Part 4 - Project Final Report

By Mithali Shashidhar

Introduction

With its distinguished lakes and rivers, summertime sand dunes, and year-round options for skiing, snowboarding, fishing, and hunting, Lewiston, Idaho is well known for its outdoor activities. The city's location alongside the Snake and Clearwater rivers makes it the perfect place for water sports including kayaking, boating, and fishing. The city provides thrilling recreational opportunities like whitewater rafting and exploring Hells Canyon, one of North America's deepest river gorges. Having had a lot of outdoor adventures, Lewiston is a destination with a myriad of recreational possibilities.

But weather patterns can shift, and wildfires can worsen the quality of the air, which can have a big effect on the very things that make Lewiston so appealing. These changes in the environment may bring with them challenges and changes to the region's outdoor appeal, particularly its water-based activities and picturesque landscapes. Smoke from wildfires can have a significant impact on Idaho's tourism industry. It can reduce the number of visitors to outdoor recreation areas, such as national parks, forests, and wilderness areas, as well as impact the health of both humans and wildlife.

Problem Statement

In this research study, I want to analyze and quantify the impact that smoke from wildfires have on the tourism industry and outdoor recreation in Lewiston. The research uses several data sources that represent the tourism industry and outdoor recreation and studies the correlation between these indicators and wildfire smoke. The findings will contribute to a better understanding of the tourism implications of wildfire smoke exposure in a localized context, informing strategies to minimize the impact

Motivation

Given the strong link that exists between outdoor recreation and the identity, well-being, and economic stability of the region, the study that I am proposing is important. The goal of this study is to present a thorough understanding of the potential effects of wildfires on the tourism industry and outdoor recreation in Lewiston, Idaho. I aim to discover adaptation techniques and policy recommendations by exploring the implicit benefits of outdoor recreation, which include gains in well-being, as well as the explicit benefits, which include monetary revenue. These insights will be critical to maintaining Lewiston's outdoor recreation industry's sustainability and strengthening the community's ability to withstand environmental setbacks. This investigation, conducted in cooperation with local communities, government agencies, and researchers, aims to produce useful data that will protect Lewiston's identity as a popular travel destination and maintain the core of its outdoor experiences for future generations.

Background/Related Work

What other research has been done in this area? How does this research inform your hypotheses, your analysis, or your system design? What are your hypotheses or research questions?

This study was inspired by the research paper titled Idaho Climate-Economy Impacts Assessment Recreation and Tourism Report by Alexander Maas and Katherine Himes which talks about how changing weather and climate patterns affect the recreation and tourism industry in Idaho. The paper discusses how the duration of natural snow cover in Idaho changed over time and some potential opportunities for Idaho's winter sports industry in light of climate change as well as How smoke from wildfires impact Idaho's tourism industry.

The research done by this paper illustrates that smoke and wildfire have a direct impact on tourism and outdoor recreation. Specific examples include its threats to mountain biking recreation, resorts that are located in areas of high wildfire risk and overall summer recreation growth.

I could not find any existing models in this space, however, I came across several datasets that I could use to answer my research questions and perform analysis. I chose to analyze 3 datasets which all represented the tourism industry in Lewiston and the larger Idaho region to conduct my analysis. The datasets are:

1. Bureau of Transportation - Airline On-Time Statistics and Delay Causes
    a. Description: This dataset is published by the Bureau of Transportation and contains information about flights that are delayed versus those that are on time. For the purpose of this analysis, I am concerned with the total number of flights arriving at the Lewiston airport and not particularly with delay statistics and causes. Therefore, this dataset can be used for that purpose.
    b. Reason behind choosing: I chose the aircraft activity dataset as this provided information about the number of flights both departing and arriving into Lewiston and this represents the tourism interest in Lewiston to some degree and would be interesting to see if there was any correlation between this dataset and the smoke estimate.
2. Idaho Park Visitation and Recreation Statistics - Visitation Statistics
    a. Description: This dataset is published by the Department of Parks and Recreation of Idaho. The data resides in a small section of the link posted above under the heading: Visitation Statistics. It counts the number of visitation totals representing combined usage between day-use and overnight camping for state parks in Idaho.

   b. Reason behind choosing: I chose the Idaho park visitation data as this represents the onsite count of people visiting and recreating in the outdoors in the State. With increased smoke in the region, I wanted to understand if this hinders individuals from venturing out into the outdoors and recreating.
3. U.S. Bureau of Labor Statistics - Employment Data for the Leisure/Hospitality Industry (1990 - 2023)
   a. Description: This dataset is published by the US Bureau of Labor Statistics. It records employment numbers (in thousands) for each month from 1990 to 2023.
   b. Reason behind choosing: The employment numbers (in thousands) for the Leisure/Hospitality industry was used to understand the correlation between increased wildfire smoke in the leisure industry. Employment numbers provide a quantitative estimate of the industry and the amount of work going on.

