Recreational Visits to National Park Statistical Analysis

Christopher Flowers

College of Information Technology, Western Governors University

C772 Data Analytics Graduate Capstone

Dr. Shaw

June 7, 2022

**A: Research Question**

Does Great Smokey Mountain National Park have more monthly recreational visitors than Yellowstone National Park?

**Hypothesis**: 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.

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 this paper will address is 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.

**B. Data Collection**

Data for this question was collected by the National Parks Service from visitors to the National Parks for a year. This study looks at different types of visits to the National Parks. <https://irma.nps.gov/STATS/SSRSReports/National%20Reports/Query%20Builder%20for%20Public%20Use%20Statistics%20(1979%20-%20Last%20Calendar%20Year)> From the National Parks Service.

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.

This paper at the beginning of the process was originally to compare recreational visits for the last five years at the Yellowstone and Great Smokey Mountains National Parks, however; that changed to look at the numbers at a monthly level instead of yearly. This change was to give more numbers to compare. Challenges in this process were overcome by adapting the research or program to work in the best way possible. For example, a specific example we can look at how the data set changed completely to satisfy the need to have more numbers to compare in the analysis, The methodology used to collect data was to use the query builder on the National Parks Service website then exporting that table to excel to have the data to use. The figure below will show an example of how the National Parks Service displays the data used.

Graphical user interface, application, table

Description automatically generated

**C. Data extraction and Preparation**

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. Is to create a new SAS program file to use. Before adding libname’s to the new SAS program file, we created a folder called Capstone under the Home folder. In this folder, we have multiple folders like data files, Graphs, input, output, programs, and Statistical Models. Then in that SAS file, we want to create two libname’s to use in this program. These libnames are called Parks and Results. The Parks libname is to handle all the working tables and the Results holds just final data sets. Now with the libnames set for us to use we can move to import the spreadsheet. But before we import it into SAS we have to upload the spreadsheet to SAS Studio. Because a file that does not exist can’t be used. Now we can write the import procedure statement to import the spreadsheet into SAS with the DBMS specified to xlsx and the data table to be called Parks.monthlyData and then we run that section of the code to now import the data into the input folder aliased as Parks. Then we create the just created SAS table by running a print procedure statement. The last thing we do before cleaning the data is to run a content procedure on the Parks.MonthlyData to see what the variables we have in the data set and what their formats and data types are.

This data will be 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. The quality of this data is good. 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 there will be changes to the formats that the data are in. SAS initially shows as dollars instead of numeric like it should be. There will be aggregate columns created for the summary of each visit type for the five years. For example, an aggregate column could be called total Visits, this would be the sum of all visit types for each month. This is a change from the idea of using an aggregated column for each but it was changed to monthly to match the research question and hypothesis’. 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.

Since we have explained what will be done to clean the data. Now we explain the code that was written to clean the data set. The first thing done is to write the data step which will be the new SAS table created to store the changes made from the Parks.monthlySummaryData table. The next step is the set step which is where we list the table we have wanting to effect, in this case, its Parks.monthlyData. Using the rename function we change the name of the RecreationVisits column to just recreation, this was done because it is the main visit type that will be used. Next, we drop all columns which are not needed or relevant to this study and those dropped using the drop function are ParkType, NonRecreationHours, Region, State, UnitCode, and RecreationHours. Now we want to fix the formatting of the numeric columns showing as dollars in SAS will be changed to comma15.2 format. Before fixing the format of the numeric columns we tried to run a SUM function using the of operator for all the numeric columns to give the monthly data sum of all visits, but because of the dollar format showing it errored out in the SAS log. But now that we have changed those formats we can create our aggregated column as stated above. The last step in this data statement is to check each numeric variable for missing data and put the notice in the SAS log but there were no missing data for this data set. The next statement in the program to clean the data is the sort procedure statement using the Parks.monthlyData. In this sort of statement, we are removing any duplicate records by listing nodup in the procedure statement. We are sorting by the park, year, and month. The sorted data will be stored in parks.MonthlySorted table. Another data statement is used to drop all variables that are not needed in the dataset using the parks.monthlySorted and the new table will be Parks.monthlyClean. Now that we have a base clean dataset we will use another data statement to create a table in the Results libname, which is where we are holding the results. Then we will do a print procedure to make a report of the data. After that, we use another data step to filter the data set to only show the two parks which are relevant to the null hypothesis. The data step is filtered by using a where step with an in operator. Now we will do another print procedure to show the data from the last data statement. The final step in cleaning the data code is a data statement to create a results table for the filtered monthly data called Results.filteredMonthlyData then a print procedure to make a report of the data.

