Final Report

**1. Introduction**

In this report I seek to understand the correlations between Death Valley attendance, smoke estimates, and AQI in Pahrump, NV. For this project we were encouraged to inform city leaders about potential impacts of smoke in their community through the lens of a human centered data science question. I chose to examine Death Valley attendance due to the impact tourism likely has on the economic viability of Pahrump.

From Pahrump’s Wikipedia page, it would appear that the town’s economic activity centers on two wineries and two legal brothels (*Pahrump, Nevada*). While it may have been possible to infer potential impacts on Pahrump’s industries from data linking smoke and winery production generally, I wanted to study economic activity more directly. A search for Pahrump on TripAdvisor revealed multiple tours to nearby Death Valley National Park (The 15 best things to do in Pahrump - 2023 (with photos)). In addition, when one plots a course from Las Vegas to Death Valley on Google Maps, one of the two primary routes goes directly through town. From this data I conclude that tourists originating in Las Vegas likely drive through town, filling up their cars, buying snacks, and gathering provisions for the park. Each of these actions provides direct economic benefit to the town. This hypothesis is bolstered by the approximately 20% of adults in Pahrump who work in “Accommodation and Food Services”, or “Arts, Entertainment, and Recreation Industries”, well above the total national average of 9% (*Economy in Pahrump, NV)*. It is distinctly possible that the proximity of Death Valley drives additional employment in these sectors (Economy in Pahrump, NV). Beyond tourism, Pahrump locals may also be employed to work in the National Park or to maintain outside infrastructure (e.g., road crews, firefighters).

Identifying a negative correlation between smoke estimates/AQI and Death Valley attendance would allow citizens of Pahrump to diversity their industries or challenge their government to do more about fire prevention in the event that smoke/poor air quality should increase.

**2. Background/Related Work**

Research has already been conducted on the impact of wildfire smoke on National Part attendance. Clark et. al found in modeling black carbon (an element of wildfire smoke) and recreational visitation to 32 US National Parks that “The results of these models are mixed, but overall show little to no effect of ambient smoke on visitation to the 32 parks tested, even when allowing for critical thresholds at the extreme upper ranges of smoke exposure. This indicates that wildfire smoke does not greatly alter park attendance” (Clark et al., 2023). Gellman et al. executed a similar analysis of wildfire smoke and federal campgrounds in the US between 2008 and 2017 and found “[…] fire and smoke affect 400,000 and 1 million visitor-days per year, respectively […but] the magnitude of the smoke effect is small […] suggesting that smoke fails to deter most visitors to public lands” (Gellman et al., 2022).

These findings run contrary to expectations, though it is important to remember that correlation does not imply causation. While wildfire smoke may be burdensome to National Park visitors, it is possible that significant investments of time and money could make them reluctant to change or cancel plans.

For my analysis I will examine the correlation between Death Valley attendance and smoke estimates/AQI estimates in Pahrump, NV. Despite earlier research, I would hypothesize a negative correlation between smoke estimates/AQI and park attendance. Given many park visitors likely originate in the Las Vegas area, they could easily avoid the hassle of driving to the park and have many alternate entertainment options available in the event of adverse conditions. The impact of smoke on local visitors (e.g., those who live in and around the park) is less clear and will be addressed in the “Limitations” section of this paper.

There has also been significant prior work to quantify smoke from wildland fires which I use in my smoke estimate calculations. My estimates leverage the USGS Wildland Fire Data (Welty and Jeffries, 2021), which includes fields such as type of fire, acres burned, distance to town, recency of other fires in the same area, and year of fire recording.

A basic smoke estimate includes the acres burned multiplied by . I assume that the acreage burned is proportion to the amount of smoke, and that smoke diffuses in air with an inverse-squared relationship, similar to light. However, I also believe other factors such as type of burn, recency of burn, and historical data accuracy impacts smoke estimates.

Per David Frisbey's 2008 thesis "A comparison of smoke emissions from prescribed burns and wildfires", "The results suggest that the smoke impacts of a wildfire may not be any greater than a prescribed burn when compared using the methodology. This research demonstrates how a combination of the fuel load and the size of the burn may be more significant in controlling downwind concentration of PM10 than the atmospheric conditions. Even when there is a planned burn under prescribed meteorological conditions there can be significant impacts if the size of the burn and fuel loading are not also considered" (Frisbey, David, 2008). Examining “Forest Service Professionals Prepare for a Prescribed Burn” (Avitt, 2023), we can see that forestry services do take fuel moisture, forest stand characteristics, historical data, terrain, and elevation into account when creating a prescribed burn. Given David's findings, and assuming the Forest Service correctly accounts for the fuel variables to create less intense blazes, I multiply my basic smoke estimate for prescribed burns (by 0.50.

