Recreational Visits to National Park Statistical Analysis

Executive Summary

Christopher Flowers

Western Governors University

**The Problem and Hypothesis**

Every year millions of individuals and families set off to travel across this country. Among things, travelers attend to do while away from home is go sightseeing and visiting different attractions. Of the many different types of attractions, National Parks are among the most visited. People visit national parks for a million different reasons each year. Those types of visits include recreational visits, non-recreational visits, RV campers, tent campers, miscellaneous overnight campers, Concessioner Lodging, concessioner campers, and backcountry campers to name a few. The research question addressed was Does Great Smokey Mountain National Park has more monthly recreational visitors than Yellowstone National Park? Both parks have thousands of visitors every month (Stats Report Viewer, 2022). The null hypothesis is monthly recreational visits to the Great Smokey Mountains National Park occur with no statistically significant difference from monthly recreational visits to Yellowstone National Park. To break down the null hypothesis a bit more we look at what we testing to be true is there is no difference in more recreational visitors visiting the Great Smokey Mountains National Park compared to the famous Yellowstone National Park and this is compared to a monthly basis. While the alternative hypothesis is monthly recreational visits to the Great Smokey Mountains National Park occur at a statistically significant higher rate than monthly recreational visits to Yellowstone National Park. We will either accept or reject the alternative hypothesis based on the findings of the analysis of the null hypothesis. Additionally, the null hypothesis can be evaluated in many different ways. Here the linear regression statistical method will be used to determine if Great Smokey Mountains National Park has more recreational visitors than Yellowstone National Park. In 2021 national parks had “297,115,406 recreation visits” according to the National Parks Service website("Visitation Numbers (U.S. National Park Service)", 2022). There's plenty of stiff competition, but Great Smoky Mountains National Park may be the most beautiful park on our list, with acres and acres of lush green trees("The 25 most visited tourist spots in America", 2022), in the top twenty-five most visited two National Parks are included.

**Data Analysis Process**

This data set has a continuous dependent variable is Year. These correspond with the year the data is from. Listed next are the independent variables. Park, Park type, State, Unit code, and region are the only categorical variables in the data set. Miscellaneous overnight stays, recreation visits, non-recreational visits, recreation hours, non-recreational hours, concessioner lodging, concessioner camping, tent campers, backcountry campers, RV campers, and non-recreation overnight stays are all the continuous variables in this data set. This data is public information collected and owned by the National Parks Service. There are over 22,896 records in the data set as a whole. This data is limited because it only goes back five years. Data for the years before this research is available on the national parks website as well but for this paper, a limit on the number of years to look at had to occur. Using the latest five full years was perfect for this because it gives enough data to show trends of how people traveled before, in, and after the post-COVID-19 pandemic.

After the data was downloaded from the National Parks Service website, we first looked at the excel spreadsheet. In checking the excel sheet we are looking to see if there is any information that in no way pertains to the data. In this case, the spreadsheet had some information in the top two rows and a severely reduced third row. Also, the column headers were changed to SAS-appropriate formats including no spaces or hyphens. We are going to go ahead and remove all things in the top three rows of the excel spreadsheet. This will now just leave the column headers and the data itself. The first thing we do now is that the data spreadsheet is in a format that will not cause an issue when it’s imported into SAS. We create a new SAS program file with two libname statements, which are used to help store the data and results. But before we import it into SAS we have to upload the spreadsheet to SAS Studio, now we can write the import procedure statement to import the spreadsheet into SAS. The last thing we do before cleaning the data is to run a content procedure on the imported data table to see what the variables we have in the data set and what their formats and data types are.

The data was cleaned by removing all blank or missing data records, to begin with. Any missing data will be handled using the methodology according to “How to Deal with Missing Data”("How to Deal with Missing Data", 2022). Then remove any duplicated data followed by any unnecessary fields from the data set. This data set has it spans the most recent full five years. There will be no data type conversions needed for this data set as all fields are in the necessary data type already but formats will be changed. An aggregate column could be called total Visits, this would be the sum of all visit types for each month. Instead of splitting the data set to give a subset, this study will filter the data to only have records for two of the National Parks to compare. These two National Parks are Yellowstone National Park and Great Smokey Mountains National Park.

Data analysis techniques used will be linear regression and univariate and bivariate analysis. This data will be charted and analyzed using SAS. Among the charted items will be displays of the linear regression, univariate and bivariate results. Also, a table representing the data will be used for visualization.

**The Project Findings**

The univariate for the year has a student’s t-test score of 15573.81 with a p-value of less than .0001. The coefficient Variation is 0.07033894. The univariate for the month has a student’s t-test score of 20.54043 and a p-value less than .0001. Also has a coefficient variation of 53.3311786. Univariate for recreation shows a student’s t-test of 14.27913 with a p-value of less than .0001. Then has a coefficient variation of 76.7164998. The univariate total visits student’s t-test score of 15.27254 with a p-value of less than .0001. It also has a coefficient variation of 71.7264628. We also look at the univariate data graphically for each variable. Month, Park, and Year are just a bar chart of their values. But recreation and total visits are paired as the response variable of the variable park. This is needed to be able to view a clean chart.

