Final Project Proposal

The Effect of the Fitness App *Strava* on Mountain Bike Behavior in Parks and Protected Areas

Noah Creany

17 April, 2020

1 Introduction

Among fitness tracking apps Strava seems to stand atop the podium with 42 million registered users globally adding an additional million users each month (Lindsey 2019), and collecting a spatial data set of more than 13 trillion data points through its users who track more than 2 million activities per week ("Strava Labs" 2018). According to Lindsey (2019), although Strava was neither the first fitness app nor has the largest number of users, features such as the Segment Leader-board and integrated social-network platform distinguish the app from others in the fitness app category which has resulted in a dedicated following of users.

Strava is a part of a trend in digital media of self-monitoring or the "Quantified-self" ("Quantified Self" 2019), which provide affordances to users of connected devices such as smartphones, GPS watches and cycling computers to collect, analyze, and share data created through use of the technology (Lupton 2016a). Consequently, behavior and decision making are made within the frameworks and feedback mechanisms of the technology, which become the focus of and mediator of the experience of users (Lupton 2016b). A qualitative study of the social interactions of Strava users found that the majority of users always record their activity with Strava and that contributing likes and replies to other users' activities on the platform provided positive feedback mechanisms that result in users posting more frequently about their recorded activities (Stragier et al. 2018). In the context of fitness for physical health, it suggests that these platforms can encourage better health outcomes for its users; however, little research has explored how the use of self-tracking apps affects visitor behavior in the parks and protected areas (PPA) settings where many of these activities take place. The goal of this study is to understand the effect of *Strava* use on visitor mountain bike behavior, specifically the speed or velocity the visitor traveled while in the park.

1.1 Gamification

Gamification is a technique in digital media design and development that refers to the application of mini-games or challenges, called game elements, to provide motivation and persuasion to complete a task, goal, or desired behavior. Deterding et al. (2011) is widely cited for establishing a definition of gamification as "the use of game design elements in non-game contexts" (p. 2); but also trace its foundations to learning theory and the importance of play for culture, socialization, and learning. Seaborn and Fels (2015) provide a concise summary, "[g]amification has two key ingredients: it is used for non-entertainment purposes, and it draws inspiration from games, particularly the elements that makeup games, without engendering a fully-fledged game"(p.27).

Gamification techniques have been applied in a variety of different contexts, such as education, health, marketing, and social networks, which have led to a plurality of conceptual definitions and theoretical foundations (Seaborn and Fels 2015). In an attempt to direct future gamification research (Putz and Treiblmaier 2015) outline suitable social-psychology theory that has been used to explain and inform the influence of gamification techniques on attitudes, motivation, and behavior including but not limited to the Theory of Reasoned Action (Fishbein and Ajzen 1975), Theory of Planned Behavior (Ajzen 1991}), Social Learning Theory (Bandura and Walters 1963)/ Social Cognitive Theory (Bandura 1986) and Self-Determination Theory (Ryan and Deci 2000). Furthermore, Putz and Treiblmaier (2015) embrace this multi-theoretical approach because while a growing body of literature supports the effectiveness of gamification in achieving positive outcomes (Hamari, Koivisto, and Sarsa 2014), the understanding of the effect of gamification on user's behavior and attitudes is still emerging.

A significant distinction between Strava and other fitness tracking apps is the incorporation of game elements, or gamification, into the real-world context of a recreation experience (Chen 2017). Barratt (2017) suggests that while Strava has no overt objective to change the behavior of users, the gamification mechanisms embedded in the app make use of persuasion techniques and social feedback to "[tap] into the basic desires and needs of the users which revolve around the idea of status and achievement" (p. 330). Leader-boards are a well-established game element and central feature of Strava that allows users to compete for the fastest time on segments of trail crowning the fastest male or female rider King of the Mountain (KOM) or Queen of the Mountain (QOM), respectively ("What's a Segment?" 2012). Sailer et al. (2013), using Self-Determination Theory to understand the effect of gamification on motivation, found leader-boards and badges increased psychological needs satisfaction of competence, autonomy, and task-meaningfulness.

Additionally, Strava features such as trophies, challenges, performance visualizations, and Kudos, which are the Strava equivalent of a "Like" on other social platforms, are game elements that trigger motivational mechanisms (Sailer et al. 2013). Weber et al. (2018) found the workplace cycling social competition Love to Ride that draws upon the aspects of norms, values, and beliefs from the Theory of Planned Behavior and featured points and leader-boards increased levels of cycling participation among new, occasional and regular urban bike commuters in the U.S., U.K., and Australia. Finally, Seaborn and Fels (2015) conducted a meta-analysis of studies that used gamification techniques in a range of contexts and found mostly positive results within social networks and health and wellness contexts.

Barratt (2017), in a qualitative study of gamification of cycling within Strava, suggests a new dimension to the mountain bike experience is added through social interaction and competition facilitated by the features of the app. Furthermore, while gamification has been demonstrated be a useful tool to produce desired outcomes determined by the designer of the app or software, the effect on behavior in PPA settings where these activities often take place and if that behavior is consistent with the goals of management is not well understood.

2 Data

2.1 Data Collection & Preparation

Visitor surveys and GPS tracks were collected in May of 2018. Sampling in each PPA was stratified during weekdays and weekends, multiple entrances to parks, and began when the park opened at either 6:00 am or 7:00 am until approximately 5:00 pm or 6:00 pm. Visitors were intercepted at randomly selected minutes on the hour throughout the sampling period as they entered the park and were invited to participate in the study by completing a post-experience survey. Only visitors whose primary activity was mountain biking were asked to carry a Garmin eTrex 10 GPS unit.

The survey captured information from the visitor including age, how frequently they participated in their primary activity (1=<10 days/year, 2=11-25 days/year, 3= 26-50 days/year, 4= 51+ days/year), and a self-evaluation of skill or experience level (1= beginner to 5=expert).Visitors were asked how they used their smartphones during their visit, and if they selected "Used Strava", a follow-up question asked them how frequently they use the app (Never, Rarely, Sometimes, Often, Always). Valid responses from visitors who did not use Strava were coded the *Strava* use frequency variable as "Don't use *Strava*".

GPS units were programmed to record the visitor's location every 10 seconds to balance the resolution of their behavior and the size of the data set. All GPS tracks were projected in California State Plane Coordinate System Zone 6 (NAD83(2011) / California Zone 6) and processed to remove points where the GPS unit was given to the visitor and returned to researchers to include only points of the visitor's movement. A unique alpha-numeric code stored in the GPS attribute table and mountain bike survey response provided the ability to form Strava and Non-Strava groups for comparative analyses. Velocity was calculated for each point within a track from projected (X,Y) coordinates and time stamps within the attribute tables stored in the GPS track.

Maximum, median, and mean velocities for each of the 244 GPS tracks were calculated and associated with the unique survey alpha-numeric ID. Since the park where the GPS track was recorded is a nominal variable, binary dummy variable were created for the 6 parks. Additionally, meteorological data were downloaded from PRISM (Oregon State University 2004) for the dates and locations where sampling occurred in May of 2018 to add explanatory variables to this analysis. An Ordinary Least Squares (OLS) and non-parametric regression methods will be used to predict these aggregate measures of velocity from whether a mountain biker used Strava, their experience level, age, park where the GPS track was recorded, and meteorological conditions. Table 1 below provides a summary of the variables used in this analysis.

Table 1: Variables in Strava Data set with short description and measurement type.

Variable	Description	Measurement Type/Units
Response		
V.Max	The maximum velocity of the GPS track	Continuous (m/s)
V.Mean	The mean velocity of the GPS track	Continuous (m/s)
V.Median	The median velocity of the GPS track	Continuous (m/s)
Explanatory		
Activ.Days.Year	The number of days/year a visitor mountain bikes	Ordinal (1-10,11-25. . .)
Experience Level	The self-evaluated skill or ability level of biker	Ordinal (beginner,novice..)
Strava Use Freq	The frequency that a mountain biker uses Strava	Ordinal (Never,Rarely. . .)
Strava Use	Dichotomous variable, Uses Strava or not	Nominal (Yes, No)
Park	The park unit where the GPS track was recorded	Nominal
Age	The age of the participant in the study	Continuous (Years)
Ppt	Precipitation	Nominal (Rain, No Rain)
Mdt	Mean Dewpoint Temperature	Continuous (°C)
Tmin	Minimum Daily Temperature	Continuous (°C)
Tmax	Maximum Daily Temperature	Continuous (°C)
MinVP	Minimum Vapor Pressure Deficit	Continuous (hPa)
MaxVP	Maximum Vapor Pressure Deficit	Continuous (hPa)

3 Model Assumptions

Initial exploratory analysis indicated that the distributions of the dependent variables of velocity did not follow a normal distribution so we conducted a Box-Cox transformation procedure to determine the type of transformation to the dependent variables that would be better suited for linear modelling. The results of this procedure recommended transformations that were either close to zero (-0.1 for maximum and mean velocity) or a non-standard exponential transformation (0.7 for median velocity). We chose to log-transform the dependent variables to maintain consistency across our dependent variables in the analysis and make interpretation of the model more straightforward.

In order to validate and assess the performance of the model on new data the data were subsetted into a train and test datasets, withholding 30% of the data for a test dataset. Next, we performed an initial linear regression on the three log-transformed dependent variables against all of the explanatory variables to determine if the model would satisfy OLS regression assumptions. In addition to numerical and graphical diagnostics to assess the satisfaction of the assumptions of independent, identical and normally distributed residuals and constant mean and variance, we performed diagnostics to determine if the model was influenced by multicollinearity, outliers, and influential points.

3.1 Maximum Velocity

An initial linear regression was performed on maximum velocity which resulted in a significant linear model $F_{14,143} = 3.26, p = .0002$ with an R-Square 0.2419, Adjusted R-Square 0.1677. The residuals appear to satisfy the assumption of normal distribution as assessed by the fit of the observations to the expected distribution line in the QQ-plot but has a deviation from the normal bell curve in the Histogram (Figure 1). Additionally, the residuals appear to have constant variance as assessed by the studentized residuals vs predicted values.

The model's residual values were plotted in sequence which indicated no apparent serial correlation in error terms (Figure 2). A Brown-Forsythe test of constant variance, $t = 1.606, p = .110$, confirmed findings from the graphical diagnostics indicating constant variance. However, the correlation test of normality of residuals indicated that residuals are not normally distributed with a correlation of 0.985 which is slightly less than the expected correlation of 0.987 for $\alpha = .05$ with this sample size.

