Contents

Introduction

More and more frequently summers in the western US have been characterized by wildfires with smoke billowing across multiple western states. There are many proposed causes for this: climate change, US Forestry policy, growing awareness, just to name a few. Regardless of the cause, the impact of wildland fires is widespread. There is a growing body of work pointing to the negative impacts of smoke on health, tourism, property, and other aspects of society. The analysis of wildfire smoke impacts on Minot, North Dakota, is imperative due to the escalating occurrences of wildfires, resulting in significant smoke dispersion. The gravity of this study lies in its potential to uncover the adverse effects of wildfire smoke on public health, economy, and the environment. By addressing the estimations of smoke impacts over the past 60 years, the analysis aims to provide vital insights into the severity and patterns of smoke exposure faced by the city. Understanding these impacts is pivotal in guiding policy formulation for mitigating future wildfires' effects and safeguarding the well-being of Minot's residents.

Background/Related Work

Research in the domain of environmental impact on respiratory health, particularly regarding fire smoke exposure, has been extensive. Numerous studies have investigated the correlation between wildfire smoke and various respiratory illnesses, contributing valuable insights into the public health implications of prolonged exposure to smoke. These existing studies informed the hypotheses and methodologies adopted in this analysis, providing a foundational understanding of the relationships under investigation.

Research Question 1: Fire Smoke and Tuberculosis Rate Relationship

Hypothesis: Increased exposure to fire smoke is positively correlated with higher tuberculosis rates in Ward County, Minot.

Previous research, such as [ref1], suggests a link between air pollutants, including wildfire smoke, and respiratory diseases like tuberculosis. This existing body of work informs the hypothesis that increased exposure to fire smoke could contribute to elevated tuberculosis rates in the region. Utilizing statistical techniques like Pearson's correlation and OLS regression aims to assess and validate this hypothesized relationship.

Research Question 2: Fire Smoke and Influenza Rate Relationship

Hypothesis: Elevated levels of fire smoke are associated with increased influenza rates in Minot County.

Studies exploring the impact of air pollution on influenza rates have indicated a potential association between heightened pollution levels and higher susceptibility to respiratory infections. Research by [ref 2] suggests that fire smoke exposure might similarly influence influenza rates. The hypothesis is that increased exposure to fire smoke is linked to higher incidences of influenza in the region. Using regression plots and Linear Regression Models aims to validate this hypothesis and ascertain the predictive nature of fire smoke exposure on influenza rates.

Research Question 3: Gender Disparities in Hospitalization Rates

Hypothesis: Gender differences influence hospitalization rates due to variations in exposure to fire smoke in Minot County. Existing studies, such as [ref 3], highlight variations in health outcomes based on gender-specific responses to air pollution exposure. The hypothesis is that there are disparities in hospitalization rates between males and females resulting from differences in exposure to fire smoke. Employing The Mann-Whitney U Test aims to validate this hypothesis, ensuring fairness in evaluating the gender-based impacts of fire smoke exposure on hospitalizations.

During the development of this analysis, various existing models were contemplated for adaptation or adoption to suit the data and research questions. While several models offered insights into similar relationships, the choice was made based on the relevance of specific statistical techniques to the nature of the dataset and the ethical considerations inherent in the study. Additionally, extending Part 1 results involved leveraging the datasets used previously, refining and augmenting the analyses to explore the intricate relationships between fire smoke exposure and respiratory health outcomes in Minot County.

Methodology

The chosen methodology involves a meticulous aggregation of wildfire data, considering spatial proximity, fire size, and temporal dynamics. Ethical considerations significantly influenced study design, ensuring transparent and unbiased data representation. The selection of an ARIMA model for future predictions was influenced by its applicability in time series forecasting and its capacity to capture trends and patterns. The methodology prioritized human-centered data science principles, emphasizing the ethical and responsible use of data to derive meaningful conclusions.

The formula used for the fire estimate is

Fire smoke estimate = $(Assigned\ Fire\ Type\ x\ GIS\ Acres\ x\ 15.625)/(Distance)$

Where, the weights for Assigned Fire Types are:

- Wildfire: 1.0
- Unknown - Likely Wildfire: 0.9
- Likely Wildfire: 0.9
- Prescribed Fire: 0.5
- Unknown - Likely Prescribed Fire: 0.4

Additionally, an ARIMA predictive model was developed for forecasting smoke estimates over the next 25 years (2024-2049). ARIMA models are standard for time series predictions. To convey prediction uncertainty, prediction intervals with a 95% confidence level are calculated. The ARIMA model parameters, represented as (p, d, q) where p is autoregressive, d is the degree of differencing, and q is the moving average, are selected via grid search to minimize the average RMSE over 5 folds.

The methodological choices made in the extension plan analysis were not only driven by analytical objectives but also influenced by ethical considerations and a human-centered approach. The selection of statistical methods was based on the relevance of each technique to address specific research questions effectively. Pearson's correlation coefficient was chosen due to its ability to measure the linear relationship between two continuous variables, offering an initial understanding of associations between fire smoke and respiratory diseases. The use of Ordinary Least Squares (OLS) regression furthered this investigation by allowing a more comprehensive exploration of the relationship, considering potential confounding variables and providing insights into predictive aspects.

In examining the association between fire smoke and influenza rates, the decision to use regression plots and Linear Regression Models was guided by the need to visualize and quantify the potential impact of fire smoke exposure on influenza rates over time. These methods were chosen for their ability to provide a clear representation of trends and to model the predictive nature of fire smoke exposure on influenza rates.