Now we will look at the bivariate data by looking at tables and charts. Looking at month, year, recreation, and total visit variables compared for each park. The Great Smokey Mountains National park correlates month and year at a value of -5.116E-17 and a Pr>ChiSquare of 1 for the Wald test and LR test. For Yellowstone National Park it has a value of -5.116E-17 and Pr>Chi-Square of 1 for both the Wald test and LR test. Looking at the charts for Park per year there is no difference between parks for any year. Looking at the month Park Recreation chart shows all months on an average the Great Smokey Mountain National Park has more visitors for recreation. But it can also be said for all visits as well because total visits as like results. Parks by month have no difference.

Now let’s look at the linear regression statistical model. There was no difference when looking at all three numeric independent variables to the dependent variable year as looking at just month for year. The linear regression model used the independent variables of month, recreation, and total visits. Because the variable park is not a numeric value it could not be included in the linear regression. The r squared value is 0.0326, which would be approximately 3.26 percent. This means that 3.26 percent of the model variables are correlated in the model. The equation for the linear regression model is 2019.011467 – 0.010012 \* month + 0.000001698 \* Recreation - 0.000000899 \* total visits. In other terms, coefficients mean month will decrease 0.010012 units times a year. The recreation will increase 0. 000001698 unit times year. Then total visits will also decrease 0. 000000899 unit times year. The intercept has a p-value of; less than 0.0001, the month has a p-value of 0.7972, recreation has a p-value of 0.0555 and total visits has a p-value of 0.0861.

Now let's talk about the findings from our linear regression model. Technically speaking we are looking at a multiple linear regression because of the multiple independent variables. Has stated above the R square value is 0.0326, which means the variables are not very correlated with the dependent variable of year. If you look at the bivariate data it is easy to believe that the Great Smokey Mountains National Park has more recreational visitors than Yellowstone National Park. This is shown in the chart, if you at that chart every month listed for the five years has high numbers. But we want a R squared value closer to one for favorable results. Based on the research question, in short monthly which park has more recreational visitors compared to the other. With that said, with having a P value of less than 0.0001 at the intercept we can reject the null hypothesis and accept the alternative hypothesis. As a reminder here are what the hypothesis was:

H0: Monthly Recreational Visits to the Great Smokey Mountains National Park occur no statistically significant difference from monthly Recreational Visits to Yellowstone National Park.

H1: Monthly Recreational Visits to the Great Smokey Mountains National Park occur at a statistically significant higher rate than monthly Recreational Visits to Yellowstone National Park.

What we accepted is that monthly recreational visits to the Great Smokey National Park occur at a statistically significantly higher rate than monthly recreational visits to Yellowstone National Park.

**The Project Limitations**

There are a couple of different limitations in this project. The biggest limitation here is we only used the most recent five years of data, not including the current year. This is a limitation because when the data is filtered down to include only the two parks the record count is reduced to 120 records. One hundred and twenty records is a small sample size to use in this case. Changing the number of years from five years to tens will give a better record count. Another limitation is that during the years reviewed two of the years are during a worldwide pandemic. The reason this is a limitation is because it could lead to skewed data. With the pandemic there were a lot shutdowns and National Parks were some of the only places people could visit.

**Actions proposed and Expected Benefits**

A course of action to take for this research question is to find ways to increase the advertisement dollars for Yellowstone National Park. By increasing those dollars the National Park Service can research and educate more individuals. On the topic of education, the National Park Service needs to reach out to each state’s board of education and work out a partnership to help encourage today’s youth to get outside, learn and see nature while at it. Another approach the National Park Service can do is to increase activities or sights that can be seen in the winter months and this doesn’t have to be adding inside attractions necessarily but small hiking paths or more drive-through loops at parks. Things to look at for this data in the future would be to look at the last decade worth of data instead of the last five years. Also, another thing to study in the future is to see if removing the years 2020 and 2021 to see if these years skew the results because of the COVID-19 pandemic. Another would be to separate the recreational visits variable down into more categories doing this would give a more specific result on what the National Park Service needs to do to keep a high visit number and how to detail what is being done at the parks.

The outcome of this project is to show how many visitors use National Parks. Also what type of visits are visitors doing, whether it's recreational like sightseeing or hiking versus RV Camping visitors who are most likely to want to relax and hang out. This will benefit the National Parks Service in how their budget is set because with knowing which parks have more visitors they can redirect more or less Park Rangers. Also, the outcome will help determine where and how to spend advertisement dollars towards. The most famous National Parks aren’t always the most visited. Great Smokey Mountain National Park had the second most recreational visits in 2021 with 14.1 million ("Visitation Numbers (U.S. National Park Service)", 2022). This model can be reused by just changing out the years of the data or adding on more years.

Sources

How to Deal with Missing Data. (2022). Retrieved 12 May 2022, from <https://www.mastersindatascience.org/learning/how-to-deal-with-missing-data/#:~:text=When%20dealing%20with%20missing%20data,of%20missing%20data%20is%20low>.

Stats Report Viewer. (2022). Retrieved 16 May 2022, from <https://irma.nps.gov/STATS/SSRSReports/National%20Reports/Query%20Builder%20for%20Public%20Use%20Statistics%20(1979%20-%20Last%20Calendar%20Year)>

The 25 most visited tourist spots in America. (2022). Retrieved 13 May 2022, from <https://www.newsweek.com/most-visited-tourist-spots-america-disney-new-york-california-1616737>

Visitation Numbers (U.S. National Park Service). (2022). Retrieved 12 May 2022, from https://www.nps.gov/aboutus/visitation-numbers.htm