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. This data could be used by the National Parks Service to determine how they market its National Parks to prospective visitors.

Linear Regression is the correct statistical method to use for this analysis. This is correct because it will allow for the prediction of what type of visits to National Parks occur the most. With the prediction, the National Parks Service can better prioritize new additions to its parks. As mentioned this data will be analyzed using the language SAS. The code will be written in SAS Studio versus using other versions of SAS like SAS Enterprise. SAS is the correct language to use for this analysis because of the amount of data being used combined with its ability to run advanced statistical methods without overloading the interpreter. SAS is very helpful with large datasets because of how fast it can process large amounts of data. In SAS the ability to easily write procedures to provide univariate analysis, graphs, means, and more. Another reason SAS is correct to use for this study is for its ability to filter data. Filtering data is something R or Python would not be able to accomplish.

Calendar

Description automatically generated with medium confidence

Screenshot of the original excel spreadsheet to compare to the updated screenshot below.

A screenshot of a computer

Description automatically generated with low confidence

Spreadsheet with the top three rows removed.

Graphical user interface, application, Word

Description automatically generatedBlank SAS file to use

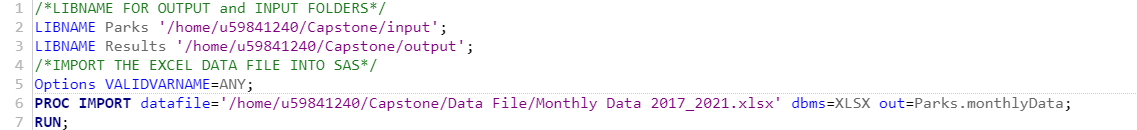


Libnames initialized

Graphical user interface, text, application, email

Description automatically generated

Showing the Monthly Data 2017\_2021.xlsx is added to SAS Studio and the folders created to store files and tables.



Proc import statement is used to import the data.

Graphical user interface, text, application, email

Description automatically generated

Proc Print to test the import data was successful

Table

Description automatically generated

Data that was imported into SAS

Graphical user interface, text, application, email

Description automatically generated

Checking the contents of Parks.monthlyData

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

Findings of the contents of Parks.monthlyData

A close-up of a document

Description automatically generated with low confidence

Parks.monthlyData showing the dollars on numeric columns.

Graphical user interface, text, application

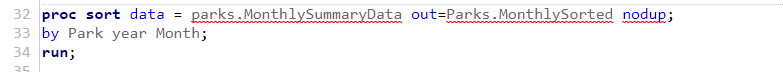
Description automatically generated

Data step in cleaning the data drop, rename, and more.

Table

Description automatically generated

Showing the Parks.monthlySummaryData



Sort statement

Table

Description automatically generated

Showing the sorted data in Parks.monthlySorted

Graphical user interface, text, application, email, Teams

Description automatically generated

Code for the drop statement, filtered data, and more.

Table

Description automatically generated with medium confidence

Parks.monthlycleandata table

Table

Description automatically generated

Parks.monthlycleandata report

Table

Description automatically generated

Report of Parks.filteredMonthlyData

Table

Description automatically generated with medium confidence

Report of Results.filterMonthlyData

**D. Analysis**

Now we will begin to look at the analysis of this research question. The first step in looking at the analysis is to look at the univariate data. 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 a 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.

