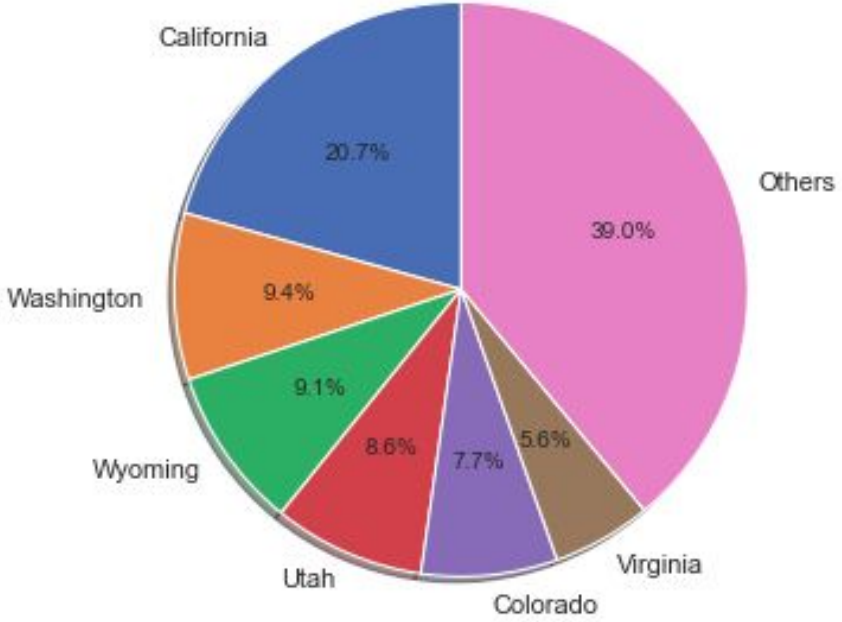
Regression Model for National Park Trails

Jeffrey Ng

Exploratory Data Analysis + Linear Regression / Modeling

https://www.kaggle.com/planejane/national-park-trails

EDA
Breakdown of National Parks by State



California 20.7%

Washington 9.4%

Wyoming 9.1%

Utah 8.6%

Colorado 7.7%

Virginia 5.6%

Other 39%

EDA Key Questions to Consider

What are the characteristics of a poor trail? **One that doesn't get visited**

Do hikers prefer certain route types? **Out n back**, **Loop**, Point to Point

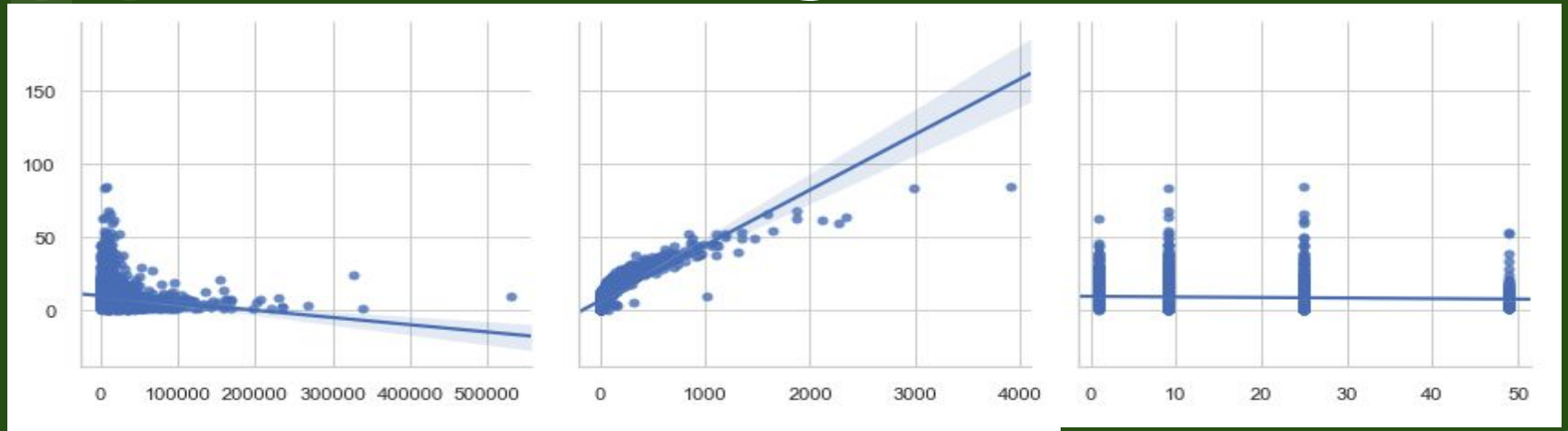
Do the length of trails vary from great National Parks, average National Parks, and Low Popularity parks? **No (ANOVA, α=.10, pvalue=.79)**

