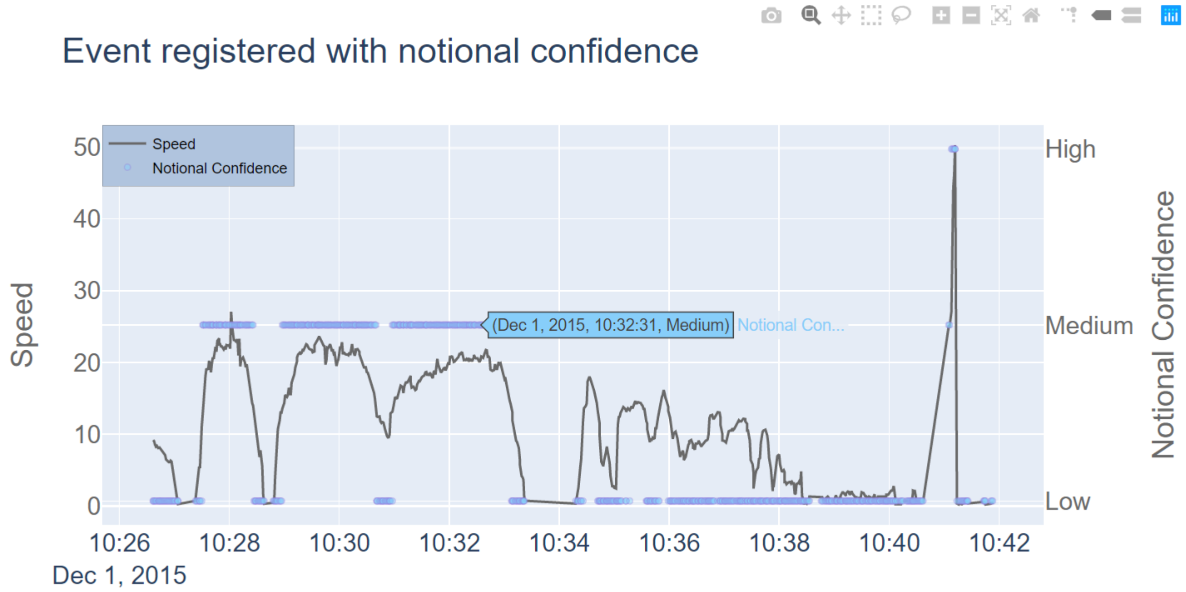


Figure 10: Speed against time with notional confidence

- For example, to change speed from 0 to 60 miles-per-hour in 10 seconds, an object must accelerate at 2.682 meters per second per second. Keeping this in mind, the 'event' that might be an accident was implemented. Also, Notional Confidence, and Notional Severity Index was allocated.

- From this figure 10, journey seems to have less accident probability, as the due is speed with which vehicle was commuting was lower. Since the data for this single journey was dense, the Opacity was reduced to get a better visualisation. As the changes of accident for this journey driver are less, he will not be securing bad source. Though, a bit of support from the FloowCoach can make the journey safest.

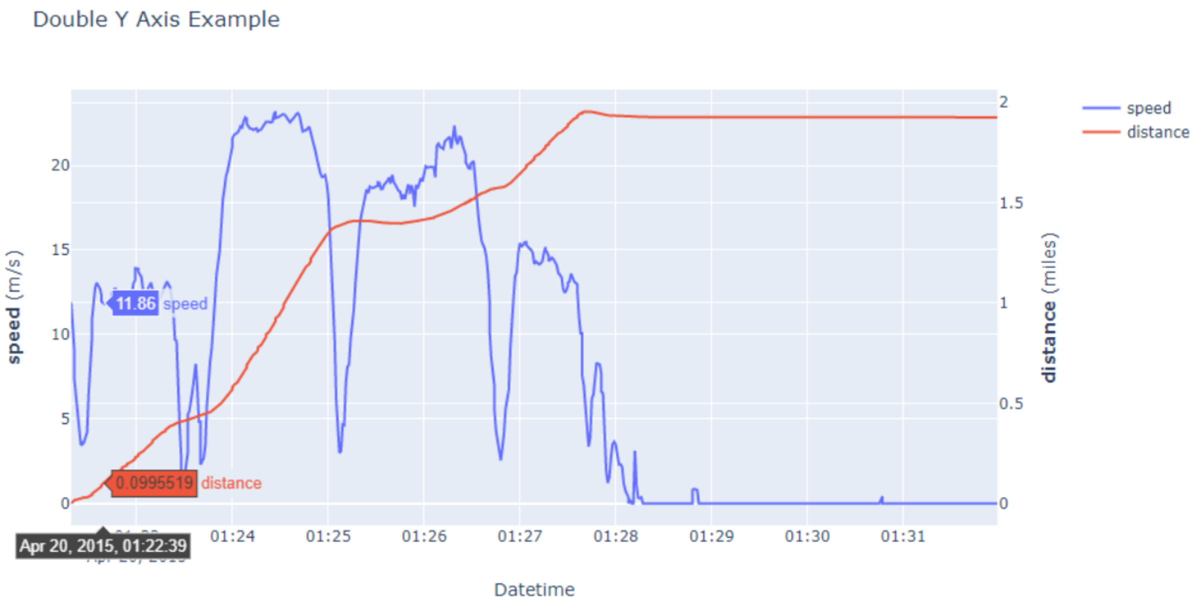


Figure 11: Speed & Distance against time

Pros (Sensor & Event)

- By gathering more granular data (using sensor) a traditional machine learning model can be able to predict an event.
- Identification of events are a better strategy for the insurers to assess insurance risk allowing them to price policies straight from the telematics data.

Cons (Sensor & Event)

- Missing of data & anomalies can inhibit predicting claims, probability of event.

## 4. Result

Using Accelerometer, a better understanding of orientation of phone during the journey was informative as shown in Table 1 and Figure 7 & 8. Having said that, Accelerometer data is not enough for identifying relationship with Speed or vice versa. GPS and Accelerometer data are not also enough to identify fatigue emotion of driver. Current data set can be used for implementing new features like velocity, travel time using the formula mentioned (Jim, 2013). Accelerometer information can be used to identify distraction during each journey.

Based on correlation coefficient information, further accelerometer data was used to implement visualise motion of sensors. Using more granular data, event functionality can be used to ease insurers decision making process and assess policyholder’s profile for risk management. It can further be informative for providing support to driving in uplifting their profile and gaining rewards to achieve safer and better transportation.

## 5. Future Direction

Based on the available features like as accelerometer coordinates (x, y, z) can be helpful to predict the motion of each journey. Monitoring driver's profile by implementing the machine learning model to predict drive skills (like speed, braking, concern, fatigue, and distraction) based on historical data.

Emotional AI can be leveraged by gathering computer vision to analyse emotion state (fatigue, distraction) of the driver behind the wheel, this elicits a safer and comfortable transportation (Craye, 2015).

## 6. Reference

Asuero, A. G., Sayago, A. and González, A. G. (2006) 'The correlation coefficient: An overview', *Critical Reviews in Analytical Chemistry*, pp. 41–59. doi: 10.1080/10408340500526766.

Jiménez-Meza, A., Arámburo-Lizárraga, J. and de la Fuente, E. (2013) 'Framework for Estimating Travel Time, Distance, Speed, and Street Segment Level of Service (LOS), based on GPS Data', *Procedia Technology*, 7, pp. 61–70. doi: 10.1016/j.protcy.2013.04.008.

Craye, C. and Karray, F. (2015) 'Driver distraction detection and recognition using RGB-D sensor', pp. 1–11. Available at: http://arxiv.org/abs/1502.00250.