1)

Predicting National Park Service Trail Popularity With Physical, Behavioral, and Reported Measures: Multiple Regression Analysis

Report by Sammy Rose Dobrozsi

October 18th, 2022

2)

With over 21,000 miles of recreational trails managed by the United States National Parks Service, it is critical to understand how many trail users are going to different trails and for what reasons in order to allocate resources and direct future development. The present data was scraped from the website AllTrails roughly three years ago and includes behavioral measures of trail users (e.g. trail usage levels), reported measures by trail users (e.g. trail user reviews), and physical attributes of trails (length, elevation). AllTrails is a trail database, hosting GPS files of trails entered by users supported by metadata, focused on a phone app that can assist with trail selection and navigation. The different types of data are advantageous for gaining a more holistic view of what makes trails popular (and presumably more desirable to users).

3)

This data was retrieved from Kaggle: https://www.kaggle.com/datasets/planejane/national-park-trails

The original data consists of 3,313 trail listings from AllTrails. Certain non-usable variables were removed from the .csv in Excel prior to importing into R, such as Name, Country, State, Lat and Long coordinates, Features, Activities, and Units. Trail ID was converted to string data, and Visitor Usage missing values were filled using rounded mean imputation. All ordinal-numeric or categorical variables (Trail Difficulty Rating, Route Type, Visitor Usage) were converted to factor before conducting regression analysis.

The study at hand seeks to predict Trail Popularity (y), measured by a compound index, using seven different physical, behavioral, or social measures of trail use. As popularity is no longer visible on AllTrails since this data was scraped, it is not possible to know how it is computed, since there may have been other trail attributes added/removed since then that contributed to its calculation. The seven measures of trails and trail use (x_i) are:

Physical Measures

Trail Length measured in meters,

Trail Elevation Gain measured in meters,

Trail Difficulty Rating in a scale of odd numbers from 1 to 7, set manually by AllTrails "based on trail condition, steepness of grades, gain and loss of elevation, and the amount and kinds of natural barriers that must be traversed,"

Route Type which describes whether trails are a loop back to the start, and out-and-back that returns along the same path to the start, or a point-to-point that does not loop or return to the start,

Behavioral Measures

Visitor Usage which rates the number of trail check-in's on a scale from 1 to 4,

Number of Reviews which is the number of reviews left by trail users on a given trail's AllTrails listing,

Reported Measures

Average Rating reported by trail users on a scale from 1 (worst) to 5 (best) by 0.5-point steps.

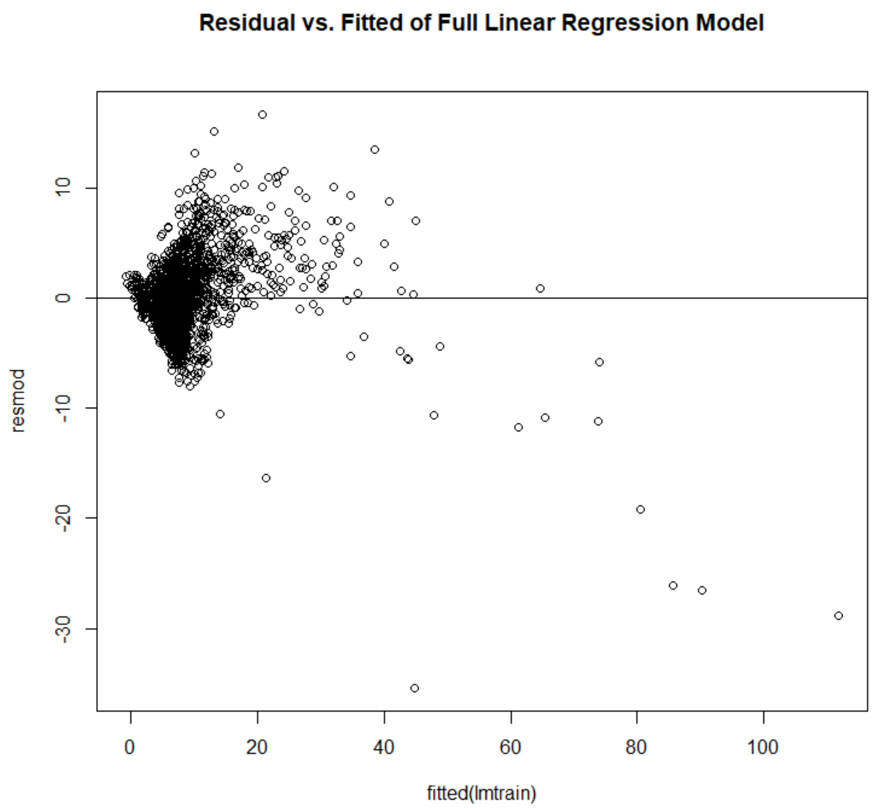
4)

Before conducting regression analysis, the data was partitioned into a training dataset (60% of observations) and a validation dataset (40% of observations). A regression model was run on the full dataset with all seven x variables.