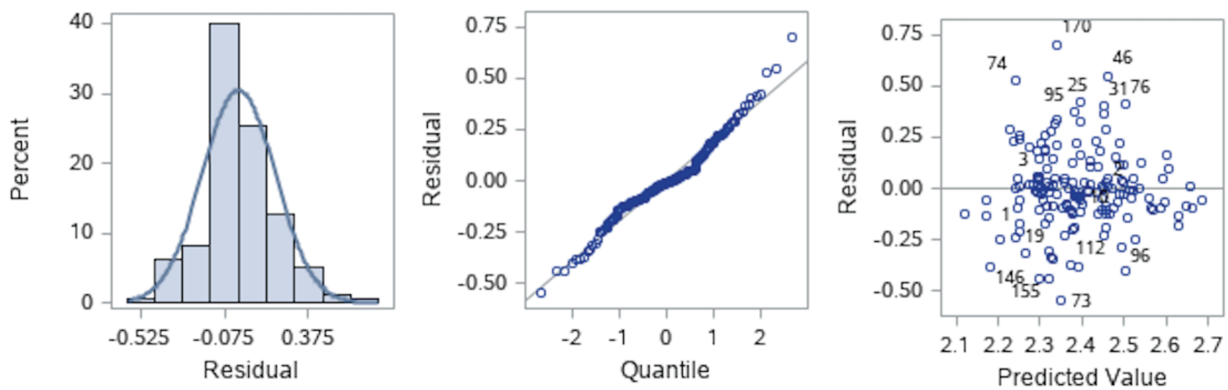

Figure 1: Graphical Residual Diagnostics for Initial Maximum Velocity Model

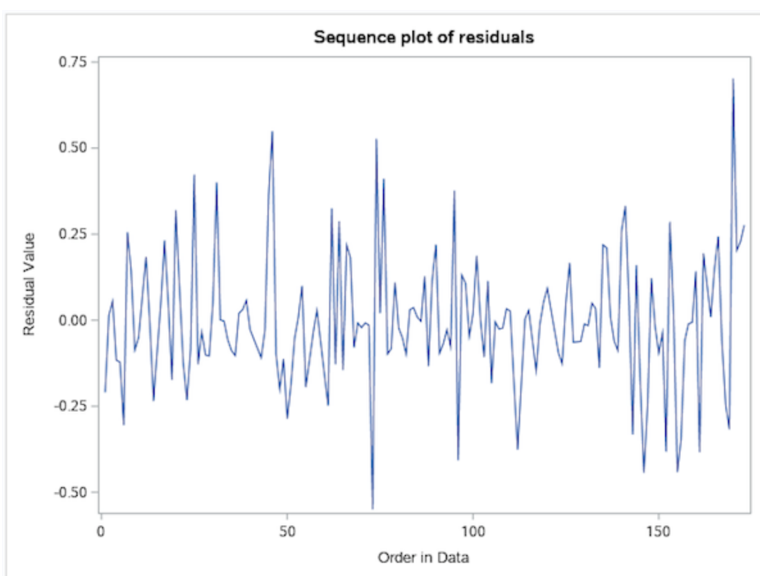

Figure 2: Sequence Plot of Residuals for Maximum Velocity

Next we evaluated if multicollinearity, or two or more highly correlated predictors, exists between the independent variables in the model. Of the 15 predictors in the model, 9 had VIF values greater than 10 and the average VIF for the model was 58.31. Next a Principal Components Condition Index was consulted to further investigate multicollinearity within the data. Condition Indices for 8 of the 16 components were greater than 10, and 2 components had more than 1 variable that accounted for more than 50% of the variance which were mostly the meteorological data. Finally, we assessed the model for influential and outlier observations that would affect the predictions from the model with DFFITS, DFBETA, Studentized Residuals, and Cook's D statistics. There appear to be a number of outlier observations that are poorly explained by the model's predictors, influential observations, or a combination of both, summarized in Figure 3 below.

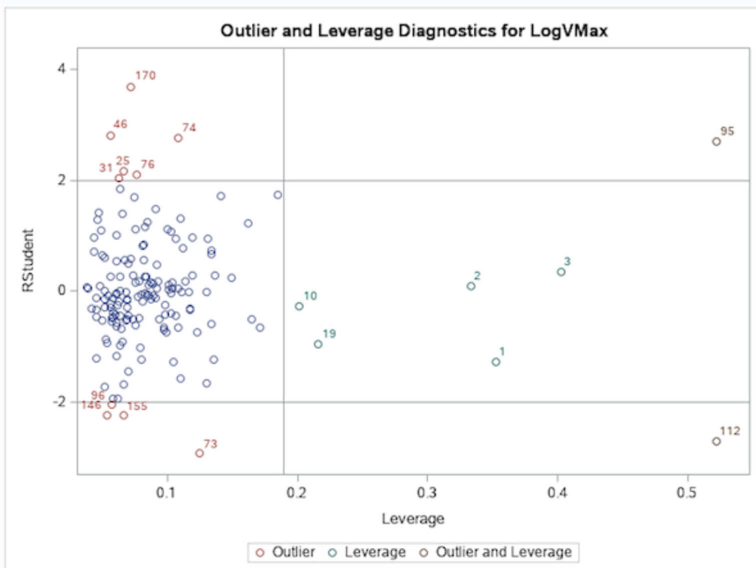

Figure 3: Outlier and Leverage Obseravations in Maximum Velocity Initial Model

3.2 Mean Velocity

An initial linear regression was performed on mean velocity which resulted in a significant linear model $F_{14,144} = 2.78, p = .0011$ with an R-Square 0.2129, Adjusted R-Square 0.1364. The residuals appear to satisfy the assumption of normal distribution as assessed by the fit of the observations to the expected distribution line in the QQ-plot and normal bell curve in the Histogram (Figure 4). The residuals appear to have met the assumption of constant variance as assessed by the studentized residuals vs predicted values with independent, identical, and zero mean distribution.

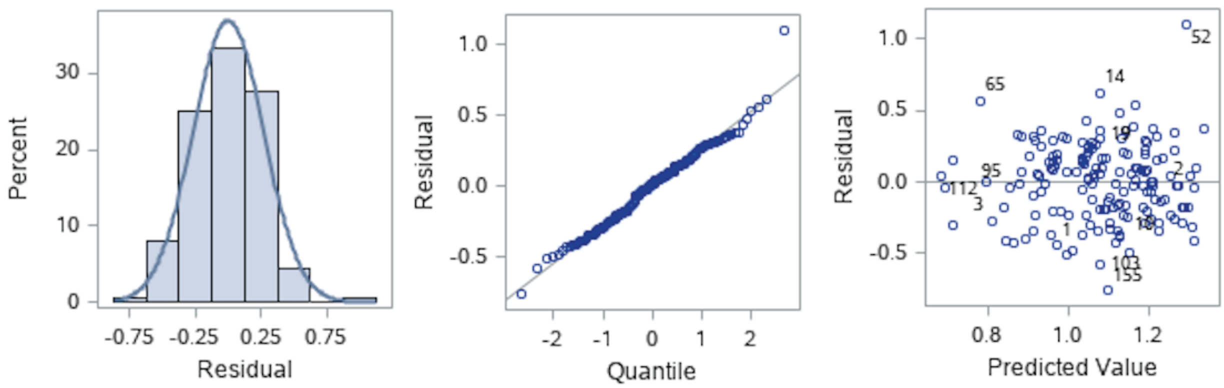

Figure 4: Graphical Residual Diagnostics for Initial Mean Velocity Model

Residuals were plotted in sequence which indicated no apparent serial correlation in error terms (Figure 5). A Brown-Forsythe test of constant variance, $t = 0.323, p = 0.747$ confirmed findings from the graphical diagnostics of no violation of the assumption of constant variance. Further, the correlation test of normality of residuals indicated that residuals are normally distributed with a correlation of 0.990 which is greater than the expected correlation of 0.987 for $\alpha = .05$ with this sample size.

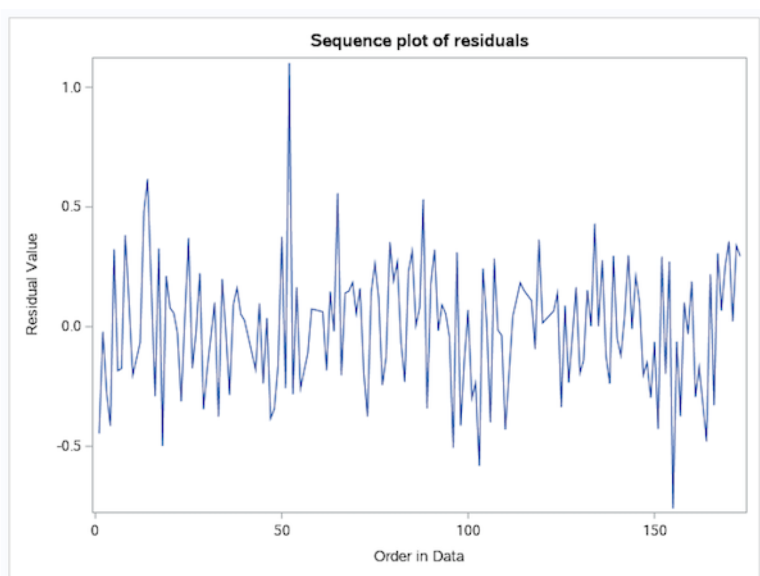

Figure 5: Sequence Plot of Residuals for Mean Velocity

Next we evaluated if multicollinearity exists between the independent variables in the model. Of the 15 predictors in the model, 9 had VIF values greater than 10 and the average VIF for the model was 56.77. A Principal Components Condition Index was consulted to further investigate multicollinearity within the data. Condition Indices for 9 of the 15 components were greater than 10, and 3 components had more than 1 variable that accounted for more than 50% of the variance. Finally, we consulted the DFFITS, DFBETA, Studentized Residuals, and Cook's D statistics to detect for outliers and influential observations in the data. Figure 6 below indicates the outliers and influential observations in the initial model.

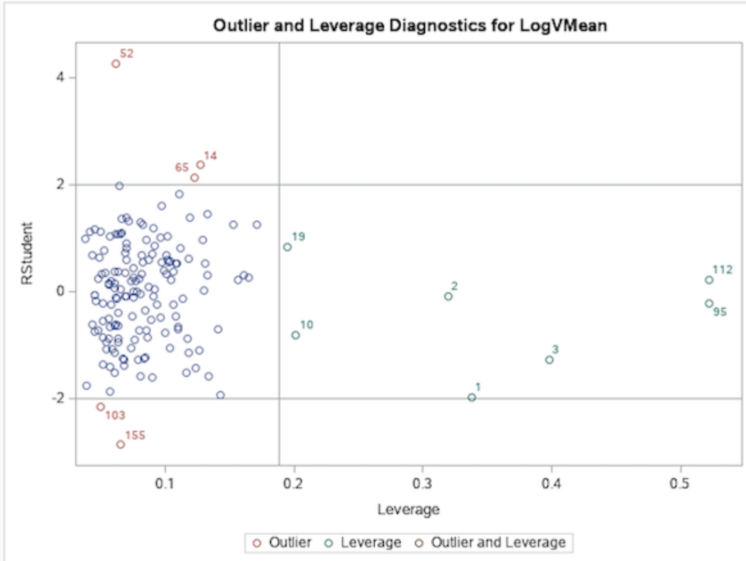

Figure 6: Outlier and Leverage Obseravations in Mean Velocity Initial Model

3.3 Median Velocity

An initial linear regression was performed on mean velocity which resulted in a significant linear model $F_{14,144} = 1.94, p = .027$ with an R-Square 0.1584, Adjusted R-Square 0.0766.The residuals appear to potentially violate the assumption of normal distribution as assessed by the fit of the observations departing in the tails from the expected distribution line in the QQ-plot and deviation from the normal bell curve in the Histogram (Figure 7). The residuals appear to have meet the assumption of constant variance as assessed by the studentized residuals vs predicted values with independent, identical, and zero mean distribution.

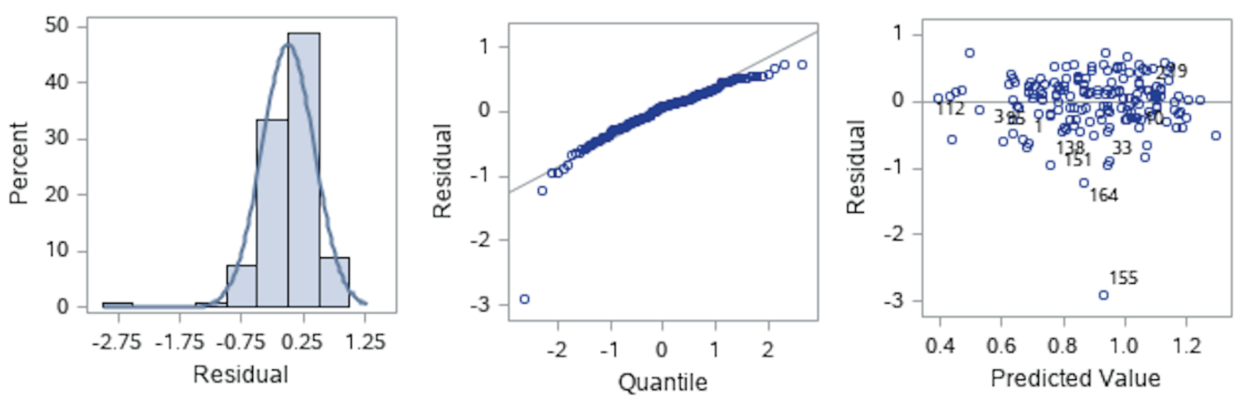

Figure 7: Graphical Residual Diagnostics for Median Velocity Initial Model

Residuals were plotted in sequence which indicated no apparent serial correlation in error terms (Figure 8). A Brown-Forsythe test of constant variance, $t = 0.8486, p = .397$, confirmed findings from the graphical diagnostics of no violation of the assumption of constant variance with the data. However, the correlation test of normality of residuals confirmed the residuals do not meet the assumption of normal distributed with a correlation of 0.923 which is less than the expected correlation of 0.987 for $\alpha = .05$ with this sample size.