I also assume that fires in areas burnt within the last 2 years should produce much less smoke than otherwise calculated. The idea of a differential burn is supported by “Burn out: Frequent fires are changing Western landscapes” (Pontecorvo, 2020). For the areas of a fire burned within the past 2 years, I multiply the basic smoke estimate by 0.20.

Finally, the USGS wildland fire metadata notes that "Areas burned prior to 1984 in this dataset represent only a fraction of what actually burned. While areas burned on or after 1984 are much more accurate and complete, errors still can and do occur" (Welty and Jeffries, 2021). Given the underestimation of acres burned, I multiply the smoke estimates for fires prior to 1984 by 1.5.

There are limitations with this smoke estimate, including no variable for fuel composition, weather, terrain, etc. which will be discussed in more detail in the “Limitations” section of this report.

AQI data is sourced directly from the EPA’s API, the code for which can be found in my “epa\_comparison” script (Ekrolen). More details on methodology can be found in Section 3 of this paper.

Death Valley attendance data is sourced from National Park annual attendance records. While the land was officially declared a National Park October 31st, 1994, the National Parks Service has kept attendance records for the site since 1933. The data consists of year of park visitation and annual total number of recreation visits. Per the National Park Service’s Visitor Use Statistics Page, a “Recreation Visit” is “The entry of a person onto lands or waters administered by the NPS except as defined above for non-reportable and non-recreation visits [e.g., entry into the park by NPS employees or contractors, commuter or through traffic, guides, government personnel with business in the park]. Funeral parties at National Cemeteries, school groups, etc. are reportable as ‘recreation’ use since their use is for the purpose for which the park was established. Visits originating on surface vehicles (trains, boats, other) and aircraft may be counted if they stop and disembark passengers on NPS administrated territory. The applicable rule is that one entrance per individual per day is countable” (U.S. Department of the Interior, 2023). Per the National Parks Service Disclaimer page, “Copyright law does not protect “any work of the U.S. Government” where “a work prepared by an officer or employee of the U.S. Government as part of that person's official duties” (See, 17 U.S.C. §§ 101, 105). Thus, material created by the NPS and presented on this website, unless otherwise indicated, is generally considered in the public domain. It may be distributed or copied as permitted by applicable law” (U.S. Department of the Interior, 2020).

**3. Methodology**

My process generally consisted of defining and calculating the smoke estimate, calculating the AQI, comparing the two, and testing correlation with Death Valley attendance. I will walk through each step in the following paragraphs. Detailed technical information can be found in my repo’s README and scr/ code files (Ekrolen).

**3.1 Smoke Estimate**

I began by downloading and unzipping the GeoJSON Files.zip “Wildland Fire Polygons Fire Feature Data Open Source GeoJSON Files” from USGS (Welty and Jeffries, 2021). Of the GeoJSON exports, I used the “combined” dataset to avoid duplicated fires from data merges. Due to GitHub file size restraints, I saved the "USGS\_Wildland\_Fire\_Combined\_Dataset.json" file to the directory above my project’s parent directory with the intent of reducing the data then saving it to the intermediate\_data folder.

For my analysis I limited fires to those occurring after 1963 and within 1250 miles of Pahrump, NV. The first script, data\_acquisition, reads in the fire GeoJSON information and finds the distance between the closest edge of each fire and the center of Pahrump. I chose to use the edge of the fire rather than the center because it may be closer to town resulting in more smoke. However, I did choose to center the Pahrump town coordinates to avoid biasing smoke measurements to one side of the city. The fires which occurred within 1250 miles of Pahrump were retained.

Next, I extracted a list of fires which occurred after 1963. This list was inner joined with the set of fires which occurred within 1250 miles of Pahrump to create my fire sample set.

Then I created the annual smoke estimate. The research behind the estimate calculation is detailed in the “Background/Related Work” section, however I will post the final smoke estimate formula below. This estimate was created for each fire occurring after 1963 within 1250 miles of Pahrump.

