DATA 512: Final Project Report

Author: Logan O'Brien

Introduction:

This project studied the present and future impact of wildfire smoke on the residents of Richland Washington, a city in Benton County. My work began by estimating the air quality near the city by gathering wildfire data and AQI data, and by constructing my smoke estimate. I then attempted to predict the future air quality of the city. Next, I turned my attention to the potential health impacts of smoke on Richland residents by retrieving population and healthcare data and performing various analyses.

This area of study is important and has significant implications for the city leaders. Quantifying and predicting the impact of wildfire smoke on the health of Richland will aid city leaders in understanding what kind of threats this poses in the coming years. Additionally, it will help support their decision-making on where to allocate the city's limited funds and whether healthcare institutions will require aid in treating related effects.

The findings of this study may also be of interest to the general population and help to keep them informed of risks to their health.

Background/Related Work:

While studying the effects of wildfires on the specific city of Richland, WA may be relatively novel, research on the impact of wildfire smoke on health, more broadly, is certainly not sparse. I surveyed a variety of relevant research publications (see e.g. references [6] through [14]) to better inform my work. For instance, I came across an article that recounted a meta-analysis of many different relevant studies which helped me build my familiarity with the research topic (Source [9].).

Through my research, I learned several important facts to help provide context. First, much work has been done to provide evidence that wildfire smoke is having a deleterious impact on human health. According to [6], "Previous epidemiologic studies of PM2.5 during wildland fire-smoke events have reported primarily positive and consistent associations with respiratory effects (Delfino et al. 2009; Liu et al. 2017; Moore et al. 2006; Mott et al. 2005; Rappold et al. 2011)". For instance, [7] found a 7.2% greater risk of admissions for respiratory illness on poor air-quality days when the air quality was below a particular threshold, due to wildfire smoke (particularly $PM_{2.5}$). These findings increased my confidence that I would likely find significant results from analyzing the connection between wildfire smoke and health conditions.

Additionally, I learned several other things that influenced my analysis. For one, I noted that in [8]., the authors divided the number of health establishment "visits by the population at risk" (pg. 106). I decided to implement this technique in my analyses, as it would help me control for the effects of changes in the population size. The articles surveyed also influenced my choice of which respiratory diagnoses I would study.

While I had a rough idea of my project focus before surveying the current literature, reading through a variety of articles clarified my approach. Ultimately, I settled on the research question "How will wildfire

smoke impact the number of hospitalizations for respiratory issues in Richland, WA in the coming years?"

My preparatory investigation also yielded several useful datasets to extend my analysis beyond the general exploration of wildfire smoke and support the further analysis needed to address the research question. First, I utilized population estimate data from the Census Bureau, a government entity [1]. This data was useful because it allowed me to control for population in my analyses. The lowest level of granularity provided was for the county level, so I exported the data for Benton County. The data was stored across multiple files ([2]. and [3].) that required separate downloading and preprocessing. While I could not locate a license or terms of use for the data on the web pages, the data files contained a suggested citation. Additionally, I received confirmation from a customer service agent that the data was provided for the public and that I was free to use it provided I cited it. Ultimately, I ended up using a range of data from 2010 – 2022, before ultimately dropping several years from consideration.

It is not easy to find openly available healthcare data (for obvious reasons), but after some searching, I discovered the Agency for Healthcare Research and Quality (AHRQ) [4]. The AHRQ provides an interactive dashboard on its website to examine various healthcare data points, called the Healthcare Cost and Utilization Project (HCUPnet) [5]. The site makes it clear that the user of the data is not permitted to make attempts to identify individuals or healthcare establishments in the data.

I extracted data on the number of patient discharges, for Benton County, for various diagnosed conditions. In particular, the data contained the following fields for the selected county and year:

- *Number of Discharges*
- *Average Length of Stay (in days)*
- *Rate of Discharges per 100,000 Population*
- *Age-Sex Adjusted Rate of Discharges per 100,000 Population*
- *Aggregate Hospital Costs (in $)*
- *Average Hospital Costs per Stay (in $)*

I exported the Benton County data from the tool for each year and then processed the data. For my analysis, I needed only the number of discharges for each year.

Methodology:

I conducted my analyses across a series of 5 notebook files (see references [15].). The first part of my study (the Part 1 Common Analysis) was dictated to a large extent by the assignment guidelines - specifically, I was tasked with determining what the smoke impacts were on Richland, WA for the past 60 years. I began by retrieving wildfire data from USGS.gov, a government website that provides various data [14]. I loaded the data using the Professor's *wildfire* Python module that he provided to us for use. Then, I filtered the wildfire data to fires occurring within 1250 miles of Richland, WA between 1963 and 2023. After retrieving the wildfire data and filtering it, I next created a smoke estimate to represent how poor the air quality was and applied it to the wildfire data. This smoke estimate incorporated my hypothesis that the value should be proportional to the size of the fire (represented in square miles burned) and inversely proportional to the fire's distance from Richland.

