**FIT 5147**

**Assignment 4**

**Data Exploration and Visualisation Project Proposal**

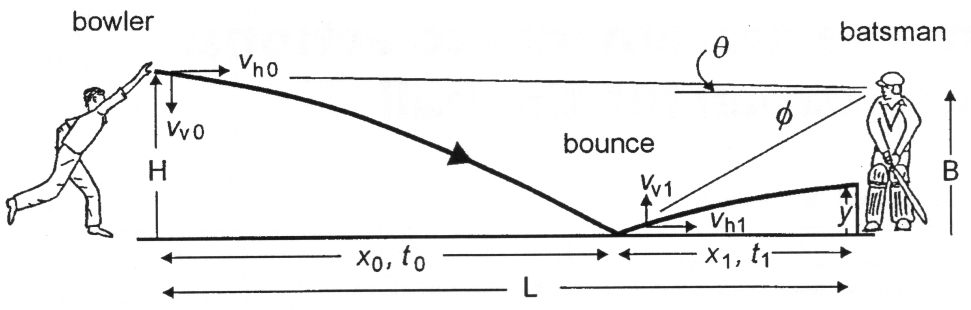
Linus McManamey

**Introduction**

This project is a web based Shiny application written in the R programming language. The application displays information from a dataset that has generated the XYZ coordinates of a cricket ball at three different positions on a cricket pitch, the positions are:

* The XYZ coordinates of the cricket ball when it was released by the bowler
* The XYZ coordinates of the cricket ball when it struck the ground
* The XYZ coordinates of the cricket ball when it arrived at the batsman's stumps

The other important data to be used by this application is the speed of the ball at each of the three location mentioned above and the angle at which the ball left the bowler's hand and the angle of the ball as it it bounced of the surface of the cricket pitch.



Using the data mentioned above this application will attempt to define physical attributes of a cricket pitch based on the differing trajectories of a cricket ball using a linear regression model.

**Problem description and motivation**

**The Problem**

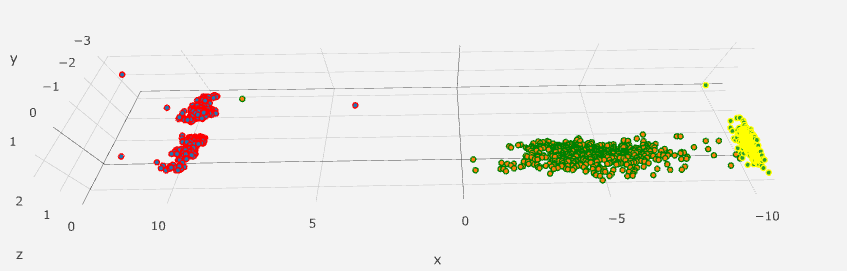
The overall use of the application described in this project is to establish the display of certain areas of a cricket pitch that generate different results when it is struck by a cricket ball.

The display of this information will take into account the speed and angle of the ball when it was released by the bowler, the speed and angle of the ball as it left the pitch and the speed of the ball as it arrives at the far end of the cricket pitch. The **motivation** for displaying this data as interactive visual information would be to discover certain characteristic of a cricket pitch. Based on the scope of this application the there are only two characteristic that are being explored. Originally the project had wider discovery parameters, however based on the allowable time frame only two characteristic were explored. They are the statistical exploration of:

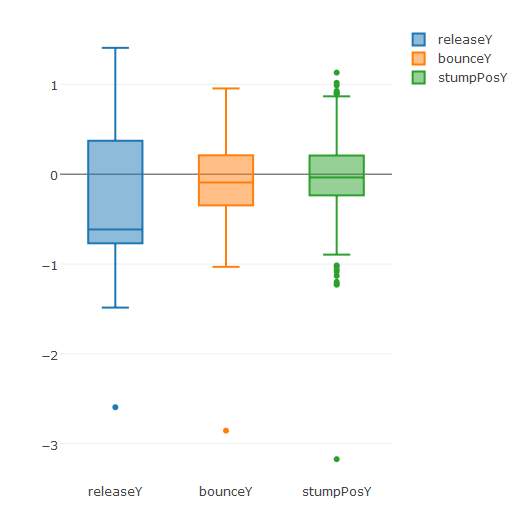
* Areas on the cricket pitch that created high bounce
* Areas on the cricket pitch that created low bounce

