Relationships Between Health Outcomes and State Park Amenities in Counties in New York State: 2020-2021 Data

By: Anne Miller

Abstract

The goal of the project was to analyze relationships between health outcomes and the availability of amenities at state parks in the state of New York. The analysis was conducted at the county level to see if health outcomes in counties where the park amenities were present differed from health outcomes in counties where the park amenities were not present. The project was divided into the two following questions, and both questions were analyzed using R:

Question 1: Do mean health outcomes change based on the state park amenities available and, if so, what is the magnitude and direction of the relationship(s)?

Question 2: Does that relationship vary between groups of counties with different ranges of health outcomes and, if so, what is the magnitude and direction of the relationship(s) within each group?

Data

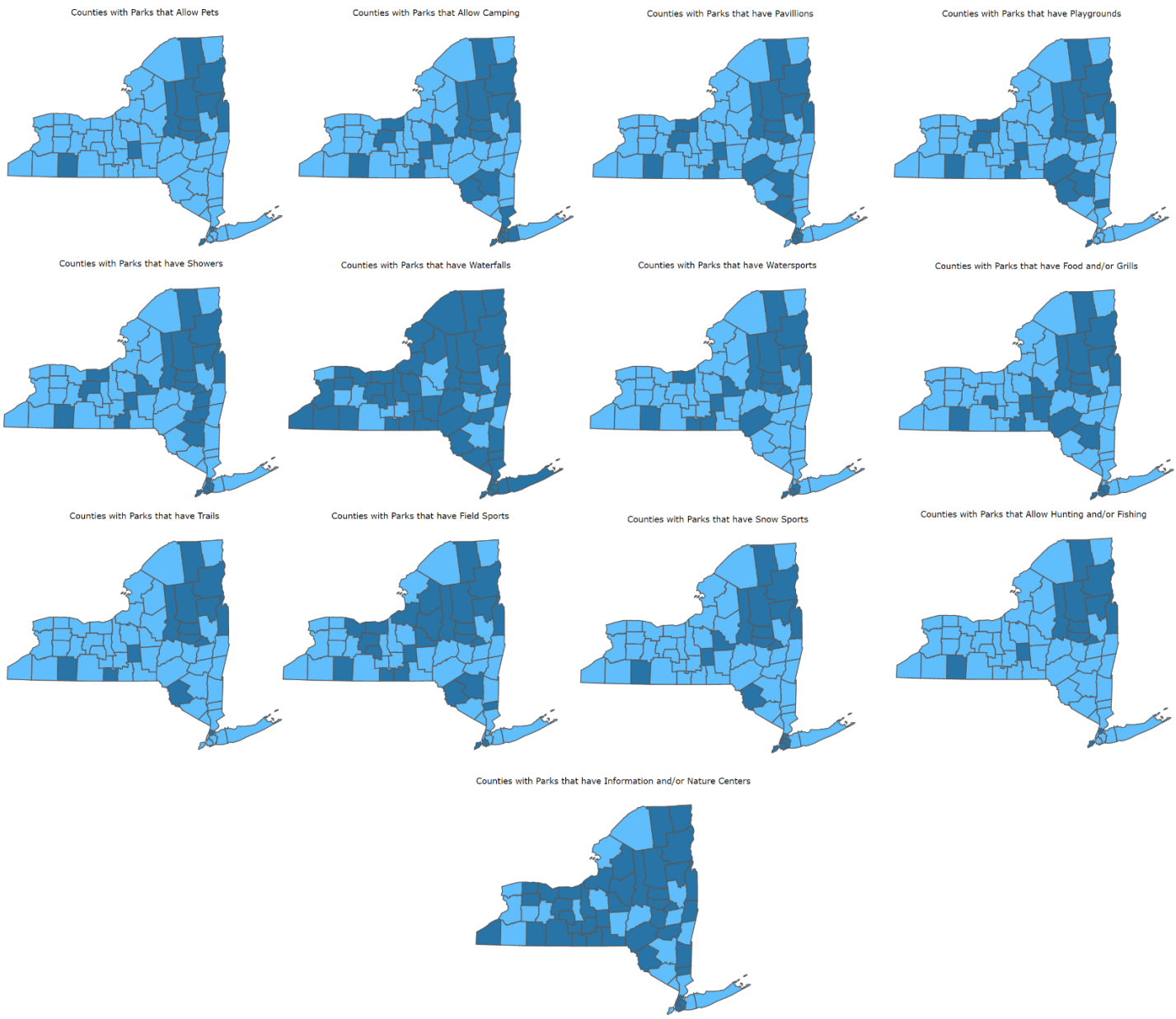
All data was collected from either www.ny.gov or www.census.gov. Park amenity data was collected for almost 200 state parks from https://parks.ny.gov/. This data was last updated at various dates within the years 2020 and 2021. The data included availability of 27 different park amenities, including nature centers, field sports, hiking trails, playgrounds, swimming pools, etc. The dataset used in the analysis included 62 observations, one for each county in New York State. The park amenity data was binary; if there was at least one state park within that county with a specific park amenity, that county received a value of 1 for that amenity. Likewise, if there were no state parks within that county with the amenity, the county received a value of 0 for that amenity.

After collecting the park amenity data, the dataset for the analysis included observations with 27 predictors each--nearly half as many predictors as observations. A high ratio of predictors to observations could lead to overfitting, which occurs when the analysis is fit too closely to the data on which it was conducted and therefore cannot be accurately applied to other data. Too many

predictors could also lead to issues with multicollinearity, meaning some of the predictors are highly correlated to each other. Multicollinearity in the predictors could lead to high variance in the results, which could result in an inability to identify which predictors are responsible for the changes in the responses.

To reduce this ratio, the predictors were grouped into broader parent categories. For example, the following park amenities were aggregated into one predictor titled *water sports*: boat rentals, boat launch, canoeing/kayaking, stand-up paddleboarding, surfing/windsurfing, swimming pools, and swimming beaches. If any state park within a given county had any one of these amenities, the county received a 1 for *water sports*. Likewise, if no state parks within that county contained one of these amenities, the county received a 0 for *water sports.* This reduction resulted in 13 predictors for 62 observations, a much more suitable ratio for analysis.

The maps below show the availability of the state park amenity variables by county. Counties colored in dark blue do not have any state parks with the amenity shown in the title, and counties colored light blue have at least one state park with that amenity.