Next, I retrieved a more official measure of air quality, the U.S. EPA's AQI. I utilized an API to retrieve the AQI data [19]. First, I checked to ensure that there was at least one sensor near Richland. Then, I

retrieved daily AQI measurement data from sensors within a geographic bounding box of 12.5 by 12.5 miles (if I am recalling/interpreting the configuration correctly). I retrieved AQI data as far back as 2001. To simplify the measurement data, for each day of AQI measurements, in cases where there was both a PM_{10} and $PM_{2.5}$ measurement, I extracted the larger of the two - at the recommendation of my classmate, Emily Rolen. After this, I converted the daily AQI values into an annual summary by taking the average of each year. It is important to note, that throughout my efforts to work with the wildfire data from USGS and retrieve and manipulate the AQI data, I heavily relied on various Python code examples provided by the professor and used them with permission. After generating values for my smoke estimate and retrieving the AQI data, I prepared these data for comparison by aggregating my smoke estimate by taking the average of all smoke estimate values for fires that occurred each year. This produced an annual smoke estimate value for each year from 1963 to 2020.

After acquiring the air quality estimates, I proceeded with my analysis. I began by utilizing a linear regression model to predict future annual values of my smoke estimate based on historic data (year and corresponding smoke estimate value). Then, I conducted an exploratory visual analysis. I created a plot depicting the distribution of fires that occurred at various distances from Richland, examined the number of acres burned over time, and compared my smoke estimate with the annual AQI summary I gathered. I also calculated the correlation between my smoke estimate and the AQI. (Note: I returned after the class presentation, while writing this document, and created another plot comparing my smoke estimate with the AQI after removing a couple outliers and dropping some years of data. I also recalculated the correlation between the variables because I removed the outliers).

After conducting the initial analysis for the Part 1 Common Analysis, I started my extension plan. I gathered population estimate data from the U.S. Census Bureau [1]. for Benton County for 2011 – 2022 [2]. and [3]. I exported the data from the website and then merged and preprocessed the data files. Also, I gathered healthcare data from the AHRQ, as described above for Benton County for 2011 through 2020 (see also [4]. and [5].). Before exporting the healthcare data, I had to determine which discharge diagnoses were relevant to my analyses. Through my literature survey discussed above, I observed that the researchers typically examined multiple respiratory or cardiovascular conditions, and some were used in multiple papers (pneumonia for example – see [6], [7], and [10].). Ultimately, after checking to see which conditions [6], [7], and [10] studied, and cross-referencing them with what information was available on ARQ's dashboard tool [5]. for all years, I selected the following conditions to examine: Pneumonia, Acute Bronchitis, Chronic Obstructive Pulmonary Disease (COPD), and Aspiration Pneumonitis. Note: there were a few intricacies and assumptions here. First, there was a diagnosis coding change in the data in 2015, and as a result, there is no data for that year. Additionally, I assumed that the selected conditions are equivalent between the two coding systems, even though the codes (and occasionally the exact name) differ. I decided to exclude 2011 from my analysis since not all conditions had data for that year. Additionally, I also removed 2020 from consideration due to the abnormal events of that year – leaving me with healthcare data from 2012 – 2019 (excluding 2015). After downloading the healthcare data (each year and condition individually) I manually pieced the datasets together into groups where each dataset contained all years of data for a given condition. Then, I loaded the data into my notebook and added the number of discharges of each condition together to arrive at the total number of patient discharges for each year.

After collecting the new datasets, I performed a new analysis. I began by doing a visual exploration of trends in the AQI and population estimate data. Next, I calculated the Pearson correlation between the

air quality (AQI) and the patient discharges per 1,000 people in Benton County. And finally, I used a couple of linear regression models to predict future patient discharge rates.

Findings:

Recall that in the first part of my analysis, I focused on studying the presence of wildfires and smoke near Richland, WA. I predicted future values (for the years 2024-2049) of my smoke estimate via a linear regression model. The results suggest a slow but steadily increasing smoke estimate in the coming years.

For the supporting exploratory visual analysis, I arrived at the following results. Out of the fires studied (within 1250 miles of Richland and between 1963 and 2023), they occurred over a range of distances from Richland (see Figure 1). In particular, it appears that the majority of them occurred within about 1,000 miles of Richland and a significant number were between 150 to 450 miles from the city.