A picture containing shape

Description automatically generated

Univariate procedure code

Univariate of year is the next two screenshots

Table

Description automatically generated

Table

Description automatically generated

Univariate of month is the next two screenshots

Graphical user interface, application, table, Word

Description automatically generated

Table

Description automatically generated

Univariate of recreation is the next two screenshots

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

Univariate for total visits are the next two screenshots below

Graphical user interface, table

Description automatically generated

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Univariate month chart and code for the chart

Graphical user interface, application, Word

Description automatically generated

Univariate year chart and code for the chart

Graphical user interface, application, Word

Description automatically generated

Univariate chart for park and code for the chart

Graphical user interface, text, application, email

Description automatically generated

Code for charts of univariate recreation and total visits.

Chart

Description automatically generated

Recreation Univariate for the two parks

Chart

Description automatically generated

Total Visits univariate for the two parks

Graphical user interface, text, application, email

Description automatically generated

Code for bivariate data

Graphical user interface, table

Description automatically generated

Table statistics for Great Smokey Mountain NP

Chart, histogram

Description automatically generated

Charts for variables correlation

Graphical user interface, application, table

Description automatically generated

Shows the Yellowstone NP table statistics

A picture containing chart

Description automatically generated

Charts for variables correlation

Graphical user interface, text, application, email

Description automatically generated

Code for charts of recreation and total visits compared to the dependent variable of year.

Chart, bar chart

Description automatically generated

Recreation per year for each Park

Chart, bar chart

Description automatically generated

Total Visits per year for each park

Text

Description automatically generated

Code for month per year per park

Chart, bar chart

Description automatically generated

Month park per year

Graphical user interface, text, application

Description automatically generated

Code for parks per year

Chart, bar chart

Description automatically generated

Chart for park per year

Graphical user interface, text, application, email

Description automatically generated

Code for year, total visits, and recreation grouped by month per park.

Chart, bar chart

Description automatically generated

Month Park Recreation chart

Chart, bar chart, histogram

Description automatically generated

Month Park Total visits

Graphical user interface, text, application, email

Description automatically generated

code for Linear Regression

Graphical user interface, application, table

Description automatically generated

Part one of linear regression

Table

Description automatically generated

Part two of the linear regression

**Chart

Description automatically generated**

Charts for observed predicted for the year

Diagram

Description automatically generated

FIT year diagnostics

A picture containing graphical user interface

Description automatically generated

Residuals

**E. Data Summary and Implications**

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.

Now let's talk about the findings from our linear regression model. As stated earlier in there was no difference when looking at the linear regression model using three independent variables versus using only one independent. 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. 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.

In conclusion, this study was to look at visits to the Great Smokey Mountains National Park compared to Yellowstone National Park. This study was performed using SAS for everything from importing the data into SAS to being able to evaluate. To clean the data and separate it down to just these two parks. SAS was the appropriate technology to use because of its ability to filter data along with running advanced statistical analysis. Based on the findings of this report we can reject the null hypothesis because the p-value is 0.0001, which is less than .05. This means monthly the Great Smokey Mountains National Park had more recreational visitors than Yellowstone National Park. Ways this study can be used is to look at what attractions are being used at what time of the year and find ways to increase visits in the slow months. A further course of study would be to break the recreational visit variable into more categories.

Sources

4 Types of Data in Statistics. What Are They?. (2022). Retrieved 13 May 2022, from <https://www.chi2innovations.com/blog/discover-data-blog-series/data-types-101/>

5 Min Tutorial – SAS Linear Regression with PROC REG Procedure. (2022). Retrieved 13 May 2022, from <https://data-flair.training/blogs/sas-linear-regression/#:~:text=Linear%20regression%20in%20SAS%20is,is%20referred%20to%20as%20Y>.

D, V. (2022). Multivariate analysis: an overview - Students 4 Best Evidence. Retrieved 13 May 2022, from <https://s4be.cochrane.org/blog/2021/09/09/multivariate-analysis-an-overview/#:~:text=Multivariate%20analysis%20is%20defined%20as,and%20their%20structure%20are%20important>

Graph Samples Gallery. (2022). Retrieved 13 May 2022, from <https://support.sas.com/en/knowledge-base/graph-samples-gallery.html>

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>.

SAS Help Center. (2022). Retrieved 13 May 2022, from <https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_reg_gettingstarted01.htm>

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

What is Linear Regression? - Statistics Solutions. (2022). Retrieved 13 May 2022, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/