Methodology

This study comprised two phases. Phase I involved analyzing data from wildfires within a 1250-mile radius of Lewiston between 1963 and 2023 to create a smoke estimate. Phase II involved analyzing the impact of this smoke on the leisure/hospitality industry and outdoor recreation in Lewiston. At both these stages human centered considerations were applied. In both stages, principles such as data integrity, license considerations, and data acquisition were implemented.

Phase I

In Phase I, we utilized data from two sources after being assigned a city in the United States for analysis from this sheet. The other two data sources include a geojson file dataset collected and aggregated by the US Geological Survey. This dataset contains fire polygons for wildfires that have occurred across the USA and an API that returns a response to obtain the Air Quality Estimates for regions in the USA.

The wildfire polygon dataset was used to filter out polygons in a 1250 radius from Lewiston, Idaho between 1963 and 2023. Once the wildfires had been filtered out, the goal was to create a smoke estimate that would quantitatively capture the impact of these fires and the smoke released from them. This estimate was created using some of the attributes from the wildfire data - distance, area of the fire, and the fire_type.

The distance from the fire was chosen as it would intuitively be inversely proportional to the smoke generated in that area. The area of the fire is indicative of the expanse and intensity, directly proportional to the amount of smoke it generates. Lastly, the fire type is a field in the dataset specifying if the fire is a wildfire or a prescribed fire. Weights are assigned based on this categorization, as prescribed fires are perceived as more controlled and mild compared to wildfires.

The fields used to create the smoke estimate are as follows:

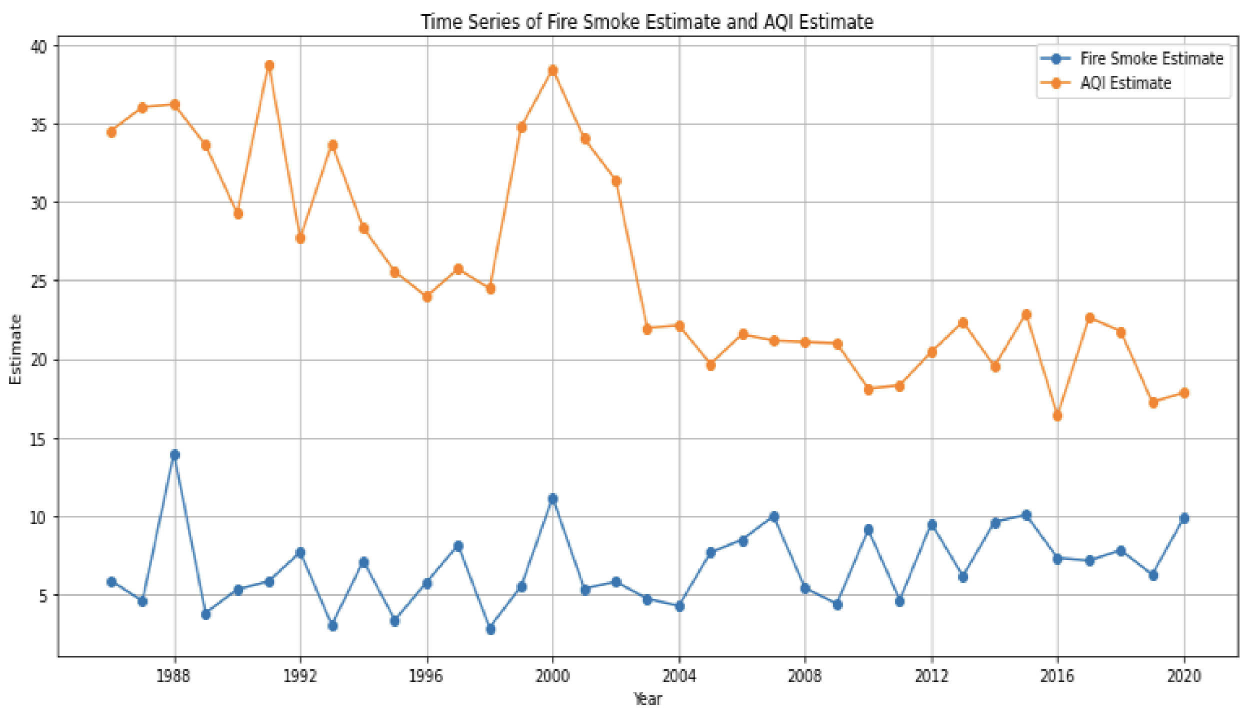
1. GIS_Hectares: The GIS-calculated hectares of the fire polygon, calculated using the Calculate Geometry tool in ArcGIS Pro.
2. Distance: The average miles from the fire boundary to Lewiston, Idaho.
3. Assigned_Fire_Type: An attribute that can take 5 values.

