Written Report

STA 210 - Project

Yay-Stats - Ali Raich, Manny Mokel, Yifu Wang, Chenyu Wang

Introduction

We are investigating how different amenities affect the overall quality of parks. We are motivated to explore this because of the surge in popularity in parks during the COVID-19 pandemic. Parks are vital for maintaining good mental and physical health, especially during a pandemic. Access to safe outdoor areas and green space can reduce stress, anxiety, depression, obesity, and mortality. We are interested in exploring the disparities in parks based on geographic areas, high/low income areas, and neighborhoods mostly of color or white. Based on the Trust for Public Land, 100 million people do not have access to a park within 10 minutes walking distance of their home. The Park Score index rating system measures how well 100 US cities are meeting the increasing need for easy access to parks. Parks individually monitor their metrics for the ParkScore index and report them back. Though The Trust for Public Land has been collecting data since 2011, we will focus only on data from the year 2020. However, not every city has the technology and resources to collect this information so equity is added to the ParkScore evaluation. The data uses "points" to help normalize values. Thus, the higher points the better. The key variables that we will be using are basketball_points, dogpark_points, playground_points, restroom_points, splashground_points, park_benches and total_points. The definitions of the key variables are as follows:

`basketball_data`: basketball hoops per 10,000 resident

`dogpark_points`: dog parks per 100,000 residents points

`playground_data`: playgrounds per 10,000 residents

`restroom_points`: restrooms per 10,000 residents points

`splashground_points`: splashgrounds and splashpads per 100,000 residents point

`park_benches`: number of park benches

`total_points`: total points

`capital`: mutated variable that says yes if given city is a state capital and says no if it is not.

`Total points` is the total of points from amenities, investment, access and acreage. In this model we will just be looking at amenities.

Please note that in the next section will make `basketball_data` and `playground_data` into categorical variables, `basketball_points` and `playground_points`. The cutoff values for basketball and playground are both 4.

Additionally, we use the built-in R dataset `cities`, from the `maps` package to add to our data. By performing appropriate transformations, and utilizing the left join function in dplyr, we are able to identify whether or not a city is a state capital, and what impact it may have on the total score.

Source: https://www.tpl.org/parks-and-an-equitable-recovery-parkscore-report

The general research question that we are going to explore is:

How do different amenities of a city's park environment (`basketball_points`, `dogpark_points`, `playground_points`, `restroom_points`, `splashground_points`, `park_benches`) affect the overall quality (`total_points`) of its parks?

Our Hypothesis that we will be exploring are:

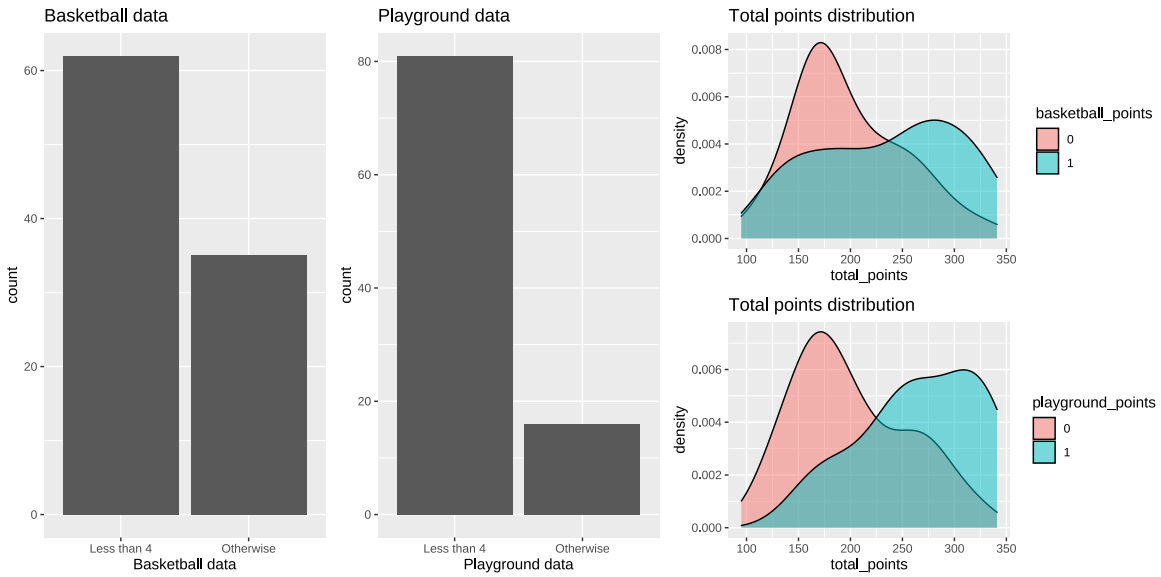
HO: The given variables (basketball hoops, dog parks, playgrounds, restrooms, splashgrounds, capital and benches) have little to no effect on the overall quality of a park.

