**Model Trees for Identifying Exceptional Players in the NHL Draft**

Paper Track (Cambria, 14pt)

Paper ID

**Abstract**

Recruiting strong players is crucial for a team’s success. We describe a new data-driven interpretable approach for assessing draft prospects in the National Hockey League. Successful previous approaches have 1) built a *predictive model* based on player features [1], or 2) derived performance predictions from the observed performance of *comparable players* in a cohort [2]. This paper develops model tree learning, which incorporates strengths of both model-based and cohort-based approaches. A model tree partitions the feature space according to the values of discrete features, or learned thresholds for continuous features. Each leaf node in the tree defines a group of players, easily described to hockey experts, with its own group regression model. Compared to a single model, the model tree forms an ensemble that increases predictive power. Compared to cohort-based approaches, the groups of comparables are discovered from the data, without requiring a similarity metric. [results]

**1. Introduction: [Identifying Predictive Player Cohorts?]**

* Two approaches :
  + Cohort-based, model-based
  + Discuss strengths, weaknesses of each.
* Model-ensemble addresses strengths, weaknesses of each. Get the best of both worlds.
* Model tree:
  + Interpretable
  + Allows for variety in time and league
  + Increases predictive power
* Tingle on team draft quality
* Short summary of results
* Two Goals
  + Build predictive model
  + identify cohorts (cite Weissbock, Sony)
  + Unification: identify cohorts s.t. we can build good predictive models.
* More accurate predictions, non-linearity
* Predictively valid cohorts.
* Problem: Cohorts should be interpretable. Interpretability even more important than accuracy.
* Solution: Equation tree (LMT). What’s the terminology for these tree models.
* Get good results in terms of tree, Spearman.
* Extract explainability

Compared to previous player assessment work, our model tree learning approach has several advantages, including:

* Interpretable cohorts: the nodes in tree model provide clear metrics to cluster players. For example, CSS\_rank is more valuable than Height in determining which cohort the player belongs to. Meanwhile, the coach or hockey expert can find similar players for a prospect easily by inspecting the tree model based on his draft performance data.
* Predictive performance values. Within each cohort, the probability weather a player can play a game in NHL or not can be calculated through the cohort own logistic regression model. This means player can be ranked with comparable players in each cohort, e.g. strongest player against strongest cohort, thus, top players is obvious and visible in the tree model.
* Exceptional mining. In tree model learning, we compute the weighted feature difference for each player, compared to its cluster mean value. An expert is capable of looking at the strongest and weakest point of a player. This explains the overall ranking by providing a drill-down way to inspect prospect specifically.

*Paper Outline.*

**2. Previous Work**

* Regression models: shuckers (gam), salford guys:poisson, Wilson
* Cohort-based review
* Model-based clustering
* Exception Mining

*Regression Approaches*. For evaluating ice hockey players, regression models are widely adopted in previous work. [Edward.H] extended classic Poisson model to assess players at the NHL level based on a series of historical performance metrics. [3] used generalized linear model to predict a prospect career GP. [1] shows that “numbers game” can beat scouting by applying generalized additive model in drafting process. However, with respect to draft data, the regression model can aggravate the zero-inflation problem, since almost half of the draftees never play a game in NHL. As a result, it can lead the model to fitting towards GP equals zero, which lowers the predictive power.

Similarity-Based Approaches. Others also have used clustering approaches to group similar players. [2] created Prospect Cohort Success to generate comparable NHL player for draftees from various major junior leagues to support drafting decisions. However, this clustering method is inferior to regression model in terms of predictive power. To address these issues, an intuitive approach would be to combine tree induction to cluster players with building linear regression in leaf node to predict cohort performance.

In our current work, the player cohort is generated through LogitBoost algorithm[Jerome Friedman] to build a tree with logistic regression model in leaf nodes. Our tree model uses over 10 player features as input and it automatically chooses the most relevant of them to differentiate players. As a result, each player can be easily assigned to his own cohort along with a predictive logistic model.

Wilson used several machine learning models to predict whether an NHL player achieves a career total of more than 160 games in NHL, given aggregate statistics for his four four NHL seasons [3]. In contrast, we use junior league data for predicting future NHL performance.

**3. Dataset**

Our data was obtained from public-domain on-line sources, including nhl.com, eliteprospects.com, and draftanalyst.com. We are also indebted to David Wilson for sharing his NHL performance dataset [Wilson 2016]. The full dataset is posted on the worldwide web [Draft Dataset 2017]. Following [Shuckers 2016], we took as our dependent variable **the total number of games *gi*** played by a player after 7 years under an NHL contract. Our dataset includes also the total time on ice after 7 years. The results for time on ice were very similar to number of games, so we discuss only the results for number of games. The independent variables include demographic factors (e.g. age), count metrics for the year in which a player was drafted (e.g., goals scored), and the rank assigned to a player by the NHL Central Scouting Service (CSS). We consider players drafted into the NHL between 1998 to 2008. If a player was not ranked by the CSS, we assigned (1+ the maximum rank for his draft year) to his CSS rank value. [is t his what Shuckers did?] Another preprocessing step was to pool all European countries into a single category. Table 1 lists all data columns and their meaning. Table 2 shows an excerpt from the data set.

|  |  |
| --- | --- |
| Variable Name | Description |
| id | nhl.com id for NHL players, otherwise Eliteprospects.com id |
| DraftAge | Age in Draft Year |
| Country | Nationality. Canada -> 'CAN', USA -> 'USA', countries in Europe -> 'EURO' |
| Position | Position in Draft Year. Left Wing -> 'L', Right Wing -> 'R', Center -> 'C', Defencemen -> 'D' |
| Overall | Overall pick in NHL Entry Draft |
| CSS\_rank | Central scouting service ranking in Draft Year |
| rs\_GP | Games played in regular seasons in Draft Year |
| rs\_G | Goals in regular seasons in Draft Year |
| rs\_A | Assists in regular seasons in Draft Year |
| rs\_P | Points in regular seasons in Draft Year |
| rs\_PIM | Penality\_in\_Minutes in regular seasons in Draft Year |
| rs\_PlusMinus | Goal Differential in regular seasons in Draft Year |
| po\_GP | Games played in playoffs in Draft Year |
| po\_G | Goals in playoffs in Draft Year |
| po\_A | Assists in playoffs in Draft Year |
| po\_P | Points in playoffs in Draft Year |
| po\_PIM | Penalty\_in\_Minutes in playoffs in Draft Year |
| po\_PlusMinus | Goal differential in playoffs in Draft Year |
| sum\_7yr\_GP | Total NHL games played in player's first 7 years of NHL career |
| sum\_7yr\_TOI | Total NHL Time\_on\_Ice in player's first 7 years of NHL career |
| GP\_7yr\_greater\_than\_0 | Played a game or not in player's first 7 years of NHL career |

Table Player Attributes listed in dataset

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | Player  Name | Draft  Age | Coun  try | Hei  ght | Wei  ght | Posi  tion | Over  all | CSS\_  rank | rs  \_GP | rs\_  G | rs\_  A | rs\_  P | rs\_  PIM | rs\_  Plus  Minus | sum\_  7yr\_  GP | sum\_  7yr\_  TOI | GP\_  7yr > 0 |
| 8469455 | Jason  Speeza | 18 | CAN | 75 | 210 | C | 2 | 1 | 63 | 46 | 76 | 122 | 45 | 0 | 322 | 5714 | yes |
| 8469474 | Colby  Armstrong | 19 | CAN | 74 | 185 | R | 21 | 8 | 76 | 36 | 43 | 79 | 168 | 0 | 199 | 3397 | yes |
| 8469454 | Ilya Kovalchuk | 18 | EURO | 75 | 230 | R | 1 | 1 | 53 | 43 | 24 | 67 | 141 | 0 | 466 | 9905 | yes |
| 8470105 | Scottie Upshall | 19 | CAN | 72 | 200 | R | 6 | 5 | 68 | 35 | 54 | 89 | 149 | 0 | 230 | 2998 | yes |
| 8471675 | Sidney Crosby | 18 | CAN | 71 | 200 | C | 1 | 1 | 73 | 78 | 110 | 188 | 94 | 0 | 434 | 9144 | yes |
| 8474141 | Patrick Kane | 19 | USA | 71 | 177 | R | 1 | 2 | 65 | 67 | 87 | 154 | 94 | 44 | 515 | 9927 | yes |
| 8473419 | Brad Marchand | 18 | CAN | 69 | 181 | L | 71 | 80 | 68 | 29 | 37 | 66 | 83 | 40 | 218 | 3418 | yes |
| 27 | Yared Hagos | 18 | EURO | 73 | 218 | C | 70 | 24 | 43 | 11 | 26 | 37 | 24 | 1 | 0 | 0 | no |

Table Sample Player Data. rs = regular season. We use the same statistics for the playoffs (not shown).

## 4. Model Tree Construction

Model trees are a flexible formalism that can be built on the basis of any regression model. An obvious candidate for a regression model would be linear regression; an alternative would be a generalized additive model [Shuckers 2016], or a Poisson regression model specially built for predicting counts [Turrash]. We introduce a different approach: a logistic regression model to predict whether a player will play any games at all in the NHL (*gi>0)*. The motivation is that many players in the draft never play any NHL games at all (up to 50% depending on the draft year). This poses an extreme zero-inflation problem for any regression model that aims to predict directly the number of games played. In contrast, for the *classification problem* of predicting whether a player will play any NHL games, zero-inflation means that the data set is balanced between the classes. This classification problem is interesting in itself. For instance, a player agent would be keen to know what chances their client has to participate in the NHL. The logistic regression probabilities *pi=P(gi>0)* can be used not only to predict whether a player will play any NHL games, but also to *rank players* such that the ranking correlates well with the actual number of games played. Intuitively, if we can be confident that a player will play at least one NHL game, we can also expect the player to play many NHL games. Our method is therefore summarized as follows. [why does the list font look funny? Updated still looks funny]

1. Build a tree whose leaves contain logistic regression model.

2. The tree value splits assign each player *i* to a unique leaf node *li*, with a unique logistic regression model *m(li)*.