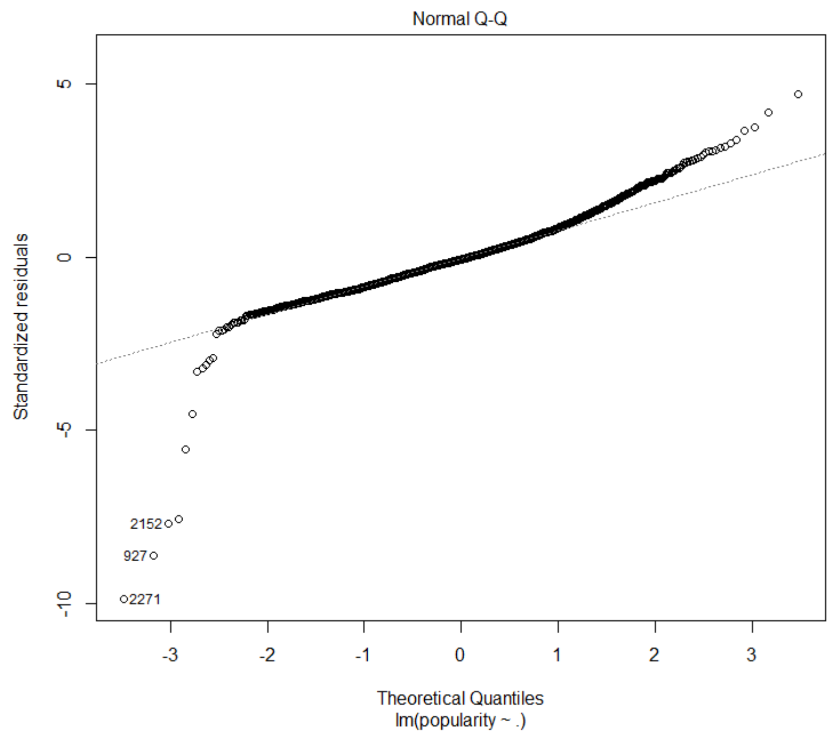
```
Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)    
(Intercept)             1.79623251  0.52295941    3.43             0.00061 ***
length                 -0.00001464  0.00000656   -2.23             0.02574 *  
elevation_gain          0.00025676  0.00018412    1.39             0.16330    
difficulty_rating3     -0.44307888  0.20735394   -2.14             0.03274 *  
difficulty_rating5     -0.63002706  0.27589715   -2.28             0.02250 *  
difficulty_rating7     -1.13009389  0.44998772   -2.51             0.01211 *  
route_typeout and back  0.26397744  0.17832091    1.48             0.13894    
route_typepoint to point -0.60586955 0.32415697  -1.87             0.06176 .  
visitor_usage2          0.69018822  0.18984511    3.64             0.00028 ***
visitor_usage3          3.35898176  0.29562196   11.36 < 0.0000000000000002 ***
visitor_usage4          1.21267766  0.81687287    1.48             0.13783    
avg_rating1            -0.32500116  1.86901472   -0.17             0.86197    
avg_rating1.5          -2.17650475  3.65454942   -0.60             0.55154    
avg_rating2             0.44758534  1.54624960    0.29             0.77226    
avg_rating2.5          -0.37742615  1.14597915   -0.33             0.74193    
avg_rating3             0.23530787  0.72525021    0.32             0.74563    
avg_rating3.5           1.70039267  0.54992635    3.09             0.00202 ** 
avg_rating4             3.22971367  0.48985864    6.59       0.000000000055 ***
avg_rating4.5           5.61068157  0.47368547   11.84 < 0.0000000000000002 ***
avg_rating5             4.58522691  0.50004754    9.17 < 0.0000000000000002 ***
num_reviews             0.03457308  0.00052188   66.25 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.62 on 1966 degrees of freedom
Multiple R-squared:  0.812,	Adjusted R-squared:  0.81 
F-statistic:  425 on 20 and 1966 DF,  p-value: <0.0000000000000002
```