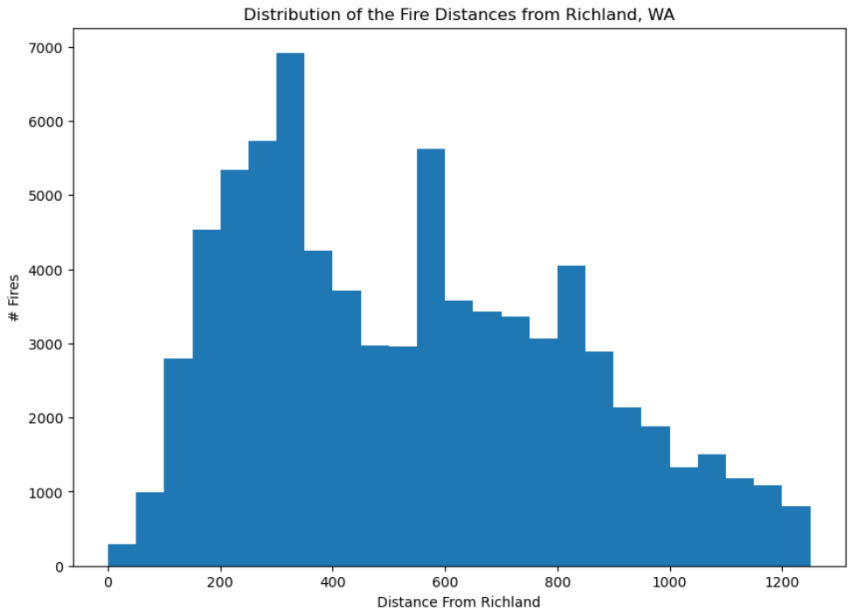

Figure 1

My visual analysis also found that the total acreage burned by wildfires is not monotonically increasing, but it is growing significantly over time (Figure 2).

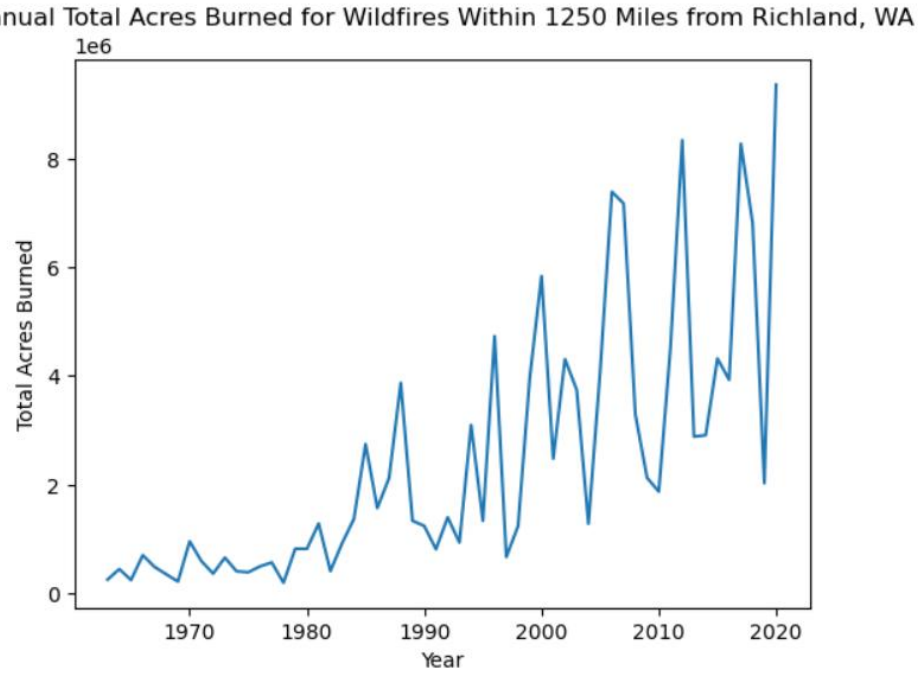

Figure 2

I also compared my smoke estimate to the AQI estimate for the air quality each year by plotting both on a line chart (Figure 3). As you can see, the scope of the AQI data is significantly more limited than my smoke estimate. This is because my smoke estimate was applied to all years of fire data from the USGS dataset [14]., while the AQI data from sensors within close proximity to Richland was absent for many years. Based on the figure, it appears that apart from a few outliers in the AQI estimate data, my smoke estimate is in the same order of magnitude as the AQI data – which is a good sign.

