**Next steps:**

1. Please clean up your R scripts, honestly they are messy right now.
2. Produce a plot of the player coeffs!
3. Make this doc into a report.

**Data acquisition**

Where is data from?

Data is basically from cricsheet.org , a website which deals with cricket statistics. It was having files of all the IPL seasons that happened.

The dataset contains these variables:

* Strike batsman (This Column contains the name of the batsman facing bowler at the time of play. Near about 436 factor variables are there)
* Ball (Not of any significance .Needed to be removed. Not a useful column for analysis)
* Inning ( Inning column denotes which inning is there when the ball was bowled out .Used as a factor variable with 4 factors)
* Over (Usually denotes the serial number of over. An over is a cumulative of 6 balls . Used as a factor variable with 20 factors)
* Team (Denotes the name of a team which is batting at the moment. It is a factor variable with 12 factors)
* NonStrikeBatsman (Denotes the name of Batsman which is acting as Non Striker at the moment. Factor variable with 435 factors)
* Bowler(Denotes the name of bowler.Factor variable with 385 factors)
* RunsScored ( Gives the number of Run Scored on the ball bowled out.Target column of prediction of our model)
* Extras( Denotes the run which are expensed by Bowler , but not counted in the total of Batsman.Counted in the total of Team batting at the time.Column with integer values)
* DismissalMethod ( Method of dismissal of a Batsman batting i.e . Strike Batsman )
* BatsmanOut (Name of the batsman got out. Not useful in our analysis)

**Description of Files -:**

* Inferences – This folder contains various Inferences made from various models and tests applied.
* Model Results – Various results such as Coefficients and summary of models are saved here.
* Raw Data – Contains the raw data used in model.
* RScripts – It contains all the R Scripts through which data cleaning, predictions, tests were performed.
* BindIPLData – It is a csv file that has 1,36,000 rows with 12 columns depicting various variables . RunsScored is the target column which we have to predict.

**Work Completed -:**

* Refined raw data and bind it in one csv file named BindIPLData.
* Completed univariate and bivariate Numerical analysis for various variables containing numerical values.
* Visualised data using common pictorial representations such as histogram of R.
* Applied poisson regression with different combinations of predictors to predict Runs Scored.
* Checked over dispersion of data using dispersion test of AER package and also using dispersion ratio (Residual Deviance/Degree of Freedom).
* Applied Chi Square test between various models to find out the significant predictor variables.
* On the basis of results of Chi Square test, formulated a model with various predictor variables.
* Applied zero inflated regression to avoid over dispersion of data.

**Future plans -:**

* Will apply Vuong Non-Nested Hypothesis test to decide best model between zero inflated and poisson regression model.
* Will try to draw out various inferences from different models based on the results they are producing.

**Questions -:**

* Kindly, suggest any suggestions to consider in the work flow of this case study.
* Suggest me an authenticated source to refer for steps being taken in case study as now I am just googling all the stuff and on the basis of various articles, I am applying things required.
* What should be the basis of checking the appropriateness of our prediction model? I would be happy , if you can brief me the parameters to be taken care of and their significance.
* R is taking much time as factor variables in Strike Batsman column is near about 436 and thus, I am not able to run step function with two or more predictors as time taken is too large. Suggest a solution in this regard.