Based on the Adjusted R^2 this model explains 81% of the variance in Trail Popularity. The R^2 and Adjusted R^2 are nearly identical; there are few extraneous variables in the model that penalize its Adjusted R^2 for complexity. Using a cutoff of 5% for the p-value of regressors, a revised model based on the full model would contain as x variables: Length, Difficulty Rating, Visitor Usage (only levels 2 and 3 were significant), Average Rating (only 3.5-5 were significant), and Number of Reviews. Elevation Gain and Route Type would be excluded.

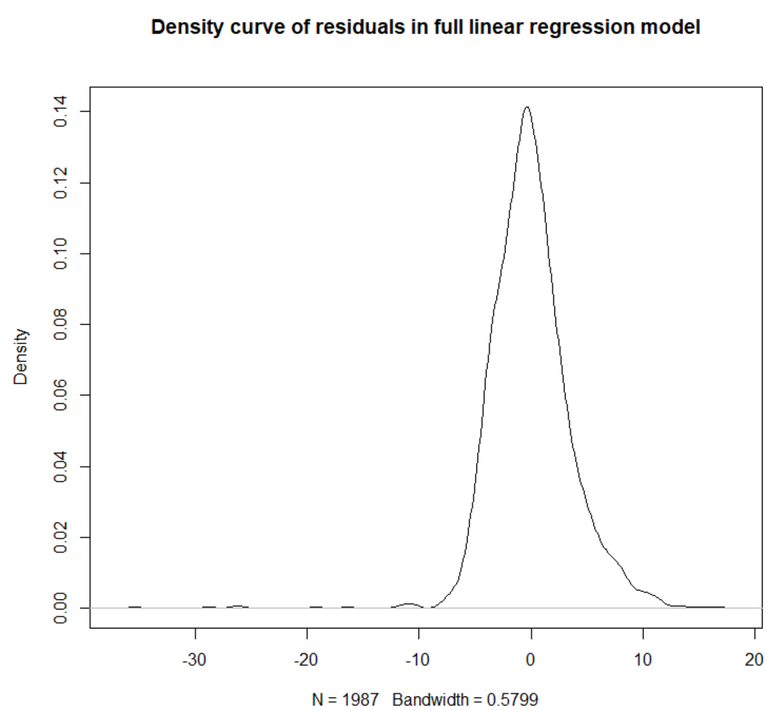
Of note, Average Rating has a higher estimate at level 4.5 (5.6) than it does at level 5 (4.59). Similarly, Visitor Usage at level 3 is higher than at Level 4 (3.36, vs 1.21), nor is level 4 significant even at the 10% level.



The residuals of the data are relatively closely clustered around the 0 line. This indicates a moderately good fit of the model. Some outliers, 10 outside the -10 mark, are present in the model.



The Q-Q Norm Plot shows the data to be closely in line with the theoretical quantiles, indicating that the assumption of normality is valid for this model and interpretations of the estimates and fit indexes are reliable. The same outliers are present, though there are very few in relation to the size of the overall dataset.