To create an annual smokiness estimate I averaged the Final Smoke Estimate across all fires per year. This method was chosen initially because it would be compared to AQI measurements over the same season, thus no amortization was necessary to “extend” the smokiness to the rest of the year. However, it is worth noting that my Death Valley attendance numbers are given annually, and thus a kind of “amortization” to estimate the annual smokiness may have been more appropriate. It's also worth noting that initial results contained a summed annual smokiness estimate rather than an averaged value due to an accidental code change. While updates to reflect the average annual smoke estimate have been propagated through all code, changes were not made to project documents submitted prior to 12/11/23. Additionally, I continue to use the linear smoke prediction model and code as was written for the summed annual smokiness estimate and submitted for Part 1.

**3.2 AQI Estimate**

AQI data is pulled from the EPA AQS API. I began by pulling information about different air quality measures to identify the 5-digit code corresponding to AQI. I then retrieved the 5-digit codes which represented individual AQI elements (e.g., 42101 represents carbon monoxide). These 5 digits codes were combined into gaseous and particulates lists for later querying. Then I created a dictionary with information on Pahrump, including the center of town’s latitude/longitude coordinates from the smoke estimate. A call to the API returned all sensors within Nye County (contains Pahrump).

The crux of AQI data acquisition lay in repeated API calls to the EPA AQS API requesting the daily summary of AQI information (gaseous or particulate) from the sensors in Nye County. While some stations produced more granular AQI information the EPA indicated “The Air Quality Index is based on daily air quality summaries, specifically daily maximums or daily averages. It is not valid to use shorter-term (e.g. hourly) data to calculate an AQI value.” (AirNow.gov, 2018). Thus only 24-HR BLK AVG AQI measurements were collected for each gas/particulate where available. Data was collected over fire season (May 1st - Oct 31st) for consistency of comparison with the annual final smoke estimate.

Typically, AQI would be calculated using the max particulate value, however my sensors only measured a single particulate, PM10. The daily per sensor values for PM10 were averaged to create a single AQI estimate per year.

I was interested in how closely my final smoke estimate mirrored AQI and would have expected a strong positive correlation. However, in graphing the scaled smoke estimate against the scaled AQI, the R2 value was 0.01 with a p-value of 0.78 - far above the threshold required for statistical significance. However, there were also limitations on my AQI calculation including location of sensors and limited data which will be discussed in my “Limitations” section. The residuals were not normally distributed in the relationship, which could result in inaccuracies in the p-value.

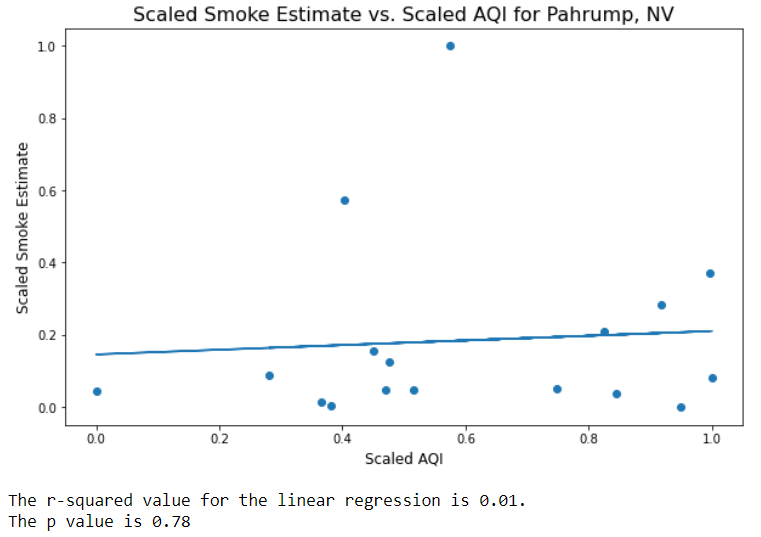


Figure 1. Scaled smoke estimates and AQI measurements do not appear to be correlated. Scaled smoke accounts for almost none of the variation in scaled AQI, and the p-value of the correlation coefficient indicates a failure to reject the null hypothesis (Beta = 0).

**3.3 Comparison with Death Valley Attendance**

Bringing in Death Valley data was fairly straightforward as it contained only the collection year and annual attendance. Some cleaning was required to remove the first 3 human-readable title rows so Pandas could process it correctly. Once stored in Pandas, the attendance data was merged into a larger table containing AQI and final smoke estimates to allow for easier graphing.