3. Use *m(li)* to compute a probability *pi=P*(player *i* plays more than 0 NHL games in his first 7 years).

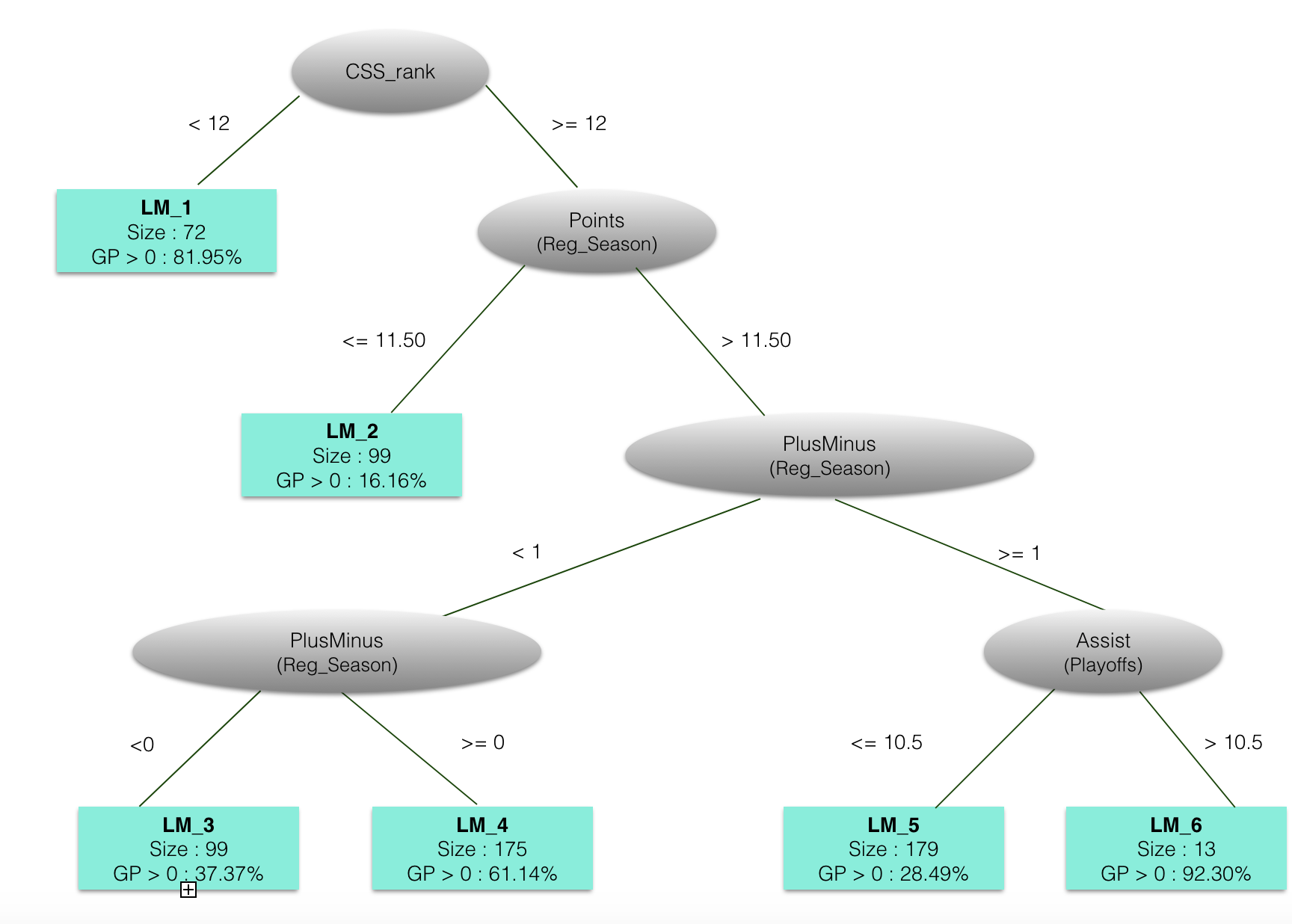


Figure Logistic Regression Model Tree for the 2004,2005,2006 cohort. The Tree was built using the LMT package of the Weka program [Hall et al. 2009; Witten 2016].

*Alternative Approaches*. Below we show the results of model tree learning based on standard learning regression. We also tried a more complicated hierarchical design where we first assign players to the non-zero-game class using the logistic regression model tree, then predict the number of games played for the non-zero-game players using linear regression. It is also possible to build a model tree whose leaves contain linear regression models. The logistic regression tree produced better predictions than the alternatives; for more information please see the appendix.

## 5. Results: Predictive Modelling

Following [Shuckers 2016], we evaluated the predictive accuracy of the LMT model using the Spearman correlation between two player rankings: i) the performance ranking based on the actual number of NHL games that a player played, and ii) the ranking of players based on the probability *pi* of playing at least one game. We also compared it with iii) the ranking of players based on the order in which they were drafted. The draft order can be viewed as the ranking that reflects the judgement of NHL teams. We provide the formula for the Spearman correlation in the appendix; the scale of the correlation is [what, 0 to 1?]. We also obtained good results using Kendall’s tau rank correlation, detailed in the appendix. Table 3 shows the Spearman correlation for different rankings.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Training Data NHL Draft Years | Out of Sample  Draft Years | Draft Order Spearman Rank Correlation | Schucker’s Approach Spearman Rank Correlation | LMT classification accuracy | LMT Spearman Rank Correlation |
| 1998, 1999, 2000 | 2001 | 0.43 | 0.46 | 84.5379% | **0.86** |
| 1998, 1999, 2000 | 2002 | 0.30 | 0.50 | 87.6832% | **0.90** |
| 2004, 2005, 2006 | 2007 | 0.46 | 0.52 | 75.5463 % | **0.83** |
| 2004, 2005, 2006 | 2008 | 0.51 | 0.56 | 68.2191 % | **0.75** |

Table Predictive Performance (our model, over all draft ranking) using Spearman Rank Correlation. Bold indicates the best values.

The rank correlation with actual number of games played was excellent for the LMT model. Compared to the actual draft order, the smallest improvement was 0.25. As one would expect, the rank correlation is higher the more accurately the model predicts which players will play at least one game. For the generalized additive model (*gam*), the reported correlations were 2001:0.53, 2002:0.54, 2007:0.69, 2008:0.71 [Schuckers 2016]. Our correlation is not directly comparable to the generalized additive model because of differences in data preparation: the *gam* model was applied only to drafted players who played at least one NHL game, and the CSS rank was replaced by the Cescin conversion factors: for North American players, multiply CSS rank by 1.35, and for European players, by 6.27. [need reference for Cescin] The Cescin conversion factors represent an interaction between the player’s league and the player’s CSS rank. A model tree offers another approach to representing such interactions: by splitting on the player location node, the tree can build a different model for each location. Whether the data warrant building different models for different locations is a data-driven decision made by the tree building algorithm. The same point applies to other sources of variability, for example the draft year. In the next section we examine the interaction effects captured by the model tree in the different models learned in each leaf.

[@yeti: what were Shuckers’ numbers? Can we reproduce his gam result? Round down to 2 significant digits.

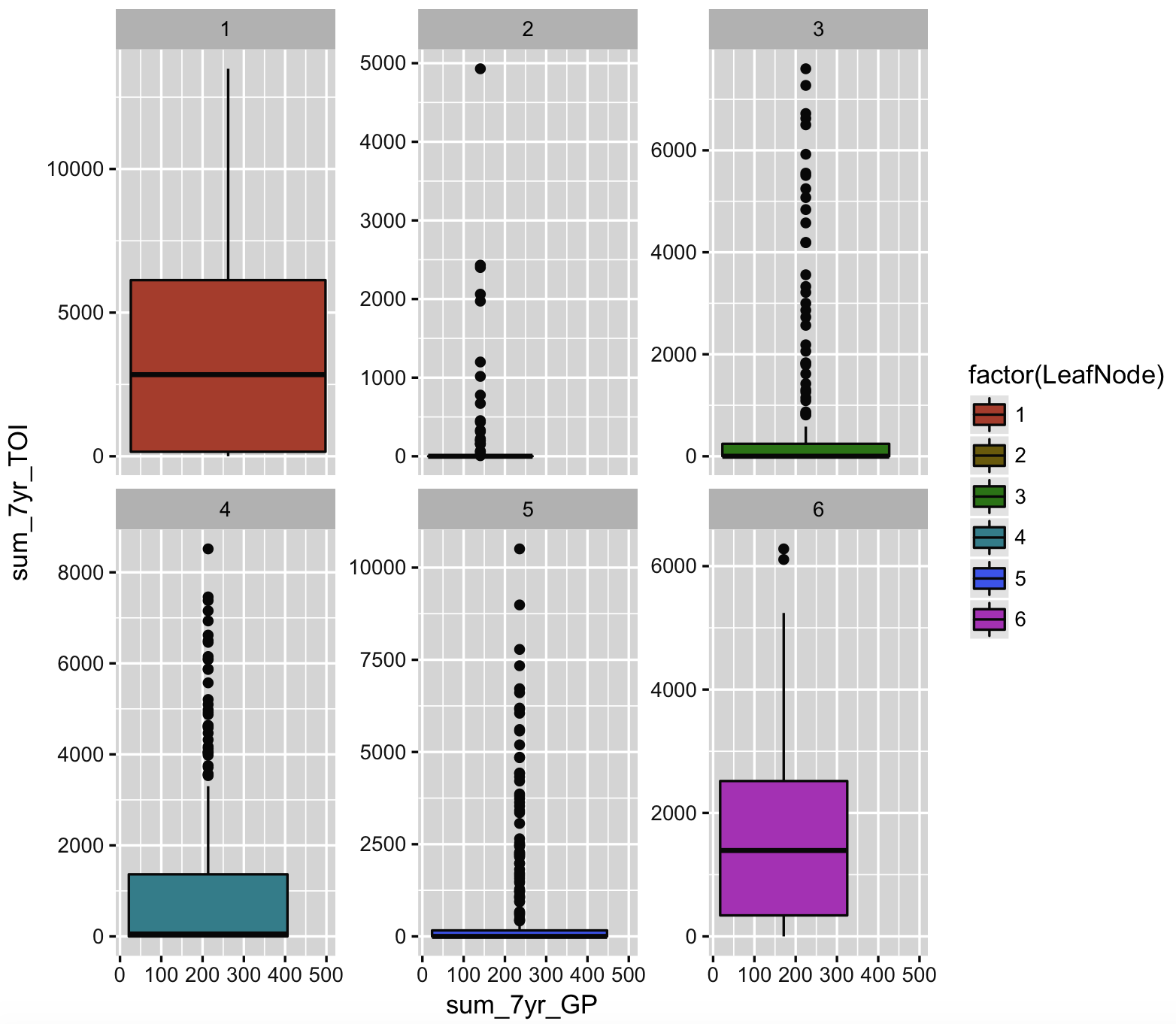
Answer: updated. We can got a lower result for now. The formula we are using basically is as following:

gam(sum\_7yr\_GP ~ Height\_norm + Weight\_norm + s(position) + cescin\_rank\_norm + rs\_Goals\_against\_avg\_norm + rs\_Points\_per\_Game\_norm + po\_Goals\_against\_avg\_norm + po\_Points\_per\_Game\_norm )]

**6. Results: Learned Groups and Logistic Regression Models**

We show boxplots for the distribution of our dependent variable *gi* the total number of NHL games played after 7 years under an NHL contract. The strongest groups are, in order, 1, 6, and 4. The other groups show weaker performance on the whole, although in each group some players reach a high number of games.

1. [why does it say TOI in the plot?
2. The fonts need to be much bigger
3. Any way you can put them on the same scale so it’s easier to compare? I know some of them will look bad.
4. Can we repeat the definitions, maybe on the right? Or even repeat the tree?
5. Maybe mark players that we discuss later?
6. Maybe order groups by strength?



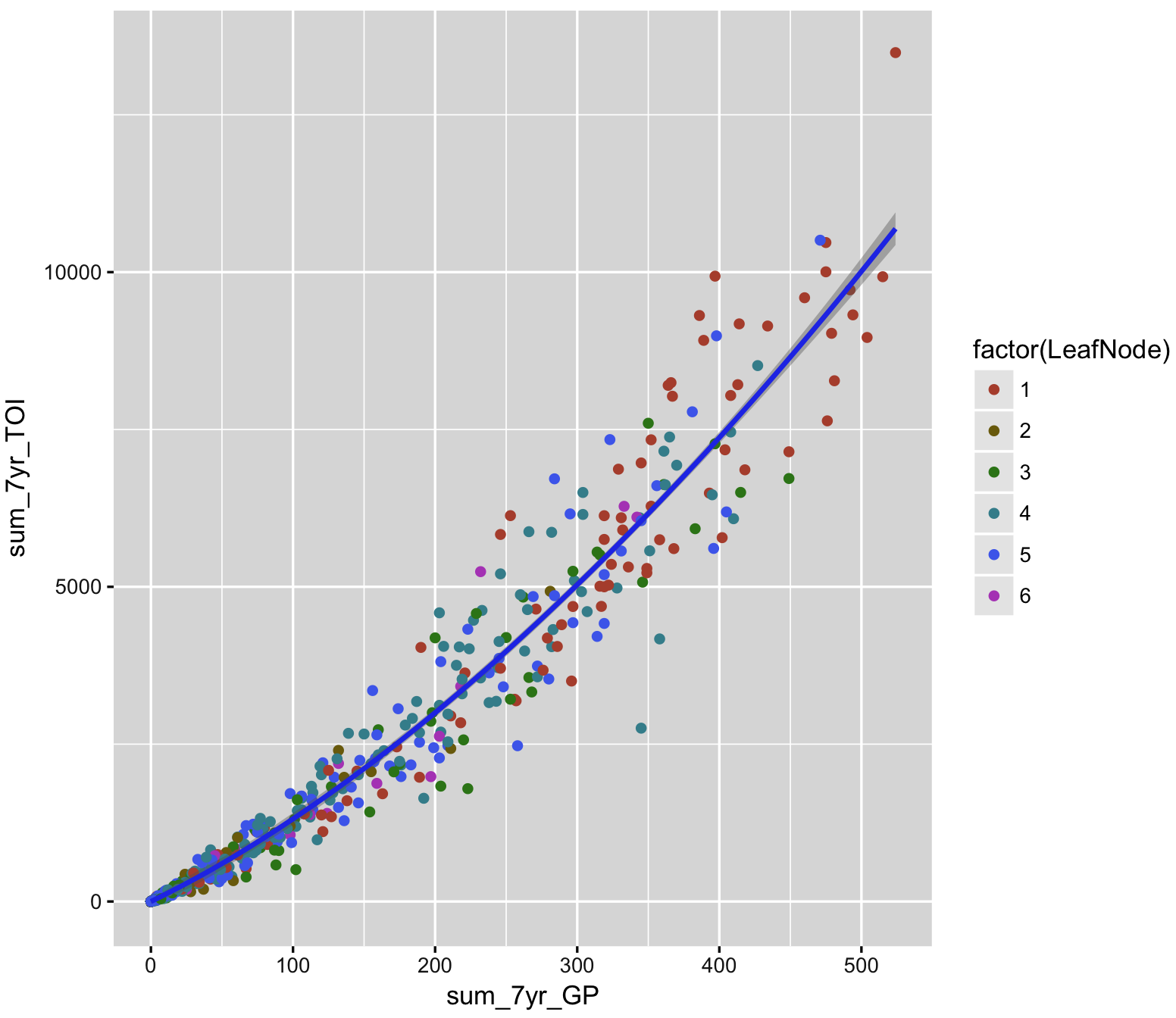
## 6.1. Figure for Independent Variables

Show TOI and GP. Show in 2d plot or in boxplots, one for each cluster and one overall?

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Met-  rics  Group | CSS\_  rank | Draft  Age | Height | Weight | Country  \_group | Position | Games  \_played | Goals | Assists | Points | Penalty\_  in\_  Minutes | Plus  Minus |
| 1 | **-19.0** | 0.65 | -0.91 | **3.06** | E: -0.25  C: 0.73  U: 0 | **D: -0.36**  L: 0  R: 0 | rs: 0  po: 0 | rs: 0  po: 0 | rs: 1.42  **po: 2.58** | rs: 2.5  po: 0 | rs: 0  po: 0 | rs: 0 |
| 2 | -1.70 | 0.65 | -1.61 | 1.19 | E: -0.38  C: 0  U:-0.24 | D: -0.24  L: 0  **R: 0.65** | rs: 0.62  po: -0.7 | rs:-0.4  po: 0 | **rs: 1.42**  po: 0 | **rs: 9.2**  po: 0 | rs: -0.47  po: 0.76 | rs: -0.52 |
| 3 | -1.91 | **1.26** | **-2.21** | 2.16 | E: -1.23  C: 0  U:-0.81 | D: 0.11  L: 0.43  R: 0 | **rs: 1.57**  po: 0 | rs:-0.4  po: 0 | rs: 0.42  po: 0 | rs: 0  po:-2.7 | rs: -0.80  po: 0.76 | **rs: 2.98** |
| 4 | -1.49 | **1.26** | **-2.21** | 0.23 | E: 0.46  C: 0  U: 0 | D: 0.11  L: -0.06  R: 0 | rs: 1.05  po: 0 | **rs:-0.42**  po: 0 | rs: 0.42  po: 0 | rs: -0.6  po:-0.3 | rs: -0.80  po: 0.76 | **rs: 2.97** |
| 5 | -0.35 | -1.11 | 0.19 | 1.19 | E: -0.95  C: 0  U: 0 | D: 0.31  L: 0.12  R: 0.19 | rs: 1.05  po: 0.32 | rs: 0.17  po: 0 | rs: 0.42  po: 0 | rs: 0.54  po: 0 | rs: -0.80  po: 1.46 | rs: -0.80 |
| 6 | -0.35 | 0.09 | -0.67 | 1.19 | E: **-8.86**  C: 0  U: 0 | D: 0.11  L: 0.43  R: 0 | rs: 1.05  **po: 1.09** | **rs: -0.42**  po: 0 | rs: 0.42  po: 0 | rs: 0  po: 0 | **rs: -0.80**  **po: 1.47** | rs: -0.80 |