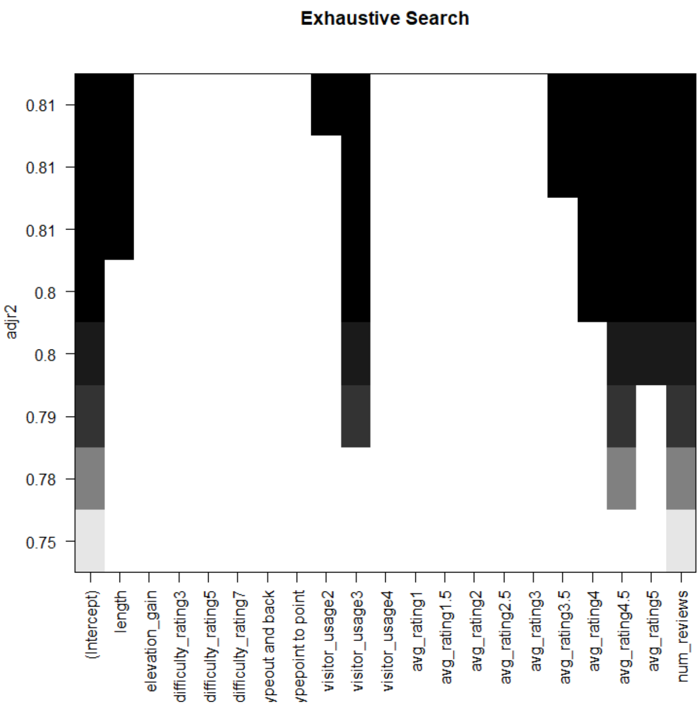


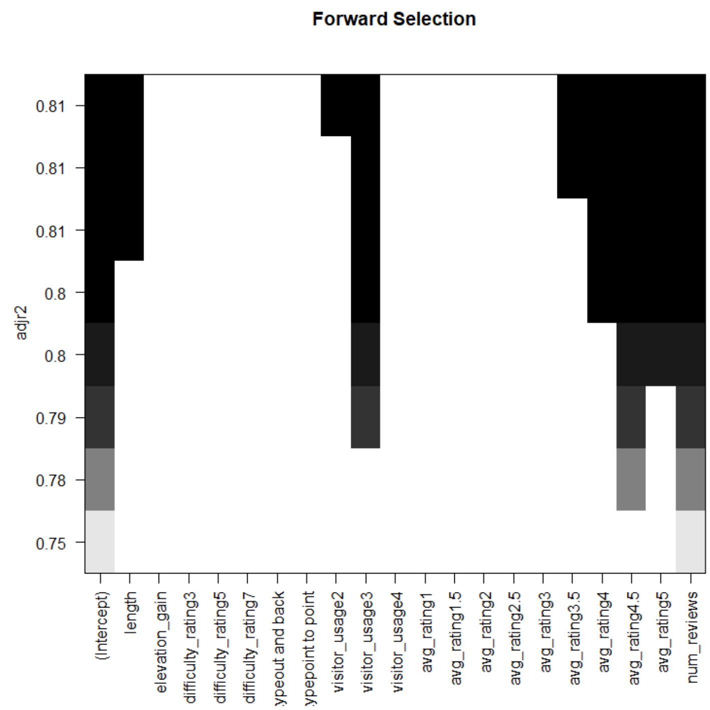
The Density Curve shows a rough bell curve shape that is evenly distributed around the zero point with few outliers. This further confirms the normality assumption of this dataset and model to be true. The outliers seen in the prior graphs do not have significant density in relation to the rest of the model, therefore their influence on the model can be assumed to be minimal.

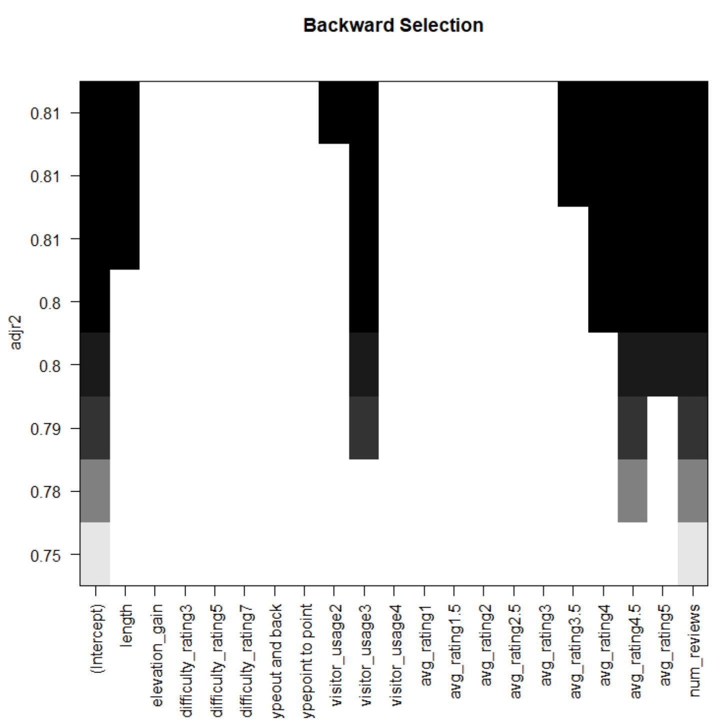
```
> accuracy(lmtrain.ic, train$popularity)
                           ME RMSE  MAE  MPE MAPE
Test set -0.000000000000632  3.6 2.58 -Inf  Inf
> accuracy(lmpred, test$popularity)
            ME RMSE  MAE  MPE MAPE
Test set 0.103 3.68 2.56 -Inf  Inf
```