The values where the attribute has the word "likely" also leave room for a scenario where this fire may not have occurred at all. However, when it takes values like "Prescribed Wildfire" and "Wildfire," there is some assurance that this fire occurred.

Based on this, I believe that the values this attribute takes can be weighted based on the assurance of a fire taking place:

Assigned_Fire_Type Value	**Description**	**Score/Weight Assigned**
Likely Wildfire	The fire was likely a wildfire, but there is no way to confirm this.	15 (The assurance of the wildfire occurring is not given)
Wildfire	The fire is a confirmed wildfire based on available attributes.	20 (Assured wildfire)
Prescribed Wildfire	The fire is a confirmed prescribed fire based on available attributes.	12.5 (Assured prescribed wildfire, which burns fewer acres than compared to wildfires)
Unknown - Likely Wildfire	This polygon came from MTBS and was labeled "Unknown." Research by	20 (Research indicates that this is an assured wildfire)

	Karen Short indicates that the fire is likely a wildfire.	
Unknown - Likely Prescribed Wildfire	This polygon came from MTBS and was labeled "Unknown." Research by Karen Short indicates that the fire is likely a prescribed fire.	12.5 (Research indicates that this is an assured prescribed wildfire)



Once I created the smoke estimate and extracted values of AQI for Lewiston between the years 1963-2020, I found that the two exhibit a similar trend. Specifically, when the smoke estimate spikes, so does the AQI, but this is not always the case.

Summary from Phase I:

Phase I aided in the creation of a smoke estimate for the Lewiston region using wildfire data. Additionally, it allowed me to predict the future impact of the smoke by forecasting the smoke estimate in the future through time series analysis.

Phase II

In phase II, the idea was to extend this analysis and to help inform the city council, city manager/mayor, and city residents about the potential future impacts of smoke on their community. Using the smoke estimate created in Phase I the idea was to estimate a specific social or economic impact on the city. Since Lewiston is known for its scenic and picturesque beauty and has many opportunities for outdoor recreation, I chose to study the impacts on tourism and services that cater to tourism.

The datasets that I used to extend this analysis have been described above and the research questions that I chose to answer using this are:

- How does the Leisure/Hospitality industry in Lewiston, Idaho, respond to heightened levels of smoke in the area?
- Do flights in and out of Lewiston correlate with the amount of smoke in the area?
- Is national park visitation in Idaho impacted by smoke?

The idea was to first identify any correlations between these data points and the smoke estimate.

Findings

I. Correlations

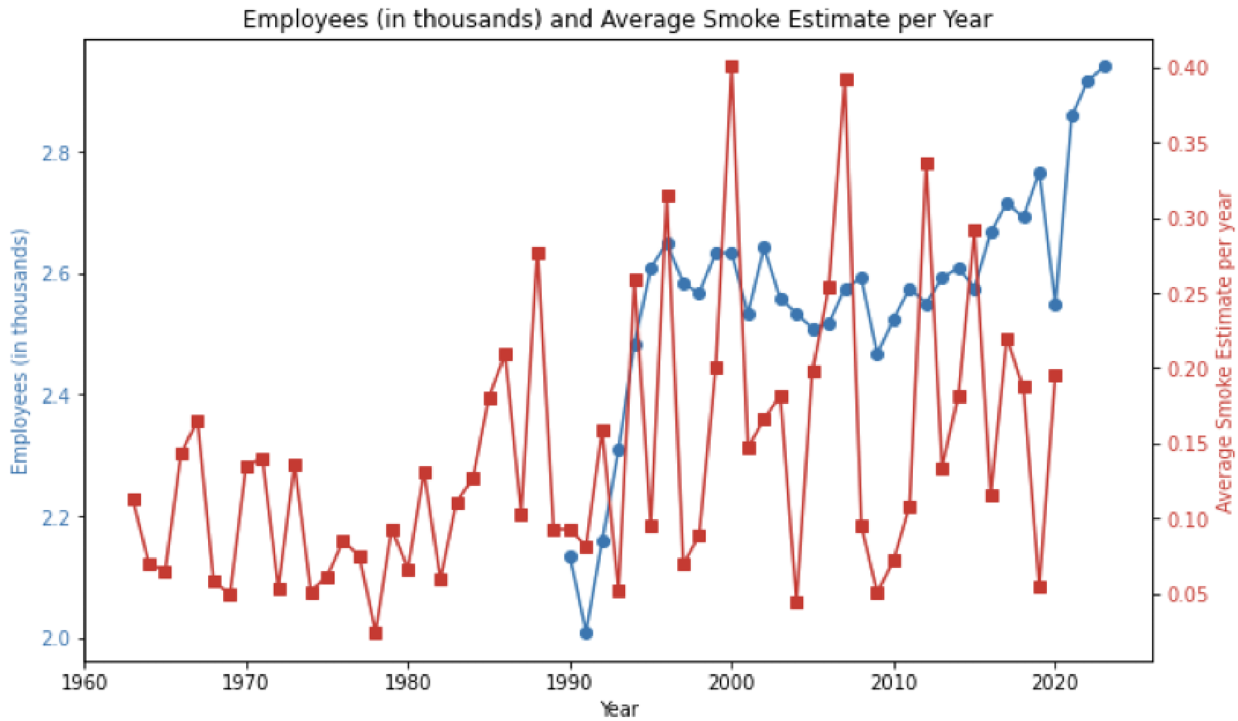
Research Question 1: How does the Leisure/Hospitality industry in Lewiston, Idaho, respond to heightened levels of smoke in the area?