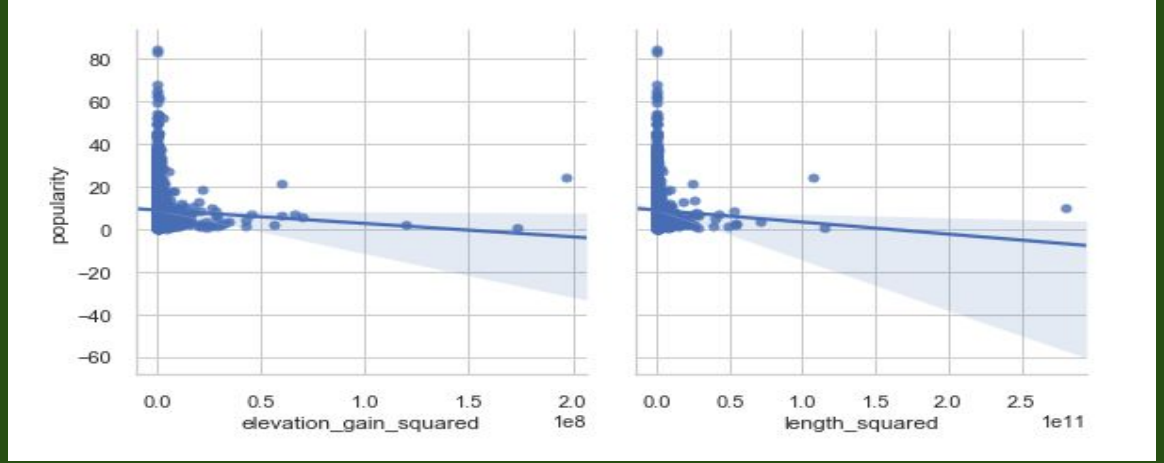
Is there a difference in trail level of difficulty between great, average and obscure National Parks? **Yes, highly rated parks have higher difficulty (ANOVA, α=.10, p value=.015)**

Does elevation gain play a role in popularizing a trail? **Yes (ANOVA, α=.10, pvalue=.096) and Somewhat No**