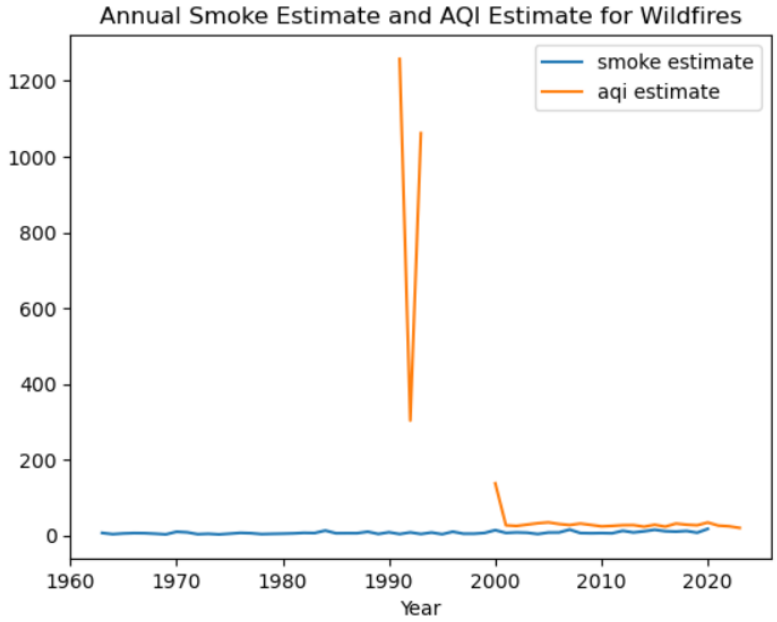

Figure 3

Note: more recently, after presenting my project in class, I decided to modify the chart depicted in Figure 3 by removing the outliers from the AQI and older years from the smoke estimate, which produced Figure 4. I also recalculated the Pearson correlation between the AQI and my smoke estimate which I had previously calculated but needed to redo after removing the outliers. I found a very weak correlation of about 0.077 which surprised me as the smoke estimate seems to match the behavior of the AQI fairly well in Figure 4.

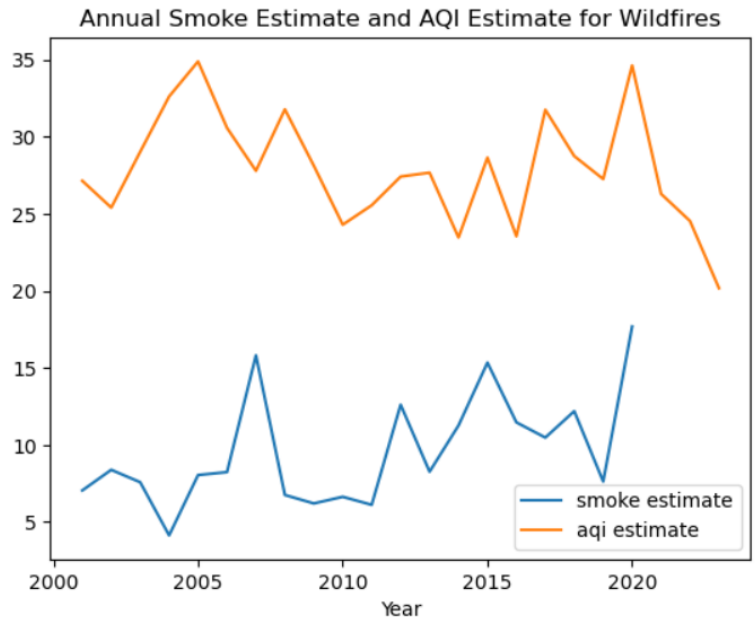

Figure 4

My project extension produced a variety of findings. I found that the population of Benton County has been steadily increasing (Figure 5), and that surprisingly, the number of patient discharges for the selected conditions per 1,000 residents of Benton County appears to have a decreasing trend (Figure 6). I also found that the average AQI estimate fluctuated between about 24 to 32 (Figure 7).