Identified Correlation Coefficient: 0.25

Observations:

The graph below features the years from 1960 to 2020 on its x-axis. It incorporates a dual y-axis setup, with the left axis labeled in blue representing employment in the leisure industry (in thousands) and the right axis labeled in red representing the average smoke estimate per year. The graph includes two lines: the blue line tracks employment in the leisure industry over the years, while the red line represents the average smoke estimate over the same period.

In general, elevated levels of smoke appear to coincide with an increase in employment and vice versa, but the correlation (0.25) is weak.



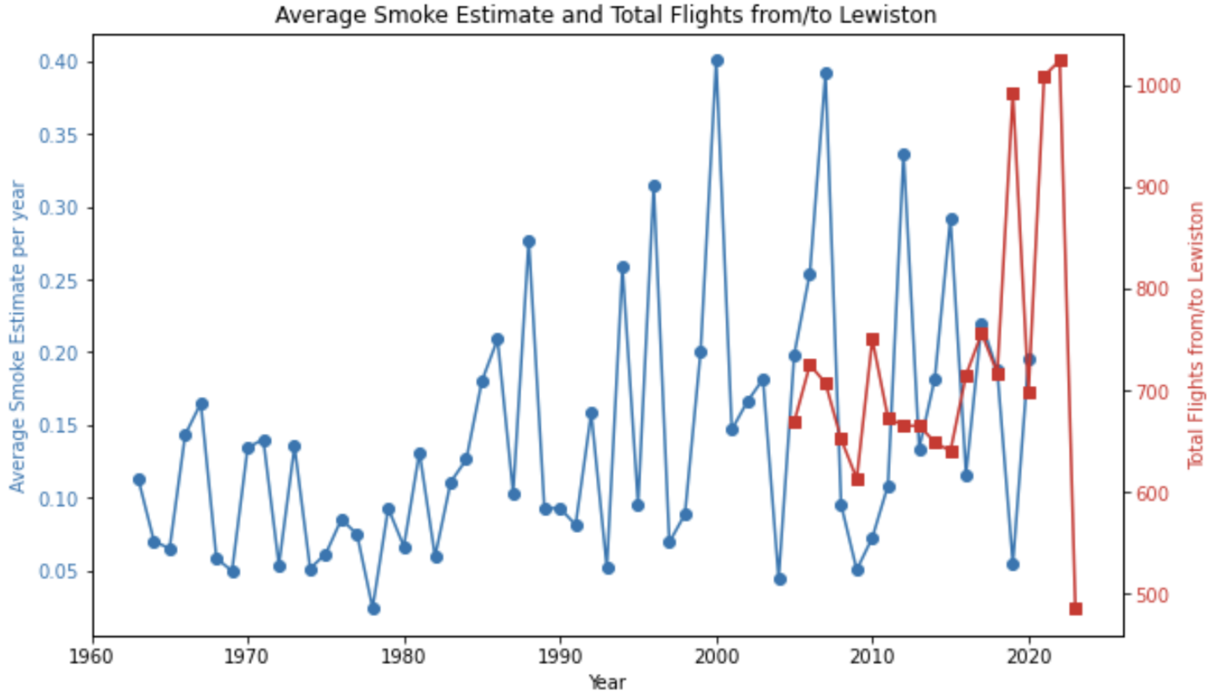
Research Question 2: Do flights in and out of Lewiston correlate with the amount of smoke in the area?