**The Data Source Used and Data Cleaning** **Data Transformations**

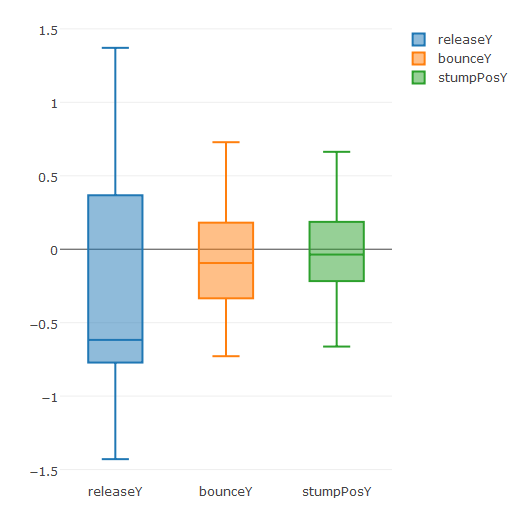
The data source is a csv file of approximately 1750 records. The data wrangling process, the data cleaning and data transformation process is attached in a python Notebook [file name]. There were some final data checking it was revealed that after the initial data cleaning process some outliers still existed in the data set as seen below.



The outliers were plotted as shown below



The outlier were removed in an attempt to normalise the data set as shown below.



**Description of the data source used**

**rX** The X position of the ball as it is released on the XY horizontal plane

**rY** The Y position of the ball as it is released on the XY horizontal plane

**rZ** The height of the ball when it was released from the horizontal plane

**rSpeed** The speed of the ball as it was released from the bowler's hand

**bX** The X position of the ball as it is bounced on the XY horizontal plane

**bY** The Y position of the ball as it is bounced on the XY horizontal plane

**bZ** The height ball of the ball as it when it bounced on the horizontal plane

**bSpeed** The speed of the ball as it hit the horizontal plane

**sX** The X position of the ball when it arrived at the stumps on the XY horizontal

**sY** The Y position of the ball when it arrived at the stumps on the XY horizontal

**sZ** The height of the ball when it arrived at the stumps from the XY horizontal

**sSpeed** The speed of the ball as it arrived at the stumps

**dAng** The angle of trajectory from the horizontal plane when the ball was released

**bAng**  The angle of trajectory from the horizontal plane when the ball bounced

**bowlerEnd** The end of the cricket pitch from which the ball was bowled

**Some Preliminary Data Exploration**

Initially the model was based on the book “Physics of Cricket” by Mark Kidger certain values will be used to calculate the different reactions the the ball when it strikes the cricket pitch.

“The initial velocity of the ball is 160km/h, but **air resistance slows it by 12%...**” (Kidger, 2011). The global variable PRE\_BOUNCE\_DRAG = 0.88 will be used to calculate this physical effect of air drag on the ball.

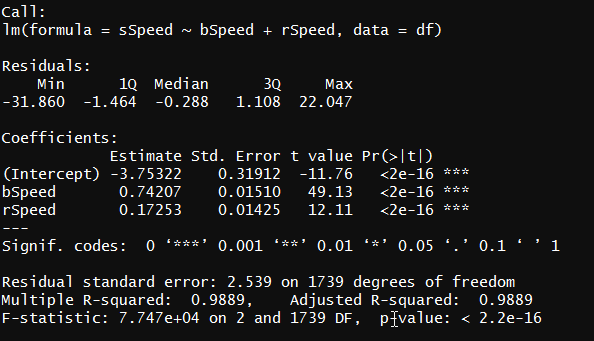
“On bouncing, the ball loses kinetic energy and thus its speed will drop.  **It will rebound with about 55-60%** of its original speed. Air resistance will slow the ball by a further 7% before it reaches the batsman” (Kidger, 2011) The global variables REBOUND\_SPEED\_COEFF = 0.45 and AFTER\_BOUNCE\_DRAG = 0.93 will be be used to calculate the above mentioned effects on the ball.