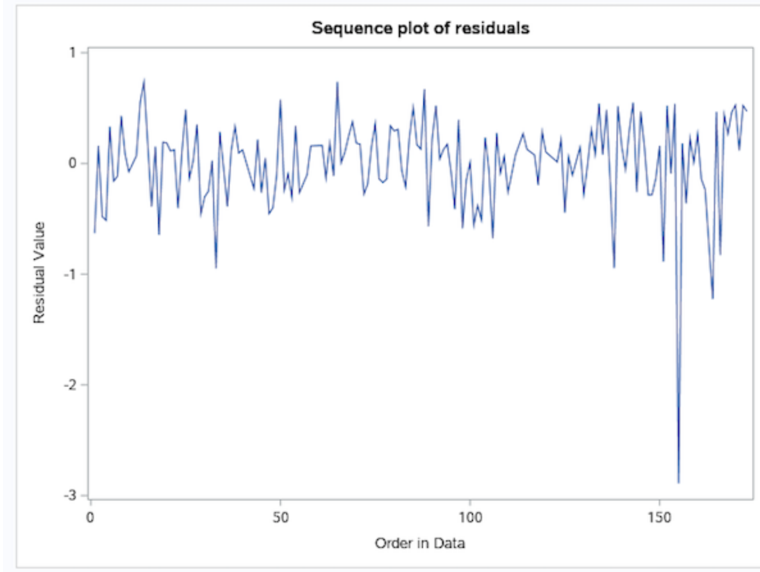

Figure 8: Sequence Plot of Residuals for Median Velocity

Next we evaluated if multicollinearity exists between the independent variables in the model. Of the 15 predictors in the model, 9 had VIF values greater than 10 and the average VIF for the model was 56.77. A Principal Components Condition Index was consulted to further investigate multicollinearity within the data. Condition Indices for 8 of the 16 components were greater than 10, and 3 components had more than 1 variable that accounted for more than 50% of the variance.Finally, we consulted the DFFITS, DFBETA, Studentized Residuals, and Cook's D statistics to detect for outliers and influential observations in the data. Figure 9 below indicates the outliers and influential observations in the initial model.

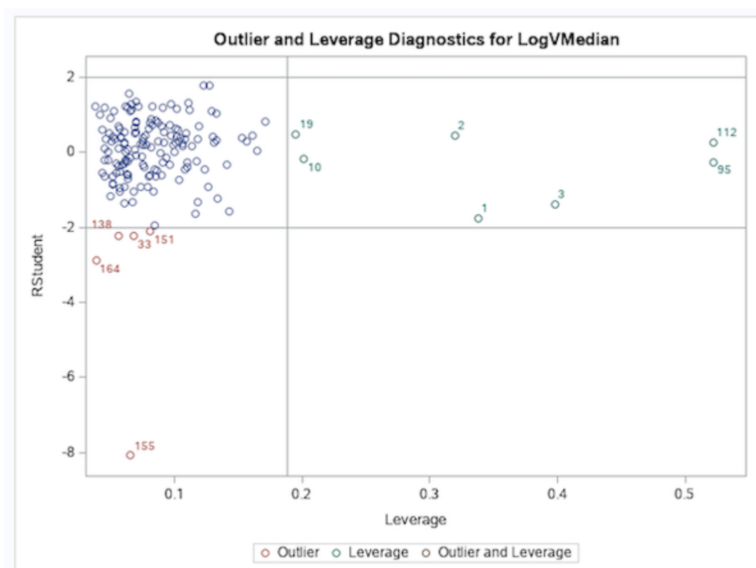

Figure 9: Outlier and Leverage Observations for Median Velocity Initial Model

3.4 Remedial Measures

From the initial OLS regressions we performed, it became clear that there were violations of the assumptions of OLS regression in all models.The maximum velocity model indicated a violation of the assumption of normal distribution of residuals from the correlation test of normality, evidence of multicollinearity, and the presence of influential and outlier observations. The mean velocity model satisfied the assumption of normal distribution and constant variance, but there are indications of multicollinearity between the explanatory variables and the presence of outliers and influential observations.The median velocity model indicated a violation of the assumption of normal distribution of residuals, evidence of multicollinearity between the explanatory variables, and the presence of outlier and influential observations.

In order to resolve some of these violations, a variable selection procedure will be used to select the best predictors in the model which should reduce the issue of multicollinearity in all three models. While all three models had influential or outlier observations, we reviewed the cases and did not see any clerical errors or unusual observations so we concluded that these are valid observations but are poorly explained by the model. We chose not to remove these observations to prevent over-fitting the model to the training data. Finally, the assumptions of constant variance and normal distribution will be re-assessed after variable selection techniques are performed on the data which will help determine if these data are suited for OLS regression.

4 Variable Selection

Because this model has 15 explanatory variables and evidence of multicollinearity, multiple variable selection techniques were performed to achieve the most parsimonious model that could explain the greatest amount of variance in the response variable. First, we performed a multiple selection criteria regression which ranked combinations of the explanatory variables by their Adjusted R-Square, Mallow's C(p), Akaike Information Criterion (AIC), and Schwartz's Bayesian Criterion (SBC or BIC) for maximum, mean, and median velocity. Additionally, we performed a regression with stepwise variable selection to determine the best model.

4.1 Maximum Velocity

The best performing model according to the multiple selection criteria had 4 variables (Experience Level, *Strava* use Frequency, Park 1 (ALWO), Park 3 (RIPA)) with an Adjusted R-Square of 0.2095. Next we consulted the stepwise selection regression model whose parameters were set to enter variables at $\alpha = .1$ and stay at $\alpha = .1$. The final model iterated three steps resulting in three predictors (Park 3 (RIPA), Experience Level, and *Strava* use frequency) with and Adjusted R-Square of .2063.The model summary for the stepwise selection is shown in Table 2 below.

Table 2: Summary of Stepwise Model for Maximum Velocity.

Step	Label	Partial R-Square	Model R-Square	C(p)	F Value	Pr >F
1	Park 3 (RIPA)	0.1143	0.1143	13.0554	20.14	<.0001
2	Exp. Level	0.0673	0.1817	2.3579	12.75	0.0005
3	*Strava* Use Freq.	0.0398	0.2214	-3.1436	7.87	0.0057

4.2 Mean Velocity

The best performing model according to the information criteria had 9 variables (Activity Days per Year, Experience Level, *Strava* use Frequency, Precipitation, Maximum Temperature, Minimum Temperature, Maximum Vapor Pressure, Minimum Vapor Pressure, and Park 3 (RIPA)) with an Adjusted R-Square of 0.1520. Next we consulted the stepwise selection regression model whose parameters were set to enter variables at $\alpha = .1$ and stay at $\alpha = .1$. The final model iterated four steps resulting in four predictors (Experience Level, Park 3 (RIPA),*Strava* use Frequency, and Activity Days per Year) with a Adjusted R-Square of 0.1343.

The model summary for the stepwise selection is shown in Table 3 below.

Table 3: Summary of Stepwise Model for Mean Velocity.

Step	Label	Partial R-Square	Model R-Square	C(p)	F Value	Pr >F
1	Exp. Level	0.0674	0.0674	15.6144	11.35	0.0009
2	Park 3 (RIPA)	0.0392	0.1067	10.4371	6.85	0.0097
3	*Strava* Use Freq.	0.0302	0.1369	6.9091	5.43	0.0211
4	Activ.Days/Year	0.0194	0.1563	5.3611	3.54	0.0618

4.3 Median Velocity

The best performing model according to the information criteria had 8 variables (Experience Level,*Strava* use Frequency, Age, Precipitation, Maximum Temperature, Maximum Vapor Pressure, Minimum Vapor Pressure, and Park 3 (RIPA)) with an Adjusted R-Square of 0.1025. Next we consulted the stepwise selection regression model whose parameters were set to enter variables at $\alpha = .1$ and stay at $\alpha = .1$. The final model iterated four steps resulting in four predictors (Park 3 (RIPA),Experience Level, *Strava* use Frequency, and Minimum Temperature) with an Adjusted R-Square of 0.0883. The model summary for the stepwise selection is shown in Table 4 below.

Table 4: Summary of Stepwise Model for Median Velocity.

Step	Label	Partial R-Square	Model R-Square	C(p)	F Value	Pr >F
1	Park 3 (RIPA)	0.0395	0.0395	9.3441	6.46	0.012
2	Exp. Level	0.0382	0.0777	4.813	6.46	0.012
3	*Strava* Use Freq.	0.0169	0.0946	3.9152	2.9	0.0906
4	Min. Temp	0.0168	0.1114	3.0479	2.9	0.0904

4.4 Interaction Terms

Next we explored if an interaction between two variables beyond the additive effects of the variables improved the performance of the model to explain variation in the dependent variable. Because Experience Level and *Strava* use Frequency appeared in all three stepwise models, we hypothesized that a combination of the two may have some synergizing effect because the more skilled or experienced the mountain biker and the more they use *Strava* we would expect them to travel at higher velocities. Additionally, we hypothesized that Activity days per year and *Strava* use frequency may have a synergy effect because the more frequent a mountain biker uses *Strava* and the more frequent they participate in mountain biking we would expect them to be more likely to be travelling at higher velocities given the nature of gamification in the *Strava* App. We defined two new higher-order variables, "Experience Level x *Strava* use Frequency" and "Activity Days Year x *Strava*", and included these new variables in the stepwise models to evaluate if the new interaction terms were significant and added to the explanatory power of the model.

For Maximum Velocity, the 'Experience Level x *Strava* use Frequency' and 'Activity Days Year x *Strava*' variables were not significant at $\alpha = 0.1$ and introduced issues of multicollinearity with other predictors in the model. Similarly, the two higher order terms failed to be significant at $\alpha = 0.1$ in the Mean Velocity model and introduced issues of multicollinearity with other predictors in the model. Finally, in the median velocity model the two new higher order terms the also failed to be significant at $\alpha = 0.1$ and introduced multicollinearity into the model. Furthermore, these higher order terms appear did not appear to be significant, do not demonstrate any multiplicative effect of explaining additional variance in the response variable variable, and will create issues with inference of the lower order term so we have chosen to exclude them from future models. We suspect that because these three ordinal variables are all measured on a 1 to 5 scale, the inverse of values on Experience Level and *Strava* use Frequency (i.e., Exp Lev 1 x *Strava* use 5 = 5 / Exp Lev 5 x *Strava* use 1

= 5) would be the same in new the higher order variable and as a result the interaction terms would not improve the model.

5 Final OLS Models

After performing variable selection procedures on the dataset we re-evaluated the OLS models produced by the two selection procedures. Upon review of the models, we chose to use the stepwise models for all three measures of velocity because they contained fewer predictors, but more predictors informed by theory of recreation behavior in the models than the models recommended by the multiple selection procedure with only a slight reduction in model Adjusted R-Square. With the exception of Median velocity, the models shown in Tables 2 and 3 were accepted for the final OLS models. When the final OLS regression on the stepwise model for Median Velocity was performed Minimum Temperature failed to be a significant predictor at $\alpha = 0.1$ and was subsequently removed from the model.

5.1 Maximum Velocity

The final OLS model for maximum velocity was a statistically significant linear model $F_{3,164} = 16.05, p < .001$ with an R-Square of 0.2270. The residuals appear to be depart from a normal distribution as assessed by the QQ plot and Histogram but have constant variance according to the residual vs predicted plot (Figure 10). Residuals were also plotted in sequence which indicated no apparent serial correlation in error terms (Figure 11) The Brown-Forsythe test of constant variance, $t = 1.662, p = .098$, confirms the assessment from the graphical diagnostics that the residuals are independent, identical, and zero mean however the Correlation Test of Normality indicates the residuals do not satisfy the assumption of normal distribution with a correlation of 0.982, which is slightly less than the expected correlation 0.987 for $\alpha = .05$ with this sample size.

