**Data Analytics Capstone Topic Approval Form**

**Student Name:** Christopher Flowers

**Student ID: 001448877**

**Capstone Project Name:** Recreational Visits to National Park Statistical Analysis

**Project Topic**: Bivariate Analysis of National Parks Visits

x**This project does not involve human subjects research and is exempt from WGU IRB review.**

**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.

**Context:** 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 different reasons each year. Those types of visits include recreational visits, non-recreational visits, RV campers, and tent campers. The linear regression statistical method will be used to determine if national parks have more recreational visits compared to other types of visits. 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:** 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)> -> National Parks Service

This data set has a continuous dependent variable is Year. These correspond with the year the data is from. Listed below are the independent variables from the data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field | Type | Field |  | Type |
| Park | Categorical | Miscellaneous Overnight Stays |  | Continuous |
| Recreation Vistors | Continuous | Non-Recreational Vistors |  | Continuous |
| Recreation Hours | Continuous | Non-Recreational Hours |  | Continuous |
| Concessioner Lodging | Continuous | Concessioner Camping |  | Continuous |
| Tent Campers | Continuous | RV Campers |  | Continuous |
| Backcountry Campers | Continuous | Non-Recreation Overnight Stays |  | Continuous |
| Unit Code | Categorical | Park Type |  | Categorical |
| Region | Categorical | State |  | Categorical |
| Monthly | Continous |  |  |  |

This data is public information collected and owned by the National Parks Service. There are over two thousand records in the data set. This data is limited because it only goes back five years.

**Data Gathering:** This data will be cleaned by removing all blank 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. 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 Recreational Visits 2021, this would be the sum of all recreational visits for 2021. 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.

**Data Analytics Tools and Techniques**: 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.

**Justification of Tools/Techniques:** 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.

**Project Outcomes**: 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.

**Projected Project End Date**: 6/25/2022

**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/

**Course Instructor Signature/Date:**