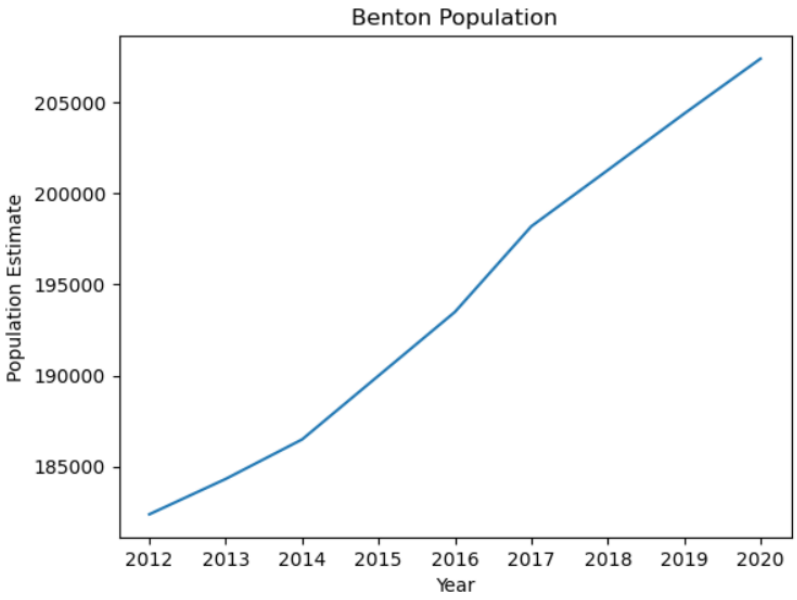

Figure 5

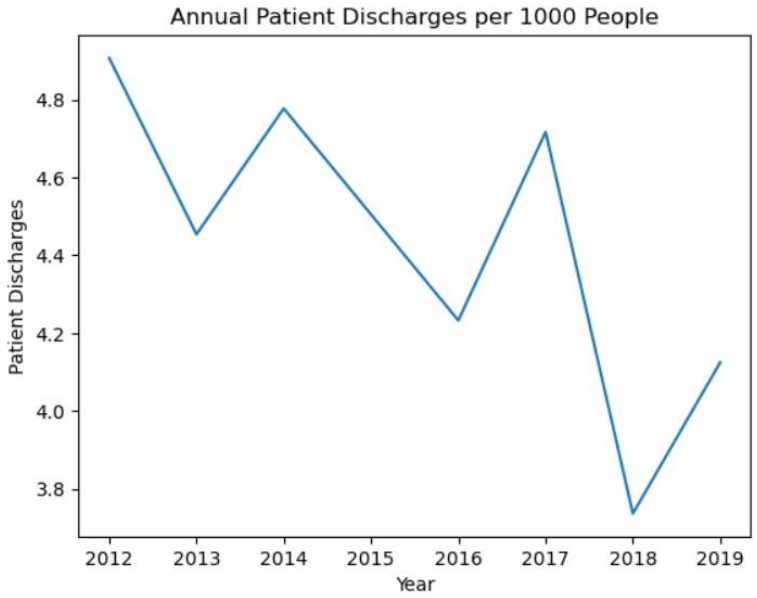

Figure 6

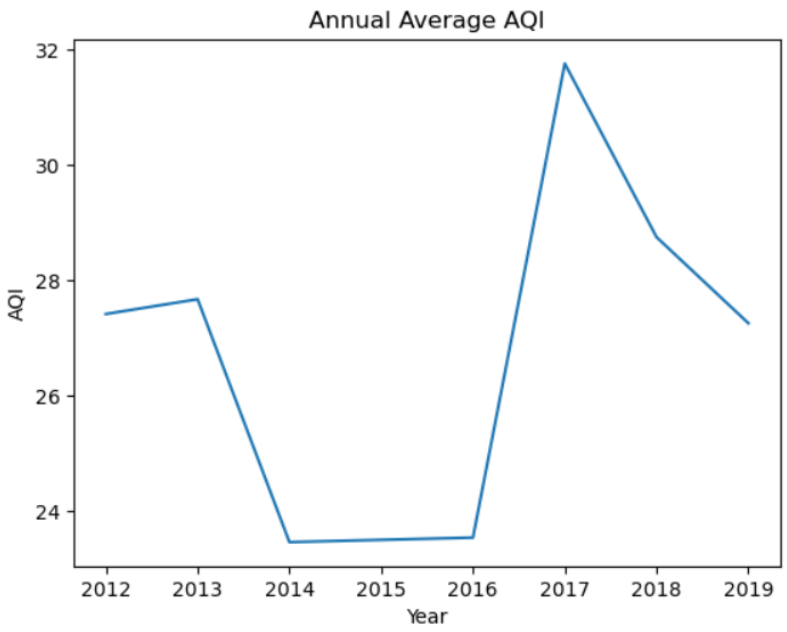

Figure 7

In my quest to examine the relationship between air quality and hospitalization rates, I created a scatter plot depicting the average AQI estimate and the patient discharge rate per 1,000 people for the selected conditions (Figure 8). From the figure, it is difficult to determine the relationship between the variables. I calculated the Pearson correlation between the variables and returned a value of about -0.03, indicating that there is essentially no correlation between the variables. The fact that there is no correlation is surprising to me, as I expected to find that the two are positively correlated and that years with higher AQI would have a higher number of patient discharges.