HA: The given variables (basketball hoops, dog parks, playgrounds, restrooms, splashgrounds, capital and benches) have a substantial impact on the overall quality of a park.

Data Description & Exploratory Data Analysis

As mentioned in the introduction, the predictor variables that we are going to use are `basketball_data`, `dogpark_points`, `playground_data`, `restroom_points`, `splashground_points`, `park_benches`, which will be used to predict `total_points`. Since all the predictor are numerical and we want to add some categorical variables into our model, we decided to convert the numerical variables `basketball_data` and `playground_data` into categorical variables `basketball_points` and `playground_points`.

The cutoff values we used for basketball and playground are both 4. A observation whose `basketball_data` / `playground_data` is less than 4 will get a `basketball_points` /`playground_points` equal to 0, otherwise it is equal to 1. We chose the value of 4 as a cutoff because we felt it was somewhat of a "middle ground" for parks having a low or high number of basketball courts and playgrounds. By the plots below, we can observe that the total points distribution of corresponding two categories (0/1) of `basketball_points` /`playground_points` are highly different, so `basketball_points` and `playground_points` appear to have a significant effect on `total_points`, which also indicates that our converting is meaningful.

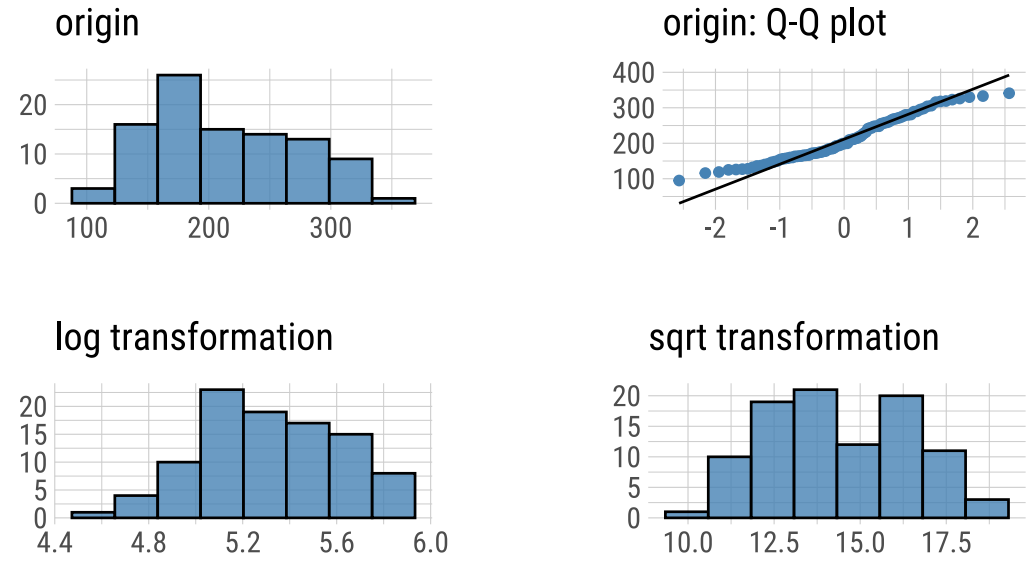


Next, we generate a mutated variable that says yes if given city is a state capital and says no if it is not (using the`maps` package).