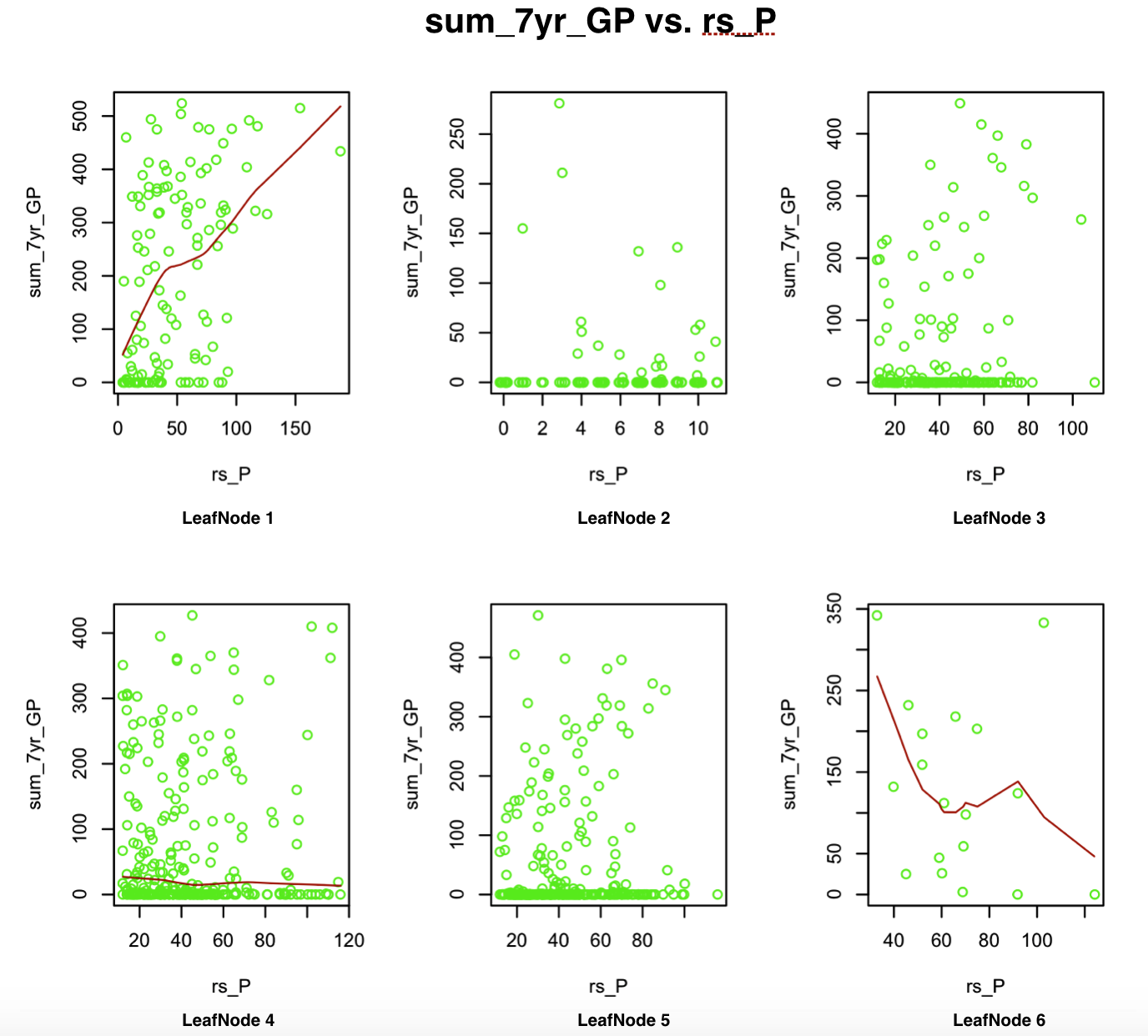
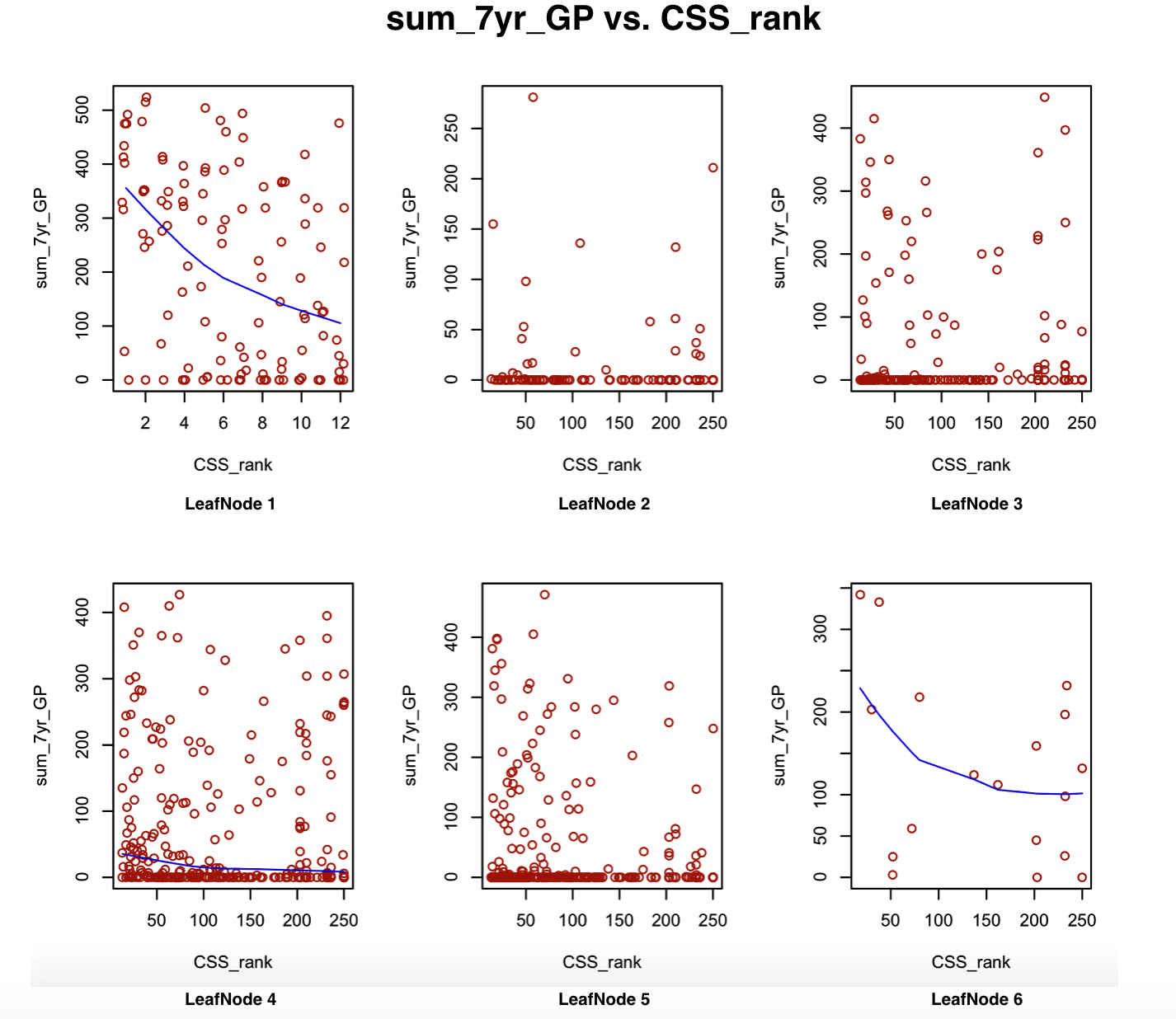
Table Group 200(4 + 5 + 6 + 7 + 8)Weights Illustration. E = Europe, C = Canada, U = USA, rs = Regular Season, po = Playoff