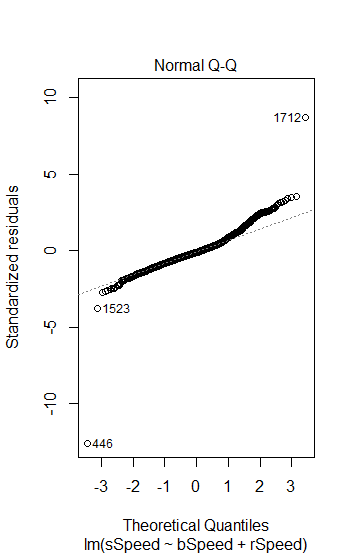
“A **cricket ball** bounces to about one third of that height (0.67 m), in which case it rebounds at a speed of 3.61 m/s. The ratio of these two speeds is 3.61/6.26 = 0.58 and is called the **coefficient of restitution** (COR)”. (Physics of Cricket, 2005) The global variable C\_O\_R = 0.42 Will be used to calculate the height of the ball should bounce back to.

The model to demonstrate different reactions the cricket ball has on the cricket pitch was initially used the above mentioned coefficients on each ball. The projected outcome versus the real out come were compared allowing the user to decided if there are any patterns that may useful or give insight into the makeup of the cricket pitch.

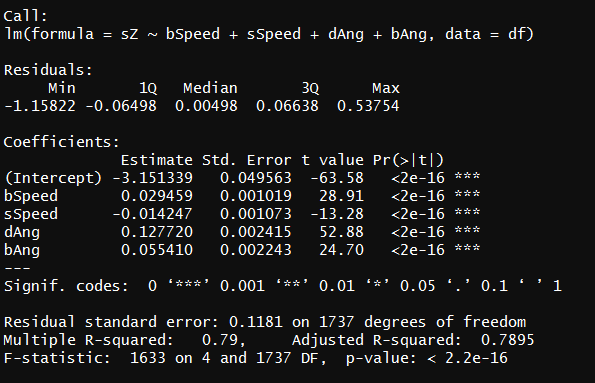
The initial data structure started out as using the physics of cricket as discussed above. The second phase of development of this project use a linear regression model to test the two different relationships. The first was the height of the ball reached at the end of the cricket pitch and the second relationship was the speed of the ball when it arrived at the batsman.

That model was found to be reasonably accurate however a linear regression model was run on the data to model the speed of the ball when it arrived the stumps, listed below:





The above speed model can predict 0.9889 (Adjusted R-squared value) of the variance in the data set making it extremely accurate as a model. It has a well distributed residuals and a mean very close to zero. “*Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis” Rego 2015.* The model has a F-Statistic of 7.747e+04 or 77470 allowing us to reject the null hypothesis that.. (H0 : There is no relationship between the speed of the ball when it arrives at the stumps and the speed of the ball when it bounces and the speed of the ball when it is released)*.*  The model also has an extremely low P- value of 0.00000000000000022, the usual value of less than 0.05 provides 95 percent confidence interval which is used to prove statistical significance. The QQ- plot above s an indication that the data has a normal distribution.



The above summary is the regression model of the stump height which is also a very strong model and will be used to represent the bounce model in predicting the high and low bounce zones.

**Precise description of what message you want your narrative visualisation to convey and what the intended audience is.**

The applications proposed specification has only allowed for a static data source, however a dynamic data source implementation would allow the intended audience for this application - broadcast media the opportunity to use the application in a historic (previous day) or real time (previous ball) tool whilst live broadcasting.