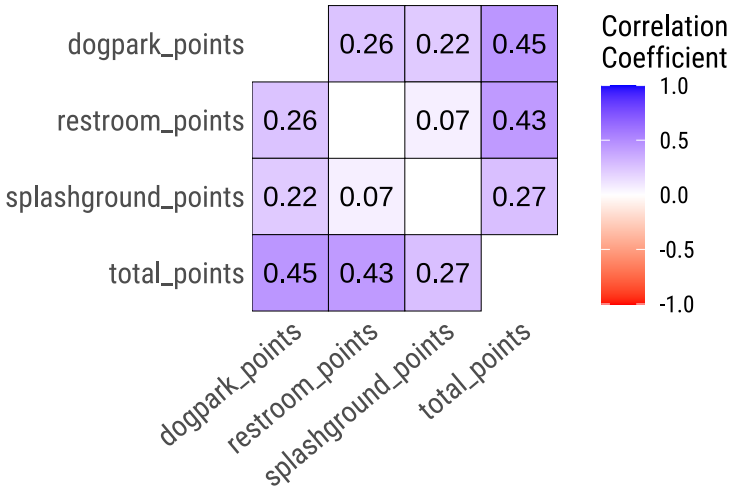
Then we moved to the numerical variables, `dogpark_points`, `restroom_points`, `splashground_points`, `park_benches`, `total_points`. Since there are too many NAN in `park_benches`, we decided to get rid of it. We will discuss this further in our conclusion section. First, we wanted to know the mean, sd, IQR, and other information of each of them.

mean	sd	se_mean	IQR	skewness	kurtosis
46.258	32.935	3.344	57	0.386	-1.143
53.495	30.249	3.071	51	0.162	-1.218
51.237	35.826	3.638	74	0.189	-1.385
211.866	60.946	6.188	95	0.360	-0.878

Then we visualized the distributions of each of them by Normality Diagnosis Plot, which includes origin distribution, log transformation distribution, sqrt transformation and Q-Q plot. It is worth mentioning that in Q-Q plot, the closer curve and straight line are, the closer the data distribution and normal distribution are, which implies that the more randomly our data is picked. Below is the Normality Diagnosis Plot of our response variable, and plots of numerical predictor variables are attached in the appendix.



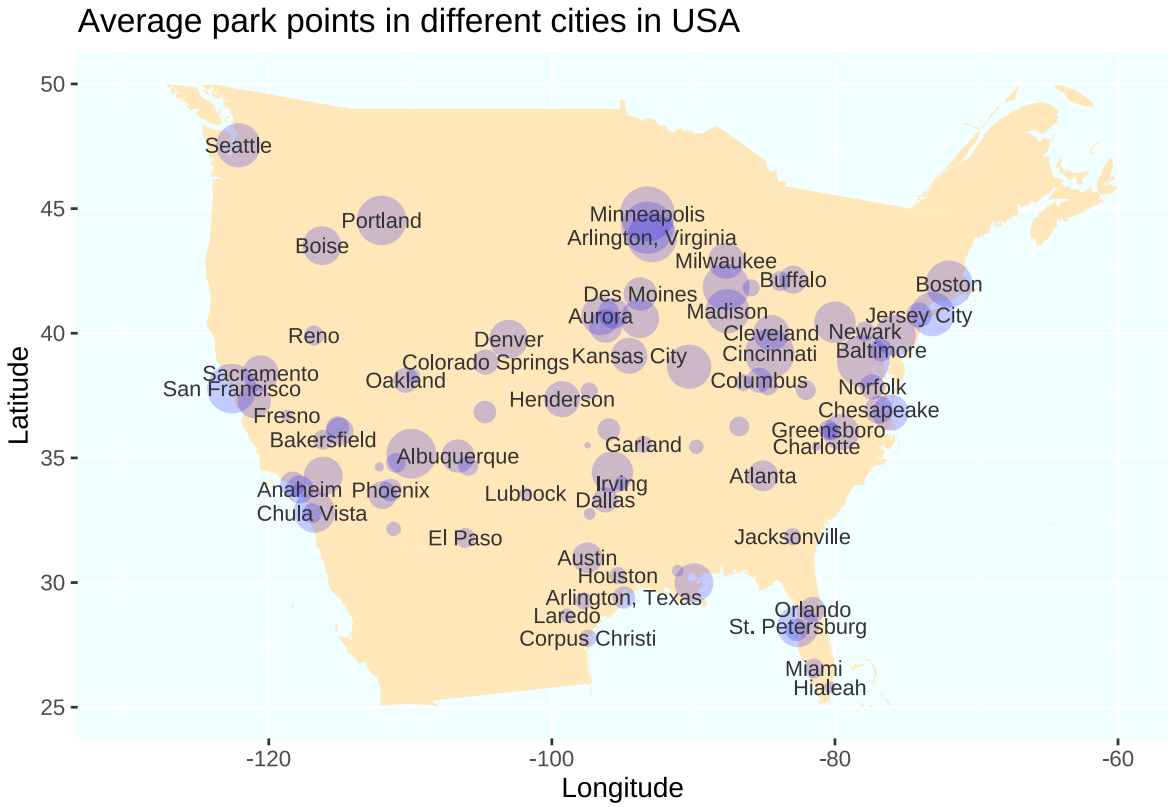
We are also interested in the correlations between each numerical variables. Below is the correlation matrix that shows correlation between coefficients.