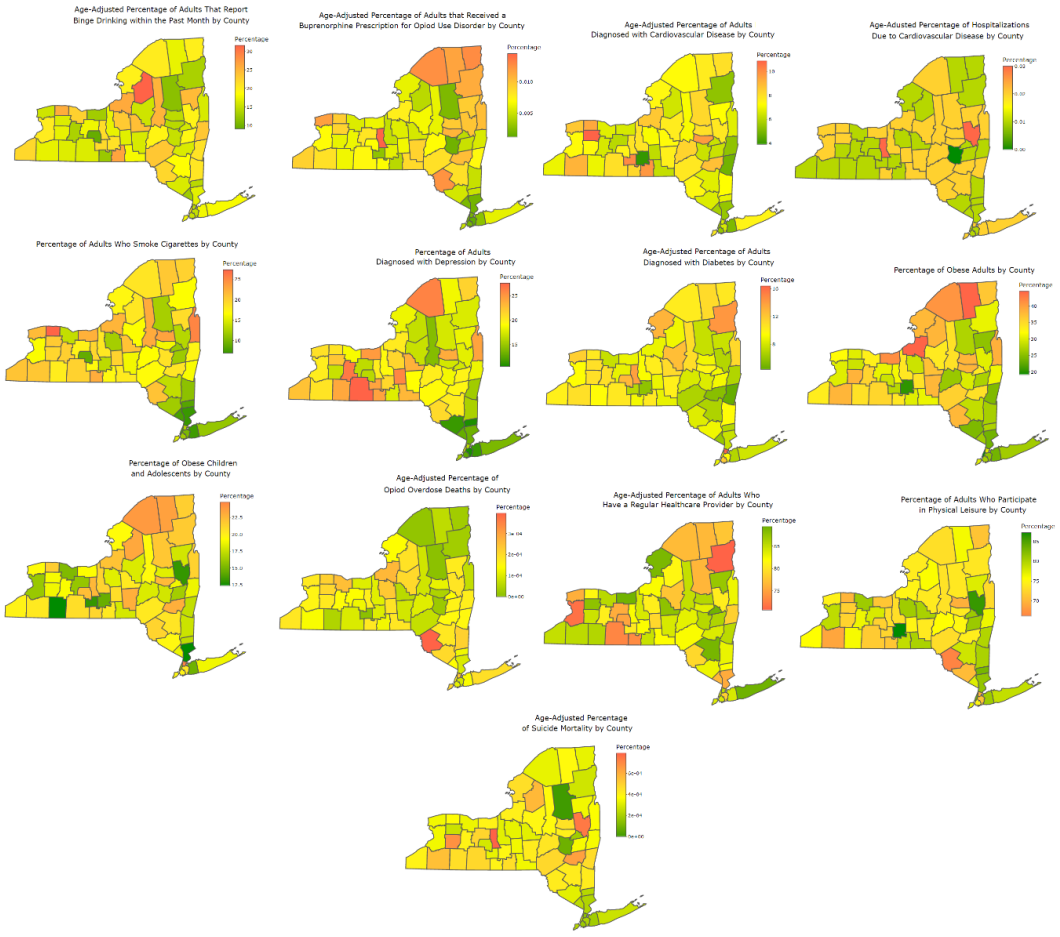


Data for the response variables was collected from www.heath.data.ny.gov and was last updated at various dates between 2020 and 2021. The majority of the response variables were collected from the *Prevention Agenda 2019-2024 Tracking Indicators* data, which consists of two datasets maintained by the New York State Department of Health (DOH) that together contain almost 100 health indicators for all counties in New York State. This data is used by both state and county organizations to track and improve health indicators over time. The most recent data in this

dataset was updated in February 2022, but the data used was from the years 2020 to 2021 so that the predictors and responses were collected during the same timeframe.

A few of the response variables were collected from the *Community Health Indicator Reports (CHIRS)*, also available at www.health.data.ny.gov and maintained by the DOH. This data contains over 300 health indicators for each of the 62 counties in New York. It was last updated on August 19, 2020.

Finally, one of the response variables was collected from the *Behavioral Risk Factor Surveillance System (BRFSS)*, which is again available via www.health.data.ny.gov and is maintained by the DOH. This dataset contains survey data for adults in New York State. The most recent data in this dataset was updated on February 9, 2023, but the data that was used for the analysis was collected in 2020 so that all the predictors and responses were obtained during the same time period. The health outcomes variables used in the analysis are shown in the maps below. The color of the county indicates the value of the variable.



From the three datasets, 13 total response variables were collected. The response variables included health outcomes such as percentage of obese adults and children, prevalence of cardiovascular disease diagnoses and depression diagnoses, and measures of substance abuse related outcomes such as ratios of opioid overdoses and the number of adults who report binge drinking within the last month.

The health data used was collected on various scales ranging from a raw count, to percentages, to a per capita value. Having multiple scales of data in one analysis can lead to certain variables with large values being regarded by the algorithms as more important than variables with smaller values when in actuality, they are just on a larger scale than the other data. The majority of the data was represented as a percentage, so the remaining health variables were transformed to a percentage as well. To do this, the population of each county was required. Population data was called from *County Population Totals: 2020-2022*, located at www.census.gov. The population total estimate used was for 2020, which is in line with the dates when the park amenity and health outcome data were collected.

In addition to the health outcomes variables collected from the DOH, a new response variable was created for the analysis. This variable was titled *overall health rating*, and it was created based on the other response variables to act as an overall indicator of health outcomes in each county compared to the health outcomes in other counties.

All response variables for each county were standardized. Responses in which a higher value indicated worse health outcomes (such as percentage of adults who smoke cigarettes) were given a negative sign, and responses where a higher value indicated better health outcomes (such as adults with a regular healthcare provider) were given a positive sign. The new standardized scores for each health outcome were summed for each observation, resulting in an *overall health rating* that represented how each observation compared to other observations within the dataset.

Question 1 Methodology

Part 1—MANOVA of All Health Outcomes

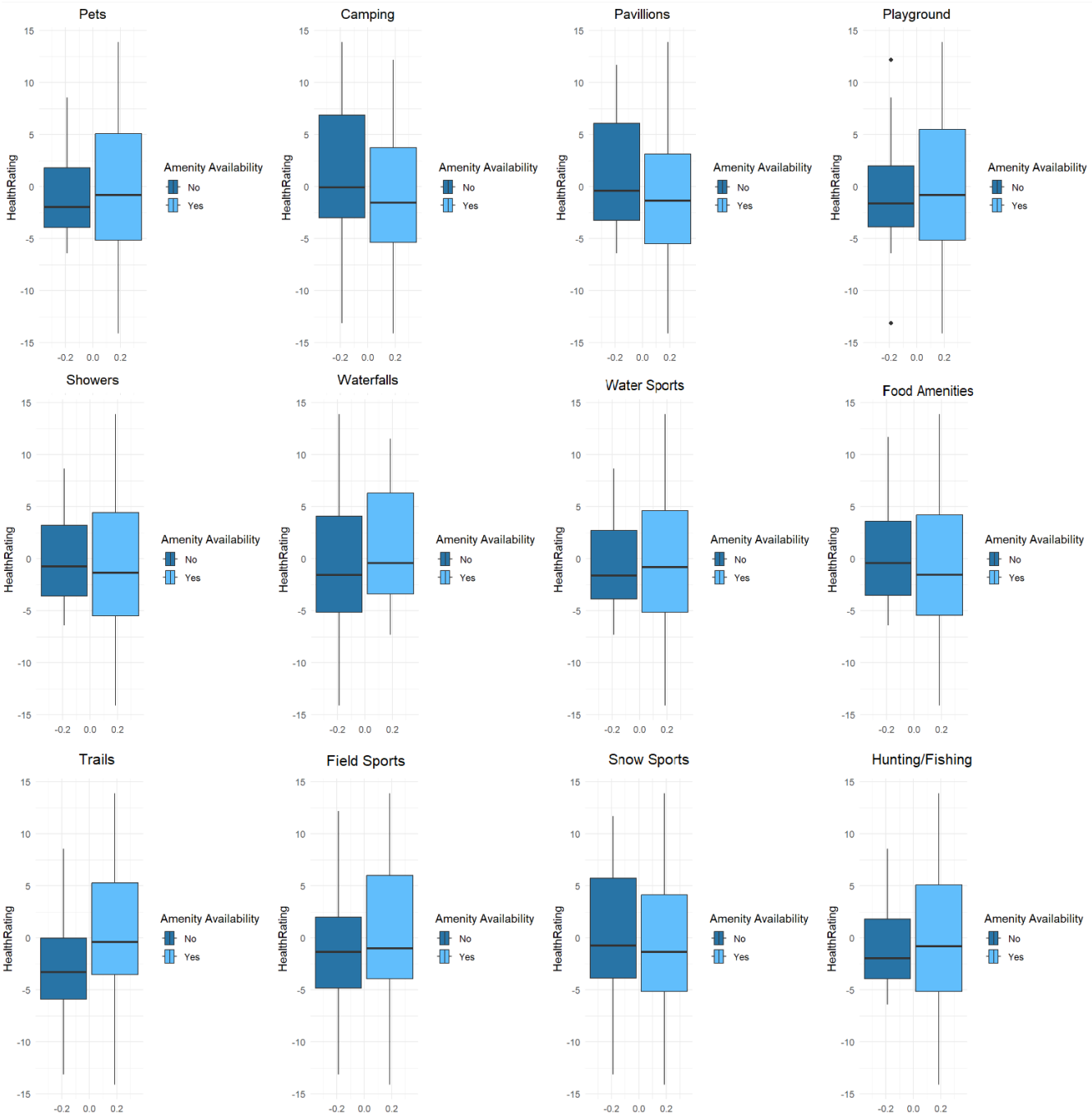
The first goal of the project was to compare mean health outcomes between counties that have each state park amenity and those that do not. This was done in two ways. First, all health outcomes were used as the response variables. Second, only the newly created *overall health rating* was used as the single response variable. Part 1 of this paper focuses on the first part of this analysis. Because there were 13 health outcomes variables to be assessed, a multivariate analysis of variance (MANOVA) test was the most appropriate approach.

MANOVA is used to test for differences in multiple continuous response variables between observations in different groups. It accounts for potential Type 1 errors that could arise from conducting multiple tests for each health outcome individually and then aggregating the results. The continuous response variables used in the MANOVA were all health outcomes other than overall health rating (as overall health rating was created based on the other health outcomes, therefore its inclusion in the analysis would be redundant), and the two groups were counties with each park amenity and counties without that amenity.

MANOVA was used to compare the average health outcomes in counties with the park amenities versus the average health outcomes in counties without the park. A difference in variance between the groups could indicate a possible association between the health outcomes and the presence of the park amenity.