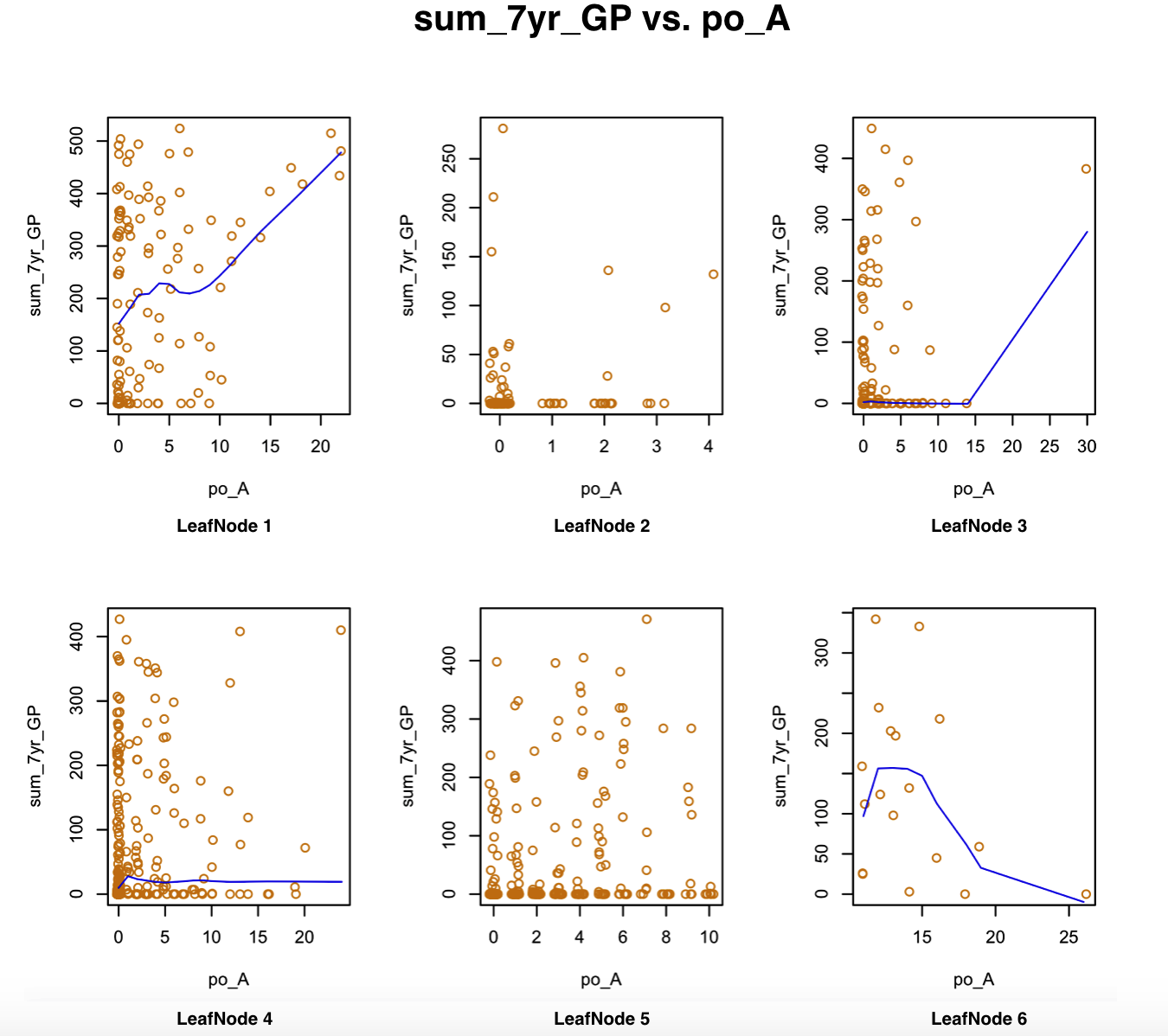
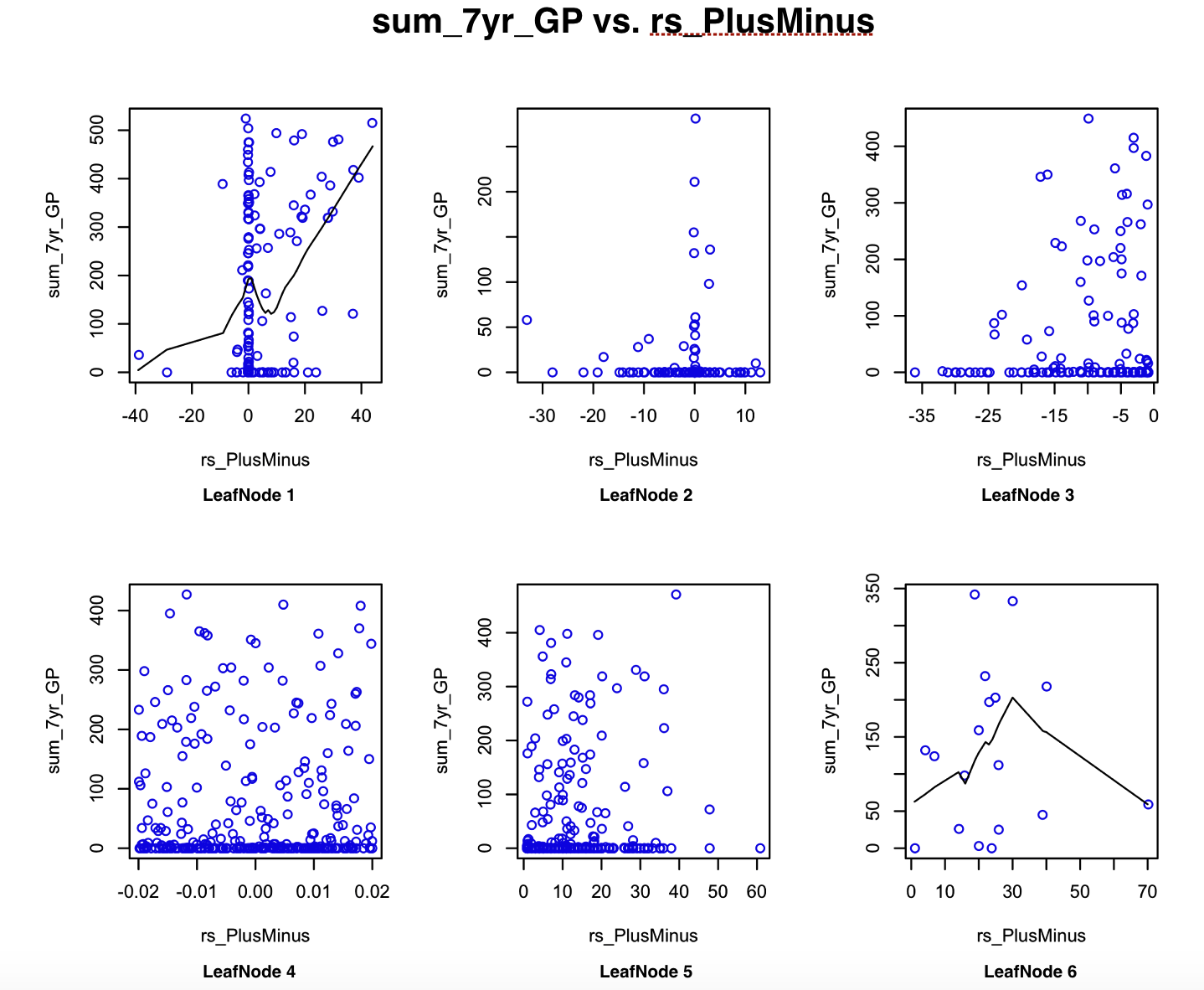
Table 5 illustrates the weights of logistic regression model in group and bolds the most relevant ones. For CSS rank, it stills plays an import role in recognizing the most excellent players in group 1, even though CSS rank of them is already smaller than 12. In contrast, its influence is invisible in leaf node 5&6 due to most players share similar ranks (larger than 50) in these two groups. Weight of Country\_group is noticeable in cohort 6, where over 80% players come from Canada and they all played at least one game in NHL. Points and Assist also contribute a lot to predicting player performance in NHL, especially in group 1&2.



## 6.2. Figure for Independent Variables (Player Statistics)

This is tricky. Show 2D-figures using PCA? Show most important predictors vs. GP? What is most important – following the tree or following the individual equations? Maybe we do the tree next.

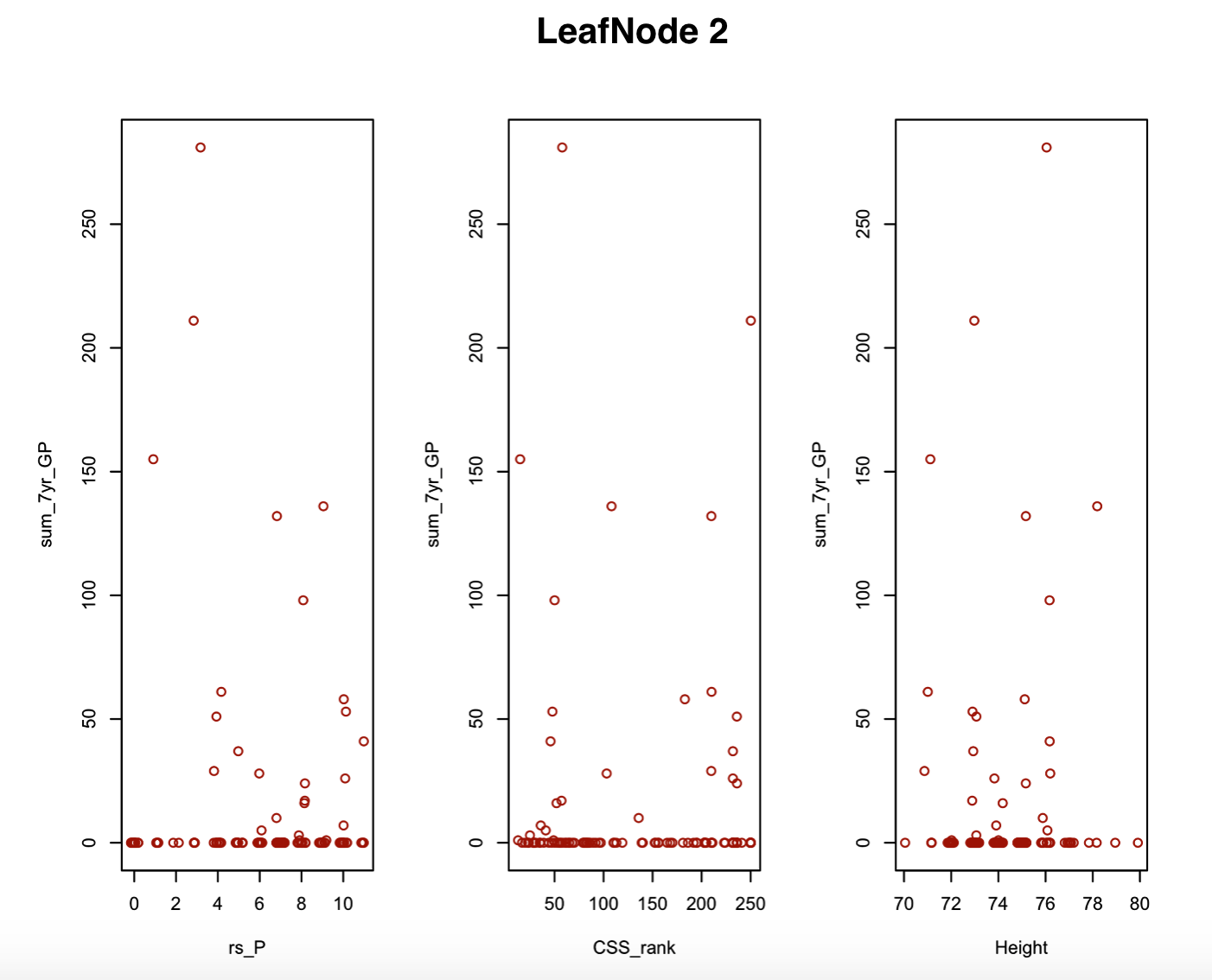
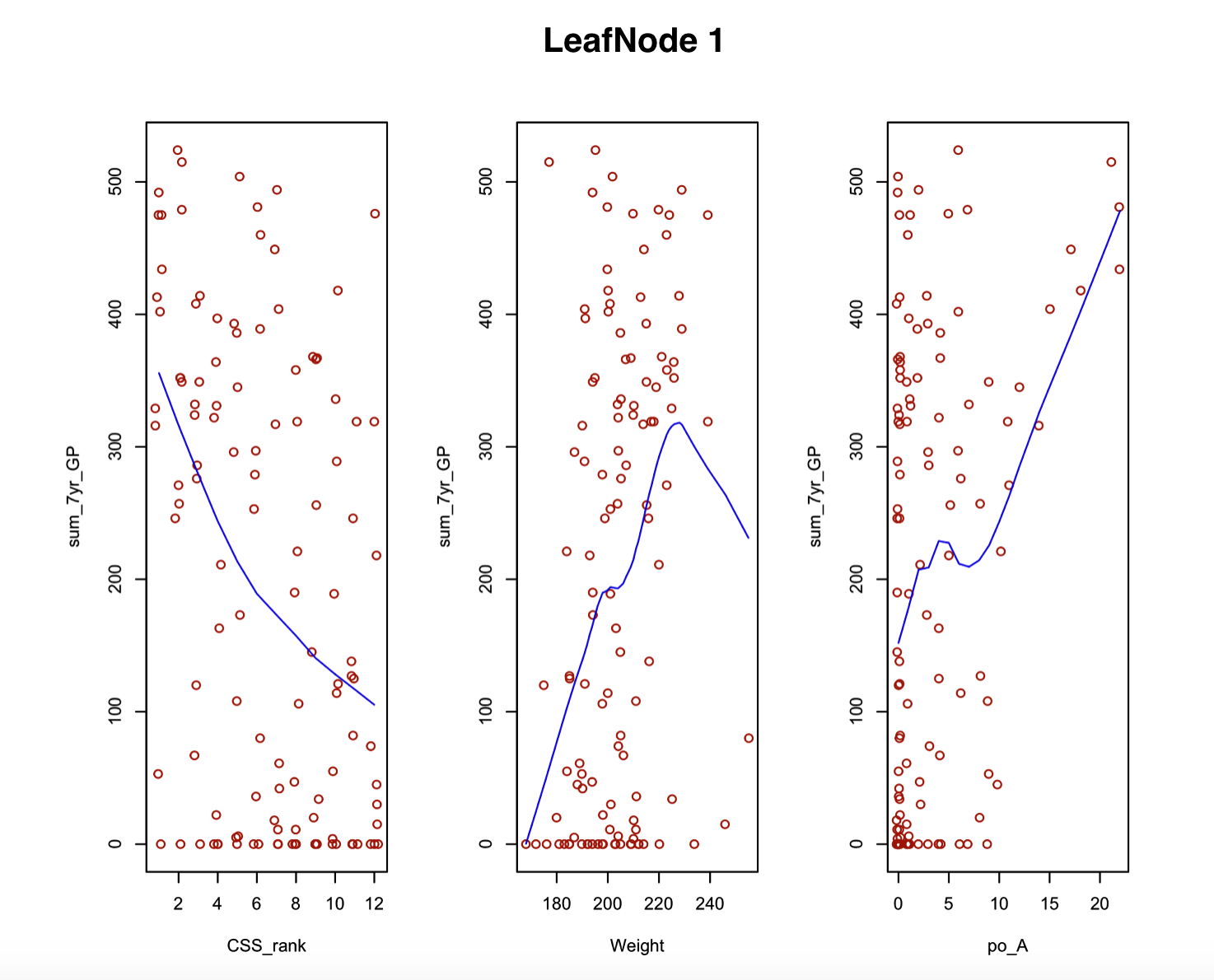