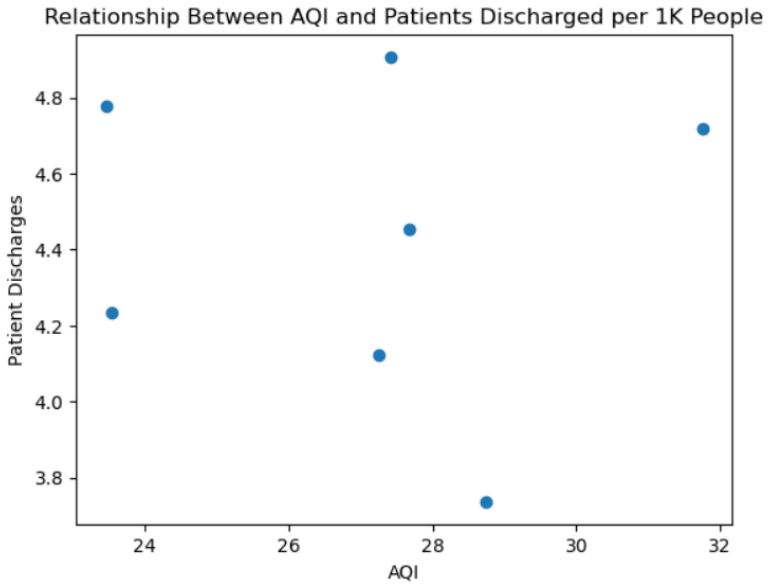

Figure 8

For the final component of my analyses, I attempted to predict the number of patient discharges (per 1,000 people in Benton) for future years. However, this presented significant challenges.

I began by first generating predictions for the AQI for 2020 – 2045, based on a linear regression model fit to the observed AQI values and corresponding years. The model reported an R-squared value of 0.077, which, as explained in [16]. means that the model is a very poor fit for the data. I plotted the results in Figure 9, and observed that the model predicted increasing values of AQI.

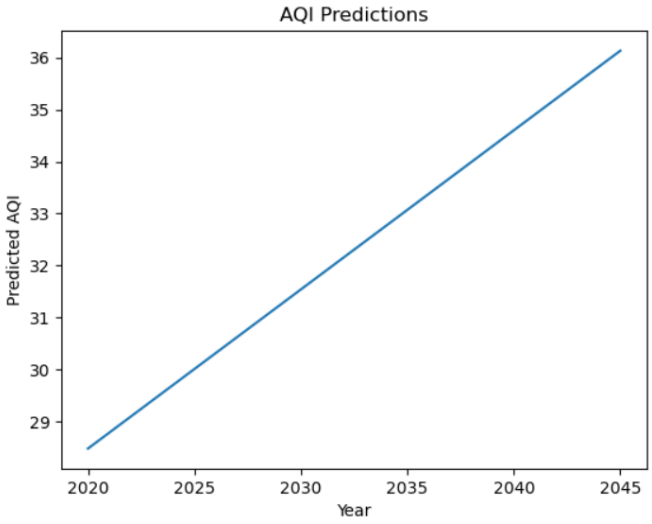

Figure 9

After predicting the AQI, I then fit a new linear regression model. This time, I used the AQI and the year as my predictors and the number of patients discharged (per 1,000 people in Benton) as my response. The training data corresponded to the years 2012 – 2014 and 2016 - 2019. This model offered a decent fit (see [17]. for an opinion on what constitutes a "good" R-squared value) to the data with an R-squared value of 0.540. I then took the previously predicted AQI values and the corresponding years and predicted the rate of patient discharges for the same time span, using this model. The model predicted that the patient discharge rate would decrease in the future (see Figure 10).