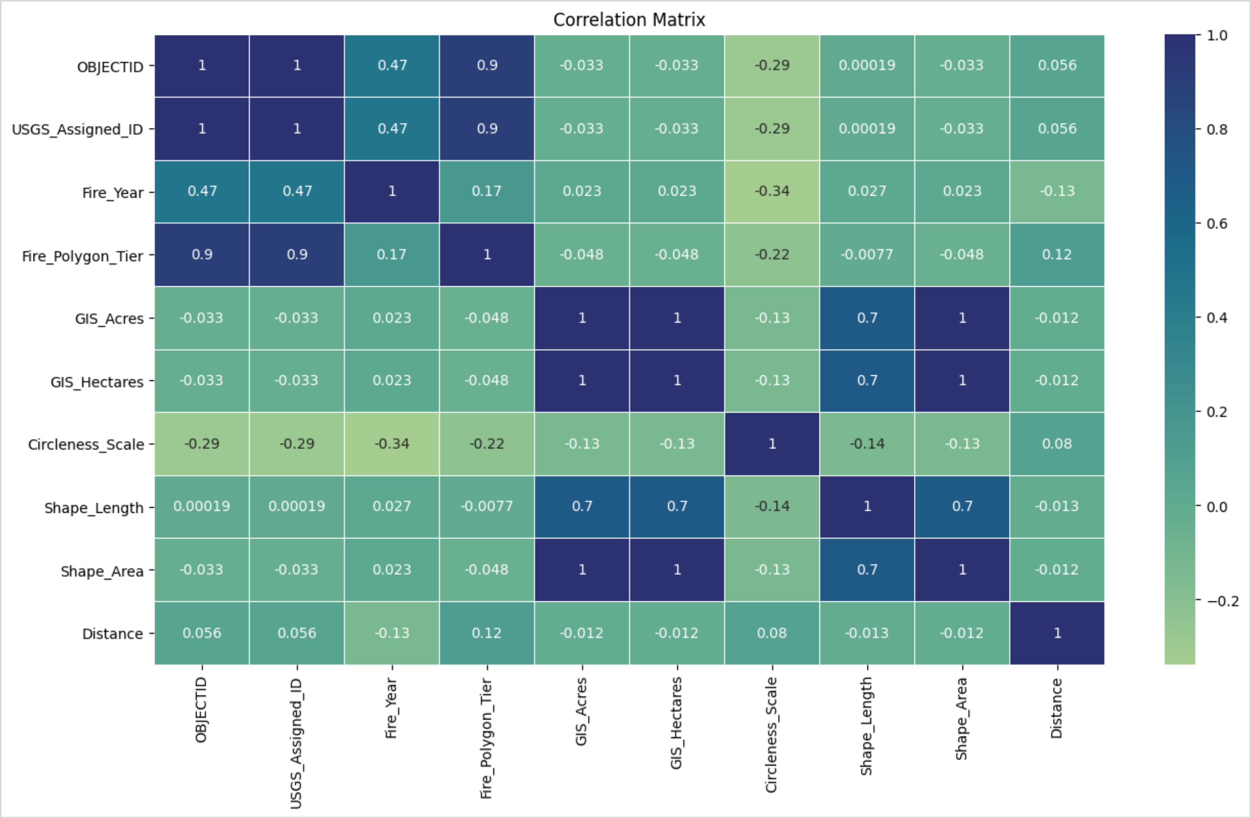
Regarding gender-based disparities in hospitalization rates, the utilization of The Mann-Whitney U Test was rooted in ethical considerations. This non-parametric test was preferred due to its robustness in handling non-normal distributions and the ethical imperative to ensure fairness in comparing hospitalization patterns between males and females based on exposure to fire smoke.

The human-centered approach was integral in the selection of methodologies that not only addressed the research questions effectively but also considered ethical implications. By choosing statistically sound methods that accommodate potential biases and maintain fairness in comparing gender-based outcomes, the study aimed to uphold ethical standards while delivering meaningful insights into the impact of fire smoke on respiratory health in Ward County, Minot.

Findings

Exploratory Data Analysis

Visualization 1



The heatmap visualization illustrates the correlations between the different factors with values relevant to the designated city, Minot, covering a span of up to 1250 miles.

The correlation matrix provided furnishes a comprehensive overview of the interrelationships among the variables under consideration, forming a critical foundation for the deliberative process in formulating a novel smoke estimate variable. A salient observation resides in the existence of a perfect positive correlation between OBJECTID and USGS_Assigned_ID, indicative of a direct but redundant association because of the context of our case-study.