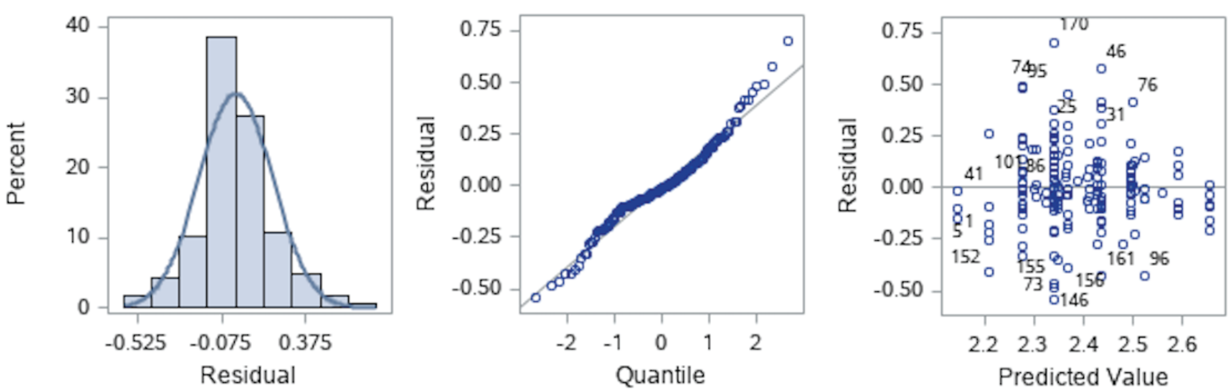

Figure 10: Graphical Residual Diagnostics for Maximum Velocity Final Model

We evaluated the model for multicollinearity and found low VIF values for the predictors with a model average VIF of 1.02 and only one component in the Principal Components Condition Index had a value above 10. This component with a condition index of 10.24, had two predictors that shared more than 50% of the variance in the model, the intercept and experience level, but because of the low VIF and condition index values we determined multicollinearity not to be an issue for this model. Finally, we consulted the DFFITS, DFBETA, Studentized Residuals and Cook's D statistics to detect outliers and influential observations in the data. Figure 12 indicates there are still outliers and influential points in the dataset which will affect the beta coefficients and $\hat{Y}$ predictions of the model, however we have chosen to include these observations so as not to overfit the model to the training data. The final model for maximum velocity satisfies the assumption of constant variance of residuals but the model appears to have a small violation of the assumption of normal distribution of residuals likely due to these residuals. Nevertheless, the model appears to otherwise perform quite well on the training data and has no issues of multicollinearity so despite the small deviation from a normal distribution we will accept this model as satisfactory.

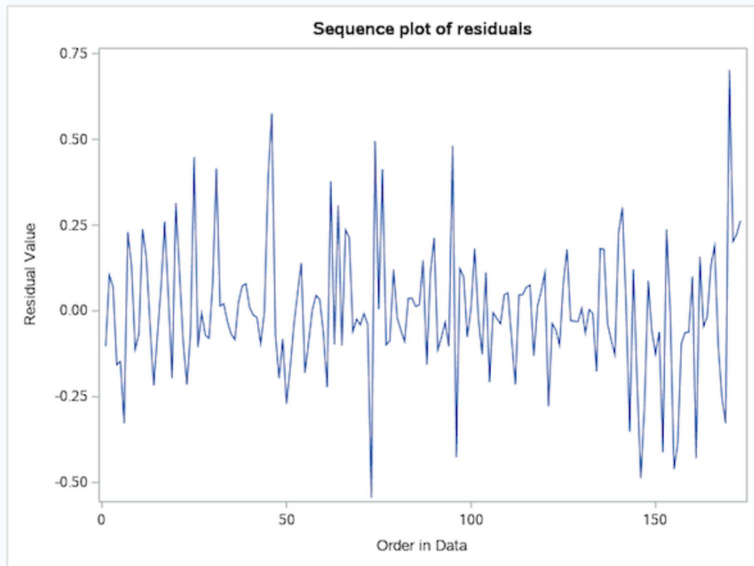

Figure 11: Sequence Plot of Residuals for Maximum Velocity Final Model

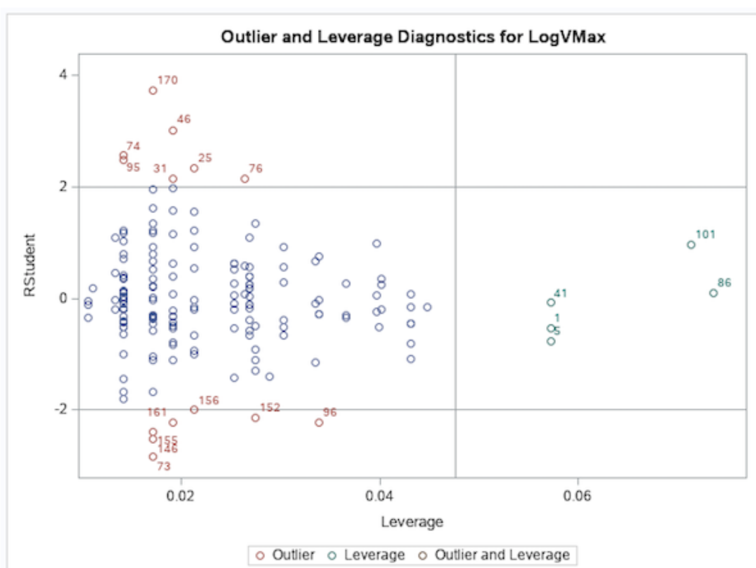

Figure 12: Outlier and Leverage Observations for Maximum Velocity Final Model

5.1.1 Model Interpretation

The final OLS model summary for Maximum Velocity is shown in Table 5 below. The variables in the model explain 22.7% of the variance in the response variable with an R-Square of 0.227. All of the predictors are statistically significant at the $p = .005$ level or below. The standardized beta coefficients allow for comparisons of the strength of association with the response variable between the predictors in the model. Ridge Park (Park = RIPA) is the largest standardized coefficient in the model with a beta of 0.033. This Park is situated on a coastal mountain range ridge with very steep and hilly terrain which is likely strongly correlated with higher mountain bike velocities. Because the response variable is log-transformed, the parameter estimates are in log (m/s) units. Back transforming the parameter estimates, this model suggests that we would expect the average maximum velocity for mountain bikers in Ridge Park to be $e^{0.153} = 1.165\ m/s$ faster than mountain bikers in other parks, holding all other variables constant. Experience level of the mountain biker is the next most important predictor in the model with a standardized beta coefficient of 0.018. For every unit increase in experience level we would expect the average maximum velocity to increase by $e^{0.066} = 1.068\ m/s$, holding all other variables constant. Finally, for every unit increase in *Strava* use frequency, we would expect a $e^{0.024} = 1.024\ m/s$ increase in average maximum velocity, holding all other variables constant.

Table 5: Final OLS Model for Maximum Velocity

Variable	DF	Parameter Estimate	Standardized Estimate	Standard Error	t Value	Pr>
Intercept	1	2.055	0.000	0.065	31.86	<.0001
Park=RIPA	1	0.153	0.317	0.033	4.61	<.0001
Exp. Level	1	0.066	0.256	0.018	3.68	0.0001
Strava Use Freq.	1	0.024	0.197	0.008	2.83	0.005

5.1.2 Model Validation

Next we evaluated our final OLS maximum velocity model to see how it performed on new data by calculating the mean squared prediction error (MSPR) on the test dataset. The MSPR for the test data was 0.0424 which we will compare to the mean squared error (MSE) of the final model on the test data to check for overfitting of the training data. We calculated the MSE for three models, a full model with all the variables in the analysis, the stepwise model, and an intercept only model to determine if we improved the overall fit of the data with our variable selection and improved predictions from an intercept only model. The MSE values for the three models are shown in Table 6 below.

Table 6: Mean Square Error (MSE) for Maximum Velocity Models.

Model	N	Mean	Std. Dev	Minimum	Maximum
Full Model	71	0.0437311	0.0727738	5.71E-08	0.4441684
Stepwise Model	73	0.0423593	0.069014	0.000056037	0.3782521
Intercept Only Model	75	0.0539772	0.0755496	7.94E-06	0.3510564

The Stepwise model has a smaller mean squared error than both the full model and the intercept only model which suggests it is the best model fit for the data. When we compare the MSE of the Stepwise model (0.0424) to the MSPR of the model on the test data we find them to be very similar which suggests that the model is reasonable and not overfit for the training data.

5.2 Mean Velocity

The final OLS model for mean velocity was a statistically significant linear model $F_{4,164} = 8.45, p < .0001$ with an R-Square of 0.1709. The residuals appear to be normally distributed as assessed by the QQ Plot and

Histogram and have constant variance according to the residual vs predicted plot (Figure 13). Residuals were also plotted in sequence which indicated no apparent serial correlation in error terms (Figure 14). The Brown-Forsythe test of constant variance, $t = 0.059, p = .952$, confirms the assessment from the graphical diagnostics that the residuals are independent, identical, and zero mean and the Correlation Test of Normality indicates the residuals satisfy the assumption of normal distribution with a correlation of 0.991, which is greater than the expected correlation 0.987 for $\alpha = .05$ with this sample size.

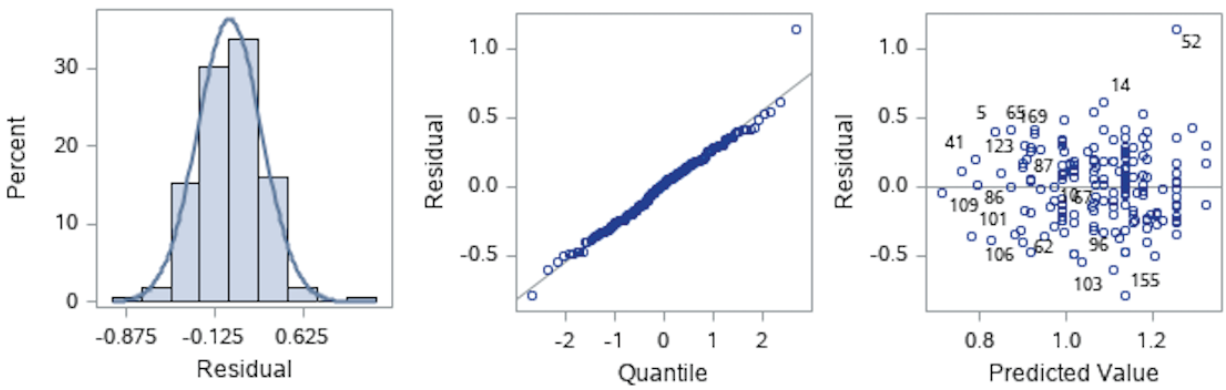

Figure 13: Graphical Residual Diagnostics for Mean Velocity Final Model

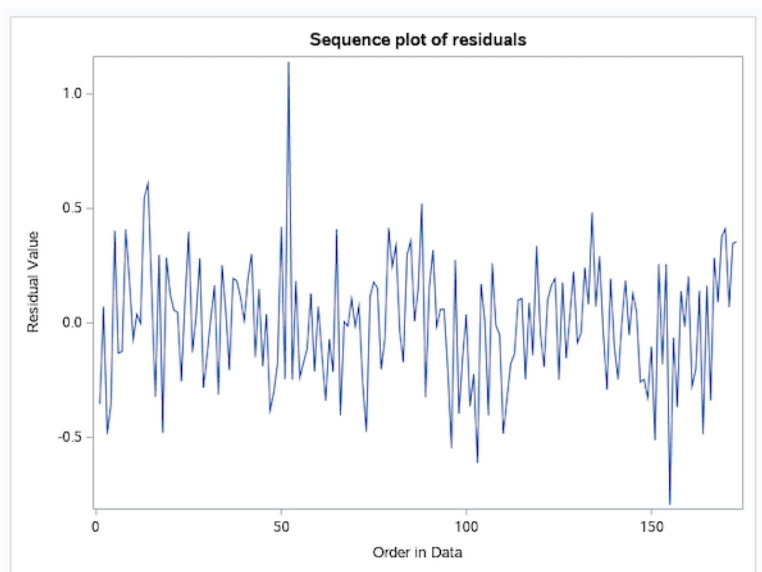

Figure 14: Sequence Plot of Residuals for Mean Velocity Final Model

We evaluated the model for multicollinearity and found low VIF values for the predictors with a model average VIF of 1.06 but two components in the Principal Components Condition Index had a values above 10. However, only one component with condition index 10.000, had two predictors that shared more than 50% of the variance in the model, Experience Level and Activity Days per Year. Nevertheless, we determined multicollinearity not to be an issue for this model with the low VIF values and relatively low values in the condition index. Finally, we consulted the DFFITS, DFBETA, Studentized Residuals and Cook's D statistics to detect outliers and influential observations in the data. Figure 15 indicates that there are still outliers and influential points in the dataset which will affect the beta coefficients and $\hat{Y}$ predictions of the model, however we have chosen to include these observations so as not to overfit the model to the training data. The final model for mean velocity satisfies the assumption of constant variance and normal distribution of residuals and has no issues with multicollinearity so we will accept this model as satisfying the assumptions of OLS regression.