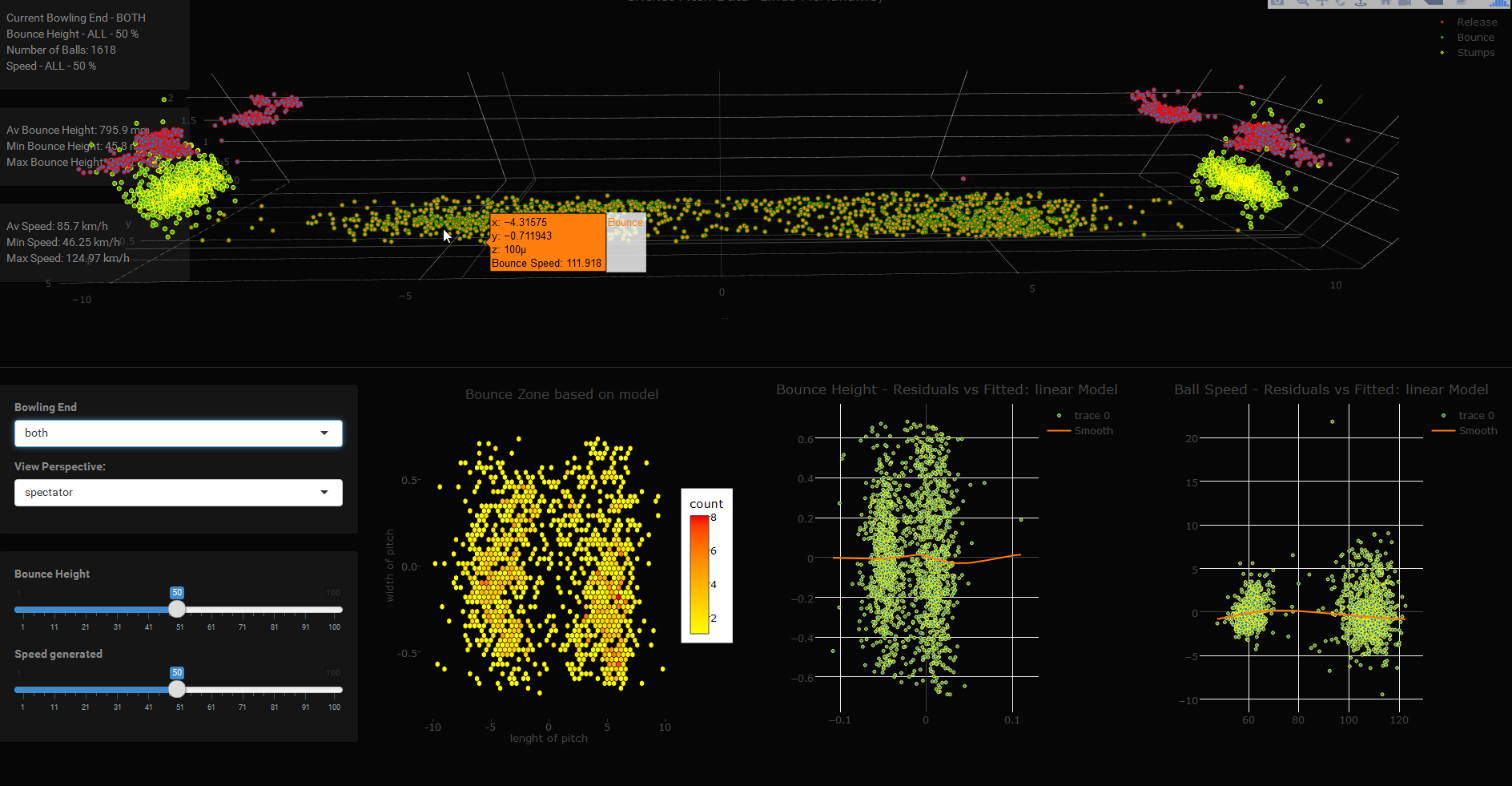
The coaching staff of a cricket team would also be able to utlise this application for analysis of bowler and cricket pitch. It could also be used by the curators of a particular ground in identifying area of the pitch that need special attention to maintain a good playing surface.