Identified Correlation Coefficient: -0.24

Observations:

The graph below features the years from 1960 to 2020 on its x-axis. It incorporates a dual y-axis setup, with the left axis labeled in blue representing the average smoke estimate per year and the right axis labeled in red representing the overall or total flight activity in Lewiston. The graph includes two lines: the blue line tracks average smoke estimates over the years, while the red line represents the total flight activity over the same period.

The calculated correlation coefficient of -0.24, coupled with the visualization, suggests a weak negative correlation between flight activity and smoke. This implies that an increase in smoke might cause a decrease in flight activity.



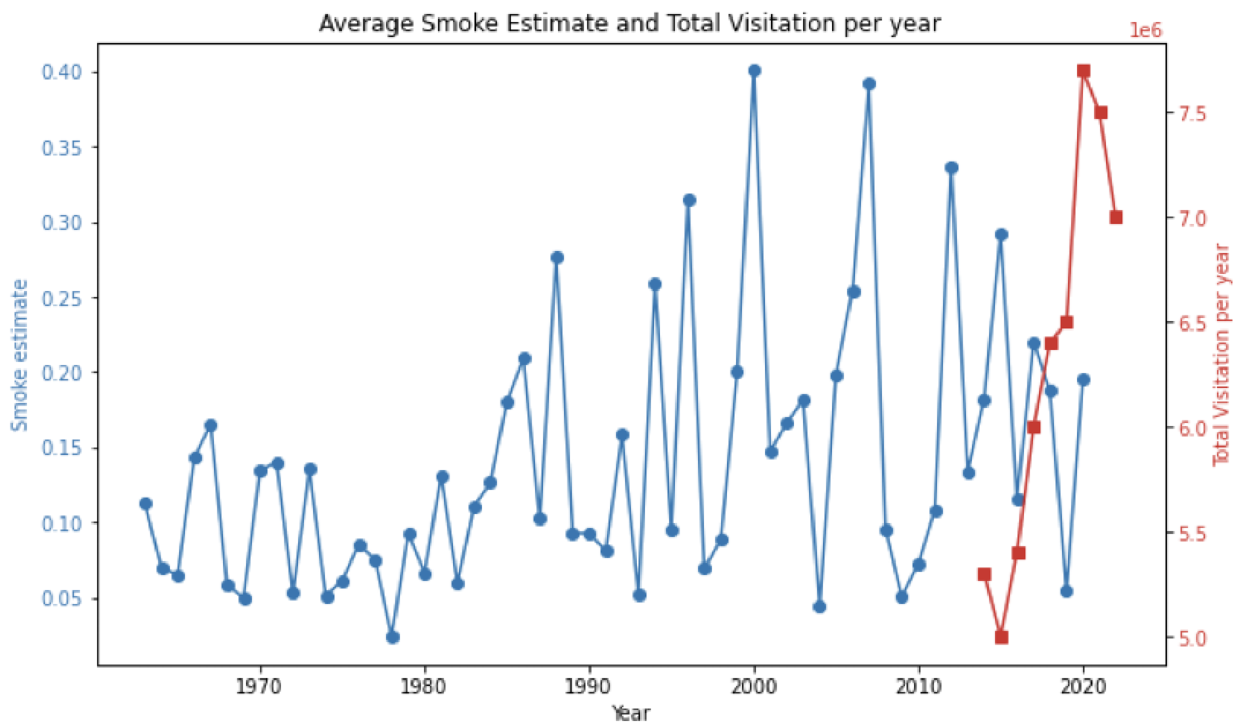
Research Question 3: Is national park visitation in Idaho impacted by smoke?

Calculated Correlation Coefficient: -0.26

Observations:

The graph below showcases the years from 1960 to 2020 on its x-axis. It features a dual y-axis setup, with the left axis labeled in blue representing the average smoke estimate per year, and the right axis labeled in red representing the total visitation in Idaho parks (in ten to the power six). The graph includes two lines: the blue line tracks average smoke estimates over the years, while the red line represents park visitation over the same period.

The park visitation appears to decrease with an increase in smoke, and vice versa, but the correlation (-0.26) between the variables is weak.



II. Modeling

With the correlations indicating an undeniable link between the smoke estimate and various tourism industry indicators, my interest turned towards modeling this relationship, predicting future impacts, and gauging the effectiveness of such modeling.