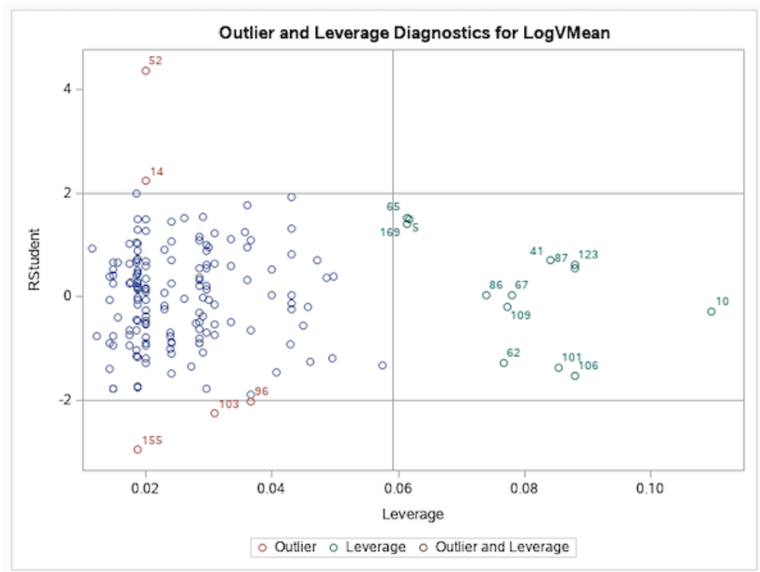

Figure 15: Outlier and Leverage Observations for Maximum Velocity Final Model

5.2.1 Model Interpretation

The final OLS model summary for mean velocity is shown in Table 7 below. The variables in the model explain 17.1% of the variance in the response variable with an R-Square of 0.1709. All of the predictors in the model are statistically significant at the $p < 0.1$ or below. The standardized beta coefficients allow for comparisons of the strength of association with the response variable between the predictors in the model. Experience level is the largest positive coefficient in the model with a beta of 0.200. This makes intuitive sense, that as the skill or experience level of mountain biker increases we would expect their velocity to increase. Because the response variable is log-transformed, the parameter estimates are in log (m/s) units. Back transforming the parameter estimates, this model suggests that we would expect that for every unit increase in experience level we would expect the average mean velocity to increase by $e^{0.070} = 1.072\ m/s$, holding all other variables constant. For every unit increase in *Strava* use frequency, we would expect a $e^{0.030} = 1.03\ m/s$ increase in average mean velocity mountain bikers travel, holding all other variables constant.

Table 7: Final OLS Model for Mean Velocity

Variable	DF	Parameter Estimate	Standardized Estimate	Standard Error	t Value	Pr>
Intercept	1	0.641	0.000	0.111	5.78	<.0001
Exp. Level	1	0.070	0.200	0.026	2.66	0.0086
Park=RIPA	1	-0.146	-0.223	0.047	-3.14	0.002
Strava Use Freq.	1	0.030	0.185	0.012	2.55	0.0115
Activ. Days/Year	1	0.046	0.130	0.026	1.75	0.0826

5.2.2 Model Validation

Next we evaluated our OLS final mean velocity model to see how well it performed on new data by calculating the mean square prediction error (MSPR) on the test dataset. The MSPR for the test data was 0.0675 which we compared to the mean squared error (MSE) of the final model on the test data to check for model overfitting of the training data. We calculated the MSE for three models, a full model with all the variables in the analysis, the stepwise model, and an intercept only model to determine if we improved the overall fit of the data with our variable selection and improved predictions from an intercept only model. The MSE values for the three models are shown in Table 8 below.

Table 8: Mean Square Error(MSE) for Mean Velocity Models.

Model	N	Mean	Std. Dev	Minimum	Maximum
Full Model	71	0.062558	0.0851791	2.22E-07	0.4366441
Stepwise Model	73	0.0675426	0.1017578	9.27E-06	0.6144009
Intercept Only Model	75	0.0678513	0.0904579	8.46E-07	0.4524791

The Stepwise model has a smaller mean squared error than the intercept only model but had a increase in MSE than the full model, which suggests there was a reduction in model fit. Nevertheless, when we compare the MSE of the Stepwise model (0.0675) to the MSPR of the model on the test data we find them to be very similar which suggests the model is reasonable and not overfit for the training data.

5.3 Median Velocity

The final OLS model for median velocity was a statistically significant linear model $F_{3,165} = 6.77, p < .0005$ with an R-Square of 0.1096. The residuals appear have a slight deviation from a normal distribution as assessed by the QQ plot and Histogram but have constant variance according the residual vs. predicted plot (Figure 16). Residuals were also plotted in sequence which indicated no apparent serial correlation in error terms (Figure 17). The Brown-Forsythe test of constant variance , $t = 0.367, p = .714$, confirms the assessment from the graphical diagnostics that the residuals are independent, identical and zero mean however the Correlation Test of Normality indicates that the residuals fail to satisfy the assumption of normal distribution with a correlation of 0.933, which is less than the expected correlation of 0.987 for $\alpha = .05$ with this sample size.

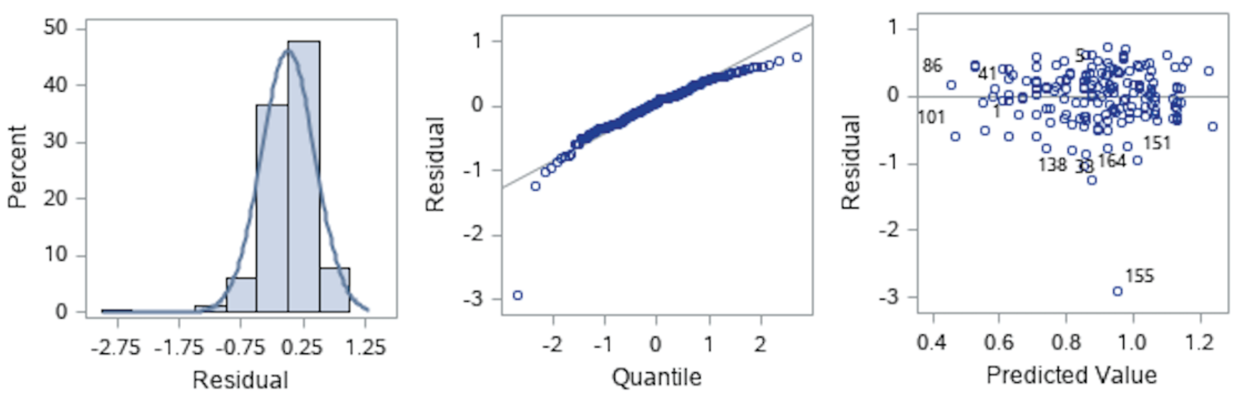

Figure 16: Graphical Residual Diagnostics for Median Velocity Final Model

We evaluated the model for multicollinearity and found low VIF values for the predictors with a model average VIF of 1.02 and only 1 component in the Principal Components Condition Index had a value above 10. This component had a condition index of 10.25 and had two predictors that shared more than 50% of the variance in the model between, the model Intercept and “Experience Level”. Nevertheless, we determined multicolinearity not to be an issue for this model with low VIF values and low condition index values. Finally, we consulted the DFFITS, DFBETA, Studentized Residuals and Cook’s D statistics to detect outliers and influential observations in the data. Figure 18 indicates that there are still outliers and influential points in the dataset which will affect the beta coefficients and $\hat{Y}$ predictions of the model, however we have chosen to

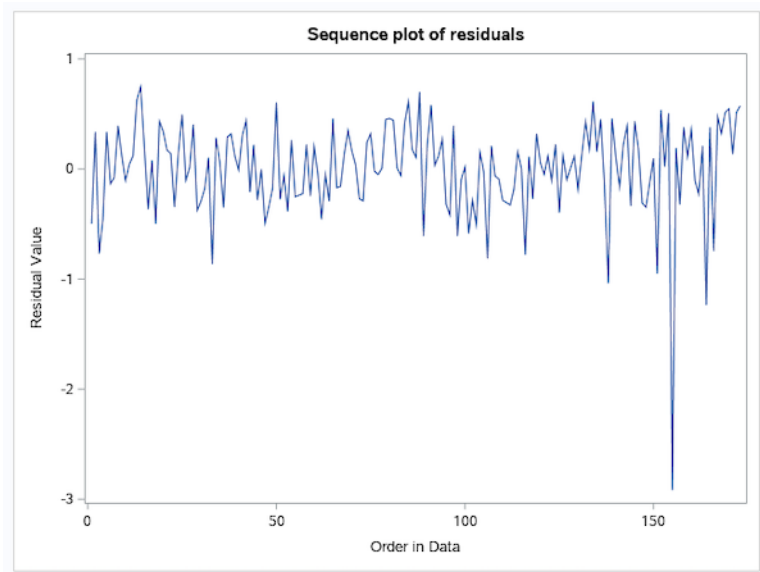

Figure 17: Sequence Plot of Residuals for Median Velocity Final Model

include these observations so as not to overfit the model to the training data. The final model for median velocity satisfies the assumption of constant variance but has a slight deviation from the assumption of normal distribution. This will affect the model's accuracy of predictions and reduce the explanatory power of the model, but we will perform a non-parametric regression on this data that does not require the assumption of normal distribution. Nevertheless, we will proceed with this model accepting that it does not fully satisfy all of the assumptions of OLS regression.

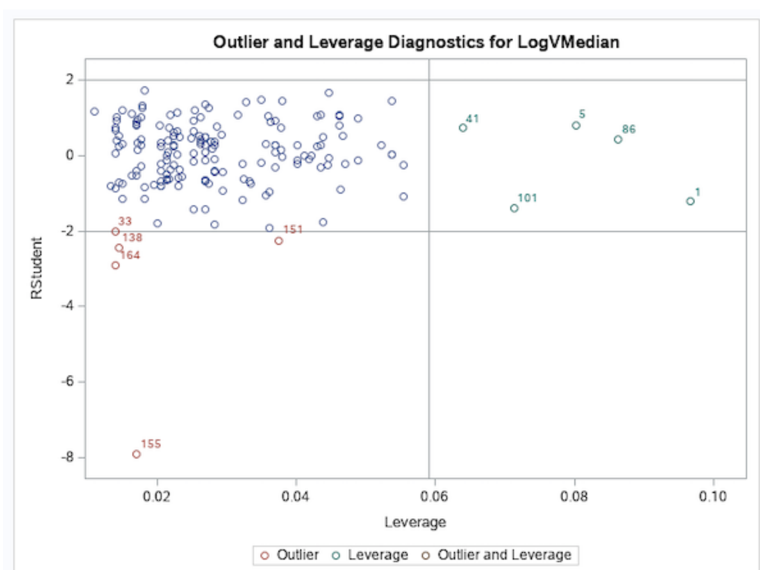

Figure 18: Outlier and Leverage Observations for Median Velocity Final Model

5.3.1 Model Interpretation

The final OLS model summary for median velocity is shown in Table 9 below. The variables in the model explain 11% of the variance in the response variable with an R-Square of 0.1096. All of the predictors in the data are significant at the $p < 0.05$ or below. The standardized beta coefficients allow for comparisons of the strength of association with the response variable between the predictors in the model. From this model we can see that like the other velocity models in this analysis, Experience Level has a greater effect on velocity than Strava Use Frequency. Because the response variable is log-transformed, the parameter estimates are in log (m/s) units. Back transforming the parameter estimates, this model suggests that we would expect that for every unit increase in Experience Level we would expect the average median velocity to increase by $e^{0.084} = 1.09\ m/s$, holding all other variables constant. For every unit increase in *Strava* use Frequency, we would expect a $e^{0.038} = 1.04\ m/s$ increase in average median velocity mountain bikers travel, holding all other variables constant.

Table 9: Final OLS Model for Median Velocity

Variable	DF	Parameter Estimate	Standardized Estimate	Standard Error	t Value	Pr>
Intercept	1	0.578	0.357	0.000	4.05	<.0001
Park=RIPA	1	-0.237	0.075	-0.237	-3.22	0.0015
Exp. Level	1	0.084	0.040	0.157	2.1	0.037
Strava Use Freq.	1	0.038	0.019	0.153	2.05	0.0423

5.3.2 Model Validation

Next we evaluated our final OLS median velocity model to see how well it performed on new data by calculating the mean square prediction error (MSPR) on the test dataset. The MSPR for the test data was 0.1312 which we compared to the mean squared error (MSE) of the final model on the test data to check for model overfitting of the training data. We calculated the MSE for three models, a full model with all the variables in the analysis, the stepwise model, and an intercept only model to determine if we improved the overall fit of the data with our variable selection and improved predictions from an intercept only model. The MSE values for the three models are shown in Table 10 below.