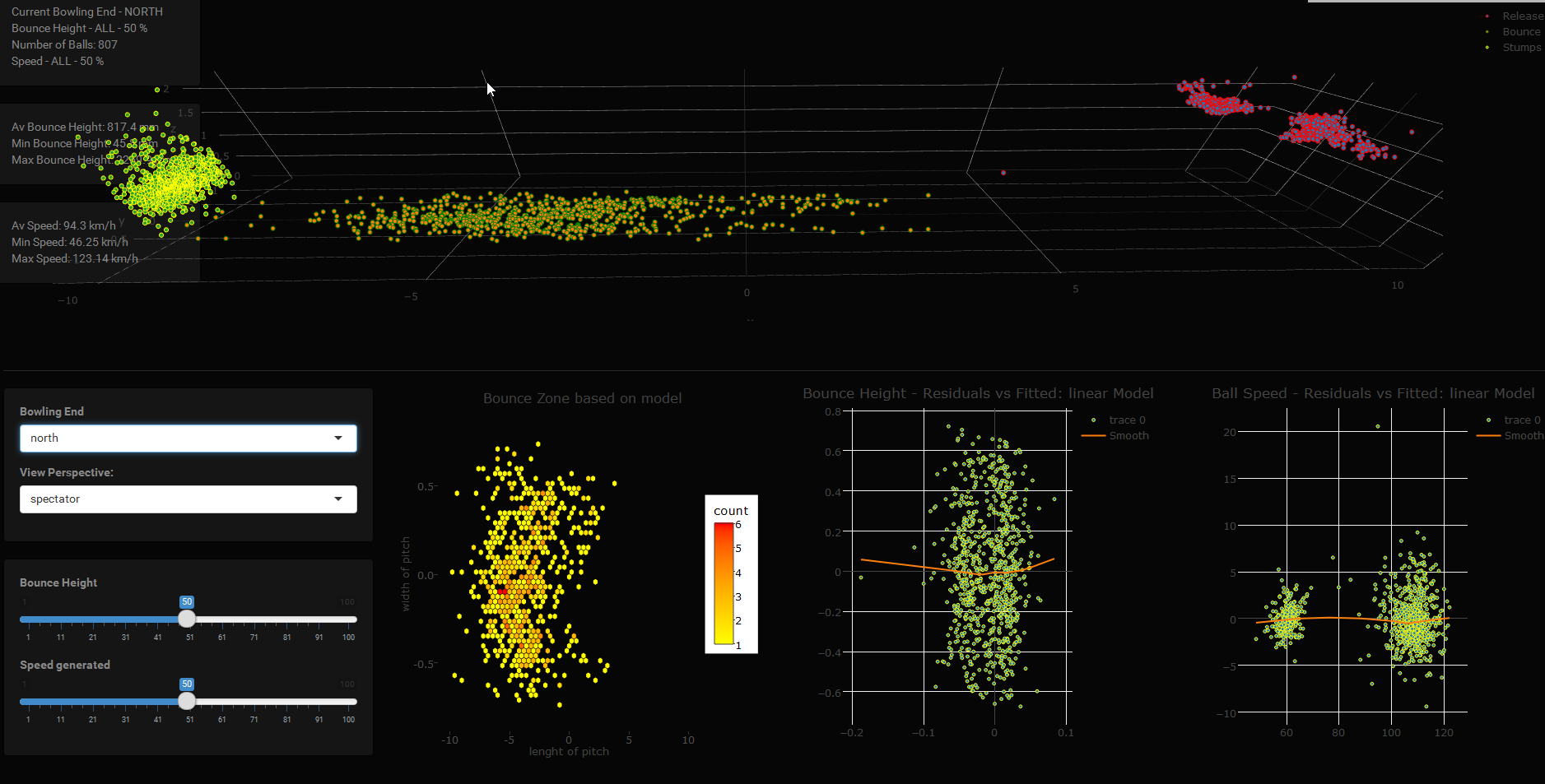
**Description of the visualisation design process including alternatives you considered and the reasons for choosing your final design.**

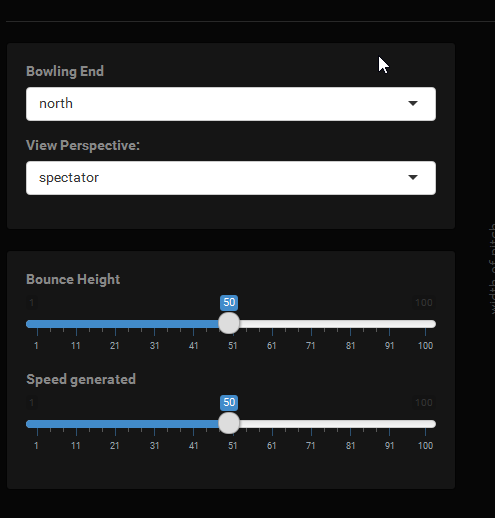
**Instructions for viewing and exploring the narrative visualisation using a standard web browser.**