To map the correlations between smoke, AQI, and Death Valley attendance I used a linear regression model with both the independent variable (smoke or AQI estimate) and dependent variable (Death Valley attendance) normalized. I normalized both measurements to ensure they were on the same scale for visual analysis. I chose linear regression not only for its simplicity, but also for its Pearsons correlation coefficient which describes how much variability in the dependent variable is attributable to the independent variable. Additionally, I calculated a p-value for the correlation coefficient which indicates its statistical significance.

**3.4 Ethical Implications**

I believe there are limited ethical concerns in using this data and modeling it in the above manner. It is possible that park attendance estimates do not accurately reflect all visitors and may leave out portions of the population (e.g., those who carpool may be more challenging to count) and fire data may be more accurate for areas with more resources (e.g., areas who could afford to send fire scouts, areas near fires, or areas with land/resources/items “worth monitoring”).

There are certainly ethical implications in how conclusions are interpreted. First, linear regression assumes 5 conditions (linearity, homoscedastic, normal distribution of errors, no/limited covariates, and no autocorrelation) to be a valid modeling technique. If these conditions are not met, I cannot assume the conclusions of correlation are valid. Additionally, it would be wrong to conflate correlation with causation. Even if I should find a strong correlation between smoke and park attendance, I could not say conclusively that smoke is what deterred visitors. Finally, I cannot state in totality how much of Pahrump’s economy would be harmed by smoke because of the correlation issue described above and because Death Valley-related business is likely only a fraction of total town income.

**4. Findings**

Visually it appeared that there was little if any correlation between the smoke estimate and Death Valley attendance. This was confirmed with an R2 of 0.0, and a coefficient which was not statistically significant (0.63>>0.05). Examining the residuals vs. fitted values it would appear that residuals may be homoscedastic, and the Q-Q plot largely adheres to the guideline. Despite my belief that smoke would impact attendance, that doesn’t appear to be the case, and my results align with that of previous research.

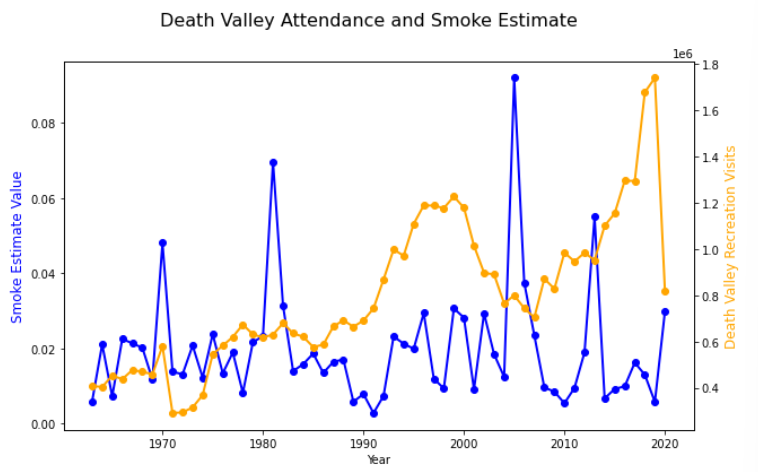


Figure 2. Annual smoke estimates and Death Valley attendance don’t appear to be visually correlated.

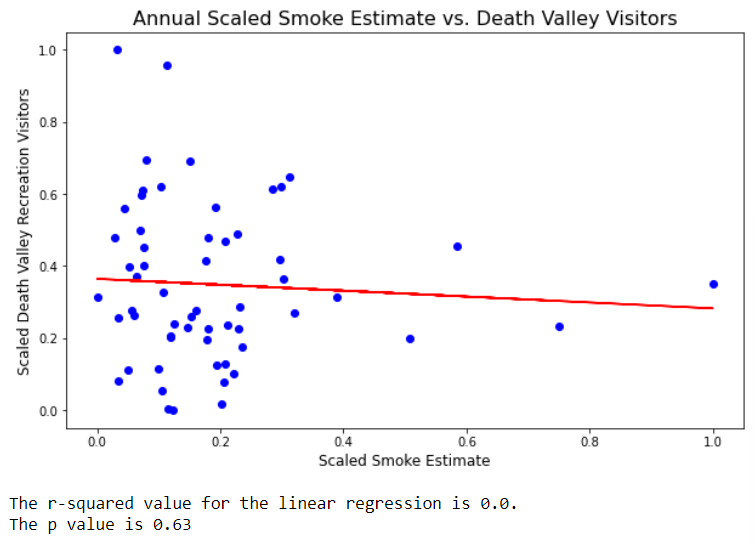


Figure 3. Lack of correlation appears to be confirmed by the linear regression model and p-value which indicates a failure to reject the null hypothesis.