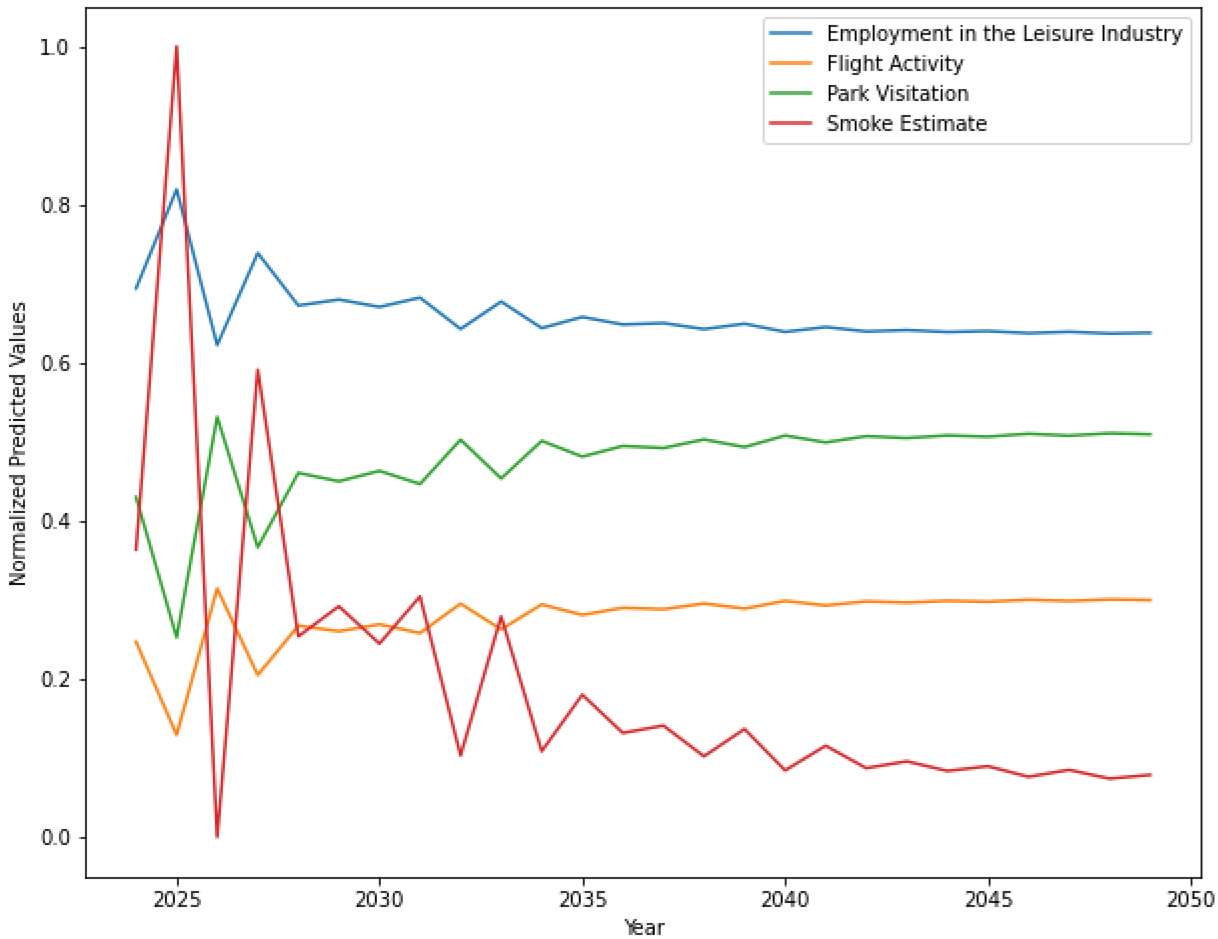
I employed a Linear Regression model, utilizing the smoke estimate as the predictor and the leisure industry indicator as the response to fit three models:

1. Model 1: employment_in_the_leisure_industry ~ B0 + B1 * smoke_estimate
    a. MSE: 0.04
    b. MAE: 0.14
    c. R-Squared: 0.06
    d. B0 (intercept): 0.62
    e. B1 (coefficient): 0.19
    f. Observations:
        i. The model exhibits a moderate performance with a relatively low MSE and MAE.
        ii. The R-squared value of 0.06 indicates that the model explains a small portion (6%) of the variance in employment in the leisure industry.
        iii. The positive coefficient (B1) for smoke_estimate (0.19) suggests a positive relationship. As the smoke estimate increases, employment in the leisure industry tends to increase, although the effect is relatively weak.