A visual depiction of the relationships between health outcomes and park amenities was created using boxplots of the overall health rating for each level of each park amenity. Reviewing boxplots for all 13 response variables would be time-consuming, so only overall health rating was used as a representation of the general health outcomes in a county. The boxplots below show the spread of the overall health rating for counties that do not have the amenity in the title in dark blue and for counties that do have the amenity in the title in light blue. The horizontal line in the middle of each box represents the median health rating within those counties. There is a visible difference in the median overall health rating based on a few of the amenities, most notably trails. This could

potentially indicate a relationship between the presence of these amenities in state parks within a county and that county's health outcomes.



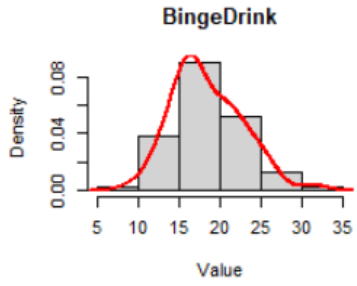
Prior to conducting the MANOVA, the assumptions of MANOVA had to be checked. MANOVA makes assumptions about the structure of the data that is fed into the analysis, and if these assumptions are violated, the results of the MANOVA could be inaccurate.

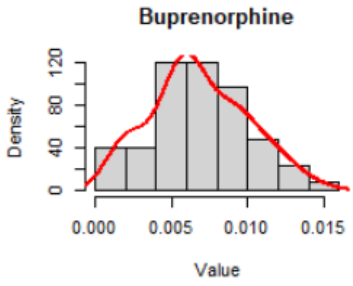
First, MANOVA assumes the dependent variables are normally distributed within each group of predictor variables. In the scope of this study, all 13 health outcome variables had to be normally distributed in all counties that had each park amenity and all counties that did not have each amenity. In addition to normal distribution, MANOVA also assumes all dependent variables have similar variance within each group of predictor variables. In the context of this study, that meant each of the health outcomes must have similar variance in counties where each park amenity is present as in counties where that amenity is not present.

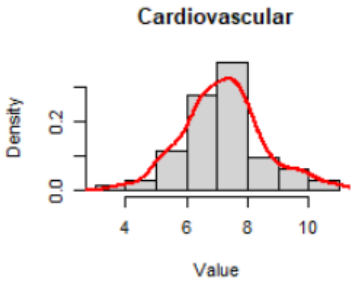
MANOVA makes inferences about parameters of the data, such as mean and standard deviation, based on the assumptions of normality and homogeneity of variance. If the data is not truly normally distributed and/or does not have equal variance, MANOVA may form incorrect inferences, which could lead to inaccurate results.

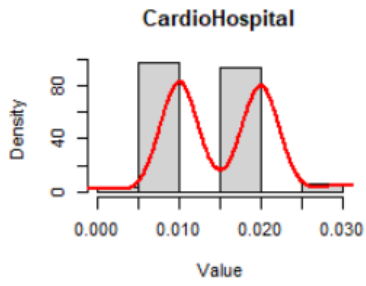
To test the assumption of normal distribution, a Shapiro-Wilk test was applied to each response variable at both levels of each predictor variable. A p-value of less than 0.05 was used as the cutoff to indicate violation of the normality assumption. If the response variable had no p-values less than 0.05 for either level of any of the predictor variables, that variable was kept in the analysis in its raw form. If the response variable had p-values of less than 0.05 for at least one level of at least one of the predictor variables, it could not be used in the MANOVA in its raw form.

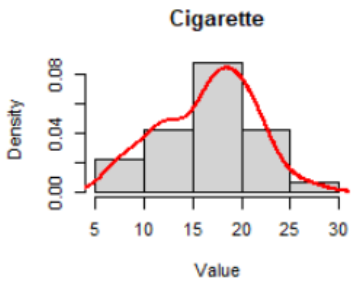
The response variables that violated the normality assumption in their raw form were transformed on three scales: logarithm, square root, and cube root. After the transformations, the Shapiro-Wilk test was conducted again. If the transformed response variable resulted in a p-value of greater than 0.05 for each level of each predictor, it was added to the analysis in its transformed form. This resulted in 8 of the original 13 responses remaining in the analysis. Histograms of each of the response variables in their raw forms are shown below.

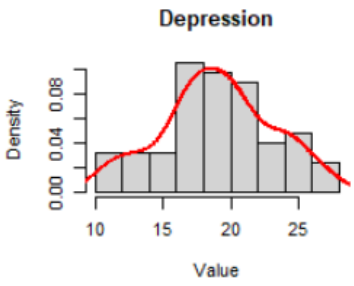

BingeDrink
Density
0.00
0.04
0.08
5
10
15
20
25
30
35
Value

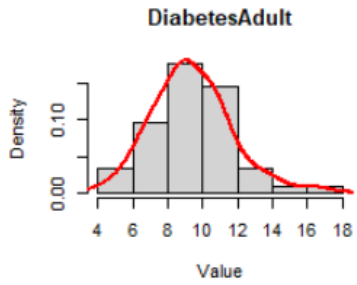

Buprenorphine
Density
0
40
80
120
0.000
0.005
0.010
0.015
Value


Cardiovascular
Density
0.0
0.2
4
6
8
10
Value

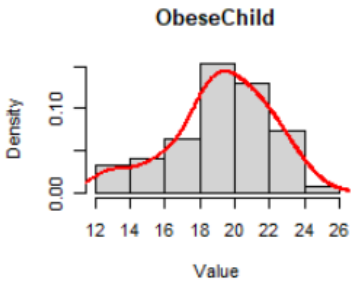

CardioHospital
Density
0
40
80
0.000
0.010
0.020
0.030
Value

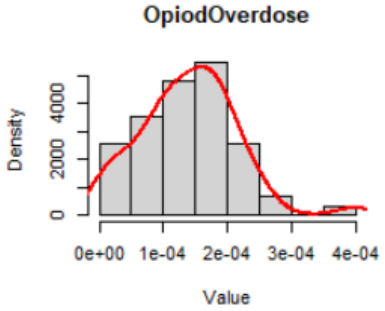

Cigarette
Density
0.00
0.04
0.08
5
10
15
20
25
30
Value

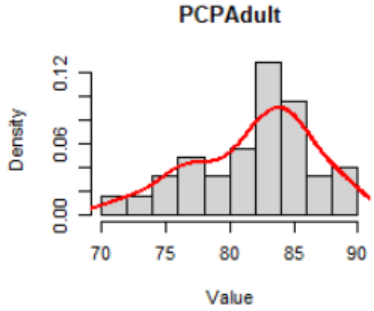

Depression
Density
0.00
0.04
0.08
10
15
20
25
Value

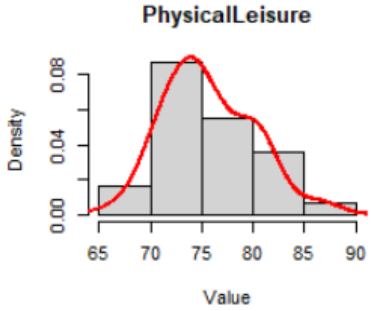

DiabetesAdult
Density
0.00
0.10
4
6
8
10
12
14
16
18
Value

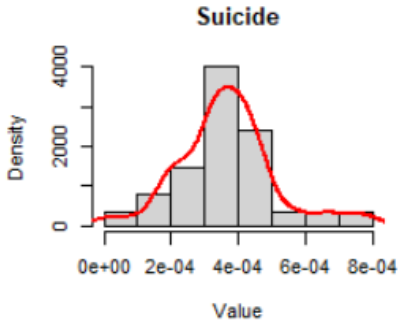

ObeseAdult
Density
0.00
0.03
0.06
15
20
25
30
35
40
45
Value


ObeseChild
Density
0.00
0.10
12
14
16
18
20
22
24
26
Value


OpiodOverdose
Density
0
2000
4000
0e+00
1e-04
2e-04
3e-04
4e-04
Value


PCPAdult
Density
0.00
0.06
0.12
70
75
80
85
90
Value


PhysicalLeisure
Density
0.00
0.04
0.08
65
70
75
80
85
90
Value


Suicide
Density
0
2000
4000
0e+00
2e-04
4e-04
6e-04
8e-04
Value

In addition to normal distribution of each of the response variables within each group, MANOVA also assumes equal variance of each response variable between the groups. This assumption was tested via Levene's test for each of the remaining variables. In other words, for each health outcome and each park amenity, Levene's test returned a p-value representing the probability that the health outcome had equal variance in counties where the amenity was present and in counties where the amenity was not present. Of the 104 combinations of remaining response variables and predictor variables, only 3 combinations showed statistically significant evidence of unequal variance:

Response Variable	Predictor Variable
Log-transformed percentage of adults who report binge drinking within the last month	Field sports
Log-transformed percentage of adults who report binge drinking within the last month	Information centers
Log-transformed percentage of adults with diabetes	Waterfalls

MANOVA is not particularly sensitive to violations of the assumption of homogeneity of variance if the variance-covariance matrices are not too dissimilar and the sample sizes are roughly equal, which is typically defined as the larger sample being less than 50% greater than the smaller sample. In the context of this analysis, that means for each of the amenities above (field sports, information centers, and waterfalls), whichever group was larger (the group of counties with the amenity versus the group without) cannot be more than 150% the size of the smaller group. Two of the combinations, log-transformed adults who report binge drinking within the last month and information centers and log-transformed percentage of adults with diabetes and waterfalls, did not meet this criterion. Therefore, the equal variance assumption was not met.