Table 10: Mean Square Error(MSE) for Median Velocity Models.

Model	N	Mean	Std. Dev	Minimum	Maximum
Full Model	71	0.1060477	0.1432166	0.000014244	0.5974618
Stepwise Model	73	0.1312111	0.188554	0.000028872	1.2394641
Intercept Only Model	75	0.1434768	0.1777056	0.000034377	0.9645178

The Stepwise model has a smaller mean squared error than the intercept only model but had a increase in MSE than the full model, which suggests there was a reduction in model fit. Nevertheless, when we compare the MSE of the Stepwise model (0.1312) to the MSPR of the model on the test data we find them to be very similar which suggests the model is reasonable and not overfit for the training data.

6 Non-Parametric Regression: Regression Trees

Next, we will explore a non-parametric regression which has fewer assumptions than OLS regression regarding normal distribution of residuals, constant variance, and has no expectation of a linear relationship between the explanatory variables and response variable. Since we have many ordinal variables in this analysis and we are predicting a continuous variable, Regression Trees appeared to be well suited for this analysis as an alternative to OLS Regression. Additionally, Regression Trees provide an intuitive visual summary of the model output that orders the predictors in terms of their importance at predicting the response variable. We will keep our training and test datasets to first fit the model to the training set and and evaluate the Regression Tree on the test dataset. Since we will be pruning the Regression Tree, we will use the full model (all variables) for each measure of velocity to determine which are the most important predictors, similar to variable selection in OLS Regression.

6.1 Maximum Velocity

A Regression Tree for maximum velocity was performed using all of explanatory variables in the dataset predicting the untransformed Maximum Velocity response variable. We performed multiple iterations to determine the optimal size of tree depth and branches. The final tree, shown in Figure 19, had 6 nodes and an Average Squared Error (ASE) of 4.8038.

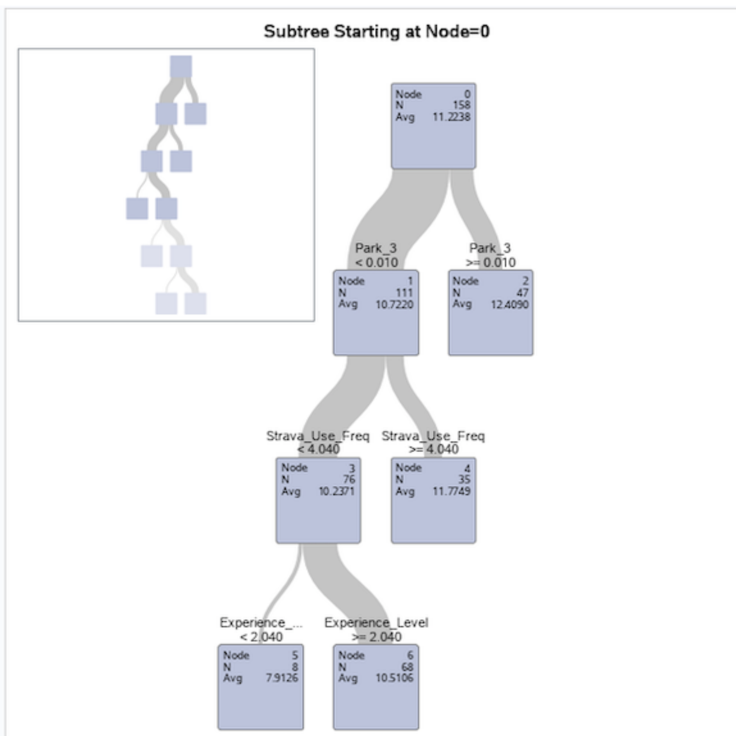

Figure 19: Maximum Velocity Regression Tree

Table 11: Maximum Velocity Variable Importance

Variable	Relative	Importance	Count
Park=RIPA	1	9.6941	1
Age	0.8762	8.494	2
Strava Use Frequency	0.7766	7.5281	1
Experience Level	0.717	6.9506	1

Table 11 summarizes the importance of explanatory variables in the tree at predicting the response variable. Whether a GPS track was recorded in Ridge Park is the most important predictor and the root node of the tree. If the track was recorded in Ridge Park the model predicts the average maximum velocity to be 12.41 m/s, and if it was recorded in some other park the predicted average maximum was 10.72 m/s. Age while important to predicting the model was pruned from the tree because it was too low in the branches at the 9th node. Next the predicted average maximum velocity for a mountain biker who uses *Strava* often or always is 11.77 m/s. However, the model predicts that the average maximum velocity a mountain biker who uses *Strava* never, rarely, or sometimes is 10.24 m/s. The last split in the tree is at the novice experience level, where the average maximum velocity of more experienced riders is 10.51 m/s and those with less experience 7.91 m/s. Finally, in order to validate our model we compared the ASE of the tree using the training data to the tree on the test dataset to compare its performance on new data. The ASE using the tree on the test data was 6.126 which is only slightly larger than the ASE of the training data (4.8038) which suggests that the model is not overfitted and will perform well on new data.

6.2 Mean Velocity

A Regression Tree for mean velocity was performed using all of the explanatory variables in the dataset predicting the untransformed Mean Velocity response variable. The default Cost Complexity pruning consistently returned a single node for the tree so we chose a Reduced Error pruning method and specified the growth of the tree to use an F statistic to split each variable and use the resulting p-value to determine the split variable for the tree. The final tree, shown in Figure 20, had 5 nodes and an ASE of 0.9339.

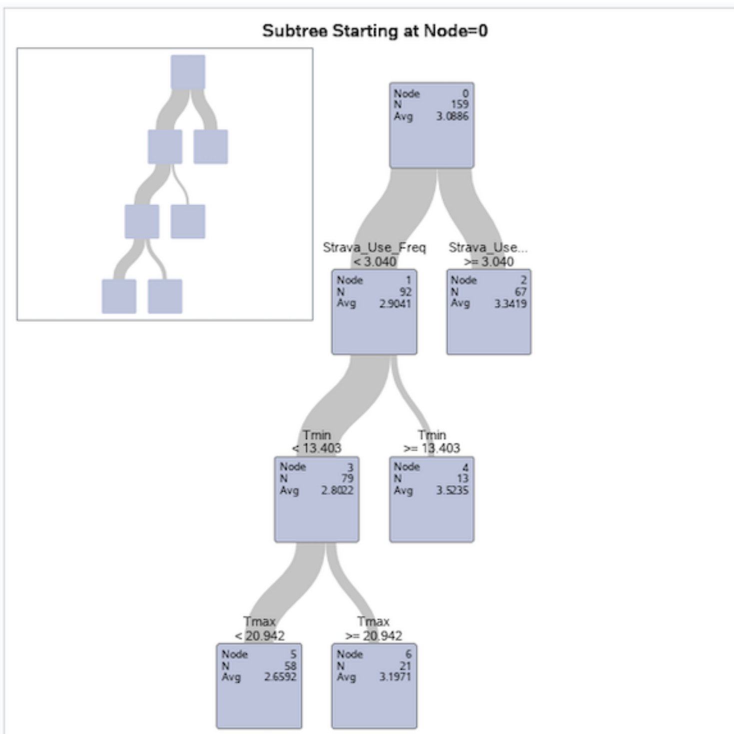

Figure 20: Mean Veloctiy Regression Tree

Table 12: Mean Velocity Variable Importance

Variable	Relative	Importance	Count
Strava Use Frequency	1	2.7255	1
Maximum Daily Temperature	0.8843	2.4101	1
Minimum Daily Temperature	0.775	2.1122	1

Table 12 summarizes the importance of each explanatory variable in the tree at predicting the response variable. The most important predictor was *Strava* Use, followed by Maximum Daily Temperature, and Minimum Daily Temperature. The predicted average mean velocity for a mountain biker who used *Strava* often or always is 3.34 m/s, while the predicted average mean velocity of those who use *Strava* less frequently is 2.904 m/s. Next, if the daily minimum temperature was greater than 13.4 °C; the predicted average 'mean velocity' is 3.52 m/s while if the minimum daily temperature was less than 13.4 °C the predicted average 'mean velocity' is 2.802 m/s. The last split in the tree is the maximum daily temperature, which if greater than 20.94 °C the average predicted 'mean velocity' is 3.197 m/s, and if less than 20.94 °C the average predicted 'mean velocity' is 2.659 m/s. Moreover, it seems that the warmer the temperature we would expect the predicted mean velocity to increase. Finally, we compared the ASE of the tree created with the training data on the test data to compare its performance and evaluate the fit of the tree. The ASE using the tree on the test data was 0.597 which is a reduction from the ASE of the tree using the training data (0.9339), likely due to the reduced error pruning parameter, which indicates that the model is not overfit and performs well on new data.

6.3 Median Velocity

A Regression Tree for median velocity was performed using all of the explanatory variables in the dataset predicting the untransformed Median Velocity response variable. Similar to the Mean Velocity model, the cost complexity pruning returned a single node for the tree so we again chose a reduced error pruning method and specified the growth of the tree using an F statistic to split each variable and use the resulting p-value to determine the split variable for the tree. The final tree, shown in Figure 21, had 10 nodes and an ASE of 0.8210.

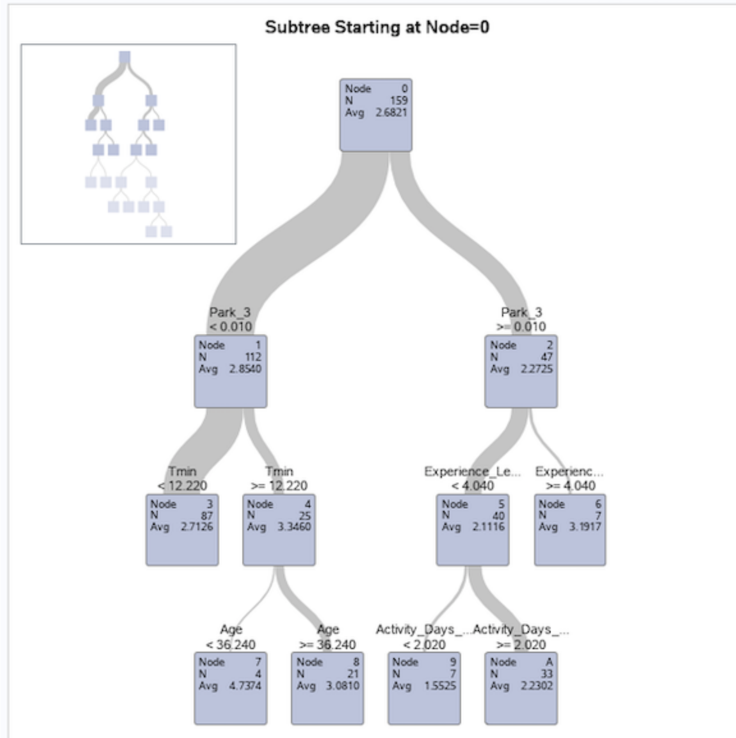

Figure 21: Median Velocity Regression Tree

Table 13: Median Velocity Variable Importance

Variable	Relative	Importance	Count
Park=RIPA	1	3.3462	1
Age	0.9073	3.0362	1
Maximum Daily Temperature	0.8342	2.7913	1
Experience Level	0.8003	2.6778	3
Activity Days per Year	0.4966	1.6616	2
Minimum Vapor Pressure deficit	0.343	1.1478	1
Strava Use Frequency	0.2036	0.6813	1

Table 13 summarized the variable importance for the predictions of the response variables. The most important predictor was Ridge Park, followed by Age, Maximum Daily Temperature, Experience Level, etc.. Although these were the most important predictors for the response variable, some were not included in the final tree because of the splitting and pruning parameters defined in the code. Nevertheless, if a GPS track was recorded in Ridge Park the model predicts the average median velocity to be 2.27 m/s, while tracks recorded elsewhere were predicted to be an average median velocity of 2.854 m/s. The split was Minimum Daily Temperature, where if the temperature was below 12.2 °C the predicted average median velocity was 2.854 m/s and if it was above 12.2 °C the predicted average median velocity was 3.326 m/s. If a mountain biker reported their experience level was advanced or expert, the predicted average median velocity was 3.191 m/s while mountain bikers with less experience were predicted to have an average median velocity of 2.112 m/s. Finally, we compared the ASE of the tree created with the training data on the test data to compare its performance and evaluate the fit of the tree. The ASE using the tree on the test data was 1.196 which is a slight increase from the ASE of the tree using the training data (0.8210) which indicates that the tree is not overfit and performs well on new data.