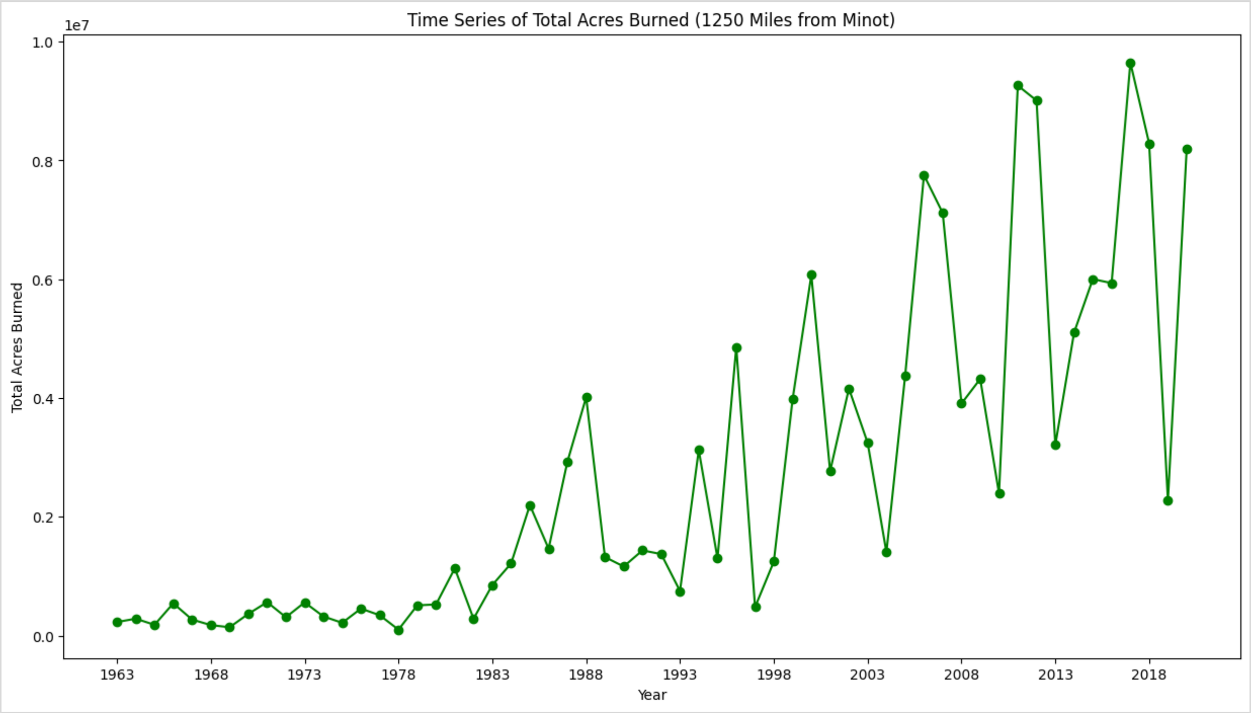
Temporal dynamics, as encapsulated by Fire_Year, display a moderate positive correlation with OBJECTID, USGS_Assigned_ID, and Fire_Polygon_Tier. This signifies a discernible temporal influence on these variables, accentuating the imperative to account for the temporal aspect in the envisaged smoke estimate. Contrastingly, the spatial variable Distance manifests minimal correlations with other variables, suggesting a limited impact on the measured parameters. In light of these spatial dynamics, on performing additional scrutiny, we did not consider them in the context of constructing the smoke estimate.

The geospatial metrics, GIS_Acres and Shape_Area, reveal a perfect positive correlation, indicating functional redundancy. Hence, we did not select these variables to avert redundancy in the envisaged smoke estimate. Noteworthy variables, such as Fire_Polygon_Tier, exhibit distinct characteristics and robust associations, thereby positing them as prospective cornerstones in the construction of the new smoke estimate.

In interpreting the graph, the color map (cmap) serves as a visual conduit delineating correlation strength and direction. The gradation of warmer hues signifies positive correlations, while cooler tones connote negative correlations. Optimization of the color map configuration enhances visual discernibility, and a judicious comprehension of the scale, ranging from -1 to 1, facilitates a nuanced understanding of the magnitude of correlations.

In summation, the presented analysis not only identifies pivotal associations and redundancies but also lays the groundwork for an informed decision-making process in the development of a bespoke smoke estimate variable. The insights gleaned from the correlation matrix, when amalgamated with domain knowledge, constitute a substantive asset as a process to learn more about the data and research.

Visualization 2

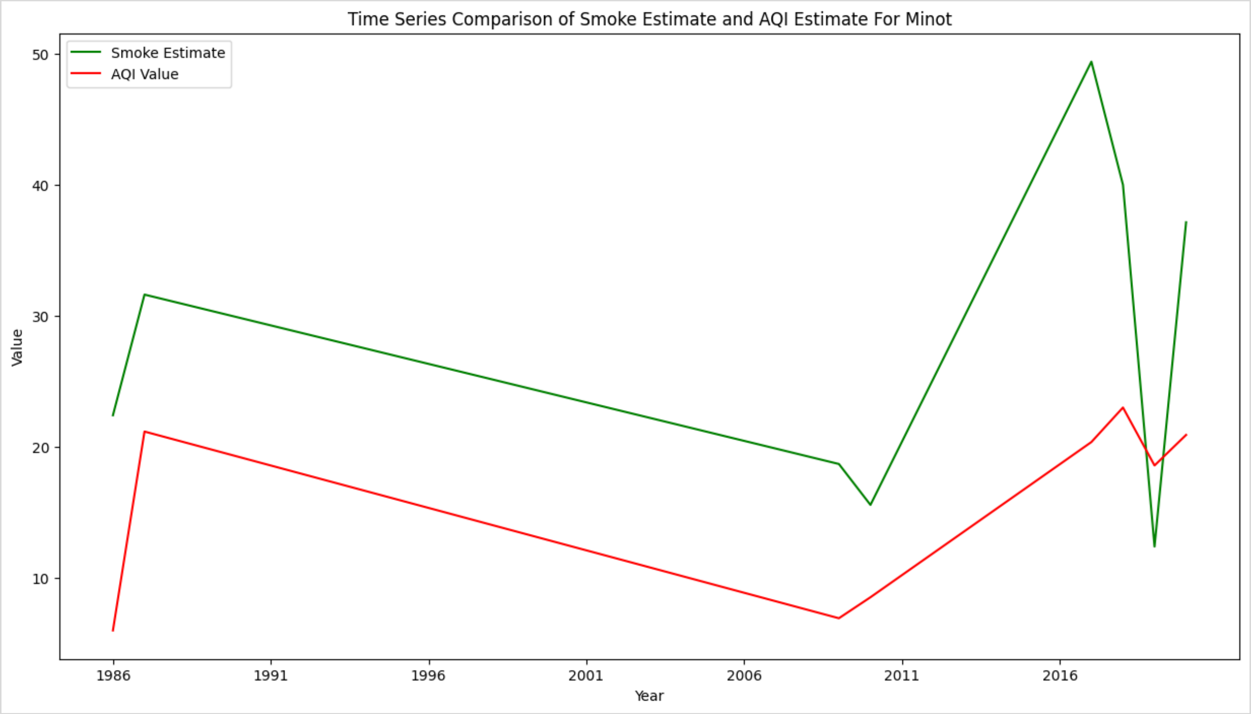


The time series graph presented above illustrates the annual progression of total acres burned by wildfires within a 1250-mile radius from Minot. Serving as a valuable analytical tool, this graph provides a clear depiction of the dynamics of wildfires in the region over a specific timeframe. The x-axis represents the years, while the y-axis indicates the total acres burned for each respective year, facilitating a straightforward comprehension of the relationship between time and the extent of wildfires.