Although the assumption was not met, the sizes of the samples provided insights into what the results of the MANOVA would be if these violating variables were left in the analysis. Combinations of responses and predictors where the larger sample also has the higher variance are robust to Type 1 errors, meaning the MANOVA is less likely to not find evidence of a difference between groups when there truly is a difference. Unfortunately, the only predictor variable in the table above for which the larger sample also had the larger variance was information centers.

Because at least one of the amenity variables for each of the two violating response variables did not meet this criterion, both violating variables were removed from the analysis. Keeping them in the analysis could have resulted in inaccurate results.

An additional assumption of MANOVA is independence of all observations, meaning the values of one observation do not affect the values of other observations. If this assumption is violated, MANOVA may result in biased inferences. To test this assumption, the *independence_test()* function in R was applied to each of the response variables. None of the tests produced a p-value of less than 0.05, indicating no statistically significant evidence of dependence among any of the response variables. Therefore, no response variables were removed from the analysis after this step of the assumption check process.

MANOVA also assumes no multicollinearity of the response variables, meaning none of the response variables are highly correlated. If two or more response variables are highly related to each other, it could cause difficulty assessing the changes in each variable. Additionally, multicollinearity can lead to unstable estimates, meaning small changes in the data could lead to large changes in the MANOVA results. The correlation between each pair of response variables was checked, and none of them were greater than 0.7, indicating multicollinearity should not be an issue for this analysis. The correlations are shown in the table below, with the darkness of the green background indicating the magnitude of the absolute value of the correlation.

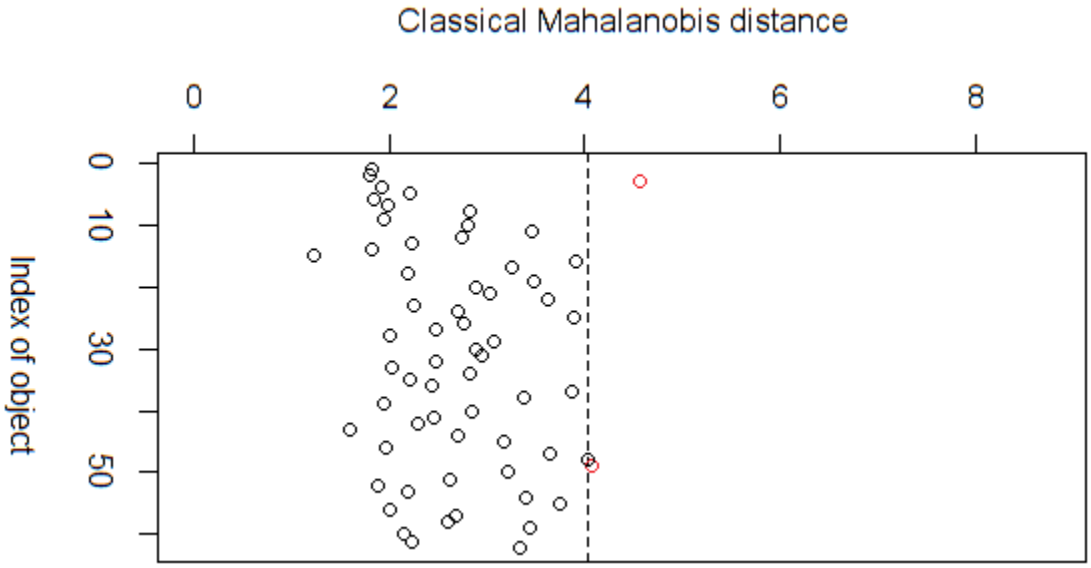
	LogBingeDrink	Buprenorphine	LogCardiovascular	Cigarette	Depression	LogDiabetesAdult	LogObeseAdult	PhysicalLeisure
LogBingeDrink	1.000	0.118	0.039	0.230	0.074	-0.160	0.019	0.242
Buprenorphine	0.118	1.000	0.352	0.580	0.446	0.196	0.495	-0.283
LogCardiovascular	0.039	0.352	1.000	0.431	0.292	0.294	0.439	-0.481
Cigarette	0.230	0.580	0.431	1.000	0.668	0.281	0.679	-0.532
Depression	0.074	0.446	0.292	0.668	1.000	0.158	0.489	-0.230
LogDiabetesAdult	-0.160	0.196	0.294	0.281	0.158	1.000	0.386	-0.443
LogObeseAdult	0.019	0.495	0.439	0.679	0.489	0.386	1.000	-0.498
PhysicalLeisure	0.242	-0.283	-0.481	-0.532	-0.230	-0.443	-0.498	1.000

Finally, MANOVA assumes there are no outliers in the data. Outliers can distort estimates of data parameters such as mean and standard deviation, which can in turn lead to distorted estimates of the effect of each predictor variable on the response variables. To test each observation

for the possibility of being an outlier, Mahalanobis' distance (MD) was used. MD measures the distance between two observations in a multivariate space and can therefore be used as a metric to test for outliers where there is more than one response variable.

R used the MD of each observation and a chi-square distribution to calculate a p-value indicating the extent to which that observation was an outlier. P-values less than or equal to 0.001 are typically considered indicative of outliers in this method. Only two of the observations--Bronx County with a p-value of 0.0001 and Seneca County with a p-value of 0.001--were identified as outliers.

Because outliers can skew the results of the MANOVA, leaving these observations in the analysis could lead to inaccurate inferences about the relationships in the dataset. However, the offending observations cannot simply be dropped from the analysis just because they are outliers; they are a part of the overall population being measured and removing them could skew the results. The analysis had to be conducted both with and without the results to see if their removal altered the results. The plot below shows a plot of the MD for each observation, and the outliers are indicated in red.



The following variables met all the assumptions of MANOVA and were used in the analysis:

Variable	Transformation
Age-adjusted percentage of adults who received one or more buprenorphine prescriptions for opioid use disorder	None
Percentage of adults who smoke cigarettes	None
Percentage of adults diagnosed with depression	None
Percentage of adults who participate in physical leisure activity	None
Percentage of adults diagnosed with cardiovascular disease	Logarithm
Percentage of obese adults	Logarithm