The highest correlation between any two numerical variables is 0.45, and it is clear we are not

experiencing serious collinearity.

Finally, we are curious about whether the response variable `total_points` are distributed independently in different regions of US. Since the `total_points` is for each park, which is too little to show on the US map, we chose to show the `total_points` of each city, which is the average `total_points` of each park in this city.



By the figure, we can say that `total_points` are distributed independently in different regions of US.

Method

We will use the Multiple linear regression model to conduct this project. Since our goal is to determine whether the selected features affect the overall quality of the park, by conducting a regression, we can determine each feature's impact on the park's quality by simply checking the coefficient of corresponding feature, and compare them with each other we would also have some understanding about which feature contributes the most to one park's quality.

In order to get the final optimal model, we would first start with some basic features of the park and use it as the baseline model. Then we would try to add some extra terms. When

we add the term every time, we will use measures of AIC, BIC, and adjusted r-squared to test whether the new added term is significant. We would not include the interaction term in this model as the final model only contains numerical variables, and the reasons for this would be elaborated later.

Note that we transform two variables about the basketball and playground into categorical variables. Reasons for this change is that we believe once a park has adequately enough place for people to do some physical exercise, the specific number of courts / playgrounds does not matter. We also created one new variable called capital, which identifies whether the park is in the capital. The reason for this is the parks in capital may have a better quality than other parks in the country due to investment and increased focus.

We start out by utilizing 5-fold cross validation for our model, by first specifying a recipe and a workflow. The performance of our model on each fold is shown below:

Fold	RMSE	R-squared
Fold1	45.970	0.494
Fold2	38.119	0.494
Fold3	47.089	0.461
Fold4	44.816	0.571
Fold5	57.311	0.138

We can observe from the cross validation that our model performs somewhat poorly. The mean R-squared value across each fold is 0.432, with a standard error of 0.0755.

The output of the model specified in our workflow is shown below:

term	estimate	std.error	statistic	p.value
(Intercept)	130.310	11.509	11.323	0.000
dogpark_points	0.600	0.141	4.261	0.000
restroom_points	0.528	0.152	3.468	0.001
splashground_points	0.243	0.125	1.949	0.054
capitalYes	9.221	12.040	0.766	0.446
basketball_points1	12.341	9.286	1.329	0.187
playground_points1	48.885	11.828	4.133	0.000

Using the `glance()` function, we can see information about the performance our model. Specifically, we look at ANOVA and P-values for each of our coefficients, as well as the R-squared, and adjusted R-squared values.

r.squared	adj.r.squared	p.value	AIC	BIC
0.469	0.437	0	1127.087	1148.47

The P-values for coefficients suggest that whether or not the city is a capital and basketball_points may not be necessary in the model. The P-value for splashground_points is approximately 0.05, so we will keep this variable in our model. The difference between the R-squared and Adj R-squared is 0.32, suggesting we have non-important variables in our model. Let us fit another model without these variables and see how our performance changes.