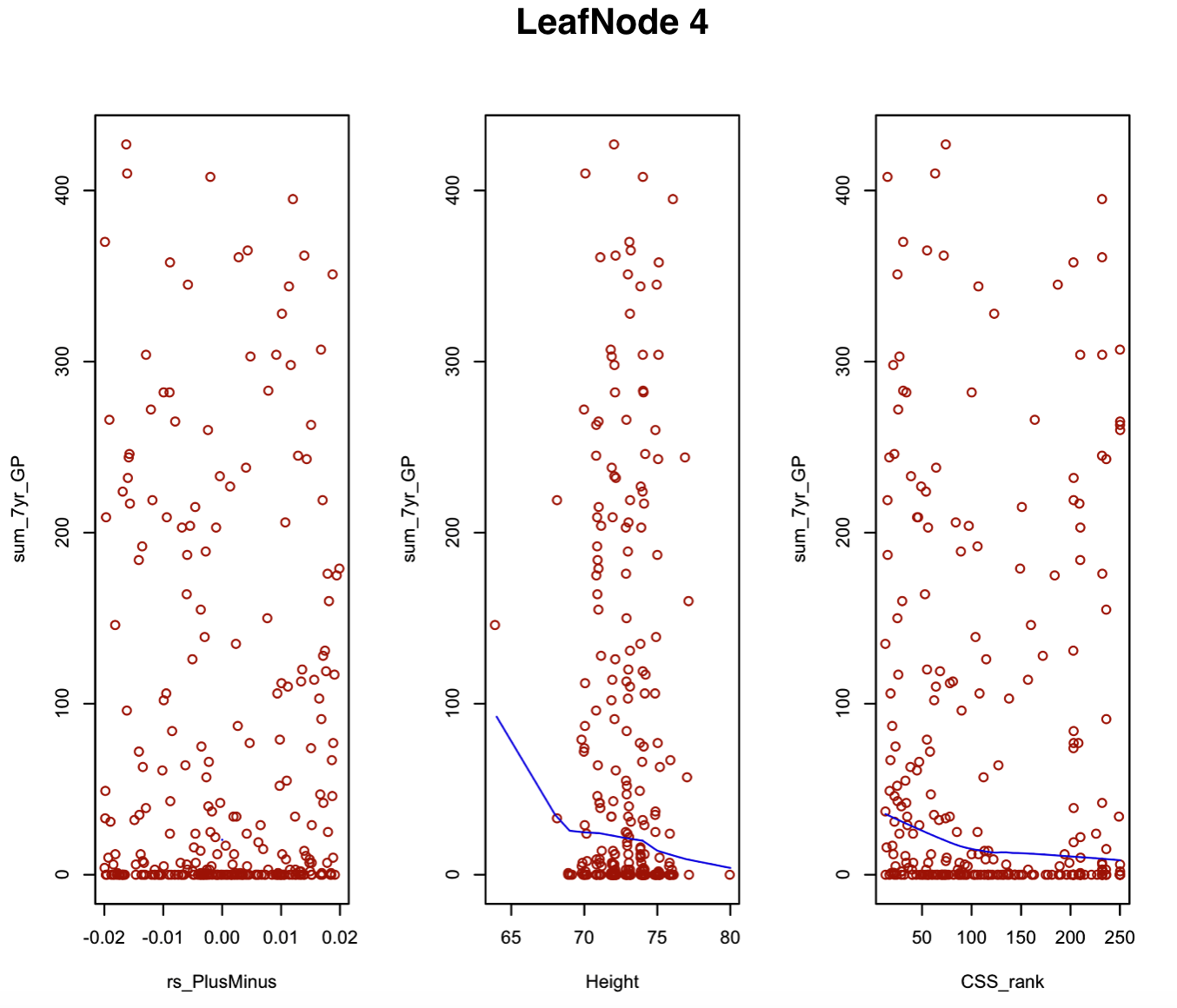
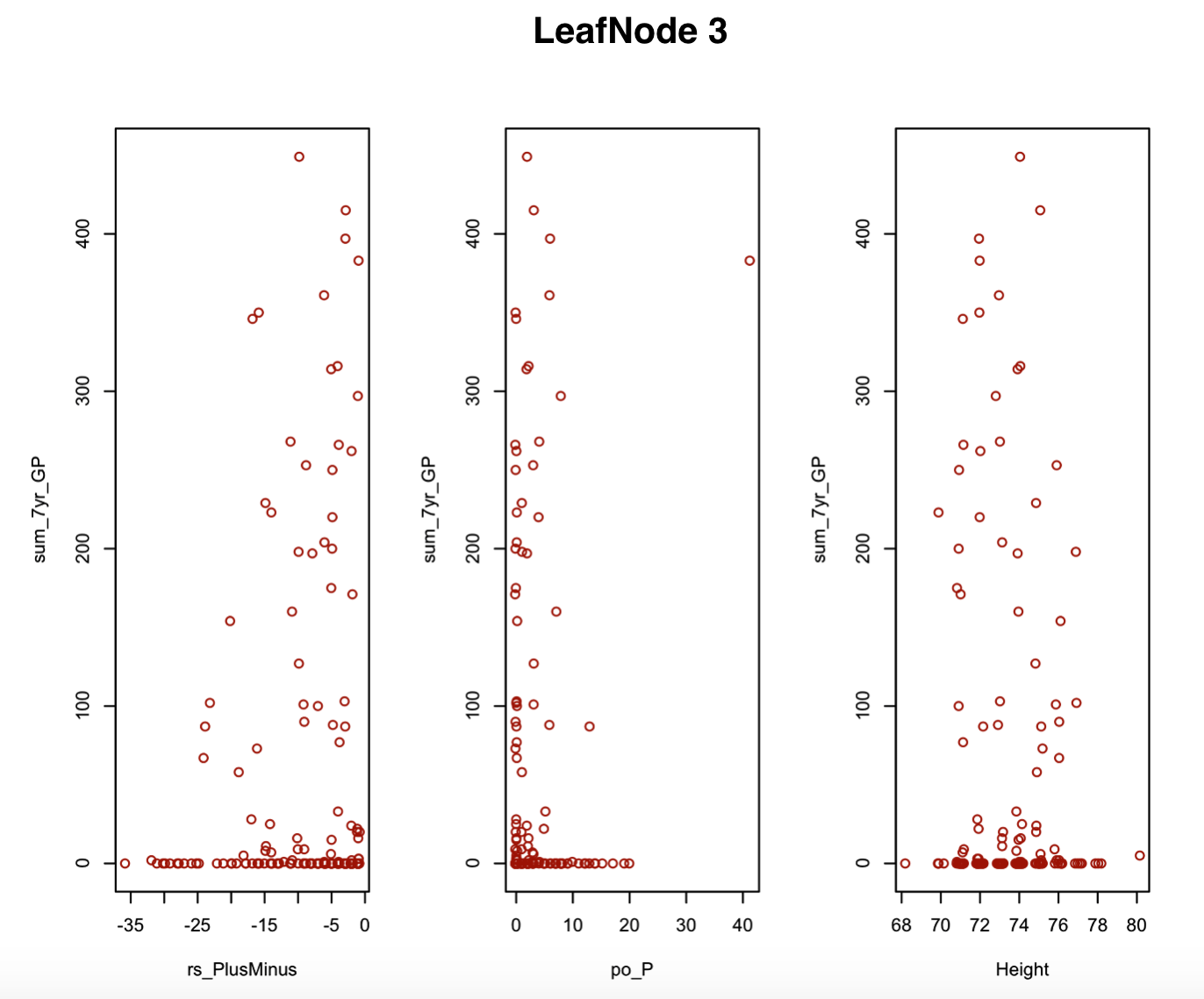