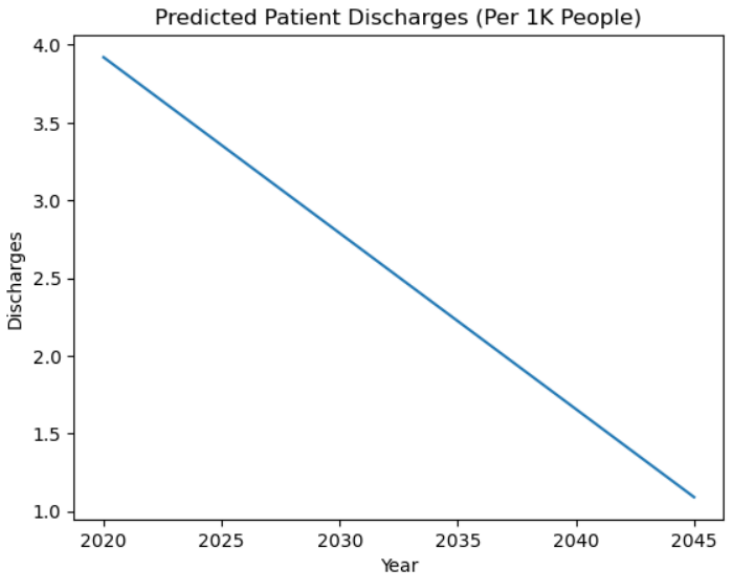

Figure 10

It is important to note that these results are subject to a variety of limitations and assumptions, some of which are enumerated in a subsequent section. However, I must point out here one particularly significant one. Because of the poor fit of the first linear model, I believe its predictions for future years of AQI must be taken with a large degree of skepticism. Therefore, it follows that my 2nd model's predictions for the patient discharge rate must also be questioned as those predictions were based, in part, on the suspect AQI predictions.

Discussion/Implications:

While my analysis did not find a strong correlation between air quality and patient health directly, I achieved a decent model fit when I added in the year for each data point too – suggesting that there is a relationship between air quality and patient health, but it is more complex. Unfortunately, I was unable to confidently predict future rates of patient discharges due to the poor fit of my first model on AQI and the corresponding year.

My recommendation to the hypothetical city council, based on my research, is to conduct a deeper study into the potential impact of poor air quality from wildfires on the city population. My findings seem to suggest that the city is not at an immediate risk of a large increase in respiratory illness (due to the absence of a positive trend in AQI and because patient discharges have been on the decline). However, I also found that wildfires appear to be growing worse based on the generally positive trend in total acreage burned. For this reason and because of the importance of promoting and protecting the health of its citizens, I encourage Richland to build off my analysis and conduct a more thorough study into the potential adverse health effects of wildfire smoke on its residents. This study would aim to establish a more concrete relationship between poor air quality from wildfire smoke and respiratory health and produce predictions of future health impacts with higher confidence. The assumptions, limitations, and pitfalls of my analysis should be considered when designing this new study.

Limitations:

My research relied on a variety of assumptions and is subject to multiple limitations, and there are likely assumptions and limitations present in my analysis that I am not cognizant of. However, here I will discuss several significant weaknesses in my analysis.

First, as noted above, I struggled to predict future rates of patient discharges for respiratory illness with high confidence. My initial plan was to create a single regression model that used the year and AQI as predictors and the discharge rate as the response variable. However, I later realized that while I could train a model of this nature, I would not be able to make predictions with it directly as I do not know the future values of AQI. For this reason, I chose to create another linear regression model to first predict AQI based on the year, for future years. Unfortunately, this model fit the data poorly, and hence I have little trust in the predicted AQI, which in turn renders the discharge rate predictions untrustworthy as well.

Another limitation of my models is that I did not consider any underlying assumptions that the models require for valid results. The underlying assumptions of these statistical techniques must be examined thoroughly before any important decisions regarding the health of Richland residents are made based on my findings. My analysis also utilized very simple models. It is quite possible that the quality and accuracy of the results would be bolstered by using more complex and sophisticated modeling techniques. For instance, I did not consider confounding effects – which is something another research study did [6]. Additionally, [10] incorporated "lagged exposure" (pg. 1418) into account in their model, which might be worth exploring.

My analyses are also limited by my data in several ways. The veracity of my models is hindered by the small number of data points (years) of healthcare and AQI at my disposal. Possessing a longer range of years of data would strengthen my findings. Additionally, when I extracted my AQI estimate from sensors near Richland, I did so in a somewhat haphazard manner, since I extracted a combination of PM_{10} and $PM_{2.5}$ and treated it as a single metric. Further, I chose to simply take an average for each year across all daily values, rather than designing a more sophisticated approach that only took data points into account for certain days or months out of the year. Also, I noticed that the AQI estimate for 1991 and 1993 exceeded the maximum value allotted [18]., which means that the error either resulted from the sensors themselves, or my API data extraction method. While this casts a bit of doubt on my process for gathering the data, I did not observe any other cases where the AQI was outside the allowable bounds, which leads me to think it is an issue with the original data source. On an unrelated note, I also assumed that the number of patient discharges is equivalent to the number of admissions (which is what my research question targets), but this is only true if you assume no one admitted dies or stays in the hospital for the remainder of their life. Lastly, it may be beneficial for me to have restricted the Benton County population to only those at risk, rather than the entire population.

There are other assumptions and limitations to my analysis that I did not discuss here, but hopefully this provides a better picture of my analysis.