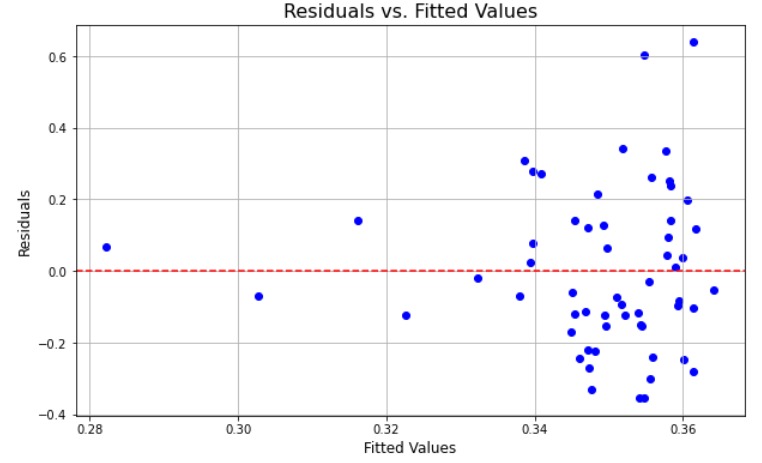
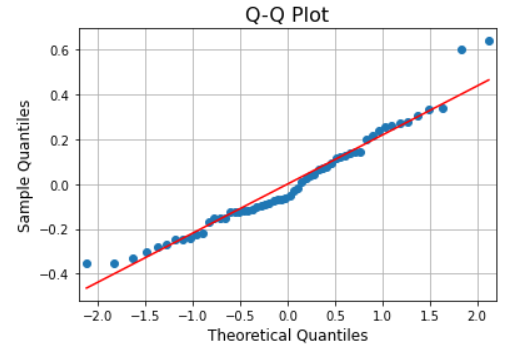
 

Figure 4. Resisudals are largely randomly distributed around 0.0, and conform to the guideline in the Q-Q plot.

Examining AQI and attendance it would appear that there is a stronger correlation. This is confirmed with an R2 of 0.33 and a p-value of 0.02. I also see that most assumptions for linear regression are met, adding confidence to the accuracy of the p-value. As with earlier research we cannot confuse correlation for causation. We cannot say conclusively that poor annual air quality leads to fewer park visitors, but it does seem to be a factor in the variation. However we can reject our null hypothesis that AQI has no correlation with park attendance.

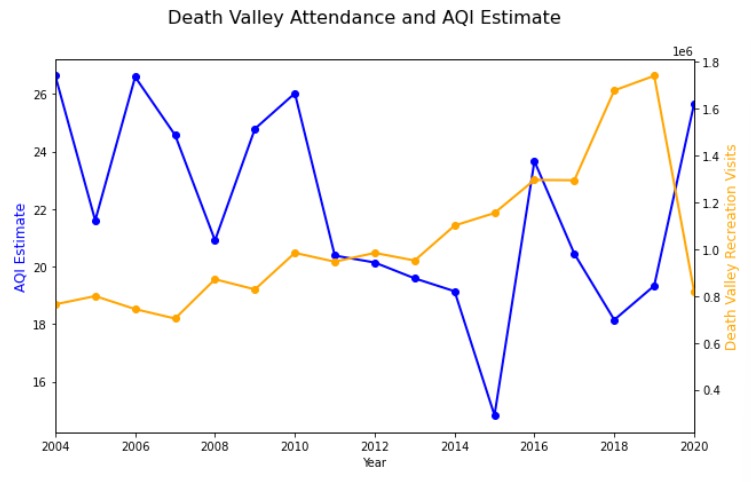


Figure 5. There may exist a negative correlation between annual AQI estimates and Death Valley attendance.