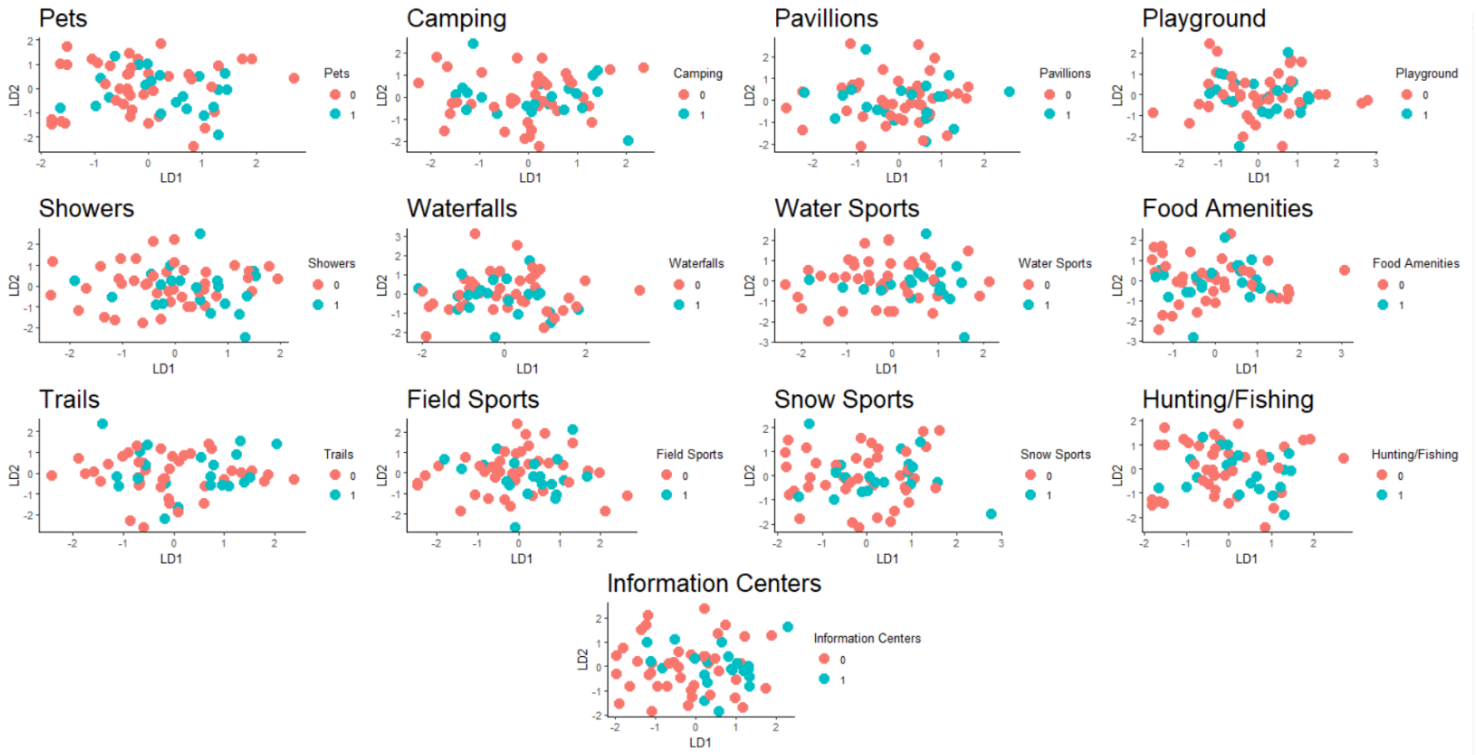
MANOVA was conducted on the responses above for all predictors on all observations in the dataset. The only predictor that showed statistically significant evidence of an association with the above responses was the presence of playgrounds with a p-value of about 0.04. The MANOVA coefficients for each of the responses are below. (Coefficients for log-transformed outcome variables have been back-transformed and are presented on the original scale.)

The coefficients represent the expected difference in each health outcome between counties that have no state parks with playgrounds and counties that have at least one state park with a playground. The coefficients are negative for each of the responses that are considered negative indicators of health and positive for the one response that is a positive indicator of health, implying the presence of playgrounds in state parks has a positive association with overall health in that county.

Outcome	Coefficient
Age-adjusted percentage of adults who received one or more buprenorphine prescriptions for opioid use disorder	-0.001%
Percentage of adults who smoke cigarettes	-1.647%
Percentage of adults diagnosed with depression	-0.254%
Percentage of adults who participate in physical leisure activity	0.764%
Percentage of adults diagnosed with cardiovascular disease	-0.290%
Percentage of obese adults	-2.758%

Linear discriminate analysis (LDA) was used as a post-hoc visual analysis of the separation (or lack of separation) in the data based on the predictors. LDA identifies combinations of the health outcomes that best separate the groups among the observations. Visible separation in the plots could indicate large differences in health outcomes between counties with a given amenity and those without. The first two linear combinations, LD1 and LD2, explain the largest amount of variance in the data.

The plots below show the results of LDA applied to all health outcomes above for each park amenity. Each point on the plots represents a different observation, or county. The color of the point indicates whether that county has the amenity stated in the title. Red indicates a county without the amenity, and teal indicates a county with the amenity. The further a point is from another point, the more those two points differ. LD1 is represented on the x-axis, and LD2 is represented on the y-axis. Each county is plotted where it falls along LD1 and LD2.



There does not appear to be any visible separation between the observations within each group, implying these groups may not differ much regarding the health outcomes used in the analysis. Although playgrounds showed statistically significant evidence of a difference in health outcomes in the MANOVA, the LDA plot for playgrounds does not display a visible difference in groups. This could be due to a few reasons. First, it could be because the differences between groups are not large enough to be captured by the LDA. Second, it could be due to confounding variables that the analysis is not controlling for. Finally, the goal of LDA is to identify a linear combination of variables that best separate the groups, while the goal of MANOVA is to test for overall differences in the groups. Although LDA can be useful for viewing relationships found in a MANOVA, ultimately each method is testing something slightly different, which may be the reason for the discrepancy.

The final step of part 1 was to repeat the analysis without the outliers. The results after removing the outliers did not change (playgrounds with a p-value of 0.04), so the outliers could be safely kept in the analysis.

Part 2—ANOVA of Overall Health Rating

The second step in question 1 was to repeat the analysis done above for all health outcomes, this time only using the overall health rating as the single response variable. For this step, an Analysis of Variance (ANOVA) was used. ANOVA is the univariate form of the MANOVA and has similar assumptions to the MANOVA: normal distribution in the response variable at each level of each predictor variable, equal variance of the response variable at each level of each predictor variable, and independence of the observations. The assumptions were tested using similar methods to those described above for the MANOVA.

The overall health rating variable met all assumptions of ANOVA, so the analysis was conducted on overall health rating for all predictors and all observations. There were no outliers for overall health rating as it was created based on standardized values, so only one test had to be done. The test resulted in statistically significant evidence of a relationship between overall health rating with camping (p-value=0.04) and trails (p-value=0.02). This implies that a county's overall health rating differs based on whether that county has state parks with camping and/or trails.

A post-hoc Tukey-Kramer was applied to overall health rating with only the two significant variables, camping and trails. The Tukey-Kramer contrast checks for differences between groups while controlling for family-wise error rates by adjusting p-values based on the number of comparisons made. The results showed moderate statistically significant evidence of an estimated increase in health rating of about 4.56 when trails are present (95% confidence interval 0.755-8.359, p-value=0.02). Although the original ANOVA showed statistical evidence of a difference in health rating based on camping and trails, the post-hoc Tukey-Kramer only showed evidence of a difference based on trails. This could be because there truly is no difference, and the ANOVA was not appropriately correcting for all the comparisons in the analysis. Additionally, it could be because the sample size is rather small (n=62), and there are quite a few predictors (p=13). This results in many p-value corrections that could be weakening the statistical power of the Tukey Kramer contrast.

Part 3--Conclusion

When all health outcomes that met the assumptions of MANOVA were considered (age-adjusted percentage of adults who received one or more buprenorphine prescriptions for opioid use disorder, percentage of adults who smoke cigarettes, percentage of adults diagnosed with depression, percentage of adults who participate in physical leisure activity, percentage of adults diagnosed with cardiovascular disease, and percentage of obese adults), there is moderate statistically significant evidence of a difference in the mean of these health outcomes in a county based on the availability of playgrounds in that county's state parks (p-value=0.04).

The health outcomes that are considered a negative contributor to a county's overall health were estimated to be slightly lower in a county that has state parks with playgrounds compared to a county that has no state parks with playgrounds (estimated decreases range from about 0.001% to 2.76%). The one health outcome that is considered a positive influence in a county's overall health was estimated to be around 0.76% higher in counties that have state parks with playgrounds versus those that do not.

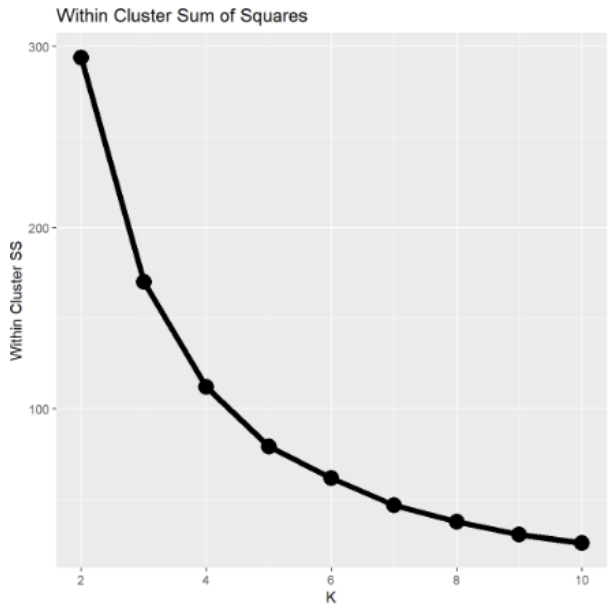
When only overall health rating was considered and the family-wise error rate was controlled, there was statistically significant evidence of an increase of about 4.557 in overall health rating in counties that have state parks with playgrounds versus those that do not (95% confidence interval 0.755-8.359, p-value=0.02).