**(on loading the full screen application both the app.R file and the global.R file need to be complied)**

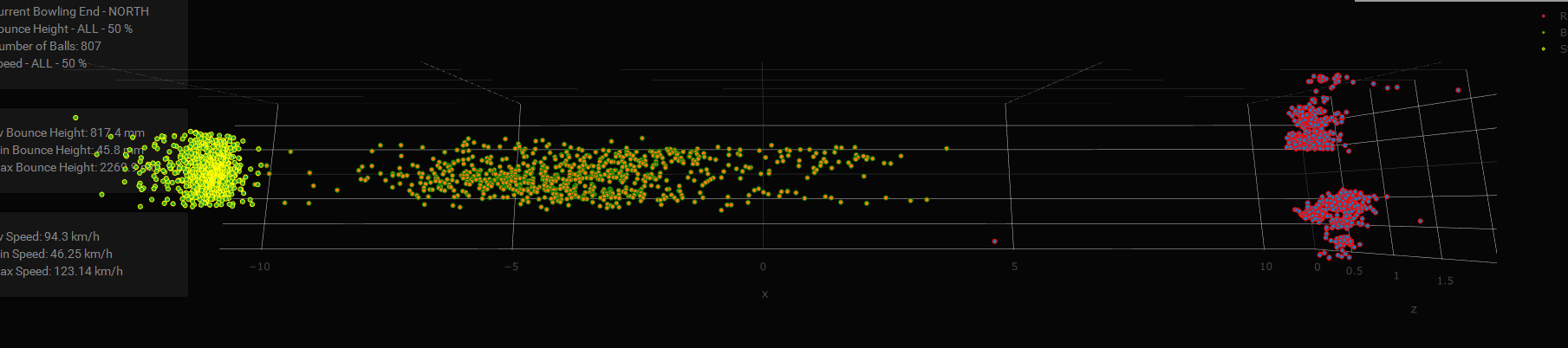
The overall interface will load with all the data loaded as shown below. From this point the user has three inputs to control.



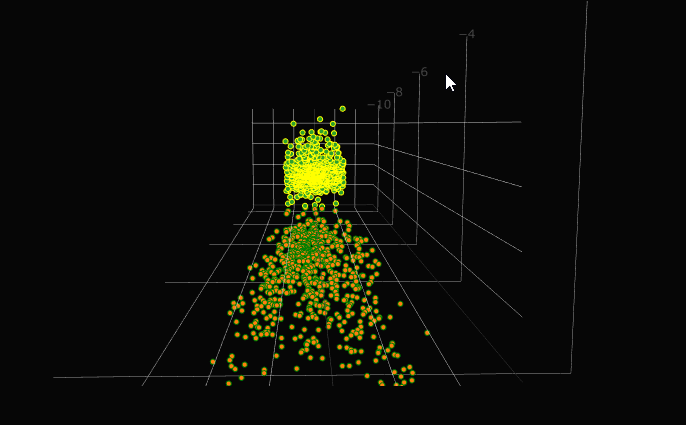
The user can alter the view as shown below, the user has selected to view all the balls bowled from the north end, the user has three choices, north south or both ends as shown above.