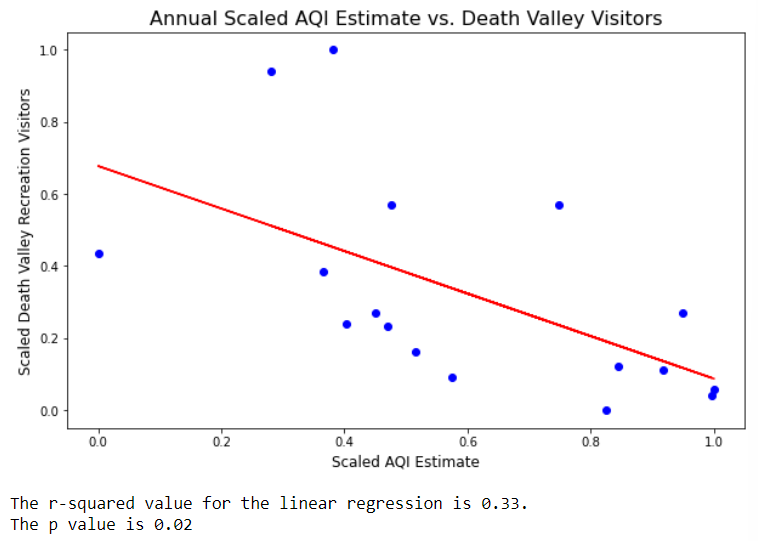


Figure 6. The apparent correlation is confirmed with a statistically significant correlation coefficient and an R2 of 0.33.

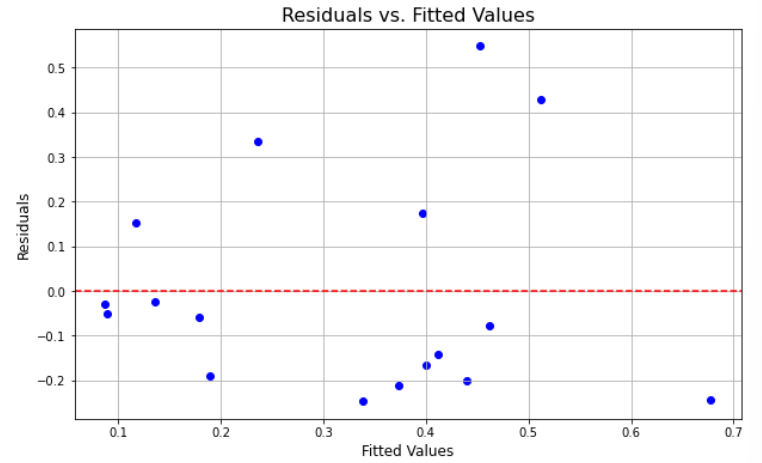
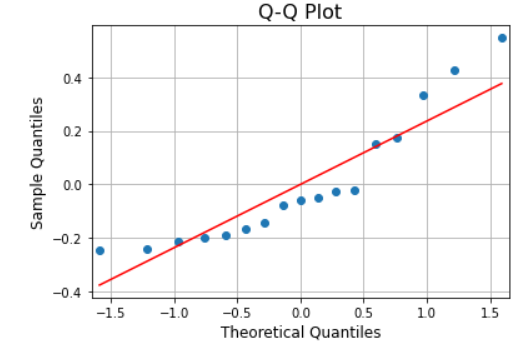
 

Figure 7. The residuals are fairly randomly distributed and largely conform with the guideline on the Normal Q-Q plot.

**5. Discussion/Implications**

**5.1 Implications of Findings**

These findings are interesting in that they align with historical work on the impacts of wildfire smoke and National Park attendance. However, they are still counter-intuitive to what most people would expect to see. It is also interesting that average AQI in Nye County during fire season seems to be correlated with park attendance in a statistically significant way. The authors of the historical works stress that correlation cannot be misinterpreted as causation, and I strongly echo that sentiment. An article summarizing Clark et. al’s findings posits that visitors have too much invested in going to the parks (finances and time) to be deterred by smoke. Given ongoing improvements in AQI (value decreasing) and the exploding interest in National Parks (attendance increasing) time may be a covariate in the apparent correlation.

At this time it would not appear that the city council should take action to curb wildland smoke to preserve park attendance. However, knowing that smoke has adverse health impacts, it may be a worthwhile effort for other human-centered reasons. Given people will travel to be outdoors in Death Valley, their exposure to smoke would be greater than if they had chosen to remain indoors. While my hypothesis may have been disproven, its falsity actually puts more visitors at risk. To prevent avoidable health impacts, Death Valley could publish smoke warnings, provide good-quality air masks, and monitor conditions around the park to direct visitors to areas with better air. I would also suggest repeating this analysis in 5 years to validate the findings still hold. It may also be worth refining the smoke estimate to better capture reality, and improve AQI measurements to capture all available particulates and gasses.