Conclusion:

Wildfires frequently make the headlines and are top of mind when it comes to the state of our ecosystems and the health of our planet. A common position is that global warming and wildfires are

increasing and that this might have negative ramifications for humanity. In this project, I examined the impact of wildfires on Richland, WA by exploring the data of wildfires near Richland and the air quality of that area. I also sought to answer the research question "How will wildfire smoke impact the number of hospitalizations for respiratory issues in Richland, WA in the coming years?" While the results of my analysis were inconclusive, it seemed to suggest that it is not likely that Richland will experience a surge in respiratory issues caused by wildfires in the near future and that the city has more time to perform additional research studies that can better predict how the health of its residents will fair with wildfire smoke in the future. Keeping the health of citizens at the forefront of data science analyses is important to ensure that the work contributes to human well-being.

References:

- [6]. *DeFlorio-Barker S, Crooks J, Reyes J, Rappold AG. Cardiopulmonary effects of fine particulate matter exposure among older adults, during wildfire and non-wildfire periods, in the United States 2008-2010. Environ Health Perspect 2019;127(3):37006. doi: 10.1289/ehp3860.*
- [7]. Liu JC, Wilson A, Mickley LJ, Dominici F, Ebisu K, Wang Y, et al.. 2017. Wildfire-specific fine particulate matter and risk of hospital admissions in urban and rural counties. Epidemiology 28(1):77–85, PMID: 27648592, 10.1097/EDE.0000000000000556. [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [8]. Moore D, Copes R, Fisk R, Joy R, Chan K, Brauer M. 2006. Population health effects of air quality changes due to forest fires in British Columbia in 2003: estimates from physician-visit billing data. Can J Public Health 97(2):105–108, PMID: 16619995. [PMC free article] [PubMed] [Google Scholar].
- [9]. Liu JC, Pereira G, Uhl SA, Bravo MA, Bell ML. 2015. A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. Environ Res 136:120–132, PMID: 25460628, 10.1016/j.envres.2014.10.015. [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [10]. Rappold AG, Stone SL, Cascio WE, Neas LM, Kilaru VJ, Carraway MS, et al.. 2011. Peat bog wildfire smoke exposure in rural North Carolina is associated with cardiopulmonary emergency department visits assessed through syndromic surveillance. Environ Health Perspect 119(10):1415–1420, PMID: 21705297, 10.1289/ehp.1003206. [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [11]. Delfino RJ, Brummel S, Wu J, Stern H, Ostro B, Lipsett M, et al.. 2009. The relationship of respiratory and cardiovascular hospital admissions to the southern California wildfires of 2003. Occup Environ Med 66(3):189–197, PMID: 19017694, 10.1136/oem.2008.041376.
- [12]. Ignotti E, Valente JG, Longo KM, Freitas SR, Hacon Sde S, Netto PA. Impact on human health of particulate matter emitted from burnings in the Brazilian Amazon region. Rev Saude Publica. 2010;44:121–130. [PubMed] [Google Scholar]. https://www.scielo.br/j/rsp/a/bp9BffF785sJmcC6hqX366d/?lang=en.
- [13]. de Mendonca MJ, et al. Estimation of damage to human health due to forest burning in the Amazon. *J Popul Econ.* 2006;19:593–610. [Google Scholar].
- [15]. Github Project Repo: https://github.com/logan-obrien/data-512-final-course-project.
- [16]. https://statisticsbyjim.com/regression/interpret-r-squared-regression/

- [17]. https://www.investopedia.com/terms/r/r-squared.asp#:~:text=What%20qualifies%20as%20a%20%E2%80%9Cgood,such%20as%200.9%20or%20above.
- [18]. https://www.airnow.gov/aqi/aqi-basics/#:~:text=Think%20of%20the%20AQI%20as,300%20represents%20hazardous%20air%20quality.

Data Sources:

- [1]. http://www.census.gov/
- [2]. Annual Estimates of the Resident Population for Counties in Washington: April 1, 2010 to July 1, 2019 (CO-EST2019-ANNRES-53). Source: U.S. Census Bureau, Population Division. Release Date: March 2020. Link: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.
- [3]. Annual Estimates of the Resident Population for Counties in Washington: April 1, 2020 to July 1, 2022 (CO-EST2022-POP-53). Source: U.S. Census Bureau, Population Division. Release Date: March 2023. Link: https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html.
- [4]. https://www.ahrq.gov/
- [5]. https://datatools.ahrq.gov/hcupnet/
- [14]. https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81
- [19]. US Environmental Protection Agency (EPA) Air Quality Service (AQS) API. Some documentation: https://aqs.epa.gov/aqsweb/documents/data_api.html. Additional details: https://www.epa.gov/outdoor-air-quality-data/frequent-questions-about-airdata.