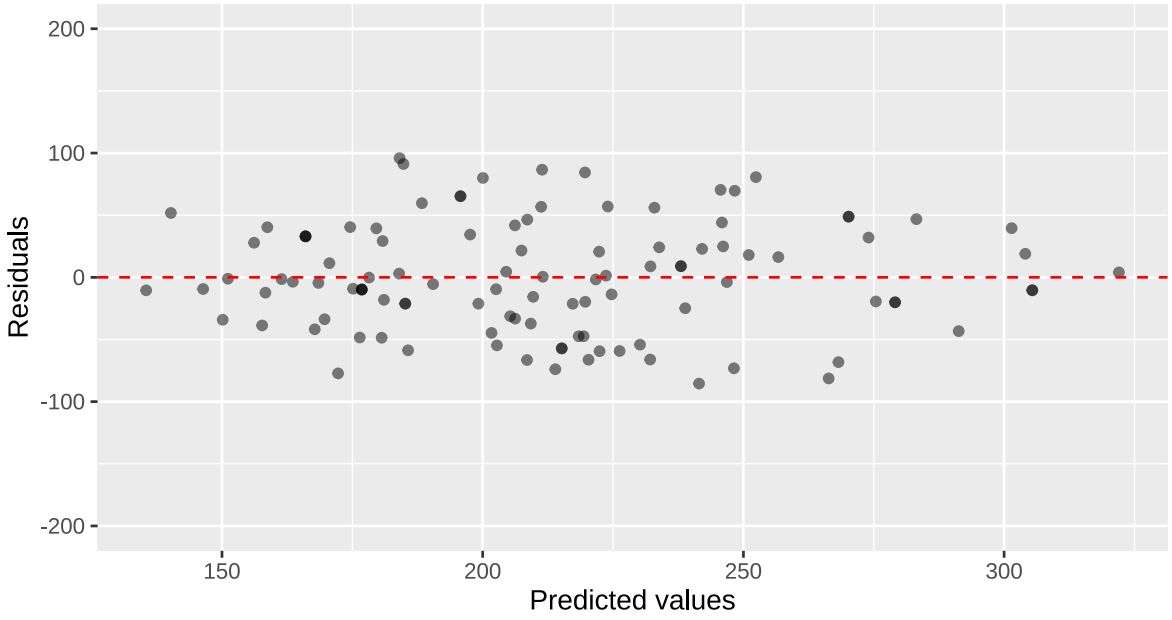
term	estimate	std.error	statistic	p.value
(Intercept)	132.297	11.088	11.931	0.000
dogpark_points	0.628	0.140	4.491	0.000
restroom_points	0.554	0.151	3.677	0.000
playground_points1	53.001	11.552	4.588	0.000
splashground_points	0.249	0.125	1.996	0.049

r.squared	adj.r.squared	p.value	AIC	BIC
0.465	0.444	0	1135.797	1151.889

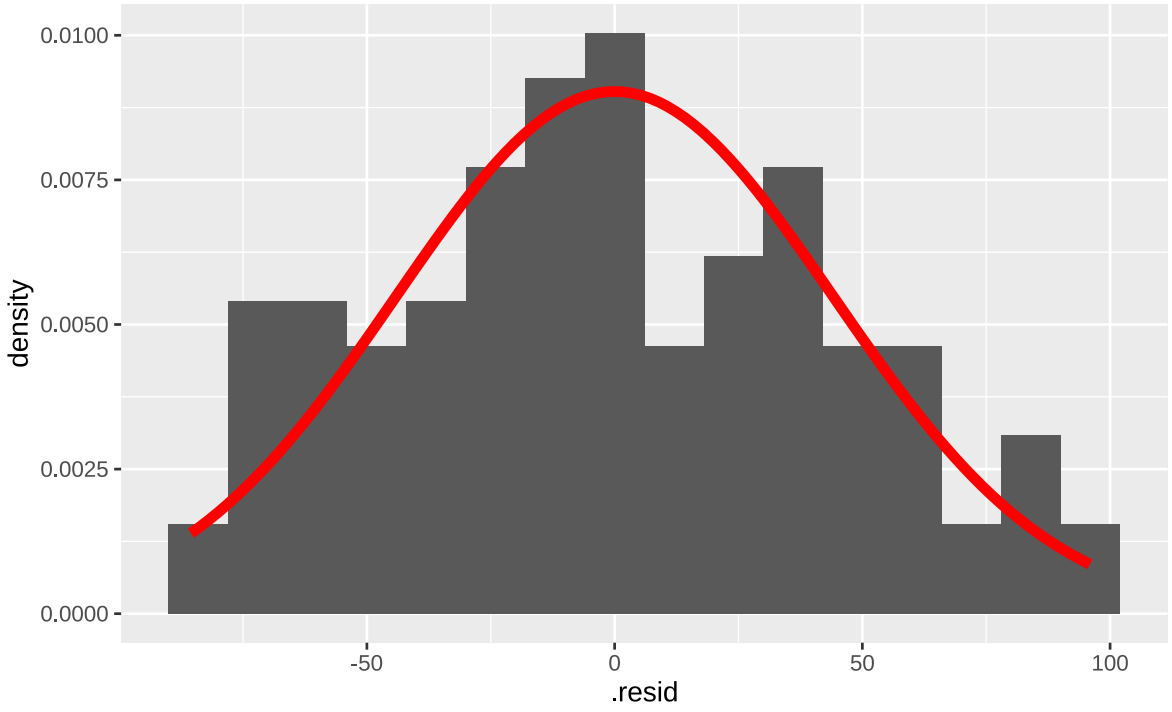
The P-values on this new model suggest each of the predictors are significant, and we also notice that our adjusted R-sqaured has gone up from 0.437 to 0.444, while R-squared, AIC, and BIC remain about the same.