7 Conclusion

The OLS regression models performed very well in explaining the maximum, mean, and median velocities with R-Square values of 0.2270, 0.1790, and 0.1096 respectively. While some might consider these values quite low, modelling human behavior is very challenging and these R-Square values represent an acceptable models. Nevertheless, OLS regression is somewhat inflexible and encumbered by very restrictive assumptions about the data and the relationship between response and explanatory variables. For these reasons, we chose to use non-parametric regression approaches as an alternative to understanding the relationships between the explanatory variables and the aggregate measures of velocity in this analysis. Interestingly, with the exception of Maximum Velocity the Regression Trees provided results that were different from the variables the Stepwise OLS models but were similar to the models recommended by the multiple selection criteria that maximized the model's explanatory power. Regression Trees treat the ordinal explanatory variables in this analysis as continuous and sometimes created splits in the data between the integers of the categories, but often close enough to an integer that it did not create any uncertainty in the decision. However, compared to other regression methods for categorical predictor variables like logit and logistic regression, Regression Trees are far more simple to perform and interpret. Furthermore, while these two methods use entirely different approaches and assumptions they converged on relatively similar conclusions about relationships within the dataset.

While this analysis provided some insight into the behavior change related to speed of travel, I would like to further analyze the GPS tracks with a network analysis that compares the functional use of the trail system between *Strava* mountain bikers and non-*Strava* mountain bikers to understand and probe differences in their spatial use of the park and trail system. Previous analysis compared kernel densities of park use between the two groups but found subtle differences in patterns of use at the scale of the whole park. Additionally, I would like to further explore how use of fitness tracking apps like *Strava* affect attitudes and perceptions about the natural environment in the locations where these apps are used. Specifically, how does use of these apps affect the perceptions of ecological impact to soils, vegetation, and wildlife that result from all recreation compared to recreationists who don't use these apps. Previous analysis using the survey where the data in this study was drawn suggest a diminished sensitivity to perceive these impacts and a reduced sense of negative affective response about these impacts (i.e. they are less likely to perceive these ecological impacts as impacts and less likely to feel that they detract from their experience or others' experience or these impacts should be avoided).

In all three OLS models of velocity, *Strava* use was a statistically significant predictor and was positively correlated with increased velocities. This demonstrates that *Strava* use does affect behavior, particularly the maximum velocities of mountain bikers. This provides support for the hypothesis that the gamification features in the *Strava* app which allow users to compete for the fastest times on segments of trail influences behavior that is significantly different than mountain bikers who don't use the app. Further, *Strava* use appears to have less influence on mean and median velocity which may suggest that the behavior is most different on the short segments of trail where *Strava* users compete, while their behavior for the rest of their ride is similar to mountain bikers who don't use *Strava*. Nevertheless, this study is just one of a growing body of research that is beginning to understand how smartphones, apps, and connected devices are changing our behavior. While the findings from this study aren't meant to be prescriptive, I would recommend to land managers of Park and Protected areas to be cognizant of apps like *Strava* because often times the trails where riders compete for the fastest time are often multi-use trails which could raise concerns about visitor safety and could potentially diminish the quality of other visitors' recreation experience.

8 Appendix: SPSS Code

```
FILENAME REFFILE '/home/u45031657/Final Project/Final_Project_Data.sav';

/* Read in the sav file using proc import*/
PROC IMPORT DATAFILE=REFFILE replace
    DBMS=SAV
    OUT=WORK.Strava;
    GETNAMES=YES;
RUN;
/**************************************************************************
MODEL ASSUMPTIONS
**************************************************************************/

proc transreg data = Strava;
    model boxcox (Vmax / lambda = -1 to 1 by 0.1)
    = identity (Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp MinVp Pa
    title1 'Box-Cox Transformation: Regressing Explanatory Variables on Maximum Velocity';
    run;
proc transreg data = Strava;
    model boxcox (VMean / lambda = -1 to 1 by 0.1)
    = identity (Activity_Days_Year Strava_Use_Freq Experience_Level Age Ppt Mdt Tmax Tmin MaxVp MinVp Pa
    title1 'Box-Cox Transformation: Regressing Explanatory Variables on Mean Velocity';
    run;
proc transreg data = Strava;
    model boxcox (VMedian / lambda = -1 to 1 by 0.1)
    = identity (Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp MinVp Pa
    title1 'Box-Cox Transformation: Regressing Explanatory Variables on Median Velocity';
    run;


/******** SEPARATE DATA INTO TRAIN AND TEST DATASETS *****/
proc surveyselect data=Strava seed=12345 out=Strava2
    rate = 0.3 outall;
run;

data StravaTrain; set Strava2;
if Selected = 0;
run;

data StravaTest; set Strava2;
if Selected = 1;
run;

proc print data = StravaTrain;
run;

proc print data = StravaTest;
run;



/**********Initial OLS model for Maximum Velocity***********/
```

```
proc reg data = StravaTrain
    plots(label) = (CooksD RStudentbyLeverage DFFITS DFBETAS);
    model LogVmax = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp MinV
    output out=vmaxout r=resid p=pred;
    title1 'Initial OLS model for Maximum Velocity';
run;
/* Initital exploratory plots */
proc sgplot data = StravaTrain;
    scatter x=LogVmax y=resid;
run;
/* Sequence plots */
data temp; set vmaxout;
    order = _n_;
proc sgplot data = temp;
    series x=order y=resid/ lineattrs=(pattern=solid);
    xaxis label = 'Order in Data';
    yaxis label = 'Residual Value';
    title1 'Sequence plot of residuals';
run;
/*Brown Forsythe test of constant variance*/
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicted varial

%resid_num_diag(dataset=vmaxout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');

/************Initial OLS model for Mean Velocity***************/
proc reg data = StravaTrain
    plots(label) = (CooksD RStudentbyLeverage DFFITS DFBETAS);
    model LogVMean = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Mi
    output out=vmeanout r=resid p=pred;
    title1 'Initial OLS model for Mean Velocity';
run;
/* Initital exploratory plots */
proc sgplot data = StravaTrain;
    scatter x=LogVMean y=resid;
run;
/* Sequence plots */
data temp; set vmeanout;
    order = _n_;
proc sgplot data = temp;
    series x=order y=resid/ lineattrs=(pattern=solid);
    xaxis label = 'Order in Data';
    yaxis label = 'Residual Value';
    title1 'Sequence plot of residuals';
run;
/*Brown Forsythe test of constant variance*/
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicted varial

%resid_num_diag(dataset=vmeanout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');

/*************Initial OLS model for Median Velocity************/
proc reg data = StravaTrain
    plots(label) = (CooksD RStudentbyLeverage DFFITS DFBETAS);
```

```
    model LogVMedian = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp 
    output out=vmedout r=resid p=pred;
    title1 'Initial OLS model for Median Velocity'
run;
/* Initital exploratory plots */
proc sgplot data = StravaTrain;
    scatter x=LogVMean y=resid;
run;
/* Sequence plots */
data temp; set vmedout;
    order = _n_;
proc sgplot data = temp;
    series x=order y=resid/ lineattrs=(pattern=solid);
    xaxis label = 'Order in Data';
    yaxis label = 'Residual Value';
    title1 'Sequence plot of residuals';
run;
/*Brown Forsythe test of constant variance*/
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicted varial

%resid_num_diag(dataset=vmedout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');


/**********************************************************************
VARIABLE SELECTION
**********************************************************************/
/** Maximum Velocity  Variable Selection **/
proc reg data = StravaTrain;
    model LogVmax = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Min
     title 'Maximum Velocity Multiple Selection Criteria';
run;

proc reg data = StravaTrain;
    model LogVmax = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Min
    title1 'Stepwise Selection';
run;

/**Mean Velocity Variable Selection**/
proc reg data = StravaTrain;
    model LogVMean = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Mi
     title 'Mean Velocity Multiple Selection Criteria';
run;

proc reg data = StravaTrain;
    model LogVMean = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Mi
    title1 'Stepwise Selection';
run;

/**Median Velocity Variable Selection**/
proc reg data = StravaTrain;
    model LogVMedian = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp 
     title 'Median Velocity Multiple Selection Criteria';
run;
```

```
proc reg data = StravaTrain;
    model LogVMedian = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp 
    title1 'Stepwise Selection';
run;
/************************************
Define Higher Order Predictors
************************************/
data StravaTrain2; set StravaTrain;
    ADY_Strava= Activity_Days_Year * Strava_Use_Freq;
    ExpL_Strava = Experience_LEvel * Strava_Use_Freq;
run;
/** Maximum Velocity Model with interaction term**/
proc reg data = StravaTrain2;
    model LogVmax = ADY_Strava ExpL_Strava Park_3 Experience_Level Strava_Use_Freq/ vif;
    title1 'Check for Interaction Effect in Maximum velocity model';
run;
/**Mean Velocity Variable Selection with interaction term**/
proc reg data = StravaTrain2;
    model LogVMean = ADY_Strava ExpL_Strava Experience_Level Park_3 Strava_Use_Freq Activity_Days_Year/
    title1 'Check for Interaction Effect in Mean velocity model';
run;

/**Median Velocity Variable Selection with interaction term**/
proc reg data = StravaTrain2;
    model LogVMedian = ADY_Strava ExpL_Strava Park_3 Experience_Level Strava_Use_Freq Tmin/ vif;
    title1 'Check for Interaction Effect in Median velocity model '
run;
/***********************************************************************
FINAL OLS MODELS
***********************************************************************/
/**********Final OLS model for Maximum Velocity************/
proc reg data = StravaTrain
    plots(label) = (CooksD RStudentbyLeverage DFFITS DFBETAS);
    model LogVmax = Park_3 Experience_Level Strava_Use_Freq/ stb vif collin;
    output out=vmaxout r=resid p=pred;
    title1 'Final OLS model for Maximum Velocity';
run;
/* Initital exploratory plots */
proc sgplot data = StravaTrain;
    scatter x=LogVmax y=resid;
run;
/* Sequence plots */
data temp; set vmaxout;
    order = _n_;
proc sgplot data = temp;
    series x=order y=resid/ lineattrs=(pattern=solid);
    xaxis label = 'Order in Data';
    yaxis label = 'Residual Value';
    title1 'Sequence plot of residuals';
run;
/*Brown Forsythe test of constant variance*/
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicted varia
```

```
%resid_num_diag(dataset=vmaxout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');

/**********************************************
Final model Validation for Maximum Velocity
**********************************************/
/****** MSPR for test set******/
data StravaTest; set StravaTest;
    LogVmaxHat =  2.055  + 0.153*Park_3 + 0.066*Experience_Level+ 0.024*Strava_Use_Freq;
    SqPredError = (LogVmax - LogVmaxHat)**2;
proc means data = StravaTest mean;
    var SqPredError;
    title1 'MSPR for Test Set';
run;

/****** Compare final model MSE to final model MSPR to check overfit ******/
/* Full model */
proc reg data = StravaTrain noprint;
    model LogVmax = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Min
store MaxModelFull;
run;
/* Stepwise model */
proc reg data = StravaTrain noprint;
    model LogVmax = Park_3 Experience_Level Strava_Use_Freq;
store MaxModelStep;
run;
/* Intercept only model */
proc reg data = StravaTrain noprint;
    model LogVmax = ;
store MaxModelIntcpt;
run;
/* Make predictions for each model */
proc plm restore = MaxModelFull;
    score data = StravaTest out = NewStravaTest1 predicted;
run;
proc plm restore = MaxModelStep;
    score data = StravaTest out = NewStravaTest2 predicted;
run;
proc plm restore = MaxModelIntcpt;
    score data = StravaTest out =NewStravaTest3  predicted;
run;
/* Estimate Error of model and calculate Means */
data NewStravaTest1; set NewStravaTest1;
ASE = (LogVmax - predicted)**2;
run;
data NewStravaTest2; set NewStravaTest2;
ASE = (LogVmax - predicted)**2;
run;
data NewStravaTest3; set NewStravaTest3;
ASE = (LogVmax - predicted)**2;
run;

proc means data = NewStravaTest1;
var ASE;
```

```
title1 'Full Model';
run;
proc means data = NewStravaTest2;
var ASE;
title1 'Stepwise Model';
run;
proc means data = NewStravaTest3;
var ASE;
title1 'Intercept Only Model';
run;

/************Final OLS model for Mean Velocity**************/
proc reg data = StravaTrain
    plots(label) = (CooksD RStudentbyLeverage DFFITS DFBETAS);
    model LogVMean = Experience_Level Park_3 Strava_Use_Freq Activity_Days_Year/ stb vif collin;
    output out=vmeanout r=resid p=pred;
    title1 'Final OLS model for Mean Velocity';
run;
/* Initital exploratory plots */
proc sgplot data = StravaTrain;
    scatter x=LogVMean y=resid;
run;
/* Sequence plots */
data temp; set vmeanout;
    order = _n_;
proc sgplot data = temp;
    series x=order y=resid/ lineattrs=(pattern=solid);
    xaxis label = 'Order in Data';
    yaxis label = 'Residual Value';
    title1 'Sequence plot of residuals';
run;
/*Brown Forsythe test of constant variance*/
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicted varia

%resid_num_diag(dataset=vmeanout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');

/*********************************************
Final model Validation for Mean Velocity
**********************************************/
/****** MSPR for test set******/
data StravaTest; set StravaTest;
    LogVMeanHat =  0.64114 + 0.0697*Experience_Level-0.14637*Park_3+0.02998*Strava_Use_Freq+0.04611*Acti
    SqPredError = (LogVMean - LogVMeanHat)**2;
proc means data = StravaTest mean;
    var SqPredError;
    title1 'MSPR for Test Set';
run;

/****** Compare final model MSE to final model MSPR to check overfit ******/
/* Full model */
proc reg data = StravaTrain noprint;
    model LogVmean = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp Mi
store MeanModelFull;
```

```
run;
/* Stepwise model */
proc reg data = StravaTrain noprint;
    model LogVmean = Experience_Level Park_3 Strava_Use_Freq Activity_Days_Year;
store MeanModelStep;
run;
/* Intercept only model */
proc reg data = StravaTrain noprint;
    model LogVmean = ;
store MeanModelIntcpt;
run;
/* Make predictions for each model */
proc plm restore = MeanModelFull;
    score data = StravaTest out = NewStravaTest4 predicted;
run;
proc plm restore = MeanModelStep;
    score data = StravaTest out = NewStravaTest5 predicted;
run;
proc plm restore = MeanModelIntcpt;
    score data = StravaTest out =NewStravaTest6  predicted;
run;
/* Estimate Error of model and calculate Means */
data NewStravaTest4; set NewStravaTest4;
ASE = (LogVmean - predicted)**2;
run;
data NewStravaTest5; set NewStravaTest5;
ASE = (LogVmean - predicted)**2;
run;
data NewStravaTest6; set NewStravaTest6;
ASE = (LogVmean - predicted)**2;
run;

proc means data = NewStravaTest4;
var ASE;
title1 'Full Model';
run;
proc means data = NewStravaTest5;
var ASE;
title1 'Stepwise Model';
run;
proc means data = NewStravaTest6;
var ASE;
title1 'Intercept Only Model';
run;
/*************Final OLS model for Median Velocity************/
proc reg data = StravaTrain
    plots(label) = (CooksD RStudentbyLeverage DFFITS DFBETAS);
    model LogVMedian = Park_3 Experience_Level Strava_Use_Freq/ stb vif collin;
    output out=vmedout r=resid p=pred;
    title1 'Final OLS model for Median Velocity'
run;
/* Initital exploratory plots */
proc sgplot data = StravaTrain;
    scatter x=LogVMean y=resid;
```

```
run;
/* Sequence plots */
data temp; set vmedout;
    order = _n_;
proc sgplot data = temp;
    series x=order y=resid/ lineattrs=(pattern=solid);
    xaxis label = 'Order in Data';
    yaxis label = 'Residual Value';
    title1 'Sequence plot of residuals';
run;
/*Brown Forsythe test of constant variance*/
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicted varial

%resid_num_diag(dataset=vmedout, datavar=resid,
   label='residual', predvar=pred, predlabel='predicted');

/**********************************************
Final model Validation for Median Velocity
**********************************************/
/****** MSPR for test set******/
data StravaTest; set StravaTest;
    LogVMedianHat =  0.57836-0.23736*Park_3 + 0.08350*Experience_Level+0.03781*Strava_Use_Freq;
    SqPredError = (LogVMedian - LogVMedianHat)**2;
proc means data = StravaTest mean;
    var SqPredError;
    title1 'MSPR for Test Set';
run;

/****** Compare final model MSE to final model MSPR to check overfit ******/
/* Full model */
proc reg data = StravaTrain noprint;
    model LogVMedian = Activity_Days_Year Experience_Level Strava_Use_Freq Age Ppt Mdt Tmax Tmin MaxVp I
store MedianModelFull;
run;
/* Stepwise model */
proc reg data = StravaTrain noprint;
    model LogVmedian = Experience_Level Park_3 Strava_Use_Freq;
store MedianModelStep;
run;
/* Intercept only model */
proc reg data = StravaTrain noprint;
    model LogVmedian = ;
store MedianModelIntcpt;
run;
/* Make predictions for each model */
proc plm restore = MedianModelFull;
    score data = StravaTest out = NewStravaTest7 predicted;
run;
proc plm restore = MedianModelStep;
    score data = StravaTest out = NewStravaTest8 predicted;
run;
proc plm restore = MedianModelIntcpt;
    score data = StravaTest out =NewStravaTest9  predicted;
run;
```

```
/* Estimate Error of model and calculate Means */
data NewStravaTest7; set NewStravaTest7;
ASE = (LogVmedian - predicted)**2;
run;
data NewStravaTest8; set NewStravaTest8;
ASE = (LogVmedian - predicted)**2;
run;
data NewStravaTest9; set NewStravaTest9;
ASE = (LogVmedian - predicted)**2;
run;

proc means data = NewStravaTest7;
var ASE;
title1 'Full Model';
run;
proc means data = NewStravaTest8;
var ASE;
title1 'Stepwise Model';
run;
proc means data = NewStravaTest9;
var ASE;
title1 'Intercept Only Model';
run;
/************************************************************************
REGRESSION TREES
*************************************************************************/
/* Maximum Velocity Regression Tree */
proc hpsplit data=StravaTrain seed=123 maxdepth=10 maxbranch=2;
    model Vmax = Activity_Days_Year Experience_Level Strava_Use_Freq Age
        Ppt Mdt Tmax Tmin MaxVp MinVp Park_1 Park_2 Park_3 Park_4 Park_5;
    output out = VMaxTree;
    code file='/home/u45031657/Final Project/VMaxtree.sas';
run;

proc sgplot data = VMaxTree;
    scatter x=Vmax y = p_Vmax/
    markerattrs=(symbol = circlefilled size = 6pt);
run;

/* Call the test data and make predictions on the tree */
data VMaxScored; set StravaTest;
%include '/home/u45031657/Final Project/VMaxtree.sas';
run;

/* Now calculate the MSPR as we did in OLS */
data VMaxtestTree; set VMaxScored;
ASE = (Vmax - P_Vmax)**2;
run;

proc means data = VMaxtestTree;
var ASE;
run;
```

```
/* Mean Velocity Regression Tree */
proc hpsplit data=StravaTrain seed = 123;
    model Vmean = Activity_Days_Year Experience_Level Strava_Use_Freq Age
        Ppt Mdt Tmax Tmin MaxVp MinVp Park_1 Park_2 Park_3 Park_4 Park_5;
    output out = VmeanTree;
    code file='/home/u45031657/Final Project/VMeantree.sas';
    grow ftest;
    prune rep ;
run;


/* Call the test data and make predictions on the tree */
data VMeanScored; set StravaTest;
%include '/home/u45031657/Final Project/VMeantree.sas';
run;

/* Now calculate the MSPR as we did in OLS */
data VMeantestTree; set VMeanScored;
ASE = (Vmean - P_Vmean)**2;
run;

proc means data = VMeantestTree;
var ASE;
run;


/* Median Velocity Regression Tree */
proc hpsplit data=StravaTrain seed = 123;
    model Vmedian = Activity_Days_Year Experience_Level Strava_Use_Freq Age
        Ppt Mdt Tmax Tmin MaxVp MinVp Park_1 Park_2 Park_3 Park_4 Park_5;
    output out = VmedianTree;
    code file='/home/u45031657/Final Project/VMediantree.sas';
    grow ftest;
    prune rep ;
run;

/* Call the test data and make predictions on the tree */
data VMedianScored; set StravaTest;
%include '/home/u45031657/Final Project/VMediantree.sas';
run;

/* Now calculate the MSPR as we did in OLS */
data VMediantestTree; set VMedianScored;
ASE = (Vmedian - P_Vmedian)**2;
run;

proc means data = VMediantestTree;
var ASE;
run;
```