The data underpinning this time series plot is sourced from the original fire dataset, meticulously filtered to include only those fires falling within a distance of up to 1250 miles from Minot and occurring between 1963 and 2023. Noteworthy is the absence of data post-2020 meeting both criteria. The filtered data was subsequently aggregated by year, with the total acres burned calculated for each year, culminating in the construction of the time series plot. To enhance readability, the x-axis intervals were configured to display years at 5-year increments.

A discerning analysis of the graph reveals a discernible trend—a marked increase in the total acres burned as the years progress. In fact, the total acres burned in 2020 register a magnitude nearly 10^3 times greater than those in 1963. While occasional fluctuations and temporary declines punctuate the trajectory, the overarching pattern remains conspicuous. Various factors contribute to this concerning trend, including rising temperatures, prolonged droughts fostering favorable conditions for wildfires, and the introduction of non-native invasive plant species altering ecosystems, rendering them more susceptible to fires. Moreover, population growth and urbanization amplify the risk of accidental ignitions and fires in the wildland-urban interface. Notably, a nuanced observation from this analysis reveals a cyclical year-on-year trend amidst an overarching upward trajectory when viewed over a more extensive timeframe.

Visualization 3

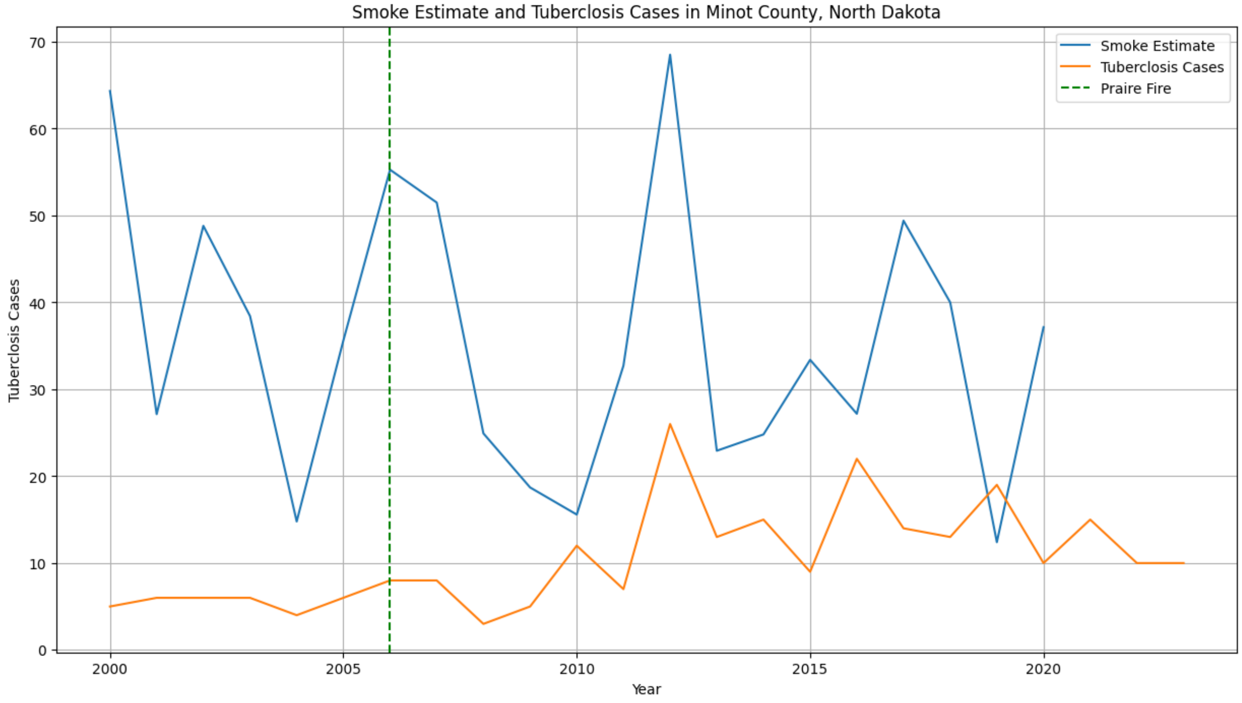


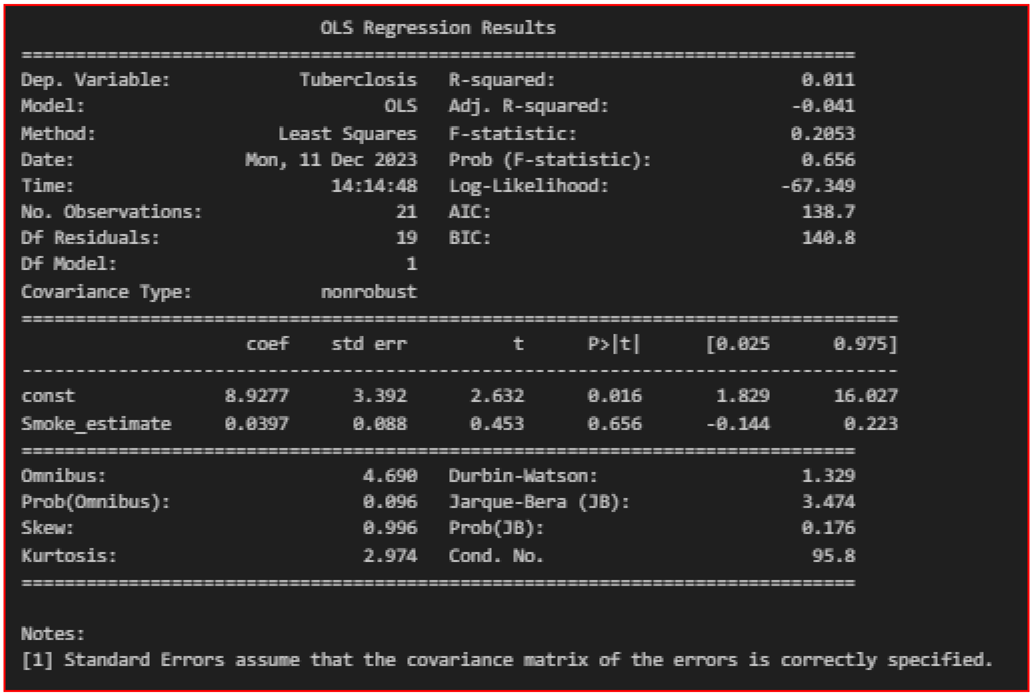
The line graph depicts a temporal interplay between the smoke estimate, derived from an ARIMA model within a 1250-mile radius of Minot, North Dakota, and the government-provided Air Quality Index (AQI). The correlation coefficient of 0.6555 suggests a moderately positive relationship between the two variables, indicating a discernible alignment in their trends. The X-axis spans the years from 1986 to 2020, offering a chronological context, while the Y-axis quantifies the respective values of the smoke estimate and AQI.

Examining the green line representing the smoke estimate, distinct temporal patterns emerge. A notable peak occurs in 2017, indicating a substantial spike in estimated smoke levels, followed by a decline in 2019 and another increase in 2020. The red line representing AQI values responds to these fluctuations, affirming the correlation. In 2017, the AQI peaks correspondingly to the elevated smoke estimate, signifying a direct impact on air quality.

In alignment with AQI standards, the majority of data points fall within the "good" air quality range (0-50). However, spikes in both the smoke estimate and AQI in 2017 and 2020 breach this threshold, indicating periods of compromised air quality. These peaks prompt a closer investigation into the specific conditions contributing to heightened smoke levels and their subsequent impact on local air quality. Understanding the causative factors behind these peaks can inform targeted interventions for effective air quality management, ensuring a comprehensive and proactive approach to environmental monitoring.

Extension Plan Analysis:

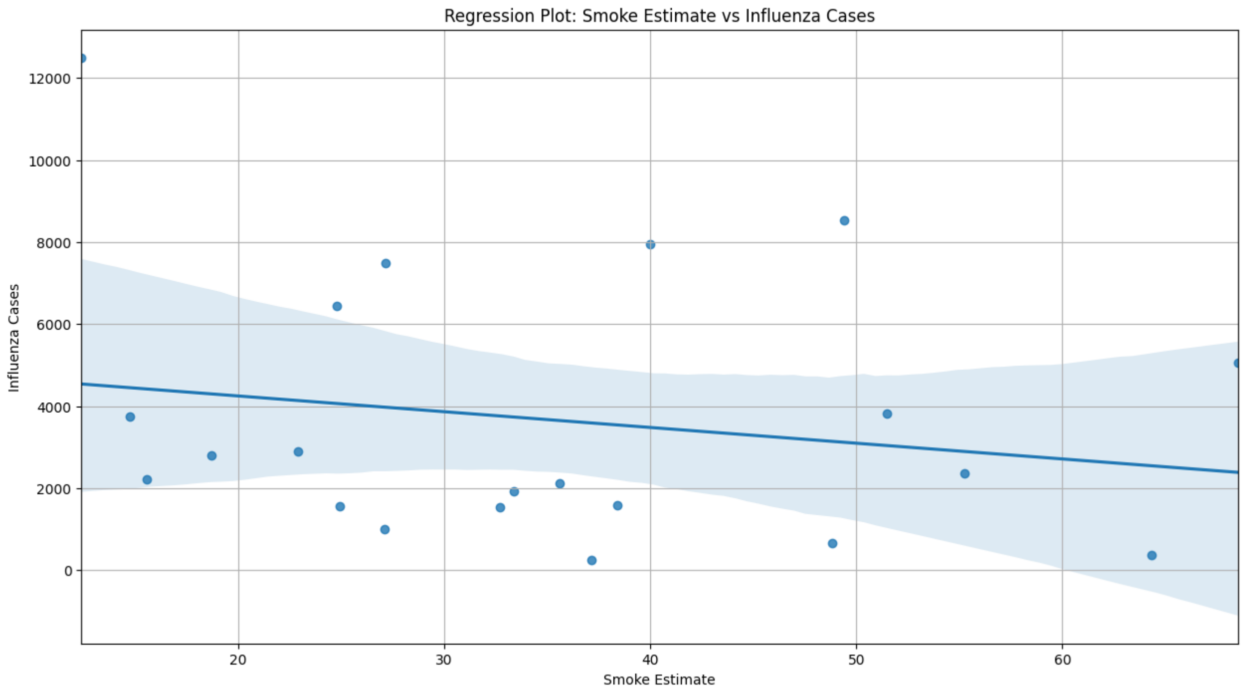


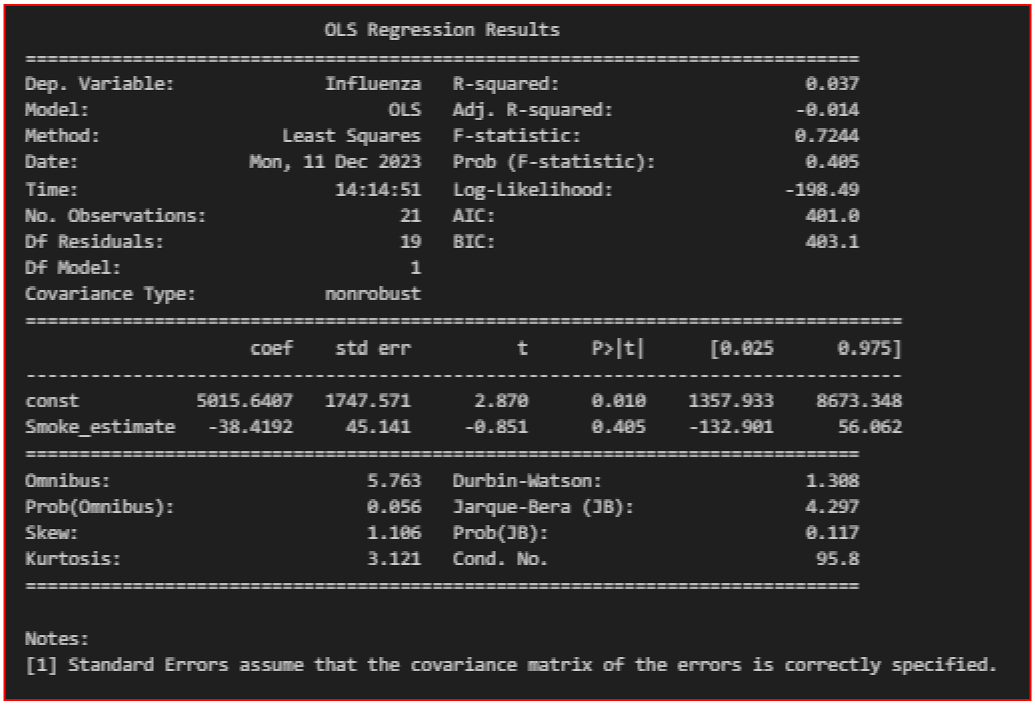


```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Tuberclosis   R-squared:                       0.011
Model:                            OLS   Adj. R-squared:                 -0.041
Method:                 Least Squares   F-statistic:                    0.2053
Date:                Mon, 11 Dec 2023   Prob (F-statistic):              0.656
Time:                        14:14:48   Log-Likelihood:                -67.349
No. Observations:                  21   AIC:                             138.7
Df Residuals:                      19   BIC:                             140.8
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              8.9277      3.392      2.632      0.016       1.829      16.027
Smoke_estimate     0.0397      0.088      0.453      0.656      -0.144       0.223
==============================================================================
Omnibus:                        4.690   Durbin-Watson:                   1.329
Prob(Omnibus):                  0.096   Jarque-Bera (JB):                3.474
Skew:                           0.996   Prob(JB):                        0.176