Since the model we finally select only include numerical values, and considering the following two facts: First, model that contains interaction between two numerical values is hard to interpret. Second, the in the EDA section we have shown that the correlation between the numerical value are relatively small. Hence, considering those two points, it is reasonably not to include interaction term in our final model.

Finally, we want to check the conditions for inference, linearity, constant variables, normality, and independence.



We can see there is no pattern in the plot of residuals vs. predicted values, and the vertical spread of the residuals is constant across the plot, so linearity and constant variables are satisfied.



By the plot, we can see the residuals are not exactly normal, but general distribution is relatively similar to the normal distribution.

Regarding the independence, as mentioned earlier below the plot titled "Average park points in different cities in USA", independence is satisfied.

Results

Our final model is displayed below:

term	estimate	std.error	statistic	p.value
(Intercept)	132.297	11.088	11.931	0.000
dogpark_points	0.628	0.140	4.491	0.000
restroom_points	0.554	0.151	3.677	0.000
playground_points1	53.001	11.552	4.588	0.000
splashground_points	0.249	0.125	1.996	0.049

$$total\_points = 132.297 + 0.628 dogpark\_points + 0.554 restroom\_points$$
$$+53.001 playground\_points + 0.249 splashground\_points$$

Conclusion

Overall, while we started with 6 variables in our model, some of which were mutated above, our final model reports the same performance with only 4 variables in the model. While we can include information about all the amenities in our model, the R-squared values appear to only increase so much. The meaning of this may be that the amenities in a park do not do a great job predicting its total score. Our R-squared value tells us that approximately 46.5% in the variance in total park score can be explained by its dogpark, restroom, and splashground scores, as well as whether or not the park has a sufficient level of playground space. While the presence of these amenities may contribute to the quality of a park, they do not determine a park's quality. In regards to our research question, these amenities do contribute to the overall quality of the park, but without further information such as amount of investment, access, and acreage, we cannot properly predict the quality (total_points) of the park.

An issue pertaining to the reliability of our data is that each city self reports its own data to the Trust for Public Land. Not every city has the same method or technology to report this type of data which can lead to skewed results. Additionally, self-reporting can lead to some missing data which further skews results. In the future, as more cities are able to collect /