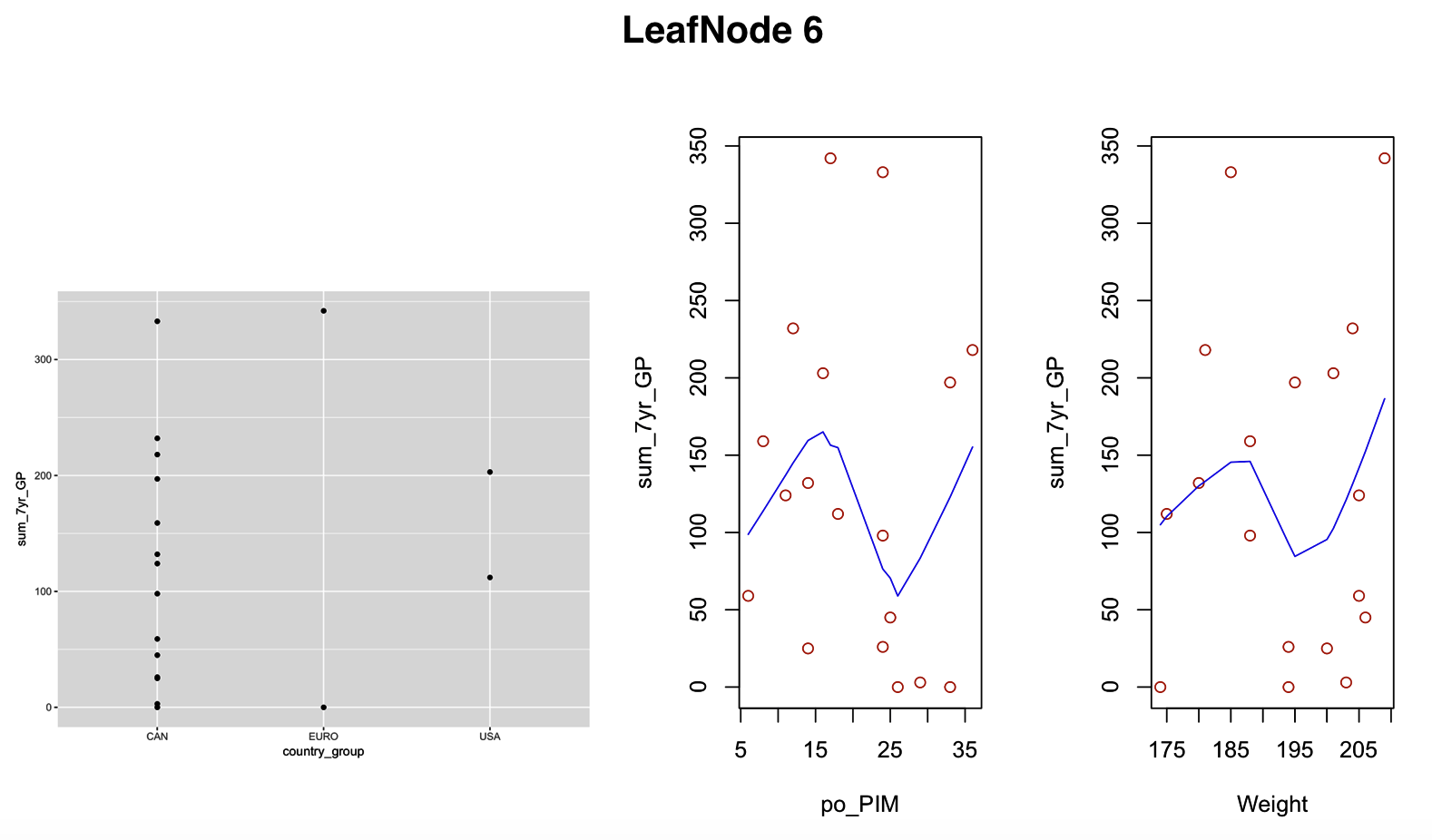
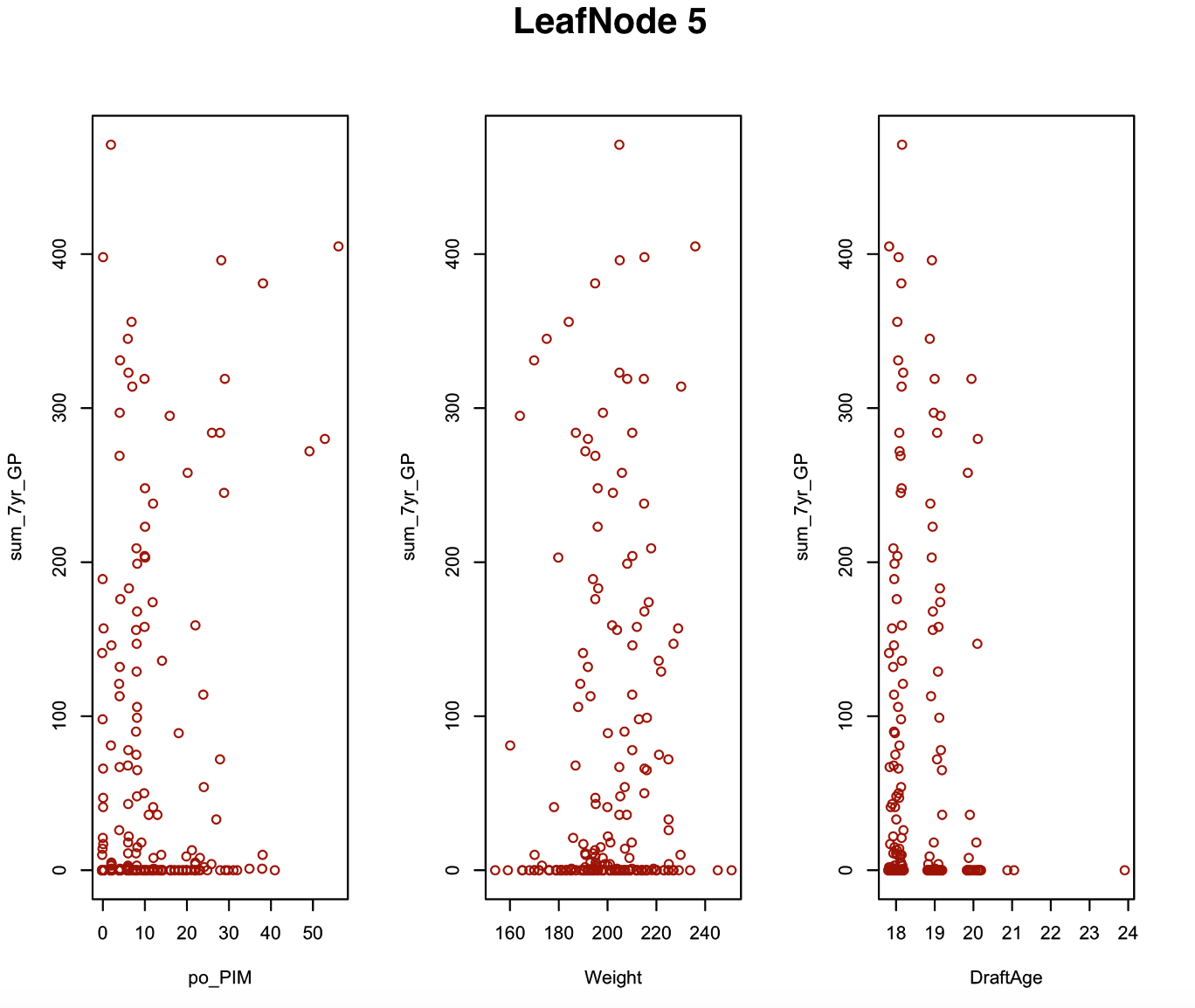
*Figure 4. Most important predictor per cluster vs. GP\_after\_7\_years.*

## 6.3. Figure for Learned Equations

Insert table with equations showing coefficients on normalized data. Bold the most important coefficient in each group







# 7. Identifying Exceptional Players

According to Florida Panthers, the numbers are often just the start of the discussion.

## Case Studies

[Figure. Maybe pick top 3 players in cluster]