Kurtosis:                       2.974   Cond. No.                         95.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The analysis unfolded significant insights corresponding to the three research questions. he linear regression analysis suggests that there isn't a statistically significant linear relationship between smoking estimates and tuberculosis cases in North Dakota. The model doesn’t effectively explain the variability in tuberculosis cases based solely on smoking estimates. An R-squared of 0.011 means that only about 1.1% of the variability in tuberculosis cases (Tuberclosis) can be explained by changes in smoking estimates (Smoke_estimate). In other words, the model doesn’t explain much of the variation in tuberculosis cases based on smoking estimates. The p-value for the coefficient of Smoke_estimate is 0.656. In regression analysis, this p-value indicates the probability of observing such a relationship between the predictor (smoking estimates) and the response variable (tuberculosis cases) if there were no actual relationship in the population. A high p-value (above 0.05, in this case) suggests that there is no significant linear relationship between smoking estimates and tuberculosis cases.

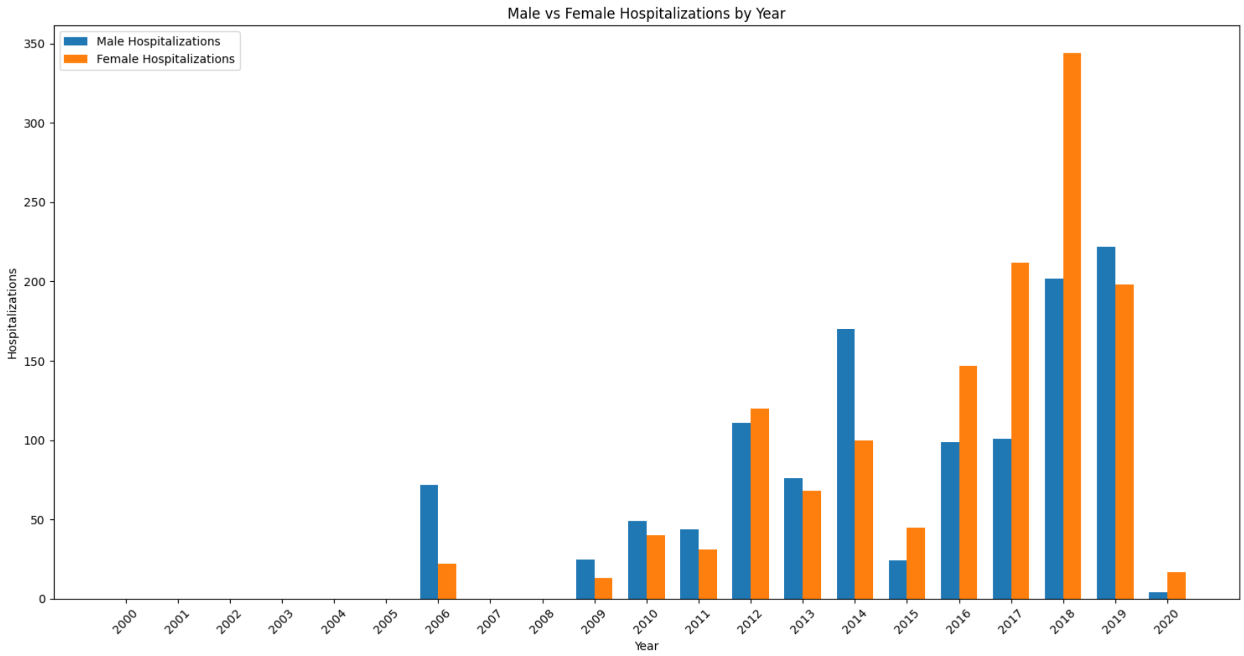