2. Model 2: flight_activity ~ B0 + B1 * smoke_estimate
    a. MSE: 0.05
    b. MAE: 0.14
    c. R-Squared: 0.06
    d. B0 (intercept): 0.31
    e. B1 (coefficient): -0.18
    f. Observations:
        i. Similar to the first model, this model demonstrates moderate performance with a slightly higher MSE.
        ii. The R-squared value of 0.06 suggests that the model explains only a small proportion (6%) of the variance in flight activity.
        iii. The negative coefficient (B1) for smoke_estimate (-0.18) indicates a negative relationship. An increase in smoke is associated with a decrease in flight activity, although the effect is relatively weak.
3. Model 3: park_visitation ~ B0 + B1 * smoke_estimate
    a. MSE: 0.09
    b. MAE: 0.24
    c. R-Squared: 0.07
    d. B0 (intercept): 0.53
    e. B1 (coefficient): -0.27
    f. Observations:
        i. This model has a higher MSE and MAE, indicating a less accurate fit compared to the other two models.
        ii. The R-squared value of 0.07 suggests that the model explains a slightly larger portion (7%) of the variance in park visitation.
        iii. The negative coefficient (B1) for smoke_estimate (-0.27) suggests a negative relationship. An increase in smoke is associated with a decrease in park visitation, and the effect is somewhat stronger compared to the other models.

Overall Comparisons:

- All three models have relatively low R-squared values, indicating that a substantial portion of the variability in the response variables is unexplained by the predictors.
- The coefficients (B1) for smoke_estimate in all models suggest some level of association between smoke and the respective response variables, though the relationships are generally weak.



Using the three models created above and the predicted values of the smoke estimate from Phase I of the analysis as my predictor in the model, I predicted Employment in the Leisure Industry, Flight Activity, and Park Visitation for the next 25 years. This graph shows some variation in estimates for the first ten years roughly, and then it smoothes out. The models are trained on a very small dataset and are too simplistic for this modeling. However, the calculated coefficients provide some degree of understanding of the association between wildfire smoke and these variables.

Discussion/Implications

These findings necessitate additional data and research to model and comprehend the association between wildfire smoke and its impact on tourism. Collecting more data to measure tourism indicators would significantly contribute to a thorough understanding of potential negative impacts. While my analysis above suggests an undeniable association between tourism industry variables and wildfire smoke, the strength of this association is not robust enough for definitive conclusions. Furthermore, obtaining more granular visitation data for national parks near Lewiston and outdoor recreation activities within Lewiston would enhance monitoring and facilitate a more comprehensive analysis.

For instance, studies have explored the use of social media to monitor and model visitation estimates at parks, employing novel models to understand recreational changes after wildfires and site closures. By employing statistical techniques like these and leveraging the power of social media, visitation estimates can be obtained at a more granular level, offering a better explanation of the impact of wildfire smoke on visitation.

Additionally, human-centered principles have been integrated at every stage of this project, serving as the core ethos of the analysis. To ensure data privacy and integrity, I carefully reviewed the licenses of all data sources used in this analysis, prioritizing those from official bodies such as the Bureau of Labor Statistics, Idaho Parks and Recreation, and the Bureau of Transportation. Recognizing the potential impact on communities, I maintained mindfulness regarding the use of accurate and fair data.

Moreover, I underscored the importance of communicating findings in an accessible manner. All results are presented through a combination of graphs and explanatory text to cater to diverse audiences. It is essential to acknowledge that this study has inherent limitations due to data constraints and time limitations, which prevented further exploration of more complex machine learning techniques.

Limitations

- The lack of tourism industry and outdoor recreation data hinders the ability to make stronger claims and generalize results.
- The data provided by Idaho Parks and Recreation includes cumulative data for parks in Idaho and the analysis of such data may not directly apply solely to Lewiston.
- The collected data covers only shorter periods, failing to span a significant number of years as intended for this project.
- I attempted a Linear Regression analysis with normalized variables, acknowledging that the assumptions made by linear regression models might have been too stringent for this analysis. Given my beginner-level approach to statistical analysis, I am uncertain whether the assumptions of linear regression are fully satisfied, and I cannot confirm the presence of a linear association between the predictor and response variables. Nevertheless, I undertook this modeling effort as a foundational step in statistical analysis.
- The wildfire dataset provided by the USGS does not include data for the years 2021, 2022, and 2023 within a distance of 1250 miles from Lewiston, Idaho.
- In the AQI API, there is no information available for gaseous matter at stations near Lewiston, Idaho, even within the bounding box.