Question 2 Methodology

Part 1—MANOVA of All Health Outcomes

For the second part of the analysis, the observations were clustered into groups based on their health outcomes, the same tests from Question 1 above were conducted within each group, and the results were compared. Bootstrap K-means clustering was used to find the optimum number of clusters for the data based on all 13 health outcomes other than overall health rating, as the overall health rating is simply a combination of the other health variables, so using it in the clustering process would be redundant.

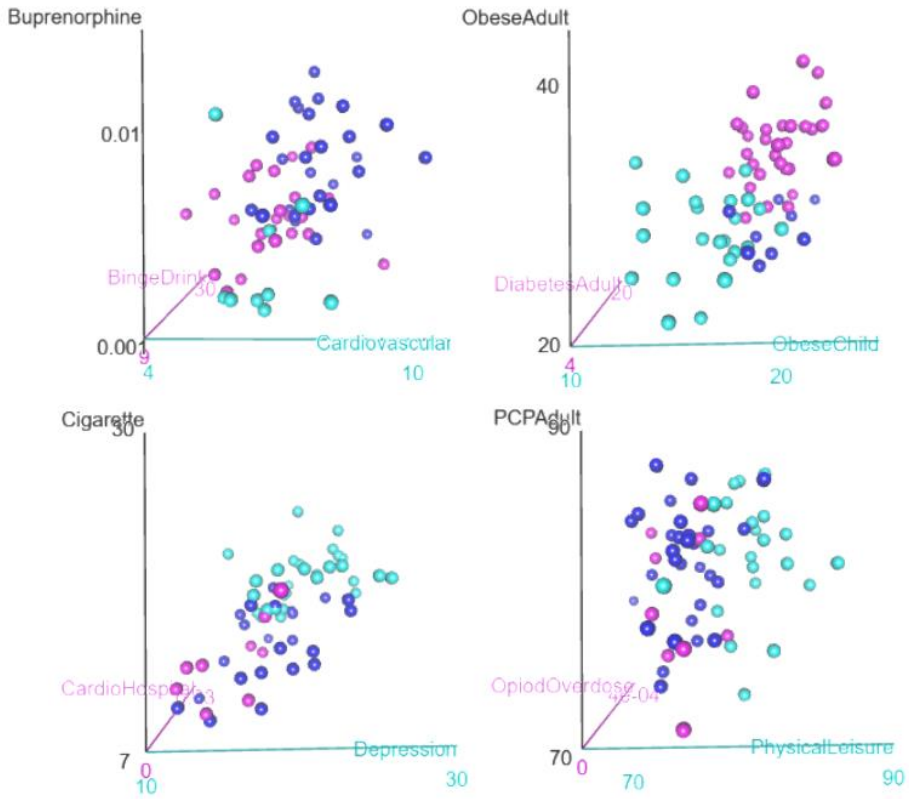
The 13 response variables were standardized prior to clustering because K-means clustering relies on distance to create the clusters, so using the raw data could cause the variables with larger values to have an undue larger effect on the results than the variables with smaller values. Values for K of 1 through 10 were used, and each value was bootstrap case resampled 50 times. The resulting within-cluster sum of squares (WCSS) for each value of K is shown in the plot below.



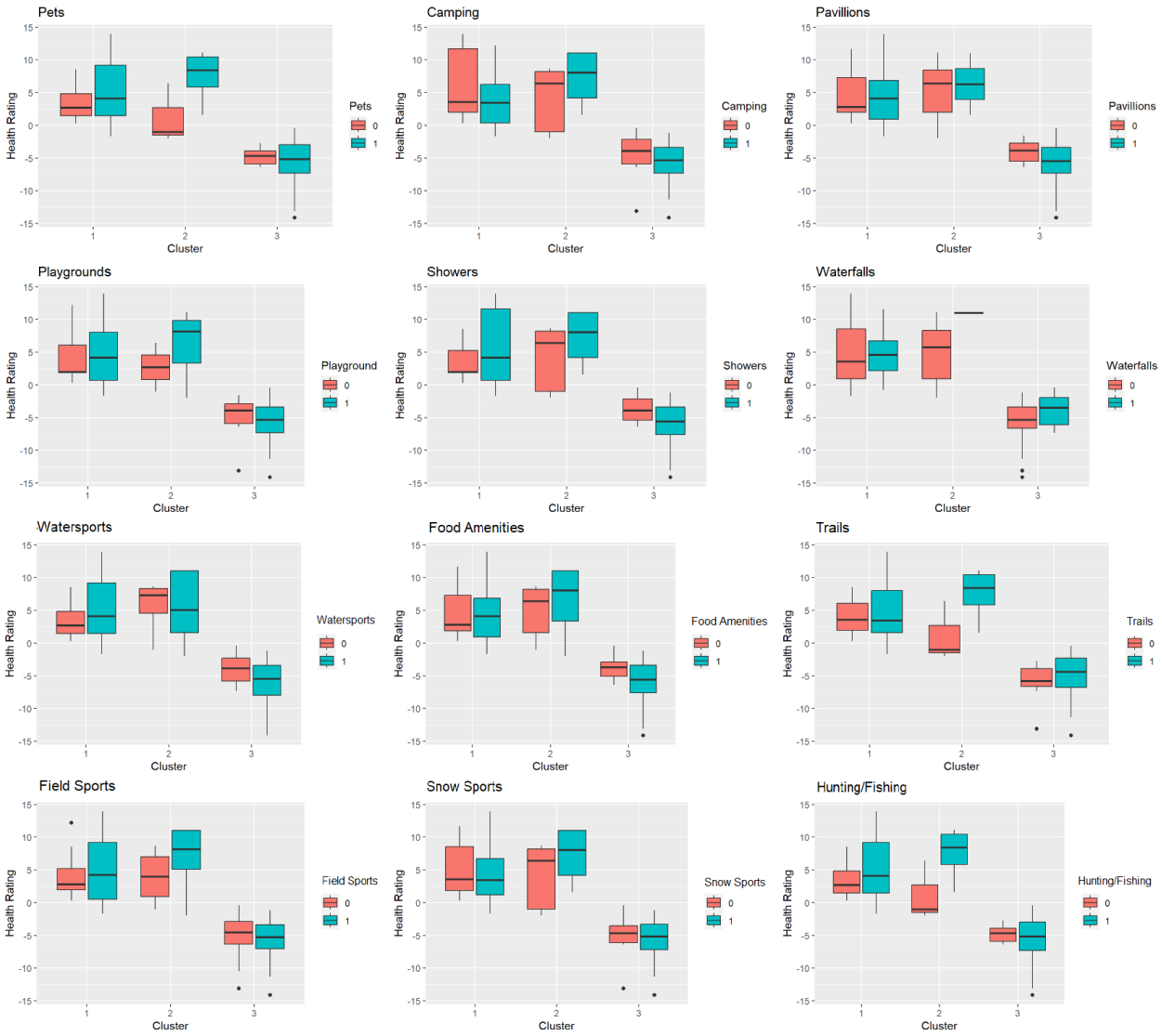
The WCSS is a measurement of the deviance from the mean in each cluster. Large values indicate the observations within each cluster are highly different from each other, and small values indicate the observations within each cluster are highly similar. The plot above shows the WCSS drops dramatically from K=1 until around K=4 and K=5, where the line starts to level off. This indicated 4 or 5 clusters would likely be a sufficient grouping for the data that resulted in observations that were highly similar to other observations within their group and highly different from observations outside of their group.

However, with only 62 observations, even 4 clusters resulted in rank deficiencies in some of the clusters that impeded further statistical analysis. To continue the analysis, the number of clusters had to be reduced to 3. Cluster 1 included counties that represented the middle, or average, health outcomes in New York State, Cluster 2 included the counties with the best health outcomes, and Cluster 3 contained the counties with the worst health outcomes. The 3D plots below show each observation color-coded to represent the cluster it belongs to. The axes of the plots are 12 of

the response variables that were used in the clustering. Each observation is plotted where it sits along those variables.