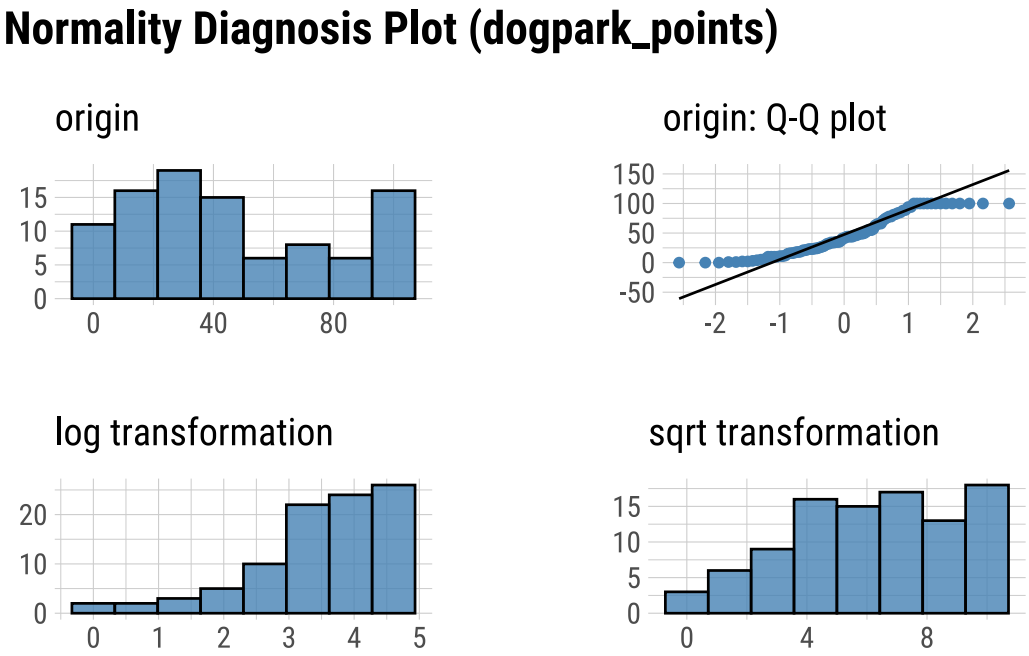
report data, we could have a more valid list, and have a more comprehensive view of park quality around the US.

Further directions we wish to take in this research would be looking into city demographics and how they impact the quality of the park. Identifying demographics which have a lack of access to quality parks is important problem, and in doing so, could improve the general livelihoods of many people if investment and focus into particular areas with lack of quality parks is carried out.

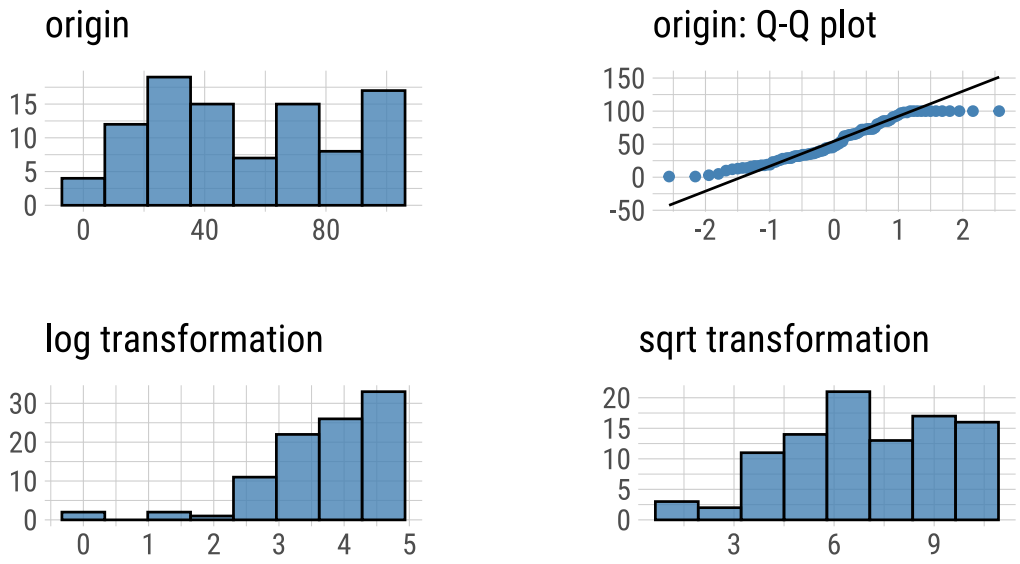
The data dictionary can be found here.

Appendix

The Normal Diagnosis Plots of numerical predictor variables (`dogpark_points`, `restroom_points`, `splashground_points`) are below.



Normality Diagnosis Plot (restroom_points)



Normality Diagnosis Plot (splashground_points)