- There is no station directly in Lewiston, Idaho, in the AQI API, and the stations within the bounding boxes also provide information for a limited period, specifically from 1986 to 2023.
- The smoke estimate values range from 3.084128e-10 to 1.435315e+02. Most values are extremely small because many fires burn very small areas located at a significant distance from the city of interest. This potentially lowers the average smoke_estimate, despite the possibility of poor air quality.

Conclusion

The objective of this study is to quantify and analyze the impact of wildfire smoke on the tourism industry in Lewiston. I used three data sources that I think could best capture the tourism industry and outdoor recreation in Lewiston - visitation data to parks, flight activity data at the Lewiston airport and Leisure industry employment. My hypotheses were to understand the correlation and association between these three aforementioned variables and wildfire smoke. The findings from calculating the correlation coefficient between the variables and performing linear regression to obtain regression coefficients state that there does exist a relationship between the two variables but their association is weak. In general, I observed that with an increase in wildfire smoke, there was:

1. A mild increase in employment in the Leisure industry
2. A slight decrease in the visitation numbers for parks in Idaho
3. A slight decrease in flight activity for the Lewiston airport

Future work for this study would include gathering more data that act as indicators of tourism and outdoor recreation as the associations made here are weak and made with minimal data as there wasn't data available for a prolonged period of time. The reader's comprehension of human-centered data science will be enhanced by the examination of fairness, privacy, accessibility, and transparency in this study.

References

a. Alexander Maas and Katherine Himes (2021). Idaho Climate-Economy Impacts Assessment Recreation and Tourism Report.
b. E. M. White, S. G. Winder, S. A. Wood (2023). Applying novel visitation models using diverse social media to understand recreation change after wildfire and site closure. https://www.tandfonline.com/doi/full/10.1080/08941920.2022.2134531.

Data Sources

The additional data sources used in this analysis are:

1. Bureau of Transportation - Airline On-Time Statistics and Delay Causes
    a. License: CC-BY
        i. DOT strongly encourages researchers to deposit data under the Creative Commons CC-BY Attribution or an equivalent license, whenever possible.
        ii. The CC-BY license allows others to distribute, remix, tweak, and build upon the work, even commercially, as long as they credit the original creator.
    b. Relevant Columns:
        i. year: the year in which the flight arrived
        ii. month: the month in which the flight arrived (values range from 1-12)
        iii. carrier_name: the name of the airline carrier of the flight that arrived at the airport
        iv. airport: abbreviation of the airport name that the flight arrived at
        v. airport_name: full name of the airport that the flight arrived at
        vi. arr_flights: the total number of flights that arrived at this airport for the specified month and carrier
    c. Note: There are 15 more columns in this dataset that specify delay causes and the number of minutes of delay. However, since this information is out of the scope and relevance of this study, it will not be used.
2. Idaho Park Visitation and Recreation Statistics - Visitation Statistics
    a. License: Custom
        i. The Idaho Department of Parks and Recreation assumes no liability for any loss that may result from the use of this data. The Idaho Department of Parks and Recreation, as a general rule, may modify or delete any of the information that is included in this data without advance notice. The Idaho Department of Parks and Recreation may also interrupt or halt access to this data when deemed necessary. The Idaho Department of Parks and Recreation shall assume no liability for any damages resulting from modification or deletion of this dataset or interruption or halting access to this data.
    b. Columns:
        i. year: the year to match with the visitation number
        ii. visitation: the visitation number for that year for Idaho state parks
3. U.S. Bureau of Labor Statistics - Employment Data for the Leisure/Hospitality Industry (1990 - 2023)
    a. License: Custom
        i. The Bureau of Labor Statistics (BLS) is a Federal government agency and everything that we publish, both in hard copy and electronically, is in the public domain, except for previously copyrighted photographs and illustrations. You are free to use our public domain material without

specific permission, although we do ask that you cite the Bureau of Labor Statistics as the source.

b. Columns:
    i. Year: the year for which employment numbers are counted
    ii. January: employment numbers for the month of January for each year
    iii. February: employment numbers for the month of February for each year
    iv. March: employment numbers for the month of March for each year
    v. April: employment numbers for the month of April for each year
    vi. May: employment numbers for the month of May for each year
    vii. June: employment numbers for the month of June for each year
    viii. July: employment numbers for the month of July for each year
    ix. August: employment numbers for the month of August for each year
    x. September: employment numbers for the month of September for each year
    xi. October: employment numbers for the month of October for each year
    xii. November: employment numbers for the month of November for each year
    xiii. December: employment numbers for the month of December for each year

The datasets used for Phase I of the project are:

1. Combined wildland fire datasets for the United States and certain territories, 1800s-Present (combined wildland fire polygons)
2. https://aqs.epa.gov/aqsweb/documents/data_api.html