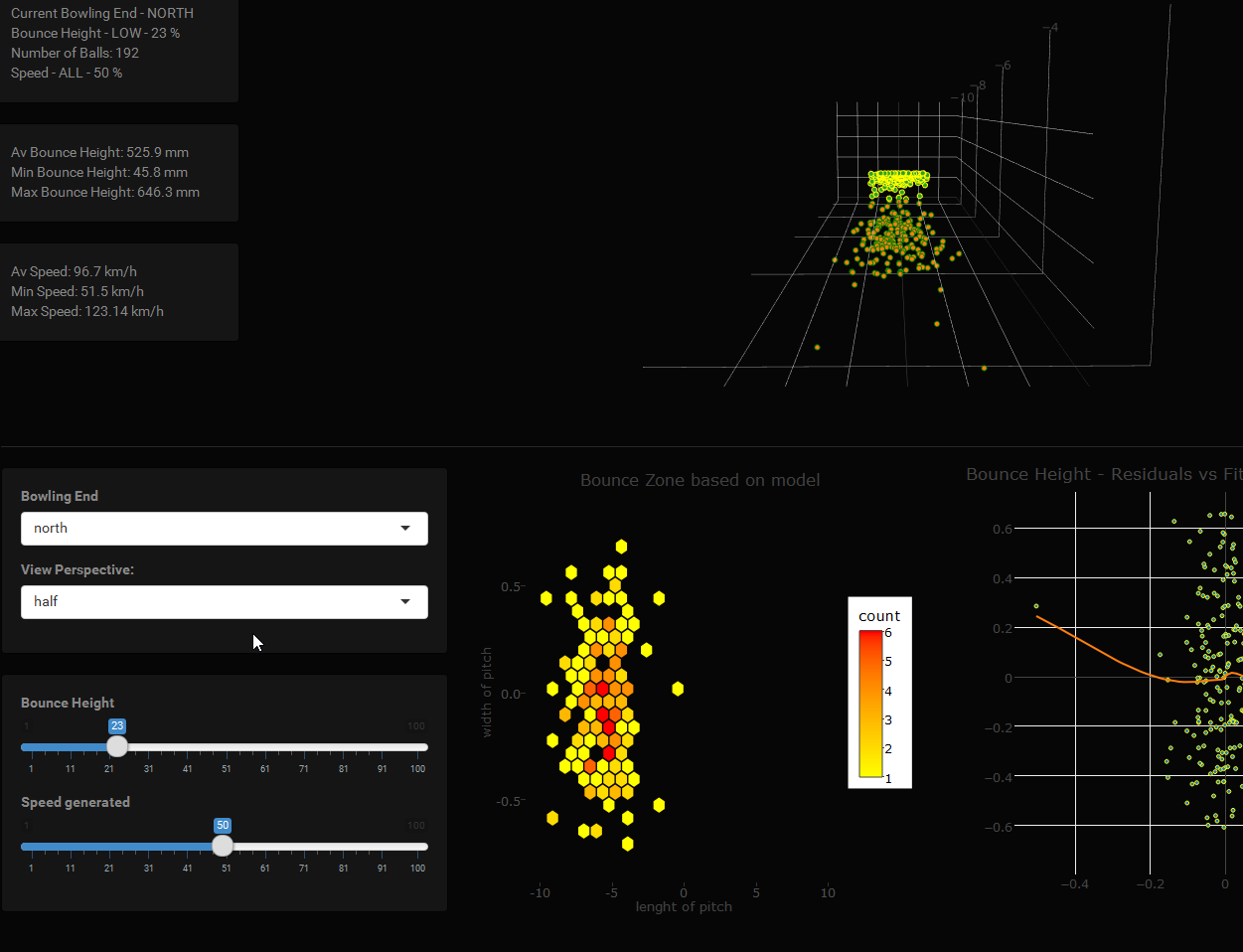
The user can change the perspective view of the data, the image below is the data displayed from above.