To test for overfitting (that the model is not fit specifically to the training data and cannot be used for inference on other data), the model produced from the training data was applied to the validation dataset. While there is a large difference between the Mean Errors of the two models, the Mean Absolute Error differs by .02, and the RMSE by .08, both of which are <1% of the value of the measures. The model had similar accuracy between the training and validation datasets and can be considered ready for deployment.

As some variables were not significant, variable selection was performed in three methods to determine a model with greater parsimony: Forward, Backward, and Exhaustive Search. Selection was determined based on the Adjusted R^2 of the resulting models:







```
> sum$adjr2
[1] 0.753 0.779 0.793 0.798 0.805 0.807 0.808 0.809
> sum2$adjr2
[1] 0.753 0.779 0.793 0.798 0.805 0.807 0.808 0.809
> sum3$adjr2
[1] 0.753 0.779 0.793 0.798 0.805 0.807 0.808 0.809
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-35.41  -2.12  -0.26   1.78  17.00 

Coefficients:
                 Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)     1.64658888  0.48913166    3.37              0.00078 ***
length         -0.00001386  0.00000345   -4.02       0.000060949909 ***
visitor_usage2  0.69920311  0.18953588    3.69              0.00023 ***
visitor_usage3  3.44712503  0.29316791   11.76 < 0.0000000000000002 ***
visitor_usage4  1.35672484  0.81427650    1.67              0.09584 .  
avg_rating1    -0.21199775  1.87336257   -0.11              0.90991    
avg_rating1.5  -1.77231307  3.66155063   -0.48              0.62842    
avg_rating2     0.42484542  1.54887892    0.27              0.78389    
avg_rating2.5  -0.37920157  1.14610588   -0.33              0.74078    
avg_rating3     0.30775037  0.72597027    0.42              0.67167    
avg_rating3.5   1.77046074  0.54949344    3.22              0.00129 ** 
avg_rating4     3.25449018  0.48940180    6.65       0.000000000038 ***
avg_rating4.5   5.56173563  0.47417282   11.73 < 0.0000000000000002 ***
avg_rating5     4.42674569  0.49500507    8.94 < 0.0000000000000002 ***
num_reviews     0.03462767  0.00051982   66.61 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.63 on 1972 degrees of freedom
Multiple R-squared:  0.81,	Adjusted R-squared:  0.809 
F-statistic:  602 on 14 and 1972 DF,  p-value: <0.0000000000000002
```

All three methods resulted in the same model: Popularity (y) predicted by Length, Visitor Usage 2, Visitor Usage 3, Average Rating 3.5-5, and Number of Reviews (x's). The actual Adjusted R^2 values are printed in the same order, Exhaustive Search, Forward Selection, then Backward Selection:

The resulting model of all three selection methods retains all the same x variables as significant:

The R^2 of the new reduced model is equal to the full model at .81, and the Adjusted R^2 is .001 lower.

The models have very similar predictive power and the improvement via parsimony is minimal.

```
> accuracy(lmtrain2.ic, train$popularity)
                               ME RMSE  MAE  MPE MAPE
Test set -0.0000000000000644 3.62 2.58 -Inf  Inf
> accuracy(lmpred2, test$popularity)
             ME RMSE  MAE  MPE MAPE
Test set 0.0999  3.7 2.58 -Inf  Inf
```

When applied to the validation set, the same anomaly with the ME occurs, however the Mean Average Error between the two are equal and Root Mean Square Error are very close with the same difference as in the full model (.08 difference).