Similar to Question 1, boxplots of the overall health rating in counties with amenities and without amenities within each cluster were created as a visual representation of group variance. The boxplots are shown below. The teal boxes indicate counties where the amenity in the title is available, and the red boxes indicate counties where the amenity is not available. There appears to be a large difference in median overall health rating for the following amenities in cluster 2: pets, waterfalls, playgrounds, trails, field sports, and hunting/fishing. These visual differences could potentially indicate a relationship between these park amenities and overall health rating within Cluster 2.



The assumptions of MANOVA were tested within each cluster using the same methods explained above. The only health outcomes that satisfied the requirements in all three clusters are shown in the table below:

Variable	Transformation
Percentage of obese children	None
Percentage of adults with a regular healthcare provider	None
Percentage of adults who smoke cigarettes	Square Root

A MANOVA was conducted on these three responses within each cluster, and the results of each MANOVA were compared. There were no p-values less than 0.05 in clusters 1 or 2, indicating no statistically significant evidence of a difference in the mean of these three health outcomes based on availability any of the park amenities within these clusters. Cluster 3, however, showed statistically significant evidence of a difference in mean health outcomes based on the presence of waterfalls (p-value=0.0002) and food amenities (p-value=0.04). The MANOVA coefficients for waterfalls and food amenities were inspected within each cluster to determine where the differences were, and the results are shown below. (The coefficients for the square root-transformed percentage of adults who smoke cigarettes have been back-transformed in the table.)

Cluster	**Amenity**	**Percentage of adults who smoke cigarettes**	**Percentage of obese children**	**Percentage of adults with a regular healthcare provider**
Cluster 1	Waterfalls	1.13%	-0.84%	2.16%
Cluster 1	Food amenities	-0.97%	-1.26%	-5.38%
Cluster 2	Waterfalls	8.33%	-3.20%	4.23%
Cluster 2	Food amenities	-1.07%	-0.22%	4.63%
Cluster 3	Waterfalls	-0.48%	-1.82%	-2.40%
Cluster 3	Food amenities	0.55%	0.06%	-1.31%

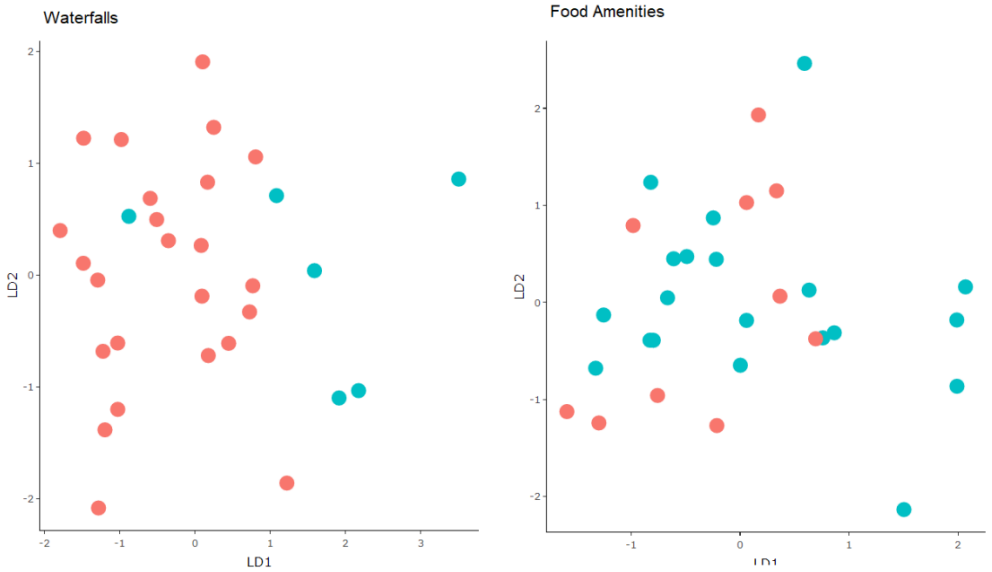
The coefficients showed a decrease in the percentage of obese children in all three clusters when waterfalls were present (estimated 0.84% decrease in Cluster 1, estimated 3.2% in Cluster 2, and estimated 1.8% in Cluster 3). The percentage of obese children also decreased in Clusters 1 and 2 when food amenities were present (estimated 1.3% and 0.2% decreases, respectively), and although it increased in Cluster 3 with the presence of food amenities, it was only very slightly (estimated 0.06%). This indicates the presence of both food amenities and waterfalls could have a positive impact on child obesity in most counties.

The direction of the estimated relationships between the other health outcomes and these two park amenities was less consistent. In Clusters 1 and 2, waterfalls had a positive impact on the percentage of adults with a regular healthcare provider (estimated 2.2% in Cluster 1 and estimated

4.2% in Cluster 2). However, in Cluster 3, the impact was negative (estimated 2.4%). The presence of food amenities only had a positive effect on the percentage of adults with a healthcare provider in Cluster 2 (estimated 4.6%), while the effects in Clusters 1 and 3 were negative (estimated 5.4% and estimated 1.3%, respectively).

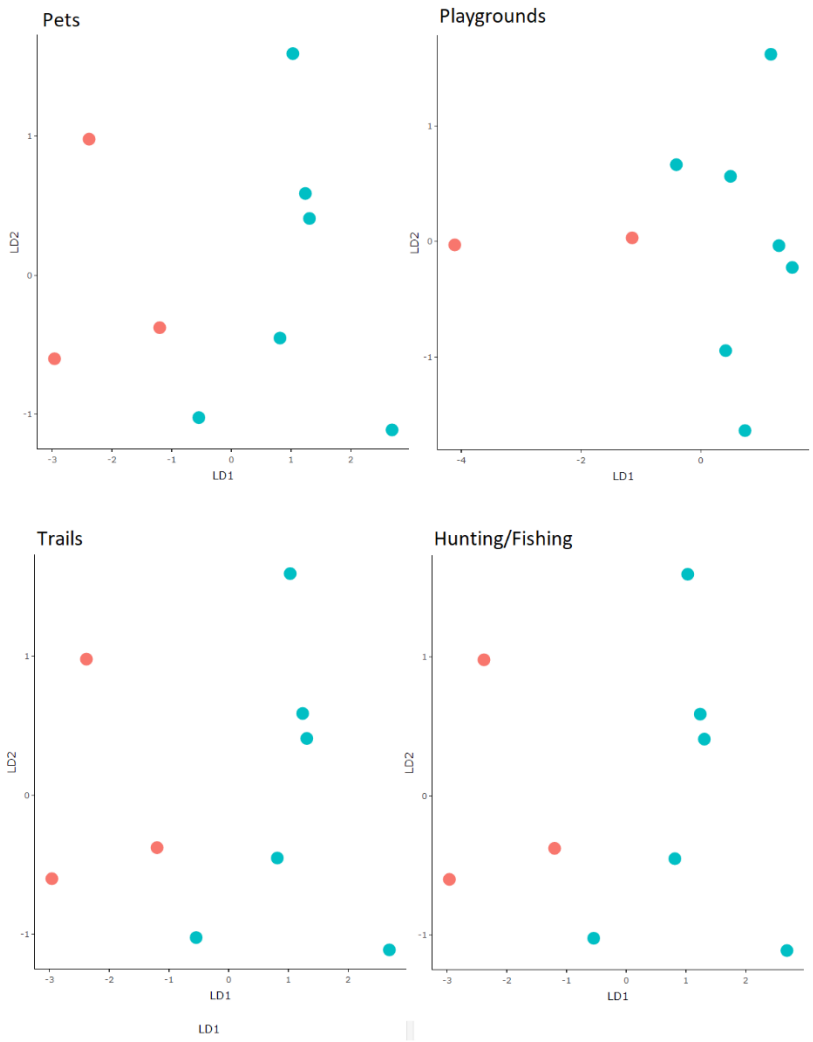
The presence of waterfalls was associated with an estimated 1.1% increase in adults who smoke cigarettes in counties within Cluster 1 and an estimated 8.3% increase in Cluster 2. However, in Cluster 3, the estimated effect was a small decrease of about 0.5%. The presence of food amenities within counties in Clusters 1 and 2 was also associated with a slight decrease in adults who smoke cigarettes (estimated 1.0% and 1.1%, respectively), but in Cluster 1, it was associated with a slight increase of about 0.6%. Although the coefficients for both predictors showed some positive health impacts in all three clusters, only Cluster 3 showed statistically significant evidence of association. Therefore, there was no statistically significant evidence of the relationships observed in Clusters 1 and 2.

LDA plots were constructed for each of the predictor variables within each of the clusters. The plots for waterfalls and food amenities in Cluster 3 (the two predictors with statistically significant evidence of an association with health outcomes in Cluster 3) are shown below. There does appear to be some visible separation between observations with and without waterfalls, but no visible separation can be seen with food amenities.