```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Influenza   R-squared:                       0.037
Model:                            OLS   Adj. R-squared:                 -0.014
Method:                 Least Squares   F-statistic:                    0.7244
Date:                Mon, 11 Dec 2023   Prob (F-statistic):              0.405
Time:                        14:14:51   Log-Likelihood:                -198.49
No. Observations:                  21   AIC:                             401.0
Df Residuals:                      19   BIC:                             403.1
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           5015.6407   1747.571      2.870      0.010    1357.933    8673.348
Smoke_estimate   -38.4192     45.141     -0.851      0.405    -132.901      56.062
==============================================================================
Omnibus:                        5.763   Durbin-Watson:                   1.308
Prob(Omnibus):                  0.056   Jarque-Bera (JB):                4.297
Skew:                           1.106   Prob(JB):                        0.117
Kurtosis:                       3.121   Cond. No.                         95.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The investigation into fire smoke and influenza rates exhibited a weak correlation in Minot County. Despite observing fluctuations in both fire smoke estimates and influenza rates across different years, the statistical analyses did not establish a substantial relationship between the two variables. The coefficient for Smoke_estimate is -38.4192. This represents the estimated change in influenza cases for a one-unit change in smoking estimates. However, as the p-value is high, this coefficient may not be reliable for making predictions or interpretations.



Moreover, the study of gender disparities in hospitalization rates yielded no significant difference between male and female hospitalization patterns concerning variations in fire smoke exposure. Statistical analyses indicated comparable hospitalization rates for both genders despite potential differences in exposure to fire smoke. With a p-value of 0.479, it indicates that if there were genuinely no difference in hospitalization rates between males and females (null hypothesis being true), there would still be a 47.9% probability of observing a difference as extreme as, or more extreme than, the one observed in the sample.

Discussion/Implications

The findings derived from this comprehensive analysis provide critical insights into the complex interplay between fire smoke exposure and respiratory health outcomes, shedding light on significant aspects that warrant consideration by stakeholders and policymakers in Minot, North Dakota.

Significance of Findings

The absence of a statistically significant relationship between fire smoke estimates and tuberculosis rates, coupled with the weak correlation between fire smoke and influenza rates, underscores the nuanced nature of these associations. While the results do not establish a direct linear relationship, they highlight the intricate factors contributing to respiratory illnesses within the region. These findings prompt a deeper exploration into the multifaceted determinants of tuberculosis and influenza, emphasizing the need for a holistic approach to public health interventions beyond singular attributions to fire smoke exposure.

Moreover, the lack of substantial gender disparities in hospitalization rates concerning fire smoke exposure challenges preconceived notions. While previous studies suggested potential gender-based variations in health outcomes due to environmental exposure, the empirical evidence from this analysis indicates comparable hospitalization patterns between males and females. This outcome necessitates a reevaluation of gender-specific health interventions, emphasizing equitable healthcare strategies catering to all demographics.

Implications for City Governance and Residents

The implications of these findings extend to actionable considerations for the city council, city managers, the mayor, and the residents of Minot. Firstly, understanding the limitations in directly attributing respiratory illness rates solely to fire smoke exposure underscores the need for a multifaceted approach to public health policies. Policymakers should prioritize comprehensive health interventions addressing various determinants of respiratory health, encompassing healthcare access, vaccination programs, and environmental policies beyond wildfire management.

The absence of pronounced gender disparities in hospitalization rates suggests the need for equitable healthcare provisions irrespective of gender. City administrators should ensure unbiased access to healthcare services, accounting for demographic variations in health-seeking behaviors and fostering inclusive health policies.

Timeframe for Action

While the findings provide invaluable insights, the urgency of addressing these implications demands prompt action. The city council, in collaboration with public health experts and environmental agencies, should initiate a strategic plan encompassing health education campaigns, robust healthcare infrastructure, and policies promoting cleaner air quality. The development of these interventions should commence within the next 12-18 months to implement proactive measures ahead of the wildfire seasons, ensuring the protection of public health against potential future threats.

Human-Centered Data Science Principles

Throughout this project, human-centered data science principles served as guiding pillars in decision-making. Ethical considerations significantly influenced study design and methodology selection, ensuring the responsible and equitable use of data. The emphasis on fairness, transparency, and inclusivity in analysis and interpretation upheld the principles of ethical data utilization, safeguarding against biases and fostering equitable insights. The findings not only reflect rigorous analytical approaches but also exemplify the ethical and human-centered considerations that underpin this study's integrity and reliability.

Limitations

The analysis conducted in this study was subject to several limitations that warrant acknowledgment, potentially impacting the robustness and interpretability of the findings.