9 References

Ajzen, Icek. 1991. "The Theory of Planned Behavior." *Organizational Behavior and Human Decision Processes* 50 (2): 179–211. https://doi.org/10.1016/0749-5978(91)90020-T.

Bandura, Albert. 1986. "Social Foundations of Thought and Action." *Englewood Cliffs, NJ* 1986.

Bandura, Albert, and Richard H. Walters. 1963. "Social Learning and Personality Development."

Barratt, Paul. 2017. "Healthy Competition: A Qualitative Study Investigating Persuasive Technologies and the Gamification of Cycling." *Health & Place* 46 (July): 328–36. https://doi.org/10.1016/j.healthplace.2016.09.009.

Chen, Chen. 2017. "Gamification in a Volunteered Geographic Information Context with Regard to Contributors' Motivations: A Case Study of OpenStreetMap." PhD thesis, University of Waterloo.

Deterding, Sebastian, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. "From Game Design Elements to Gamefulness: Defining Gamification." In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, 9–15. ACM.

Fishbein, M., and I. Ajzen. 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Reading, MA: Addison-Wesley Publishing Company.*

Hamari, Juho, Jonna Koivisto, and Harri Sarsa. 2014. "Does Gamification Work? – A Literature Review of Empirical Studies on Gamification." In *2014 47th Hawaii International Conference on System Sciences*, 3025–34. Waikoloa, HI: IEEE. https://doi.org/10.1109/HICSS.2014.377.

Lindsey, Joe. 2019. "Strava Is Booming. Just Don't Call It the Facebook of Fitness." *Outside Online*. https://www.outsideonline.com/2395489/strava-james-quarles.

Lupton, Deborah. 2016a. "Personal Data Practices in the Age of Lively Data." *Digital Sociologies*, 335–50.

———. 2016b. "The Diverse Domains of Quantified Selves: Self-Tracking Modes and Dataveillance." *Economy and Society* 45 (1): 101–22. https://doi.org/10.1080/03085147.2016.1143726.

Oregon State University. 2004. "PRISM Climate Group." http://prism.oregonstate.edu.

Putz, Lisa-Maria, and Horst Treiblmaier. 2015. "Creating a Theory-Based Research Agenda for Gamification." In *AMCIS*.

"Quantified Self." 2019. https://quantifiedself.com/.

Ryan, Richard M., and Edward L. Deci. 2000. "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being." *American Psychologist* 55 (1): 68.

Sailer, Michael, Jan Hense, Heinz Mandl, and Markus Klevers. 2013. "Psychological Perspectives on Motivation Through Gamification," 10.

Seaborn, Katie, and Deborah I. Fels. 2015. "Gamification in Theory and Action: A Survey." *International Journal of Human-Computer Studies* 74 (February): 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006.

Stragier, Jeroen, Peter Mechant, Lieven De Marez, and Greet Cardon. 2018. "Computer-Mediated Social Support for Physical Activity: A Content Analysis." *Health Education & Behavior* 45 (1): 124–31. https://doi.org/10.1177/1090198117703055.

"Strava Labs." 2018. *Strava Labs*. http://labs.strava.com.

Weber, Johann, Mojdeh Azad, William Riggs, and Christopher R. Cherry. 2018. "The Convergence of Smartphone Apps, Gamification and Competition to Increase Cycling." *Transportation Research Part F: Traffic Psychology and Behaviour* 56 (July): 333–43. https://doi.org/10.1016/j.trf.2018.04.025.

"What's a Segment?" 2012. *Strava Support*. http://support.strava.com/hc/en-us/articles/216917137-What-s-a-segment-.