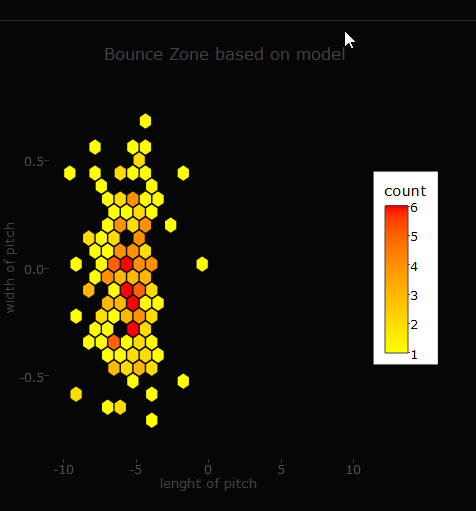
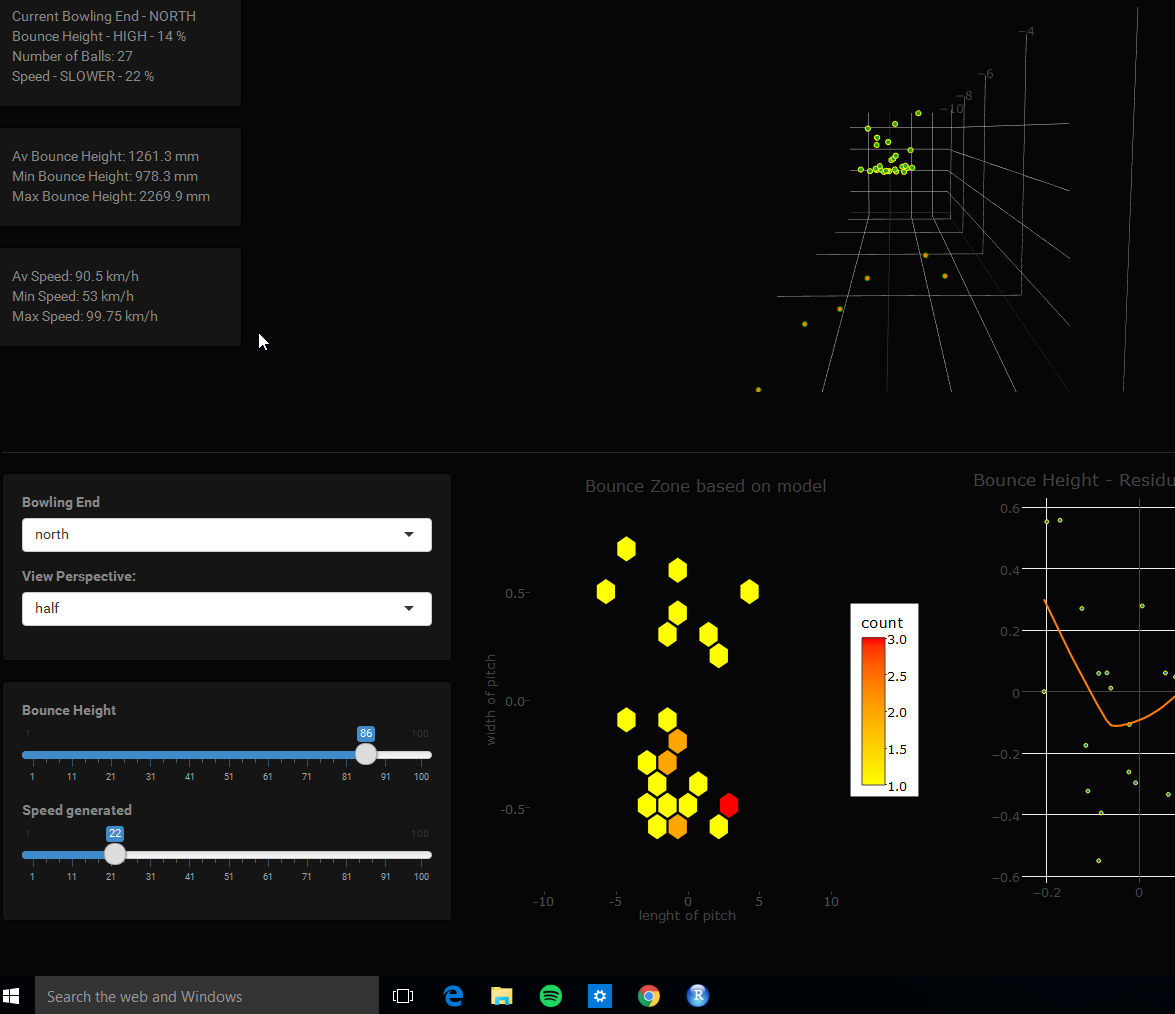


Or the user can change the perspective so they are viewing the data from the middle of the pitch.



The user has the option to then choose either the higher or lower bouncing ball.





The hex bin map indicates the areas of lower or higher bounce.

**A conclusion summarising what you achieved and also a reflection what you learnt in this project and what in hindsight you might have done differently**

In conclusion the application could do a great deal more but time constraints did not allow for any further functionality. There is a great deal more that could be developed with deviation and swing of the ball. There was an attempt made to connect the trajectories of each of the balls, however at the moment the restraints of the plotting software and the time required to circumvent them were not enough to allow it. A great deal was learnt about regression modeling, how to implement a shiny application and general R programming which was really enjoyed..

**References:**

Kidger, Mark. Physics of Cricket, edited by Mark Kidger, Nottingham University Press, 2011. ProQuest Ebook Central.

<http://www.physics.usyd.edu.au/~cross/cricket.html>.Physics of Cricket

# Quick Guide: Interpreting Simple Linear Model Output in R , *Felipe Rego,*

#### *October 2015*

<http://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R>

# **All subset regression with leaps, bestglm, glmulti, and meifly**

<https://rstudio-pubs-static.s3.amazonaws.com/2897_9220b21cfc0c43a396ff9abf122bb351.html>