An initial limitation pertains to the quality and integrity of the datasets utilized. The primary datasets employed in this analysis were comprehensive but not devoid of inherent flaws. Inherent inconsistencies and inaccuracies within the datasets posed challenges during the data cleaning phase. Moreover, the absence of uniform data standards across various sources necessitated extensive data harmonization, potentially introducing biases or oversights during the aggregation process.

The data cleaning techniques adopted, while rigorous, inadvertently omitted certain outlier observations or introduced unintended biases during the standardization and normalization processes. I tried various imputation methods including mean imputation and predictive modeling to address missing data points. However, the inherent assumptions and approximations involved in these methods introduced uncertainties or altered the distributional characteristics of the variables so I decided to reduce the number of years.

Furthermore, the statistical techniques utilized in this study, including Pearson's correlation, OLS regression, Linear Regression Models, and The Mann-Whitney U Test, are contingent upon specific assumptions. Assumptions regarding linearity, independence of observations, homoscedasticity, and normality of residuals underpinning these techniques might not have been fully met by the dataset. Violations of these assumptions could compromise the validity and reliability of the inferential analyses conducted, warranting caution in the interpretation of results.

Another critical limitation pertains to potential licensing constraints associated with the datasets used. While efforts were made to ensure compliance with licensing agreements and intellectual property rights, unforeseen restrictions or licensing changes might impact the accessibility or usability of the datasets in the future. This dependency on external data sources poses a risk of discontinuity or hindrance in replicating the study, thereby influencing the study's reproducibility and long-term viability. For most of the analysis we consider mostly Minot County or nearby places. However, it is possible that wildfires in other regions might also have a big factor on the effect of wildfire smoke in Minot County – for example: Canada because North Dakota is right on the border of US-Canada. The merging process based on the "Year" column could result in data loss due to inconsistencies or missing entries in different datasets for specific years, affecting the accuracy of the consolidated dataset. The fire smoke estimate model's assumptions regarding fire types, size, and distance might oversimplify the complex nature of smoke dispersion, potentially leading to inaccuracies in estimations. The study could have potentially neglected other respiratory conditions or health issues indirectly impacted by wildfire smoke. While the study identifies correlations between fire smoke estimates and health outcomes, establishing a direct cause-effect relationship requires more extensive experimental or longitudinal studies.

In summary, while every effort was made to mitigate these limitations, the inherent constraints associated with data quality, cleaning processes, treatment of missing values, statistical assumptions, and data source dependencies could potentially introduce uncertainties or biases, necessitating cautious interpretation and consideration of the study's outcomes.

Conclusion

Throughout this study, three core research questions were explored to ascertain the relationship between fire smoke exposure and various respiratory health outcomes in Minot County, North Dakota.

Research Question 1: Fire Smoke and Tuberculosis Rate Relationship

Hypothesis: Increased exposure to fire smoke is positively correlated with higher tuberculosis rates in Ward County, Minot.

Research Question 2: Fire Smoke and Influenza Rate Relationship

Hypothesis: Elevated levels of fire smoke are associated with increased influenza rates in Minot County.

Research Question 3: Gender Disparities in Hospitalization Rates

Hypothesis: Gender differences influence hospitalization rates due to variations in exposure to fire smoke in Minot County.

The analysis conducted in this study yielded multifaceted insights into the relationship between fire smoke exposure and respiratory health outcomes in Minot County. However, the findings did not universally support the initial hypotheses posited.

Fire Smoke and Tuberculosis Relationship: The statistical analyses did not yield conclusive evidence supporting a positive correlation between fire smoke exposure and tuberculosis rates in Ward County, Minot. The model outcomes indicated a lack of substantial variability in tuberculosis cases explained solely by smoking estimates, suggesting a minimal or negligible impact of smoke exposure on tuberculosis rates.

Fire Smoke and Influenza Relationship: Similar to the tuberculosis investigation, the analysis exploring the association between fire smoke exposure and influenza rates exhibited weak correlation patterns. Despite observable fluctuations in both variables across different years, the statistical analyses did not establish a robust relationship between fire smoke exposure and influenza incidence in Minot County.

Gender Disparities in Hospitalization Rates: Contrary to the hypothesis proposing gender-based differences in hospitalization rates due to fire smoke exposure, the statistical tests revealed no significant disparities between male and female hospitalization patterns. The analyses indicated comparable hospitalization rates irrespective of gender, suggesting a lack of distinct variations linked explicitly to exposure differences.

Contribution to Human-Centered Data Science Understanding

This study serves as an essential illustration of the complexities inherent in interpreting and deriving meaningful insights from data, especially concerning public health and environmental impact analyses. While the initial hypotheses proposed specific associations between fire smoke exposure and respiratory health outcomes, the nuanced and intricate nature of these relationships became apparent through rigorous statistical analyses.

The study emphasizes the importance of approaching data-driven inquiries with a critical lens, acknowledging the limitations and uncertainties that may arise during the process. Furthermore, it underscores the necessity of ethical considerations and human-centered principles in guiding data analysis and interpretation, ensuring transparency, fairness, and responsible use of data in deriving conclusions that inform policy and decision-making.

In essence, this study illuminates the evolving landscape of human-centered data science, highlighting the intricacies of analyzing complex data domains while navigating ethical considerations and inherent uncertainties to extract insights beneficial for public health and policy formulation.

References

[Reference 1]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9896637/

[Reference 2]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6467301/

[Reference 3]: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6466235/

Data Sources

[Data Source 1]: https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81 (USGS Wildland Fire Dataset)

[Data Source 2]: https://ghdx.healthdata.org/record/ihme-data/united-states-chronicrespiratory-disease-mortality-rates-county-1980-2014 (Respiratory Diseases and Influenza Statistics)