Interestingly, some of the LDA plots for Cluster 2 show clear separation between the groups, even though the MANOVA found no statistically significant evidence of a relationship between any of the park amenities and the health outcomes in Cluster 2. This is likely due to the small number of observations, which can lead to over-fitting with LDA. After grouping the data into clusters, Cluster 2 only contained 9 observations. With such a small number of observations, LDA can nearly always find a line of separation by chance alone. This is most likely the reason for this discrepancy between the LDA plots and the MANOVA results.

The plots for Cluster 2 showing separation between groups are shown below. The plots for pets, trails, and hunting/fishing are identical because these three amenities always occurred in the same combination within counties in Cluster 2.



Part 2—Nonparametric ANOVA of Overall Health Rating

The overall health rating was not normally distributed within the clusters and therefore could not be analyzed with an ANOVA. Instead, a rank-based nonparametric ANOVA was conducted within each cluster, and the results were compared. The nonparametric ANOVA does not make assumptions about the distribution of the data, such as normality and equal variance, as the parametric ANOVA and MANOVA do.

To conduct the nonparametric ANOVA, R transformed the overall health rating for each observation into sample ranks that were used in the analysis in lieu of the original values. To create the sample ranks, the values of overall health rating were sorted in ascending order, and the overall health rating for each observation was replaced with the corresponding value of that observation's order number.

The nonparametric ANOVA resulted in statistically significant evidence of a relationship between overall health rating and snow sports in Cluster 2 (p-value=0.02). However, when a post-hoc Tukey-Kramer contrast was conducted, the p-values for snow sports in all the clusters were greater than 0.05. Again, this could be because there truly is no difference in means, or it could be due to the small number of observations in the analysis. The number of observations was low to begin with but decreased even more after clustering, making the small sample size even more likely to affect the results.

Part 3-- Conclusion

In counties contained in Cluster 3 (counties in New York with the worst health outcomes) there is statistically significant evidence of an association between the presence of waterfalls (p-value=0.0002) and food amenities (p-value=0.04) with the square root-transformed percentage of adults who smoke cigarettes, the percentage of obese children, and the percentage of adults with a regular healthcare provider. The presence of waterfalls is associated with a decrease in the percentage of obese children in counties within all three clusters.

However, the direction of the relationships between waterfalls and food amenities with the other health outcomes were not as consistent. Food amenities were associated with a decrease in the percentage of obese children in counties within Clusters 1 and 2 (counties with the best health

outcomes), and only a slight increase in counties within Cluster 3 (counties with the worst health outcomes).

In Clusters 1 and 2, waterfalls were associated with an increase in the percentage of adults with a regular healthcare provider. Food amenities in Clusters 1 and 2 were associated with a slight decrease in the percentage of adults who smoke cigarettes but a slight increase in Cluster 3. Overall, the presence of waterfalls and food amenities within state parks seem to positively impact health outcomes in counties that already have the best health outcomes (Clusters 1 and 2) more than they positively impact counties that have the worst health outcomes (Cluster 3).

When counties are clustered into 3 groups but only the overall health rating is considered and the family-wise error rate is controlled for, there is no statistically significant evidence of an association between any of the state park amenities and overall health rating in any of the clusters.

Overall Conclusion

Question 1: Do mean health outcomes change based on the state park amenities available and, if so, what is the magnitude and direction of the relationship(s)?

When all counties are considered, there is statistically significant evidence (p-value=0.04) of an estimated 0.0001% decrease in the age-adjusted percentage of adults who received one or more buprenorphine prescriptions for opioid use disorder, an estimated 1.65% decrease in adults who smoke cigarettes, an estimated 0.25% decrease in adults diagnosed with depression, an estimated 0.29% decrease in adults diagnosed with cardiovascular disease, and an estimated 2.76% decrease in obese adults in a county that has state parks with playgrounds compared to a county that does not.

When all counties are considered, there is also statistically significant evidence (p-value=0.04) of an estimated increase of 0.76% in adults who participate in physical leisure activity in a county that has state parks with playgrounds compared to a county that does not.

When all counties are considered and the family-wise error is controlled, there is statistically significant evidence of an estimated increase of 4.56 in overall health rating in a county

that has state parks with trails compared to a county that does not (95% confidence interval 0.76-8.36, p-value=0.02).

Question 2: Does that relationship vary between groups of counties with different ranges of health outcomes and, if so, what is the magnitude and direction of the relationship(s) within each group?

When the counties were clustered into three groups based on their combined health outcomes, there was statistically significant evidence (p-value=0.0002) in counties contained in Cluster 3 that the presence of waterfalls is associated with an estimated decrease of 0.48% in adults who smoke cigarettes, an estimated decrease of 1.85% in obese children, and an estimated decrease of 2.40% in adults with a regular healthcare provider.

There is also statistically significant evidence (p-value=0.04) in counties contained in Cluster 3 that the presence of food amenities is associated with an estimated increase of 0.55% in of adults who smoke cigarettes, an estimated increase of 0.06% in in obese children, and an estimated 1.31% decrease in adults with a regular healthcare provider. Overall, the presence of food amenities in counties contained in Cluster 3 both decreased health outcomes that are considered negative indicators of health and increased health outcomes that are considered positive indicators of health.

Although there was no statistically significant evidence of a relationship between any of the park amenities and health outcomes in Clusters 1 and 2, the presence of waterfalls in counties within all three clusters was associated with a decrease in the percentage of obese children in these clusters as well (estimated 0.84% decrease in Cluster 1 and estimated 3.20% decrease in Cluster 2).

When counties are clustered into 3 groups but only the overall health rating is considered and the family-wise error rate is controlled for, there is no statistically significant evidence of an association between any of the state park amenities and overall health rating in any of the clusters.

Inferences

If the New York State Parks and Recreation Department was interested in improving health outcomes based on state park amenities, I would suggest doing more research into the possible associations between playgrounds and health outcomes in all counties based on the MANOVA results, which indicated statistically significant evidence of decreases in negative indicators of health and increases in positive indicators of health in all counties with state parks that have playgrounds.

Additionally, I would recommend further research into the relationship between waterfalls and child obesity, as there was a statistically significant negative relationship between these two variables in counties contained in Cluster 3, and there was a negative relationship associated with the presence of waterfalls in counties in all three clusters. Waterfalls also appeared to have a positive relationship with the percentage of adults with a regular healthcare provider in Clusters 1 and 2. However, waterfalls were also associated with an increase in adults who smoke in Clusters 1 and 2 and a decrease in adults with a regular healthcare provider in Cluster 3. Therefore, more research would need to be conducted.

The conflicting associations in the results could be due to possible confounding factors or due to the small sample size of this study. More in-depth research using a larger sample size while controlling for other contributors to health, such as demographic information, could produce interesting results. The small sample size was very limiting in this study and resulted in having to remove many of the park amenity variables and health outcome variables from the analysis. The small sample size also caused an inability to use the optimal number of groups needed to maximize similarities within clusters and differences between clusters. Conducting this study again with more states included could allow the researcher to include more variables, including confounding variables, and create more homogenous groups, which could result in more accurate results.

Data Sources

County population totals and components of change: 2020-2022. Census.gov. (2023, March 23). Retrieved April 15, 2023, from https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html

New York State Department of Health. (2023, February 9). *Behavioral risk factor surveillance system (BRFSS) health indicators by county and region: State of New York*. Behavioral Risk Factor Surveillance System (BRFSS) Health Indicators by County and Region | State of New York. Retrieved April 15, 2023, from https://health.data.ny.gov/Health/Behavioral-Risk-Factor-Surveillance-System-BRFSS-H/jsy7-eb4n

New York State Department of Health. (2020, August 20). *Community health indicator reports (Chirs): Latest data: State of New York*. Community Health Indicator Reports (CHIRS): Latest Data | State of New York. Retrieved April 15, 2023, from https://health.data.ny.gov/Health/Community-Health-Indicator-Reports-CHIRS-Latest-Da/54ci-sdfi

Parks, recreation and historic preservation. New York State Parks Recreation & Historic Preservation. (2022). Retrieved April 15, 2023, from https://parks.ny.gov/

Prevention agenda 2019-2024 tracking indicators: County trend data: State of New York. New York State Department of Health | Health Data NY. (2022, February). Retrieved April 15, 2023, from https://health.data.ny.gov/Health/Prevention-Agenda-2019-2024-Tracking-Indicators-Co/7j59-48xy/data