To see the effect of retaining more variables in the final model, stepwise regression in both directions based on model AIC was performed.

```
Start:  AIC=5133
popularity ~ length + elevation_gain + difficulty_rating + route_type +
    visitor_usage + avg_rating + num_reviews

                    Df Sum of Sq   RSS  AIC
- elevation_gain     1        25 25789 5133
<none>                           25764 5133
- difficulty_rating  3       112 25875 5136
- length             1        65 25829 5136
- route_type         2       109 25873 5138
- visitor_usage      3      1760 27523 5259
- avg_rating         9      5086 30850 5473
- num_reviews        1     57511 83274 7462

Step:  AIC=5133
popularity ~ length + difficulty_rating + route_type + visitor_usage +
    avg_rating + num_reviews

                    Df Sum of Sq   RSS  AIC
<none>                           25789 5133
+ elevation_gain     1        25 25764 5133
- difficulty_rating  3        89 25878 5134
- length             1        46 25835 5135
- route_type         2       123 25912 5139
- visitor_usage      3      1789 27578 5261
- avg_rating         9      5138 30927 5476
- num_reviews        1     57491 83280 7461
> summary(trails.df.step)

Call:
lm(formula = popularity ~ length + difficulty_rating + route_type +
    visitor_usage + avg_rating + num_reviews, data = train)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-35.42  -2.12  -0.25   1.77  16.69 

Coefficients:
                          Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)             1.74550368  0.52181803    3.35              0.00084 ***
length                 -0.00000728  0.00000389   -1.87              0.06171 .  
difficulty_rating3     -0.41636758  0.20651698   -2.02              0.04392 *  
difficulty_rating5     -0.50152896  0.26011837   -1.93              0.05399 .  
difficulty_rating7     -0.91542549  0.42294175   -2.16              0.03055 *  
route_typeout and back  0.26708892  0.17834977    1.50              0.13441    
route_typepoint to point -0.65615255 0.32222279  -2.04              0.04185 *  
visitor_usage2          0.70405798  0.18962995    3.71              0.00021 ***
visitor_usage3          3.38504725  0.29510135   11.47 < 0.0000000000000002 ***
visitor_usage4          1.24998593  0.81663073    1.53              0.12601    
avg_rating1            -0.30628864  1.86941536   -0.16              0.86987    
avg_rating1.5          -2.14506111  3.65535744   -0.59              0.55739    
avg_rating2             0.44963293  1.54662022    0.29              0.77130    
avg_rating2.5          -0.34451886  1.14601131   -0.30              0.76373    
avg_rating3             0.22607492  0.72539414    0.31              0.75533    
avg_rating3.5           1.70518368  0.55004767    3.10              0.00196 ** 
avg_rating4             3.24879704  0.48978506    6.63       0.000000000042 ***
avg_rating4.5           5.63276654  0.47353436   11.90 < 0.0000000000000002 ***
avg_rating5             4.63684499  0.49879552    9.30 < 0.0000000000000002 ***
num_reviews             0.03455072  0.00052176   66.22 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.62 on 1967 degrees of freedom
Multiple R-squared:  0.812,	Adjusted R-squared:  0.81 
F-statistic:  447 on 19 and 1967 DF,  p-value: <0.0000000000000002
```

The original full model had an AIC of 5133. The final model produced through stepwise regression has an AIC of 7461, which is an improvement over the original. It retains all variables except Elevation Gain, and has the same R^2 and Adjusted R^2 as the original full model.

```
> accuracy(trails.df.step.train, train$popularity)
                              ME RMSE  MAE  MPE MAPE
Test set -0.0000000000000637  3.6 2.58 -Inf  Inf
> accuracy(trails.df.step.pred, test$popularity)
               ME RMSE  MAE  MPE MAPE
Test set 0.103 3.68 2.56 -Inf  Inf
```

The RMSE and MAE have the same differences between the training and validation datasets as the original model (.08 in RMSE, .02 in MAE). This model attained through stepwise regression has very similar predictive power to the full model but slightly improved parsimony through the removal of an extraneous variable, Elevation Gain, and can be deployed in other analyses.

5)