**5.2 Human-Centered Decision Making**

Because I have no control of the data collection methodologies there is little I can do to drive additional equity beyond contacting the original data owners and calling out potential limitations in my write up. However, I was extremely human-centered in how I wrote my code and README and formatted my directory. I wanted my work to be understandable for future data scientists who may use this code to reassess the impact of AQI and smoke on Death Valley attendance. To that end I have provided extensive documentation and plain English explanations of what each block of code does and given runtime updates (e.g., “Processing data for 1968…”) to provide assurance that long segments of code are indeed running and not silently terminated or stuck in an endless loop.

Additionally, my README is designed to allow someone to reproduce my analysis from scratch. There are links to data sources, as well as detailed information on how to download, select, store, and preprocess information where necessary. Ideally data would be provided using a static link, but it does not appear that wildfire information or AQI information is given in that format.

Finally, I believe this report adds important contextual information for future researchers, allowing them to more easily understand my motivation and rationale, which may not be conveyed in my README or code documentation. It also lists at length the limitations of my calculations, data, and hypotheses which can help others decide if this analysis is right for them, or trustworthy in its findings.

**6. Limitations**

As with any analysis, there are a number of limitations in this work and extrapolating out to future works. In the interest of space, I have listed the four major limitations in subsections below. For data-specific concerns please see my README (Ekrolen).

**6.1 Consistency in Time Periods Evaluated**

Smoke estimates are assumed to pertain to fires which occur during fire season (May 1st - Oct 31st). However the combined fires in the USGS data make it very challenging to identify a single fire date. Thus, we cannot be sure that all fires happen during fire season. This has implications on how we measure AQI and Death Valley attendance. AQI measurements are limited to those which explicitly occur during fire season to align with smoke estimates. However, without definitive fire dates we cannot be sure if this removal of data is constraining the measurements for better or worse.

Further, Death Valley attendance is an annual measurement, unconstrained to fire season. When comparing AQI and smoke estimates against attendance, amortized annual estimates or a monthly attendance breakdown would have been more effective.

**6.2 Additional Factors Impact Smokiness**

I believe my smoke estimate was fairly robust given the data provided. However, many more elements impact how much smoke is created during a fire. For example, composition of fuel impacts how much a fire can burn and the resulting smokiness. Given the southwestern United States is relatively arid, it’s possible that the fires closest to Pahrump actually produced little smoke. However, this information is not available in our data and likely would be very complex to incorporate, especially for historical burns. Instead, I resort to assuming a linear relationship between acres burned and smoke produced.

Additionally, direction of terrain and national weather conditions would impact how much smoke reaches Pahrump from burning fires. Neither of these elements were contained in our data or brought in from external sources. Again, this would likely be an extremely complex incorporation, but could be interesting for the closest or largest fires. I assumed that smoke dispersed like light (1/distance^2 relationship), which may be overly simplistic.

**6.3 Local vs. Visiting Behavior**

Examining Death Valley attendance is interesting not only given its impact on tourism through Pahrump, but also because it likely provides an interesting place to visit for locals. Given the amount of school resources available on the Death Valley website, as well as the unique local ecology, I could image that it makes up part of local school children’s regional education. Attendance data does not distinguish between local and visitor attendance. While Pahrump’s small size likely makes it a miniscule individual contributor to park attendance, there are other towns nearby who may visit. It would be interesting to examine how local vs. out-of-town guests’ plans are impacted by smoky conditions. As stated elsewhere in this paper, those visiting from out of town may have too much invested to change plans when smoky conditions occur. However, locals can simply choose not to attend the park. These differing motivations could be captured in our data, but are not accounted for in our analysis.

**6.4 AQI Limitations**

While AQI is a well-defined metric of air quality, its measurement for Pahrump is fraught. Our AQI measurements pull from all sensors in the county. However, Nye County’s nail-like shape with Pahrump at the bottom mean many sensors are far from town. It is possible that AQI conditions in Nye County do not resemble local conditions in Pahrump.

Further, only a single particulate is captured in local sensors, far from the 7 typically used to calculate AQI. This likely impacts the accuracy of measurement given most particulates and all gasses are unaccounted for. It is possible that the inclusion of these other particulates and gasses could create differing correlations for smoke or park attendance.

