**Team Strava Review**

Section 2: You provide very detailed descriptions of the variables, which is very helpful. However, it is important to visualize the distribution of some of these variables before you jump into modeling. A scatterplot matrix of the quantitative variables would be a great way to help readers get a feel for how each variable is distributed and related to the proposed response variable.

102-108: Box cox should only be used to determine transformations for the response variable, not the explanatory variables. The use of log-transformations are fine, but the justification for doing so should be corrected in the text (example: you did it because your variables were right skewed, not because of anything Box Cox said).

176-183: Variable selection can help to improve multicollinearity (which is not technically a violation of the model assumptions), but all other violations of assumptions must be fixed BEFORE attempting variable selection. This section should include an expanded discussion of remedials measures that have (or will) be tried prior to conducting variable selection.

235-250: Interaction terms built from unscaled variables, by design, introduce multicollinearity into our models. This is not something to be worried about and this concern should be omitted from the text.

280-295: Trying to back-transform the coefficients to make 1 unit interpretations in X is not correct. The model is linear on the log-scale, not the original scale. As such, interpretations should be made on the log scale and the coefficients should not be exponentiated.

480-482: Regression should still treat categorical variables as categorical, not ordinal. This is accomplished using a “class” statement prior to specifying the model form. If a class statement was not used for categorical variables the first time, please try re-running the regression tree in a way that treats the categorical variables appropriately.

General Comment: When discussing accuracy, you provide one MSPR value and three MSE values (one for each model). Perhaps the confusion is just in word choice, but you should be reported three values of MSPR for each of the response variables, one for each of the linear models. In addition, in order to compare the error rates of your linear model to the regression tree, you will need to back-transform your linear model predictions and calculate the validation error on the original scale. Please try doing this and providing a direct comparison of the accuracy of your linear model to the accuracy of your regression tree.

General comment: By including the parks as an explanatory variable, you have made it so that your model is only relevant for these specific parks. Your model cannot be generalized to new parks. If you don’t need your model to generalize to new parks, then there is no worry. Just something to possibly keep in mind.