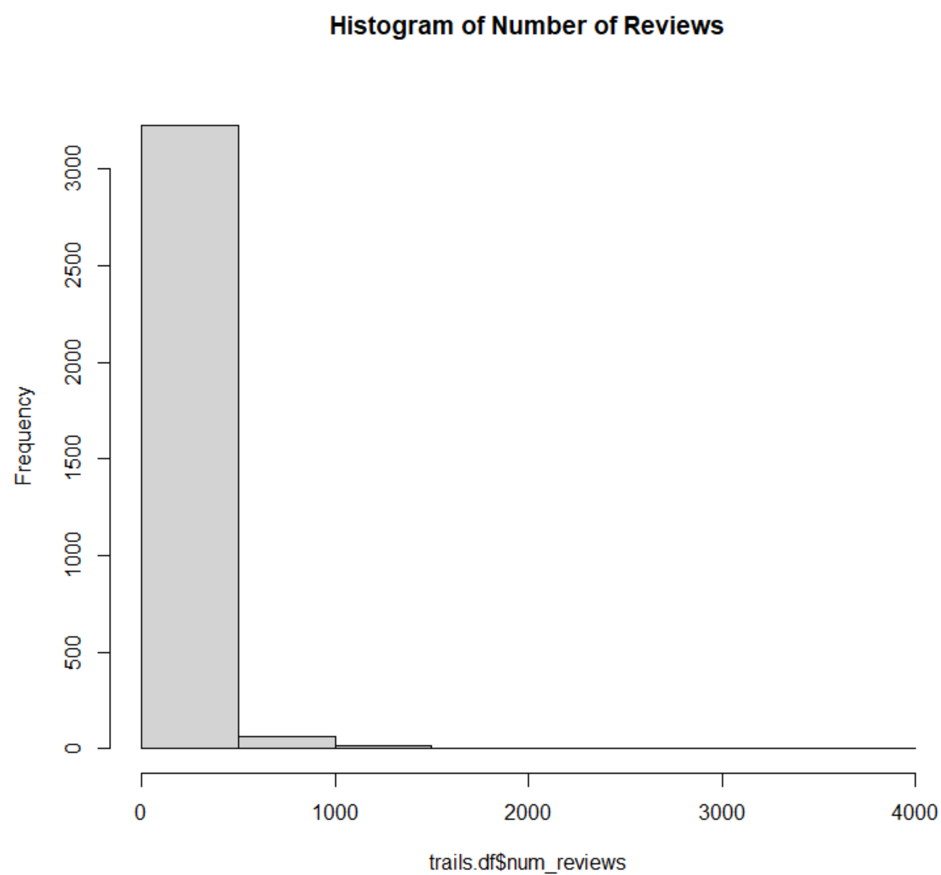
A severe limitation in the construction of this model is a lack of heterogeneity in how users rate trails. Out of the 3313 trails in the dataset, only 39 trails have a rating between 1 and 2.5. One would expect that a higher rating would have a positive coefficient estimate that would indicate a rise in popularity with a rise in average rating, and conversely a lower rating would have a negative coefficient estimate and indicate lower popularity for lower average reviews. Because average trail rating is the result of a behavior, rating a trail on AllTrails, there is clear selection bias in that users tend to only leave reviews for trails they rate highly (>2.5). If more trails were rated lower, they could become significant and produce a model with better explanatory power. More lower rated trails would also on their own be a great indicator to land and trail managers about which trails could require improvement or possibly even removal from visitor use.

The same effect of low sample size is present in Visitor Usage levels; only 39 trails have a visitor usage level of 4, less than 10% of the number of trails with a usage level of 2 or 3.

The insignificance of Elevation Gain as an indicator variable is surprising and contradicts prior research about peak experiences on trails. Trails that reach places of high elevation tend to have panoramic and excellent views of nature and far-off land and geography, which induces feelings of wonder, place, and euphoria in the trail users that see them. Based on this, one would have assumed that trails with more elevation gain would tend to be more popular, since presumably more trail users would seek out these peak experiences. In practice, there may be a cutoff point after which popularity falls off; most trail users are not exceptionally fit, experienced, or equipped to reach all the exceptional views trails with lots of elevation gain have to offer. Park infrastructure often brings these majority lower-experience trail users *to* the view they are seeking to enjoy, i.e. they never have to struggle through high elevation gain to reach them to begin with. The effect on hiker fitness or experience on trail preferences could possibly control for the lack of significance of Elevation Gain as a predictor for Trail Popularity.

Point to point route trails have a marked negative effect on the popularity of trails. Nearly 10% of trails in the dataset (320) are point to point trails. This may be the effect of behavior rather than any other variable; point to point trails may be a part of longer trekking or backpacking routes and require higher experience level and more equipment that many trail users do not have.

Number of Reviews may be particularly effected by outliers. The vast majority of trails have either zero or very few reviews:



The extreme outliers (the highest number of reviews is 3903) drive the coefficient estimate for number of reviews down. The face validity of the significance of the coefficient makes sense: trails that receive an exceptional number of reviews are serving a large number of trail users, and are leaving an impression that is likely positive.