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Top Players | Strongest Points | | | | | | | |
| 1 | Sidney Crosby | rs\_P | | | mean | rs\_A | mean | CSS\_rank | mean |
| 188 | | | 47 | 110 | 27 | 1 | 7 |
| Patrick Kane | rs\_P | | | mean | rs\_A | mean | CSS\_rank | mean |
| 154 | | | 47 | 87 | 27 | 2 | 7 |
| Sam Gagner | rs\_P | | | mean | po\_A | mean | rs\_A | mean |
| 118 | | | 47 | 22 | 4 | 83 | 27 |
| 2 | Matt Pelech | Weight | | | mean | rs\_A | mean | CSS\_rank | mean |
| 230 | | | 206 | 4 | 4 | 41 | 94 |
| Adam Pineault | CSS\_rank | | | mean | rs\_P | mean | Height | mean |
| 25 | | | 94 | 8 | 6 | 73 | 74 |
| Roman Wick | rs\_P | | | mean | CSS\_rank | mean | rs\_PlusMinus | mean |
| 10 | | | 6 | 36 | 94 | 0 | -2 |
| 3 | A.J.Jenks | CSS\_rank | | mean | | Weight | mean | Country | mean |
| 20 | | 76 | | 205 | 201 | USA | / |
| Bill Sweatt | CSS\_rank | mean | | | Position | mean | rs\_PlusMinus | mean |
| 27 | 76 | | | L | / | -1 | -9 |
| Brandon McMillan | CSS\_rank | mean | | | Position | mean | Height | mean |
| 44 | 76 | | | L | / | 71 | 73 |
| 4 | Sami Lepisto | CSS\_rank | mean | | | rs\_GP | mean | rs\_PIM | mean |
| 25 | 86 | | | 61 | 47 | 30 | 60 |
| Linus Omark | CSS\_rank | mean | | | Height | mean | DraftAge | mean |
| 55 | 86 | | | 70 | 72 | 20 | 18 |
| Oscar Moller | CSS\_rank | mean | | | Height | mean | rs\_GP | mean |
| 20 | 86 | | | 70 | 72 | 68 | 47 |
| 5 | Milan Lucic | Weight | mean | | | po\_GP | mean | CSS\_rank | mean |
| 236 | 199 | | | 23 | 9 | 58 | 71 |
| Michael Del Zotto | Position | mean | | | Country | mean | po\_GP | mean |
| D | / | | | CAN | / | 15 | 9 |
| Steven Delisle | Weight | mean | | | Country | mean | po\_GP | mean |
| 234 | 199 | | | CAN | / | 19 | 9 |
| 6 | Brad Marchand | Country | mean | | | po\_GP | mean | po\_P | mean |
| CAN | / | | | 25 | 19 | 23 | 19 |
| Mathieu Carle | Country | mean | | | CSS\_rank | mean | rs\_GP | mean |
| CAN | / | | | 53 | 107 | 67 | 65 |
| Kyle Cumiskey | Country | mean | | | po\_GP | mean | rs\_GP | mean |
| CAN | / | | | 27 | 19 | 72 | 65 |

Table Strongest points for top 3 players in each group, in comparison with the group mean value

Discuss cases

## Explaining the Rankings: identify weak points and strong points

[maybe use absolute value to identify unusual players].

|  |  |  |
| --- | --- | --- |
| Top Player | Strongest Point | Weakest Point |
|  | (e.g. age vs. average age) |  |
|  |  |  |

*Table 5. Strong Statistics and Weak Statistics for the top player in each cluster*

## Conclusion and Future Work

[predictive accuracy can be boosted]

**References**

[1] Michael E. Schuckers, Statistical Sports Consulting, LLC (2016). Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection. MIT Sloan Sports Analytics Conference (2016).

[2] Weissbock, J. (2015). Draft Analytics: Unveiling The Prospect Cohort Success Model. <https://canucksarmy.com/2015/05/26/draft-analytics-unveiling-the-prospect-cohort-success-model/>

[3] David R. Wilson (2016). Mining NHL Draft Data and A New Value Pick Chart. A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of Master of Science.

[4] <https://github.com/liuyejia/Model_Trees_Full_Dataset>

[5] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

**Appendix: Rank Correlations With Ties**

Address subtleties around ties

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data NHL Draft Years | Out of Sample  Draft Years | Draft Order Kendall  Correlation | Kendall rank correlation |
| 1998, 1999, 2000 | 2001 | 0.40 | 0.73 |
| 1998, 1999, 2000 | 2002 | 0.28 | 0.68 |
| 2004, 2005, 2006 | 2007 | 0.40 | 0.64 |
| 2004, 2005, 2006 | 2008 | 0.48 | 0.60 |

Table Our Results using Kendall-Tau's correlation

The biggest weight magnitude is assigned to Country\_group is noticeable in group 6, where no European played an NHL game. Points and Assist also contribute a lot to predicting player performance in NHL, especially in group 1&2.

[In contrast, its influence is invisible in leaf node 5&6, because most players in these two groups share similar low ranks (larger than 50).]

[discuss goals, which switches]

*[Penalty minutes* during the regular season lower the prospects of playing an NHL game, whereas during the playoff season they increase it. This is because fewer plays play relatively few playoff games, so playoff penalty minutes are a proxy for playoff games played. [check in data]]

Group 2 comprises players at CSS rank and regular season points below 12. [Group 2 consists of two major subgroups with size 83 vs. 16: the first contains players whose regular season goals are below 3, while players in the second group have regular season goals above 3. This mix leads to the surprisingly negative coefficient

-2.17]. where about. Add proportion plots for these groups.

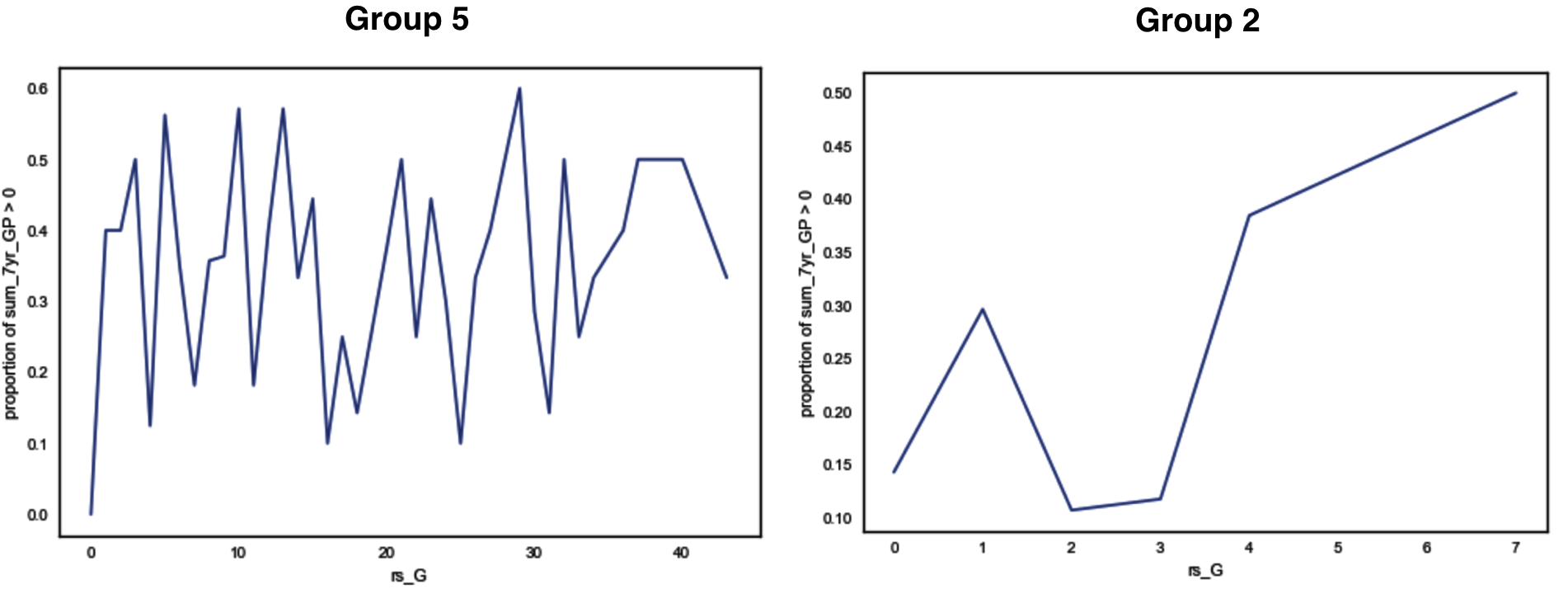


Figure Proportion\_of\_Sum\_7yr\_GP\_greater\_than\_0 vs. rs\_G in group 5&2.

In Group 5, players are also ranked above level 12, collected at least 12 regular season points, but show a positive plus-minus score above 0. This group contains two major subgroups with size *148* vs. *31*: the first with plus-minus between 1 and 20, where the proportion of above-zero players *decreases* with plus-minus score, and the second with plus-minus above 20, where the proportion of above-zero players *increases* with plus-minus score. This mix leads to an overall small negative weight of -0.73. The proportion plot appears in plotted in appendix **Error! Reference source not found.**.

Stralman was drafted in place 216, managed 265 games in the NHL his first seven years, won world championship gold with Sweden in 2017, and reached the Stanley Cup final with the Rangers in 2014. [add strong points? Don’t mention stralman?]

**Kendall-Tau Rank Correlation**

Kendall Rank correlation is a statistic used to measure the similarity of data orderings ranked by their quantities [11]. The Kendall-Tau coefficient is defined as following:

1. , where n = number of rank pairs.

Our drafting approach compared with draft order using Kendall-Tau rank correlation is illustrated in Table 7.

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data NHL Draft Years | Out of Sample  Draft Years | Draft Order Kendall  Correlation | Tree Model Kendall correlation |
| 1998, 1999, 2000 | 2001 | 0.40 | 0.73 |
| 1998, 1999, 2000 | 2002 | 0.38 | 0.68 |
| 2004, 2005, 2006 | 2007 | 0.44 | 0.64 |
| 2004, 2005, 2006 | 2008 | 0.48 | 0.60 |

Table 7 Our Results using Kendall-Tau's correlation

**Proportion & Scatter Plot for Group 5**

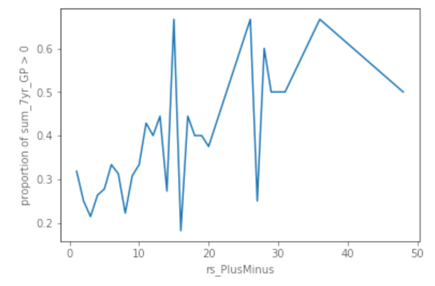
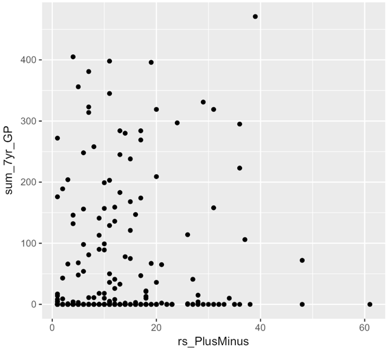
**** ****

Figure 3 Proportion Plot of sum\_7yr\_GP\_greater\_than\_0 vs. rs\_P and scatter plot of sum\_7yr\_GP vs.rs\_P in Group 5.