Building & Testing My Linear Regression

Model / Linear Regression



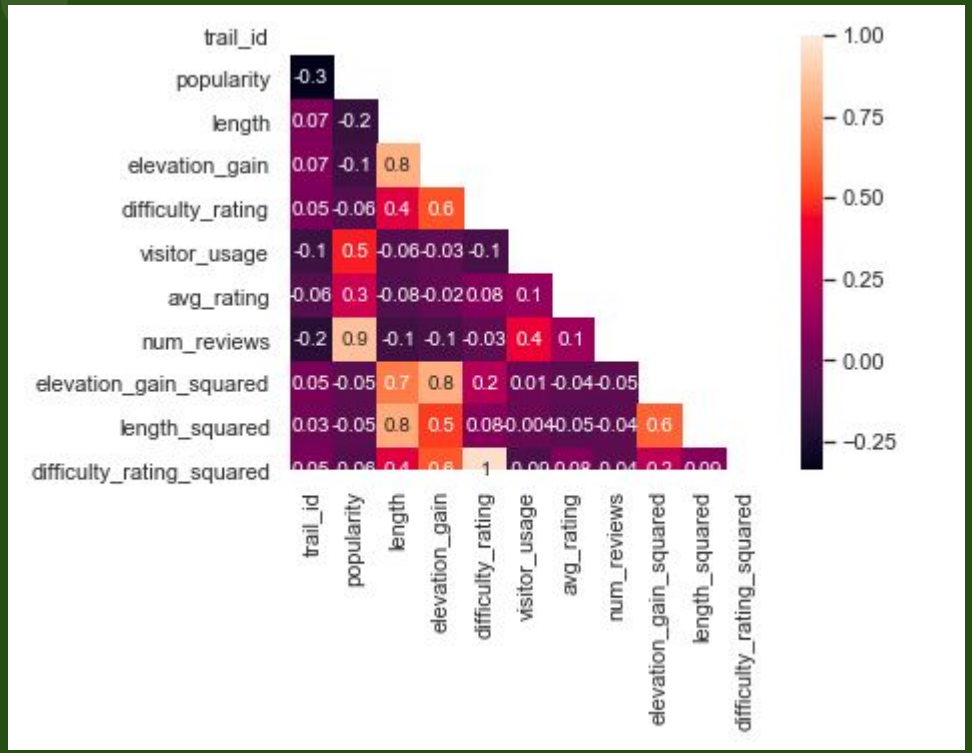


Dependent variable:

Popularity/Usage

1. Length
2. No. of reviews
3. Difficulty Rating^2
4. Elevation gained ^{2}
5. Length^2

Correlation Heatmap



R^2= **.743**

Dependent variable:
Popularity/Usage

Independent variables:

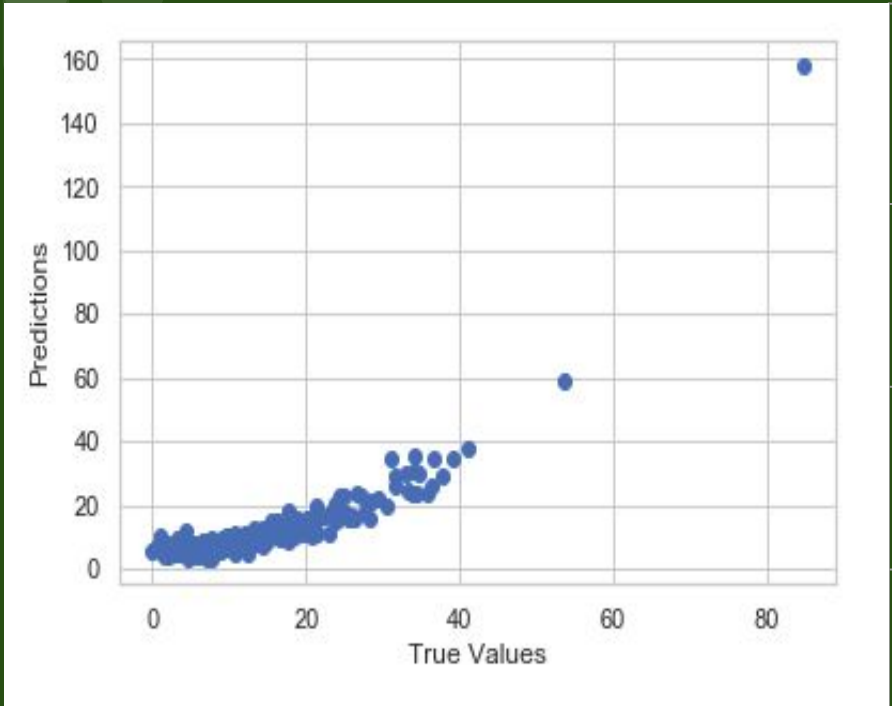
- Length ✔
- Elevation gain
- Difficulty rating
- Number of reviews ✔
- Elevation gain squared ✔
- Length squared ✔
- Difficulty rating squared ✔

Feature Selection

All Features
Length
Num_reviews
Difficulty Rating
Elevation Gained
Difficulty Rating^2
Elevation gained^2
Length^2

My Features	RFECV features
Length	Length
Num_reviews	Num_reviews
Difficulty Rating^2	Difficulty Rating ^2
Elevation gained ^2	Elevation gained
Length^2	Difficulty rating

Model Evaluation



	Train (5 feat)	Test (5 feat)	Train (RFECV)	Test (RFECV) *
MAE	2.98	2.97	2.97	2.99
MSE	16.06	21.11	16.07	21.11
RMSE	4.01	4.59	4.00	4.59

*Ridge & Lasso had similar no difference

Discussion

- Can this model be used to assess the usage rate of trails in smaller parks or camping areas?
- The model worked well but didn’t predict hugely popular busy trails.
- Slight overfitting (Train vs Test Data)
- Didn’t take in consideration certain natural landmarks such as Grand Canyon, Grand Teton, Mt. Ranier, etc.