**7. Conclusion**

In this work I sought to answer a human-centered data science question: is there a correlation between smoke estimates/AQI and Death Valley attendance? Given the outsized role of tourism in Pahrump’s economy, answering this question would allow citizens to diversity their industries or challenge their government to do more about fire prevention in the event that smoke/poor air quality should increase.

I find that my smoke estimate and park attendance have no correlation which confirms others’ results. However, I do see that AQI and park attendance appear to be correlated with statistical significance, though AQI measurements may be unreliable for Pahrump (see “Limitations” section for more information).

This study should illustrate reproducible standards key to human-centered data science work. Code and documentation are thorough and explained in detail. Data is also easily accessible, though third party sourcing limits our ability to provide a static link.

**8. References**

AirNow.gov. (2018, September). Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI). https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf

Avitt, A. (2023, April 4). *Forest Service Professionals prepare for a prescribed burn*. US Forest Service. https://www.fs.usda.gov/features/professionals-prepare-for-prescribed-burn.

Clark, M., Killion, A., Williamson, M. A., & Hillis, V. (2023, June 7). *Increasing wildfire smoke has limited impacts on national park visitation in the American West*. Ecological Society of America. https://esajournals.onlinelibrary.wiley.com/doi/10.1002/ecs2.4571.

*Economy in Pahrump, NV*. Pahrump, NV. (n.d.). https://www.bestplaces.net/economy/city/nevada/pahrump

Ekrolen. (n.d.). *Ekrolen/data-512-wildfire-project*. GitHub. https://github.com/ekrolen/data-512-wildfire-project

Frisbey, David, "A comparison of smoke emissions from prescribed burns and wildfires" (2008). Master's Theses. 3558. DOI: https://doi.org/10.31979/etd.ve25-7wu5

Gellman, J., Walls, M., & Wibbenmeyer, M. (2022). Wildfire, smoke, and outdoor recreation in the Western United States. *Forest Policy and Economics*, *134*, 102619. <https://doi.org/10.1016/j.forpol.2021.102619>.

*Pahrump, NV Reviews*. Niche. (2023, October 26). https://www.niche.com/places-to-live/pahrump-nye-nv/reviews/

Pontecorvo, E. (2020, September 10). *Burn out: Frequent fires are changing Western Landscapes*. Burn out: Frequent fires are changing Western landscapes. https://www.hcn.org/articles/climate-desk-wildfire-burn-out-frequent-fires-are-changing-western-landscapes#:~:text=However%2C%20overall%2C%20re%2Dburns,just%20fine%2C%E2%80%9D%20he%20said.

*The 15 best things to do in Pahrump - 2023 (with photos)*. Tripadvisor. (n.d.). https://www.tripadvisor.com/Attractions-g45986-Activities-Pahrump\_Nevada.html

U.S. Department of the Interior. (2020, March 2). *Disclaimer (U.S. National Park Service)*. National Parks Service. https://www.nps.gov/aboutus/disclaimer.htm

U.S. Department of the Interior. (2023, March 24). *NPS visitor use statistics definitions*. National Parks Service. https://www.nps.gov/subjects/socialscience/nps-visitor-use-statistics-definitions.htm

Welty, J.L., and Jeffries, M.I.. (2021), *Combined wildland fire datasets for the United States and certain territories, 1800s-Present: U.S. Geological Survey data release*, https://doi.org/10.5066/P9ZXGFY3.

Wikimedia Foundation. (2023, November 10). *Pahrump, Nevada*. Wikipedia. https://en.wikipedia.org/wiki/Pahrump%2C\_Nevada#Economy

**9. Data Sources**

AQS API URL root: https://aqs.epa.gov/data/api

U.S. Department of the Interior. (n.d.). *Stats report viewer*. Annual Park Recreation Visitation (1904 - Last Calendar Year). https://irma.nps.gov/Stats/SSRSReports/Park%20Specific%20Reports/Annual%20Park%20Recreation%20Visitation%20(1904%20-%20Last%20Calendar%20Year)?Park=DEVA

Welty, J.L., and Jeffries, M.I.. (2021), *Combined wildland fire datasets for the United States and certain territories, 1800s-Present: U.S. Geological Survey data release*, <https://doi.org/10.5066/P9ZXGFY3>.